Queueing Models and
Service Management
Vol. 1, No. 1, page 075-094, 2018

QMSM


# A Stochastic Model for Automated Teller Machines Subject to Catastrophic Failures and Repairs

Srinivas R. Chakravarthy[1,*] and Sudha Subramanian[2]

[1]Department of Industrial and Manufacturing Engineering
Kettering University, Flint, MI 48504, USA

[2]Analytics, Sparkfish, 2024, W 15th Street, Plano, TX 75075, USA

*(Received April 2017; accepted October 2017)*

**Abstract:** In this paper we develop a queueing model useful in service industries dealing with automatic teller machines (*ATM*s) that are commonly used by people all over the world. We assume that these service systems are subject to failures due to catastrophic events such as power outage, mechanical or electrical problems. Arrivals of customers are modeled using a Markovian arrival process and the service times are assumed to be of phase type. Individual customer cash requirements are modeled using a probabilistic rule and the machine has a finite capacity for holding the cash. Assuming the failure times, repair times, and cash replenishment times to be exponential, we analyze the model using matrix-analytic methods, and present two illustrative examples to bring out the salient features. Some well-known queueing-inventory models are shown to be special cases and in some of these cases we derive explicit expressions for the steady-state probability vectors. The model studied is generic in that it can be applied in the context of queueing-inventory situations.

**Keywords:** Catastrophes, GI/M/1-type process, inventory, Markovian arrival process, matrix-analytic method, phase type distribution, queueing.

## 1. Introduction

Queueing models play a major role in service industries such as healthcare, telecommunications, food industry, banking, supply chain management, and other businesses that serve people directly or indirectly. In this paper we consider a useful queueing model in banking sector related to automatic teller machine (*ATM*). Although this paper discusses the model in the context of *ATM*s, this is applicable to a variety of other service models such as vending machines or electronic gadgets and possibly in self-service technologies. We refer the reader to [2] for details on self-service technologies that include *ATM* usage. *ATM*s are used worldwide by people on a daily basis, and are subject to failures in the form of (a) running out of money to dispense to the customers; and (b) catastrophic events such as power outage, mechanical failures, etc. Thus, to model an *ATM* service system, one has to incorporate both catastrophic and inventory-type situations.

Queueing models in which customers are removed from the system due to catastrophic events have been studied extensively in the literature. We refer the reader to a recent paper [6] dealing with a new type of catastrophic model and this paper contains some key references on queueing models with catastrophic as well as on negative arrivals. Also, there is a huge literature on queue-inventory models with applications in many service sectors such as hospitals, restaurants, and servicing automobiles. For a recent study on queue-inventory and for some references on queue-inventory type modeling we refer the reader to [5, 7, 11, 12, 13, 14, 24, 25, 28] and the references therein.

Motivated by the study performed by authors in [27] to build a model to predict the probability of *ATM* failure occurring within a specified time frame, this paper studies the efficiency and the

* Corresponding author
Email : schakrav@kettering.edu

availability of an *ATM* system under failures, repairs, and (cash) replenishment. By incorporating the *ATM* availability information into the bank's software or on the bank's mobile application, the customers can check the availability of *ATM*s and plan their trip accordingly. Further, this will help the banks to provide a better service to their customers.

The rest of the paper is organized as follows. In Section 2 we describe the model under study in detail and the steady-state analysis of the model is presented in Section 3. A few special cases of the model are presented in Section 4. A few illustrative numerical examples to bring out the qualitative nature of the model are discussed in Section 5. Finally, in Section 6 we present some concluding remarks and future research work.

## 2. Model Description

In this paper we look at a queueing model of the type *MAP/PH/1* useful in service industry related to *ATM*. Customers needing to withdraw cash arrive at an *ATM* machine according to a Markovian arrival process (*MAP*) with representation ($D_0$, $D_1$) of order $m$. A customer finding the machine down will be considered lost. The *ATM* can be down either due to shocks or due to no cash. The maximum cash the machine can hold is assumed to be *S*. An arriving customer withdraws cash in multiple of units of local currency depending on the location of the machine. For example, the units could be rupees if in India or in dollars if in USA. We assume that the customer withdraws money according to the following probabilistic rule. With probability $p_r, 1 \leq r \leq N$, the customer withdraws *r* units of cash. The upper limit, *N*, is imposed by the regulating agency governing the *ATM* usage. For example, recently due to demonetization in India, the upper limit is specified as Rs.4,500. Note that in this example, *N* will not be set as 4,500 but rather the multiples of the units of currencies that are dispensed. Thus, if the dispensing units are in Rs.500 notes, then $N = 9$. Also, the dispensing currency could be in different denominations reducing the value of *N*. However, in this paper we will simply assume that the customers are given cash in some units but not exceeding *N*. Further we assume that *S* is some multiple of *N*. That is, $S = KN$, for some finite *K*. This assumption is only for convenience and is not a restriction as we allow the possibility of the customers to withdraw the needed amount subject to the maximum cash allowed to withdraw and the cash available in the machine at that moment. Also observe that $\sum_{i=1}^{N} p_i = 1$.

The machine can be down due to (i) cash depleting to zero or (ii) catastrophic event caused by shocks. The shocks, which are independent of the arrival process, are assumed to occur according to a Poisson process with rate θ. An arriving shock will instantaneously cause the machine to fail unless the machine is already down in which case the shocks will have no bearing. The cash replenishment times and the repair times of the machine are assumed to follow exponential distribution with parameters, respectively, given by $\delta_1$ and $\delta_2$. The service times of the customers are assumed to be of phase type with representation $(\boldsymbol{\beta}, T)$ of order *n*. The mean service time is given by $\frac{1}{\mu} = \boldsymbol{\beta}(-T)^{-1}\boldsymbol{e}$ (see, e.g., [20]).

The model studied in this paper incorporates of catastrophic events and (*s, S*)-type inventory in the context of *MAP/PH/*1-type queueing model. That is, our model falls in the category of queueing-inventory system in the sense that the classical *MAP/PH/*1-type queueing model studied here is subject to (a) catastrophic events resulting in the system needing a repair after removing all customers from the system; (b) every customer is to be served with a finite number of inventory ranging from 1 to *N* with a certain probability; and (c) when the inventory becomes empty upon completion of a service, all waiting customers are removed from the system and new arrivals are accepted only after a replenishment of inventory occurs. Thus, our queueing model in this paper falls under the topic of queueing-inventory. However, to our knowledge there is no queueing-inventory model in which demands (both waiting and future ones) are lost when there is no inventory. The closest ones that consider modeling the lost demands are in [24, 25], wherein the authors consider *M/M/*1-type queueing-inventory models with all new arrivals lost during the time the system waiting is for replenishment; however, all those customers in the queue at the time the inventory becomes zero will be kept and served after replenishment occurs. That is, in these papers the authors assume that only future demands that arrive during the replenishment times are lost and that those waiting in the queue at the time of stock outs will be met and served. Also,

note that in the above referenced papers each customer requires only one inventory.

For use in sequel we need the following notation. (a) $\boldsymbol{e}$ will denote a column vector (of appropriate dimension) of 1's; (b) $\boldsymbol{e}_i$ will denote a unit column vector (of appropriate dimension) with 1 in the $i^{\text{th}}$ position and 0 elsewhere; (c) $I$ an identity matrix (of appropriate dimension); (d) The notation "′" appearing as superscript on a vector or a matrix denotes the transpose of a matrix; (e) $\boldsymbol{T}^0$ is such that $\boldsymbol{Te} + \boldsymbol{T}^0 = \boldsymbol{0}$; (f) The symbols, $\otimes$ and $\oplus$, respectively, will stand for the Kronecker product and sum of matrices. For details on Kronecker products and sums, we refer the reader to [9, 18, 26] for details and properties on Kronecker products and Kronecker sums. Note that when there is a need to emphasize the dimension of a vector or a matrix we will do so. As an example, a unit vector of dimension $S$ will be denoted as $\boldsymbol{e}_i(S)$ rather than $\boldsymbol{e}_i$.

*MAP*, a rich class of point processes that includes many well-known processes such as Poisson, PH-renewal processes, and Markov-modulated Poisson process, was first introduced as a versatile Markovian point process by Neuts [19]. Since then, this versatile process has been studied extensively in different contexts by many authors. For further details on *MAP* and their usefulness in stochastic modeling, we refer to [16, 17, 21, 22, 23] and for a review and recent work on *MAP* we refer the reader to [1, 3, 4].

Let $\boldsymbol{\eta}$ be the stationary probability vector of the Markov process with irreducible generator $D = D_0 + D_1$. That is, $\boldsymbol{\eta}$ is the unique (positive) probability vector satisfying

$$\boldsymbol{\eta} D = 0, \boldsymbol{\eta e} = 1. \tag{1}$$

Verify that the arrival rate, $\lambda$, also known as the **fundamental rate** giving the expected number of arrivals per unit of time in the stationary version of the *MAP*, is given by $\lambda = \boldsymbol{\eta} D_1 \boldsymbol{e}$.

## 3. The Steady-State Analysis

The steady-state analysis of the queueing model described in Section 2 will be described in this section. First we need to define a few notation. Let $J_1(t)$, $J_2(t)$, $J_3(t)$, and $J_4(t)$ denote, respectively, the number of customers in the system, the cash level in the machine, the phase of the service (if any), and the phase of the arrival process, at time $t$. The process $\{(J_1(t), J_2(t), J_3(t), J_4(t)): t \geq 0\}$ is a continuous-time Markov chain (*CTMC*) with state space in grouped form given by

$$\boldsymbol{\Omega} = \{\underline{*}\} \cup \{\underline{\hat{*}}\} \cup \{\underline{\boldsymbol{0}}\} \cup \{\underline{\boldsymbol{i}}, i \geq 1\},$$

where the set of states and their definitions are as follows:

- The set of states, $\underline{*} = \{k, 1 \leq k \leq m\}$, of dimension $m$ corresponds to the system being down due to zero cash and the arrival process is in one of $m$ phases.
- The set of states, $\underline{\hat{*}} = \{(j,k), 1 \leq j \leq S, \ 1 \leq k \leq m\}$, of dimension $mS$ corresponds to the system being down due to shocks with the arrival process in one of $m$ phases and the cash level being in one of $S$ states.
- The set of states, $\underline{\boldsymbol{0}} = \{(j,k), 1 \leq j \leq S, \ 1 \leq k \leq m\}$, of dimension $mS$ corresponds to the system being in idle state with the arrival process in one of $m$ phases and the cash level being in one of $S$ states.
- The set of states, $\underline{\boldsymbol{i}} = \{(i,j,r,k), 1 \leq j \leq S, \ 1 \leq r \leq n, \ 1 \leq k \leq m\}$, of dimension $mnS$ corresponds to the system being busy with $i$ customers in the system; the arrival process in one of $m$ phases; the service process in one of n phases, and the cash level being in one of $S$ states.

It is easy to verify the *CTMC* with the above state space has the infinitesimal generator matrix, $Q$, of the form:

$$Q=\begin{bmatrix} D-\delta_1 I & 0 & \delta_1 \boldsymbol{e}'(S)\otimes I & & & & & \\ 0 & I\otimes(D-\delta_2 I) & \delta_2 I & & & & & \\ 0 & \theta I & I\otimes(D_0-\theta I) & I\otimes\boldsymbol{\beta}\otimes D_1 & & & & \\ E & \theta(I\otimes\boldsymbol{e}\otimes I) & \tilde{A}_2 & A_1 & A_0 & & & \\ E & \theta(I\otimes\boldsymbol{e}\otimes I) & & A_2 & A_1 & A_0 & & \\ E & \theta(I\otimes\boldsymbol{e}\otimes I) & & & A_2 & A_1 & A_0 & \\ \vdots & \vdots & & & & \ddots & \ddots & \ddots \end{bmatrix} \tag{2}$$

where the matrices appearing in the above generator are as follows

$$A_0 = I\otimes D_1,\ A_1 = I\otimes(T\oplus D_0)-\theta I,\ A_2 = \hat{F}\ \otimes \boldsymbol{T}^0\boldsymbol{\beta}\otimes I,$$
$$\tilde{A}_2 = \hat{F}\ \otimes \boldsymbol{T}^0\otimes I,\ \ E=\textstyle\sum_{i=1}^{N}\tilde{p}_i(\boldsymbol{e}_i(S)\otimes \boldsymbol{T}^0\otimes I), \tag{3}$$

$$\tilde{p}_i=\textstyle\sum_{j=1}^{i}p_i,\quad \hat{F}=\begin{bmatrix} \hat{F}_1 & & & & \\ \hat{F}_2 & \hat{F}_1 & & & \\ & \hat{F}_2 & \hat{F}_1 & & \\ & & \ddots & \ddots & \\ & & & \hat{F}_2 & \hat{F}_1 \end{bmatrix}, \tag{4}$$

and

$$\hat{F}_1=\begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ p_1 & 0 & 0 & \dots & 0 & 0 \\ p_2 & p_1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ p_{N-1} & p_{N-2} & p_{N-3} & \dots & p_1 & 0 \end{bmatrix},\ \hat{F}_2=\begin{bmatrix} p_N & p_{N-1} & p_{N-2} & \dots & p_2 & p_1 \\ 0 & p_N & p_{N-1} & \dots & p_3 & p_2 \\ 0 & 0 & p_N & \dots & p_4 & p_3 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & p_N \end{bmatrix}, \tag{5}$$

The generator given in (2) is of *GI/M*/1-type and possesses matrix-geometric steady-state probability vector ([20]) as shown in the next subsection. Such processes have been extensively studied in the literature (see e.g., [1, 3, 10, 15, 20]).

### *3.1. The steady-state probability vector*

In this section we will look at the steady-state probability vector of the generator given in (2). Towards this end, we define x, partitioned as $\boldsymbol{x}=(\boldsymbol{x}^*,\hat{\boldsymbol{x}}^*,\ \boldsymbol{x}(0),\ \boldsymbol{x}(1),\ \boldsymbol{x}(2)\dots)$, to be the steady-state probability vector of $Q$. That is, $\boldsymbol{x}$ satisfies

$$\boldsymbol{x}Q=0,\ \boldsymbol{x}\boldsymbol{e}=1. \tag{6}$$

For use in sequel, we further partition the steady-state vector as follows. Recall that $S = K\,N$.

- $\hat{\boldsymbol{x}}^*=(\hat{\boldsymbol{x}}_1^*,\dots,\hat{\boldsymbol{x}}_K^*)$,, where $\hat{\boldsymbol{x}}_j^*, 1\le j\le K$, of order $Nm$, gives steady-state probability that the system is in down state with the cash level at $jN$ and the arrival process is in one of $m$ states.
- $\boldsymbol{x}(0)=\big(\boldsymbol{x}_1(0),\dots,\boldsymbol{x}_K(0)\big)$, where $\boldsymbol{x}_j(0), 1\le j\le K$, of order $Nm$, gives the steady-state probability vector that the system is idle with the cash level at $jN$ and the arrival process is in one of $m$ states.
- $\boldsymbol{x}(i)=\big(\boldsymbol{x}_1(i),\dots,\boldsymbol{x}_K(i)\big)$, where $\boldsymbol{x}_j(i), 1\le j\le K$, of order $Nmn$, gives the steady-state probability vector that the system is busy with $i$ customers in the system, the cash level at $jN$ , the service is in one of $n$ phases, and the arrival process is in one of $m$ states.

Since customers are cleared from the system whenever the *ATM* machine's cash level hits zero or when a catastrophic event occurs, the queue is always stable and the steady vector $\boldsymbol{x}$ satisfying (6) always exists. The steady-state equations given in (6) can be rewritten as

$$\boldsymbol{x}^*(D-\delta_1 I)+\sum_{i=1}^{\infty}\boldsymbol{x}(i)E=\boldsymbol{0}, \tag{7}$$

$$\widehat{\boldsymbol{x}}^*[I\otimes(D-\delta_2 I)]+\theta\,\boldsymbol{x}(0)+\theta\sum_{i=1}^{\infty}\boldsymbol{x}(i)(I\otimes\boldsymbol{e}\otimes \mathrm{I})=\boldsymbol{0}, \tag{8}$$

$$\delta_1\boldsymbol{x}^*(\boldsymbol{e}'\otimes I)+\delta_2\widehat{\boldsymbol{x}}^*+\boldsymbol{x}(0)[I\otimes(D_0-\theta I)]+\boldsymbol{x}(1)\tilde{A}_2=\boldsymbol{0}, \tag{9}$$

$$\boldsymbol{x}(0)[I\otimes\boldsymbol{\beta}\otimes D_1)+\boldsymbol{x}(1)A_1+\boldsymbol{x}(2)A_2=\boldsymbol{0}, \tag{10}$$

$$\boldsymbol{x}(i-1)A_0+\boldsymbol{x}(i)A_1+\boldsymbol{x}(i+1)A_2=\boldsymbol{0}, i\geq 2\,, \tag{11}$$

subject to the normalizing condition given by

$$\boldsymbol{x}^*\boldsymbol{e}+\widehat{\boldsymbol{x}}^*\boldsymbol{e}+\boldsymbol{x}(0)\boldsymbol{e}+\sum_{i=1}^{\infty}\boldsymbol{x}(i)\,\boldsymbol{e}=1. \tag{12}$$

Since the generator given in (2) is of *GI* /*M*/1−type, we can apply the well-known results from [20] to get the result in the following theorem.

**Theorem 1.** *The vectors,* $\boldsymbol{x}^*, \widehat{\boldsymbol{x}}^*,\ \boldsymbol{x}(0)$, *and* $\boldsymbol{x}(1)$, *are obtained by solving the following equations*

$$\boldsymbol{x}^*(D-\delta_1 I)+\sum_{i=1}^{\infty}\boldsymbol{x}(i)\,E=\boldsymbol{0}, \tag{13}$$

$$\widehat{\boldsymbol{x}}^*[I\otimes(D-\delta_2 I)]+\theta\,\boldsymbol{x}(0)+\theta\sum_{i=1}^{\infty}\boldsymbol{x}\,(i)(I\otimes\boldsymbol{e}\otimes \mathrm{I})=0, \tag{14}$$

$$\delta_1\boldsymbol{x}^*(\boldsymbol{e}'\otimes \mathrm{I})+\delta_2\widehat{\boldsymbol{x}}^*+\boldsymbol{x}(0)[I\otimes(D_0-\theta I)]+\boldsymbol{x}(1)\tilde{A}_2=\boldsymbol{0}, \tag{15}$$

$$\boldsymbol{x}(0)[I\otimes\boldsymbol{\beta}\otimes D_1)+\boldsymbol{x}(1)[A_1+RA_2]=\boldsymbol{0}, \tag{16}$$

subject to the normalizing equation

$$\boldsymbol{x}^*e+\widehat{\boldsymbol{x}}^*e+\boldsymbol{x}(0)e+\boldsymbol{x}(1)(I-R)^{-1}e=1, \tag{17}$$

and the rest of the steady-state vectors are obtained as

$$\boldsymbol{x}(i)=\boldsymbol{x}(1)R^{i-1}, i\geq 1, \tag{18}$$

where *R* is the minimal non-negative solution to the matrix-quadratic equation:

$$R^2A_2+RA_1+A_0=0. \tag{19}$$

**Note:** The computation of the *R* matrix can be carried out using a number of well-known methods such as (block) Gauss-Seidel iterative by exploiting the special structure of the coefficient matrices, $A_0$, $A_1$, and $A_2$, which are of dimension *mnS*. This is very important especially when *m*, *n* and *S* are significantly large. Further, the very special structure of matrix *R*, as shown below, should be exploited.

**Theorem 2.** *The matrix R, which is the minimal non-negative solution to (19), is of the form*

$$R=\begin{bmatrix} R_1 & 0 & 0 & \ldots & 0\\ R_2 & R_1 & 0 & \ldots & 0\\ R_3 & R_2 & R_1 & \ldots & 0\\ \vdots & \vdots & \vdots & \ldots & \vdots\\ R_K & R_{K-1} & R_{K-2} & \ldots & R_1 \end{bmatrix}. \tag{20}$$

**Proof.** First note that the matrices $A_0$, $A_1$, and $A_2$ are lower triangular. Hence, using the probabilistic interpretation of the rate matrix *R* (see e.g., [20]), it is obvious that *R* is also lower triangular. Let *R* be of the form

$$R = \begin{bmatrix} R_{1,1} & 0 & 0 & ... & 0 \\ R_{2,1} & R_{2,2} & 0 & ... & 0 \\ R_{3,1} & R_{3,2} & R_{3,3} & ... & 0 \\ \vdots & \vdots & \vdots & ... & \vdots \\ R_{K,1} & R_{K,2} & R_{K,3} & ... & R_{K,K} \end{bmatrix}. \tag{21}$$

Using the special structure of the coefficient matrices, $A_0$, $A_1$, and $A_2$, the matrix-quadratic equation in (19) can be rewritten as:

$$U_{i,j}F_1 + U_{i,j+1}F_2 + R_{i,j}[I_N \otimes (T \oplus D_0 - \theta I)] = 0,\ 2 \le i \le K, 1 \le j \le i-1, \tag{22}$$

$$U_{i,i}F_1 + R_{i,i}[I_N \otimes (T \oplus D_0 - \theta I)] + I_{Nn} \otimes D_1 = 0,\ 1 \le i \le K, \tag{23}$$

where $F_1 = \hat{F}_1 \otimes \boldsymbol{T^0\beta} \otimes I, F_2 = \hat{F}_2 \otimes \boldsymbol{T^0\beta} \otimes I$, and $U_{i,j}$, $1 \le i, j \le K$, is the $(i,j)^{\text{th}}$ (block) element of $R^2$. Note that $R^2$ is also lower triangular and hence $U_{i,j} = 0$, for $j > i$, $1 \le i, j \le K$. Also, it is easy to verify that $U_{i,j}$, $1 \le j \le i$, $1 \le i \le K$, is given by

$$U_{i,j} = \begin{cases} \sum_{k=j}^{i} R_{i,k}R_{k,j}, 1 \le j \le i-1, 2 \le i \le K, \\ R_{i,i}^2,\ j = i,\ 1 \le i \le K. \end{cases} \tag{24}$$

Noting that $U_{i,i} = R_{i,i}^2$ and the fact the coefficient matrices appearing in (23) do not depend on $i$, it is clear that $R_{i,i}$, $1 \le i \le K$, are identical. We will denote this common value to be $R_1$ and thus $R_1$ is the minimal non-negative solution to

$$R_1^2F_1 + R_1[I_N \otimes (T \oplus D_0 - \theta I)] + I_{Nn} \otimes D_1 = 0,\ 1 \le i \le K. \tag{25}$$

Now we will show $R_{2,1} = R_{3,2} = \cdots = R_{K,K-1}$ and the common value be denoted as $R_2$. Towards this end we look at (22) by setting $j = i - 1$ for $2 \le i \le K$. Now with the help of (24) and the fact that $R_{i,i} = R_1$, $1 \le i \le K$, the $(K-1)$ equations are given by

$$[R_{i,i-1}R_1 + R_1R_{i,i-1}]F_1 + R_1^2F_2 + R_{i,i-1}[I_N \otimes (T \oplus D_0 - \theta I)] = 0,\ 2 \le i \le K. \tag{26}$$

It is obvious that the minimal non-negative solution to (26) does not depend on $i$ and hence $R_{2,1} = R_{3,2} = \cdots = R_{K,K-1} = R_2$, where $R_2$ is the minimal non-negative solution to

$$[R_2R_1 + R_1R_2]F_1 + R_1^2F_2 + R_2[I_N \otimes (T \oplus D_0 - \theta I)] = 0,\ 2 \le i \le K. \tag{27}$$

Assuming that the result is true for $r = 1, 2, \cdots, i$, we will prove the result for $r = i + 1$. That is, assuming that $R$ is of the form,

$$R = \begin{bmatrix} R_1 & 0 & 0 & ... & 0 & 0 & 0 & ... & 0 \\ R_2 & R_1 & 0 & ... & 0 & 0 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & ... & \vdots & \vdots & \vdots & ... & \vdots \\ R_i & R_{i-1} & R_{i-2} & ... & R_1 & 0 & 0 & ... & 0 \\ R_{i+1,1} & R_i & R_{i-1} & ... & R_2 & R_1 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & ... & \vdots & \vdots & \vdots & ... & \vdots \\ R_{K,1} & R_{K,2} & R_{K,3} & ... & R_{K,K-i} & R_i & R_{i-1} & ... & R_1 \end{bmatrix}, \tag{28}$$

we will show that $R_{i+1,1} = R_{i+2,2} = \cdots = R_{K,K-i}$ and the common value be denoted as $R_{i+1}$. Towards this end we look at the following equations

$$U_{i+j,j}F_1 + U_{i+j,j+1}F_2 + R_{i+j,j}[I_N \otimes (T \oplus D_0 - \theta I)] = 0, 1 \le j \le K-i, 1 \le i \le K-1. \tag{29}$$

Now with the help of (28) along with (24), the equations in (29) are rewritten as

$$R_{i+j,j}R_1 + R_iR_2 + R_{i-1}R_3 + \cdots + R_1R_{i+j,j}]F_1 + [R_iR_1 + R_{i-1}R_2 + \cdots + R_1R_i]F_2$$
$$+R_{i+j,j}[I_N \otimes (T \oplus D_0 - \theta I)] = 0, \quad 1 \le j \le K-i,\ 1 \le i \le K-1. \tag{30}$$

Clearly, that the minimal non-negative solution to (30) does not depend on $j$ and only on the lag $i$ and hence $R_{i+1,1} = R_{i+2,2} = \cdots = R_{K,K-i}$ and let $R_{i+1}$ denote their common value. Note that $R_{i+1}$ is the minimal

non-negative solution to

$$[R_{i+1}R_1 + R_iR_2 + \cdots + R_1R_{i+1}]F_1 + [R_iR_1 + \cdots + R_1R_i]F_2 + R_{i+1}[I_N \otimes (T \oplus D_0 - \theta I)] = 0. \tag{31}$$

This completes the proof of the theorem.

The result in the following theorem is intuitively obvious and serves as an accuracy check in the computation of the steady-state probability vector.

**Theorem 3.** *We have*

$$\boldsymbol{x}^* + \widehat{\boldsymbol{x}}^*(\boldsymbol{e} \otimes I) + \boldsymbol{x}(0)(\boldsymbol{e} \otimes I) + \boldsymbol{x}(1)(I - R)^{-1}(\boldsymbol{e} \otimes I) = \boldsymbol{\eta}, \tag{32}$$

*where* $\boldsymbol{\eta}$ *is as given in* (1).

**Proof:** Post-multiplying equation (8) and (9) by $(\boldsymbol{e} \otimes I)$; the equations (10) and (11) by $(\boldsymbol{e} \otimes \boldsymbol{e} \otimes I)$, and adding the resulting equations with (8), we obtain

$$[\boldsymbol{x}^* + \widehat{\boldsymbol{x}}^*(\boldsymbol{e} \otimes I) + \boldsymbol{x}(0)(\boldsymbol{e} \otimes I) + \textstyle\sum_{i=1}^{\infty} \boldsymbol{x}(i)(\boldsymbol{e} \otimes I)]D = \boldsymbol{0}. \tag{33}$$

The stated result follows from (33) and the uniqueness of the vector $\boldsymbol{\eta}$.

The result in the following theorem is intuitively obvious since in steady-state the input rate should be equal to the output rate. This also serves as an accuracy check in the computation of the steady-state probability vector.

**Theorem 4.** *We have*

$$\textstyle\sum_{i=1}^{\infty} \boldsymbol{x}(i)(\ \boldsymbol{e} \otimes \boldsymbol{T}^0 \otimes \boldsymbol{e}) + \theta \sum_{i=1}^{\infty} i\boldsymbol{x}(i)\,\boldsymbol{e} + \sum_{i=1}^{\infty}(i-1)\boldsymbol{x}(i)E\boldsymbol{e} = \lambda(1 - P_{Loss}), \tag{34}$$

*where the probability that an arriving customer is lost due to the system being down is given by*

$$P_{Loss} = \tfrac{1}{\lambda}[\boldsymbol{x}^* D_1 \boldsymbol{e} + \widehat{\boldsymbol{x}}^*(\boldsymbol{e} \otimes D_1\boldsymbol{e})]. \tag{35}$$

**Proof.** First note the following definitions and their formulas.

- The quantity, $\boldsymbol{x}^* D_1 \boldsymbol{e} + \widehat{\boldsymbol{x}}^*(\boldsymbol{e} \otimes D_1\boldsymbol{e})$, gives the rate of loss of customers at their arrival times due to the system being down.
- The quantity, $\sum_{i=1}^{\infty} \boldsymbol{x}(i)(\boldsymbol{e} \otimes \boldsymbol{T}^0 \otimes \boldsymbol{e})$, gives the rate of customers leaving the system with a service.
- The expression, $\theta \sum_{i=1}^{\infty} i\boldsymbol{x}(i)\boldsymbol{e} + \sum_{i=1}^{\infty}(i-1)\boldsymbol{x}(i)E\,\boldsymbol{e}$, gives the rate of customers lost after getting admitted into the system either due to cash level becoming zero soon after a service completion or the system suffers from a catastrophic event.

Now post-multiply each one of the equations (7) through (11) by $\boldsymbol{e}$ of appropriate dimension. Secondly, multiplying the equation (11) by $i$ and adding this over $i$ along with the other equations that were post-multiplied by $\boldsymbol{e}$, we get

$$\textstyle\sum_{i=1}^{\infty} i\ \boldsymbol{x}(i)(\boldsymbol{e} \otimes D_1\boldsymbol{e}) + \boldsymbol{x}(0)(\boldsymbol{e} \otimes D_1\boldsymbol{e}) =$$

$$\theta \textstyle\sum_{i=1}^{\infty} i\ \boldsymbol{x}(i)\boldsymbol{e} + \sum_{i=1}^{\infty}(i-1)\boldsymbol{x}(i)E\boldsymbol{e} + \sum_{i=1}^{\infty} \boldsymbol{x}(i)(\boldsymbol{e} \otimes \boldsymbol{T}^0 \otimes \boldsymbol{e}) \tag{36}$$

The stated result follows immediately from Theorem 3 and the fact that $\lambda = \boldsymbol{\eta} D_1 \boldsymbol{e}$.

### *3.2. Stationary waiting time distributions*

In this section we will focus on deriving an expression for the Laplace-Stieltjes transforms (*LST*s) of the waiting time distributions of an admitted customer in the queue as well as in the system without any regard to whether the customer received a service or not. Observe that the *LST* s do not depend on the future arrivals and thus there is no need to track the phase of the arrival process.

Let $\boldsymbol{y}$, partitioned into vectors of dimension $Sn$ as $\boldsymbol{y} = (\boldsymbol{y}(0), \boldsymbol{y}(1), \cdots)$, denote the steady-state probability vector of the system at an arrival epoch. That is, the vector $\mathbf{y}(i)$ which is further partitioned as $\boldsymbol{y}(i) = (\boldsymbol{y}_1(i), \cdots, \boldsymbol{y}_S(i))$ is such that soon after an arrival epoch the steady-state probability of finding $i$ customers in the system with cash level being $j$, and the service is in phase $k$ is given by the $k^{\text{th}}$ component of the vector, $\boldsymbol{y}_j(i)$, $i \geq 1$, $1 \leq j \leq S$. It is easy to verify that

$$\boldsymbol{y}(i) = \begin{cases} c\,\boldsymbol{x}(0)(I_S \otimes D_1\boldsymbol{e}\boldsymbol{\beta}), \ i = 1, \\ c\,\boldsymbol{x}(i-1)(I_{Sn} \otimes D_1\boldsymbol{e}), \ i \geq 2, \end{cases} \tag{37}$$

where

$$c = \frac{1}{\lambda(1 - P_{Loss})}, \tag{38}$$

and $P_{Loss}$ is as given in (35).

Let $W$ denote the waiting time in the system of an admitted customer such that this tagged customer can leave the system with or without a service. The waiting time, $W$, can be viewed as the time until absorption in a *CTMC* with an absorbing state. Towards this end we define the *CTMC* with state space given by

$$\widetilde{\Omega} = \{*\} \cup \{\underline{\boldsymbol{i}}, i \geq 1\},$$

where $\{*\}$ denotes the absorbing state and the set of states $\underline{\boldsymbol{i}}$ contains the states: $\underline{\boldsymbol{i}} = \{(i, j, k) : 1 \leq j \leq S, 1 \leq k \leq n\}$, which correspond to the case when there are $i$ customers in the system (including the one who just arrived), the cash level is $j$, and the phase of the service process is in $k$. The generator of the *CTMC* with the above state space $\widetilde{\Omega}$ is given by

$$\tilde{Q} = \begin{bmatrix} 0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \ldots \\ \theta\boldsymbol{e} + (\boldsymbol{e} \otimes \boldsymbol{T}^0) & \hat{A}_1 & 0 & 0 & 0 & \ldots \\ (\theta\boldsymbol{e} + \hat{E}\boldsymbol{e}) & \hat{A}_2 & \hat{A}_1 & 0 & 0 & \ldots \\ (\theta\boldsymbol{e} + \hat{E}\boldsymbol{e}) & 0 & \hat{A}_2 & \hat{A}_1 & 0 & \ldots \\ (\theta\boldsymbol{e} + \hat{E}\boldsymbol{e}) & 0 & 0 & \hat{A}_2 & \hat{A}_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}, \tag{39}$$

where

$$\hat{A}_1 = I_S \otimes (T - \theta I), \hat{A}_2 = \hat{F} \otimes \boldsymbol{T}^0\boldsymbol{\beta}, \hat{E} = \sum_{i=1}^{N} \tilde{p}_i\, \boldsymbol{e}_i(S) \otimes \boldsymbol{T}^0, \tag{40}$$

and $\hat{F}$ is as given in (4).
The following theorem gives an expression for the *LST*, $w^*(s)$, of the waiting time, $W$.

**Theorem 5.** *The LST, $w^*(s)$, of the waiting time in the system of an admitted customer is given by*

$$w^*(s) = \sum_{i=1}^{\infty} \boldsymbol{y}(i)\boldsymbol{a}_i\,(s), \ Re(s) \geq 0, \tag{41}$$

*where $\boldsymbol{a}_i(s)$ is given by*

$$\boldsymbol{a}_i(s) = \begin{cases} \left(sI - \hat{A}_1\right)^{-1}(\theta\boldsymbol{e} + \boldsymbol{e} \otimes \boldsymbol{T}^0), \ i = 1 \\ (sI - \hat{A}_1)^{-1}[\theta\boldsymbol{e} + \boldsymbol{e} \otimes E\boldsymbol{e}] + \left(sI - \hat{A}_1\right)^{-1}\hat{A}_2\boldsymbol{a}_{i-1}(s), \ i \geq 2\,. \end{cases} \tag{42}$$

**Proof.** First note that the vector, $\boldsymbol{y}(i)$, as given in (37) gives the steady-state probability vector of being in level $\boldsymbol{i}$, for $i \geq 1$. Secondly, the random variable $W$ is nothing but the time until absorption with one absorbing state of the *CTMC* whose generator is as given in (39). The stated results now follows from the law of total probability.

**Corollary.** *The mean waiting time, $\mu_W$, in the system of a tagged admitted customer before leaving the system either with or without a service is given by*

$$\mu_W = \sum_{i=1}^{\infty} \boldsymbol{y}(i)\widehat{\boldsymbol{a}}_i, \tag{43}$$

where

$$\widehat{\boldsymbol{a}}_i = \begin{cases} (\boldsymbol{e} \otimes (\theta I - T)^{-1}\boldsymbol{e}), \ i = 1, \\ (\boldsymbol{e} \otimes (\theta I - T)^{-1}\boldsymbol{e}) + \left(\hat{F} \otimes (\theta I - T)^{-1}\boldsymbol{T}^0\boldsymbol{\beta}\right)\widehat{\boldsymbol{a}}_{i-1}, \ i \geq 2\,. \end{cases} \tag{44}$$

**Proof:** By definition, we have $\mu_W = -\left.\frac{dw^*(s)}{ds}\right]_{s=0}$. Denoting by $\hat{a}_i = -\left.\frac{da_i(s)}{ds}\right]_{s=0}$, $i \geq 1$, it is easy to verify the following.

$$(-\hat{A}_1)^{-1}[\theta\boldsymbol{e} + (\boldsymbol{e} \otimes \boldsymbol{T}^0)] = \boldsymbol{e}$$

$$\hat{F} \otimes \boldsymbol{T}^0 = (\boldsymbol{e} \otimes \boldsymbol{T}^0) - (\boldsymbol{e} \otimes \hat{\boldsymbol{E}})$$

$$\widehat{\boldsymbol{a}}_1 = (-\hat{A}_1)^{-2}[\theta\boldsymbol{e} + (\boldsymbol{e} \otimes \boldsymbol{T}^0)] = (\boldsymbol{e} \otimes (\theta I - T)^{-1}\boldsymbol{e}), \tag{45}$$

$$\widehat{\boldsymbol{a}}_i = (-\hat{A}_1)^{-2}\left[\theta\boldsymbol{e} + \left(\boldsymbol{e} \otimes \hat{E}\right)\right] + (-\hat{A}_1)^{-2}\hat{A}_2\boldsymbol{e} + (-\hat{A}_1)^{-2}\hat{A}_2\widehat{\boldsymbol{a}}_{i-1}$$

$$= [\boldsymbol{e} \otimes (\theta I - T)^{-1}\boldsymbol{e}] + (\hat{F} \otimes (\theta I - T)^{-1}\boldsymbol{T}^0\boldsymbol{\beta})\widehat{\boldsymbol{a}}_{i-1}, i \geq 2.$$

The stated result follows immediately from Theorem 5 and (45).

**Note:** While there is no closed form expression for $\mu_W$, one can efficiently compute the mean as follows.

**Step 0:** Set $i = 1$; $\widehat{\boldsymbol{a}}_1 = (\boldsymbol{e} \otimes (\theta I - T)^{-1}\boldsymbol{e})$; $\boldsymbol{\xi} = \boldsymbol{y}(1)\widehat{\boldsymbol{a}}_1$; $\boldsymbol{b} = \widehat{\boldsymbol{a}}_1$.
**Step 1:** $i \leftarrow i + 1$; $\boldsymbol{b} \leftarrow \widehat{\boldsymbol{a}}_1 + (\hat{F} \otimes (\theta I - T)^{-1}\boldsymbol{T}^0\boldsymbol{\beta})\boldsymbol{b}$; $\boldsymbol{\xi} \leftarrow \boldsymbol{\xi} + \boldsymbol{y}(i)\boldsymbol{b}$.
**Step 2:** If $\boldsymbol{y}(i)\boldsymbol{e} > \varepsilon$, where $\varepsilon$ is a very small pre-specified number such as $10^{-6}$, go to Step 1; otherwise, $\mu_W = \xi$.

The following theorem gives an expression for the *LST*, $w_q^*(s)$, of the waiting time in the queue of an admitted customer before getting into service or leaving the system (due to shortage in cash level or catastrophic event). The proof is very similar to Theorem 5 and will be omitted.

**Theorem 6.** T*he LST, $w_q^*(s)$, of the waiting time in the queue of an admitted customer is given by*

$$w_q{}^*(s) = \sum_{i=1}^{\infty} \boldsymbol{y}(i+1)\boldsymbol{a}_i(s), \ Re(s) \geq 0, \tag{46}$$

*where $\boldsymbol{y}(i)$ and $\boldsymbol{a}_i(s)$ are, respectively, as given in* (37) *and* (42).

The mean waiting time in the queue of a tagged admitted customer is given in the corollary whose proof is similar to the corollary dealing with $\mu_W$.

**Corollary.** *The mean waiting time, $\mu_{W_q}$, in the queue of a tagged admitted customer is given by*

$$\mu_{W_q} = \sum_{i=2}^{\infty} \boldsymbol{y}(i+1)\,\widehat{\boldsymbol{a}}_i, \tag{47}$$

*where $\widehat{\boldsymbol{a}}_i$ is given by* (45).

**Note:** One can use Little's result to get the mean waiting times from the mean numbers in the system as follows:

$$\mu_{NS} = \lambda(1 - P_{Loss})\mu_W \ \text{ and } \mu_{NQ} = \lambda(1 - P_{Loss})\mu_{W_q},$$

where $P_{Loss}$ is as given in (35) and the mean number ($\mu_{NS}$) of customers in the system and the mean number ($\mu_{NQ}$) of customers in the queue are obtained as

$$\mu_{NS} = \sum_{i=1}^{\infty} i\,\boldsymbol{x}(i)e = \boldsymbol{x}(1)\,(I - R)^{-2}\boldsymbol{e} \ \text{ and } \mu_{NQ} = \sum_{i=1}^{\infty}(i-1)\boldsymbol{x}(i)\boldsymbol{e} = \boldsymbol{x}(1)R\,(I - R)^{-2}\boldsymbol{e}.$$

### *3.3. The system performance measures*

In this section we will list a number of system performance measures of interest along with their expressions. These are in addition to the ones mentioned earlier.

- The probability, $P_{DNCZ}$ , that the system is down due to cash level being zero is given by

$$P_{DNCZ} = \boldsymbol{x}^* \boldsymbol{e}.$$

- The probability, $P_{DNCE}$ , that the system is down due to catastrophic events is given by

$$P_{DNCE} = \widehat{\boldsymbol{x}}^* \boldsymbol{e}.$$

- The probability, $P_{Down}$ , that the system is down is given by

$$P_{Down} = \boldsymbol{x}^* \boldsymbol{e} + \widehat{\boldsymbol{x}}^* \boldsymbol{e}.$$

- The average, $\mu_{CLSD}$ , cash level when the system is down is given by

$$\mu_{CLSD} = \frac{1}{\boldsymbol{x}^* \boldsymbol{e} + \widehat{\boldsymbol{x}}^* \boldsymbol{e}} \sum_{i=1}^{S} j \widehat{\boldsymbol{x}}_j^* \boldsymbol{e}.$$

- The probability, $P_{Idle}$ , that the system is idle is given by

$$P_{Idle} = \boldsymbol{x}(0)\boldsymbol{e}.$$

- The mean, $\mu_{Cash}$ , level of cash in the system is given by

$$\mu_{Cash} = \sum_{j=1}^{S} j \left[\widehat{\boldsymbol{x}}_j^* e + \boldsymbol{x}_j(0)\boldsymbol{e} + \sum_{i=1}^{\infty} \boldsymbol{x}_j(i)\boldsymbol{e}\right].$$

- The rate, $R_{ADCL}$ , at which admitted customers are lost is given by

$$R_{ADCL} = \theta \sum_{i=1}^{\infty} i\, \boldsymbol{x}(i)\boldsymbol{e} + \sum_{i=1}^{\infty} (i-1)\boldsymbol{x}(i)E\boldsymbol{e}.$$

- The rate, $R_{ADLS}$ , at which admitted customers leave the system by getting a service is given by

$$R_{ADLS} = \sum_{i=1}^{\infty} \boldsymbol{x}(i)(\boldsymbol{e} \otimes \boldsymbol{T}^0 \otimes \boldsymbol{e}).$$

- The probability, $P_{Loss}$ , that an arriving customer is lost due to the system being in down state is

$$P_{Loss} = \frac{1}{\lambda}[\boldsymbol{x}^* D_1 \boldsymbol{e} + \widehat{\boldsymbol{x}}^* (\boldsymbol{e} \otimes D_1 \mathbf{e})].$$

## 4. Special Cases

In this section we will look at a few special cases of the model under study and in some cases (e.g., when $S = 1$) derive explicit expressions for the steady-state probability vector. While special cases involving $S = 1$ by themselves may not be of any practical value, they do play an important role in the accuracy check of numerical implementation. We also present special cases that reduce to some well-known classical queueing models.

### *4.1. Case* 1*: M/M/*1*-model with N* = 1*, S* > 1

In this case, we assume that the arrivals occur according to a Poisson process and the service times are exponential. It is easy to verify that

$$A_0 = \lambda I, A_1 = -(\lambda + \theta + \mu)I, A_2 = \tilde{A}_2 = \mu \widehat{F}, E = \mu \boldsymbol{e}_1,$$

where

$$\widehat{F} = \begin{bmatrix} \mathbf{0} & 0 \\ I_{S-1} & \mathbf{0} \end{bmatrix}.$$

With the simplified expressions for the input data matrices as given above, the matrix $R$ (see (20)) is such the (block) entries, namely, $R_i$ , $1 \le i \le K$ , are scalars and are obtained recursively as follows.

$$R_1 = \frac{\lambda}{\lambda + \theta + \mu}, \quad R_1 = \frac{\mu \sum_{k=1}^{i-1} R_k R_{i-k}}{\lambda + \theta + \mu}, \quad 2 \le i \le k.$$

The steady-state probability vector $\boldsymbol{x}$ (see Theorem 1) can explicitly be obtained as shown below.

**Theorem 7.** *In the case of M/M/1−type model for ATM system, the scalar, x*, and the vectors,* $\widehat{\boldsymbol{x}}^*$*, and* $\boldsymbol{x}(i)$*, i ≥ 0, are given by*

$$\widehat{\boldsymbol{x}}^* = \frac{\theta}{\delta_2}\boldsymbol{x}(0)[I + \lambda\widetilde{R}(I-R)^{-1}], \tag{48}$$

$$\boldsymbol{x}(0) = \delta_1 x^* \boldsymbol{e}'[\lambda I - \lambda\theta\widetilde{R}(I-R)^{-1} - \lambda\mu\widetilde{R}R\widehat{F}]^{-1}, \tag{49}$$

$$\boldsymbol{x}(i) = \lambda\boldsymbol{x}(0)\widetilde{R}R^{i-1}, \tag{50}$$

where $x^*$ is the normalizing constant and $\tilde{R} = [(\lambda+\theta+\mu)I - \mu R\widehat{F}]^{-1}$.

**Proof.** Follows immediately by substituting the simplified expressions for the input data matrices in (13) through (18).

### *4.2. Case* 2*: M/M/*1*-model with* $N = 1, S = 1$

In the case of Poisson arrivals and exponential services along with $N = S = 1$, it is easy to verify that

$$x^* = \frac{r\mu}{\delta_1(1-r)}x(0), \hat{x}^* = \frac{\theta}{\delta_2(1-r)}x(0), x(i) = x(0)r^i, i \geq 0,$$

$$r = \frac{\lambda}{\lambda+\theta+\mu}, x(0) = \frac{(\theta+\mu)\delta_1\delta_2}{\lambda\mu\delta_2 + \delta_1(\theta+\delta_2)(\lambda+\theta+\mu)},$$

$$\mu_W = \frac{1}{\mu+\theta}, \quad \mu_{W_q} = \frac{x(0)r}{(1-P_{Loss})(1-r)(\mu+\theta)},$$

where $P_{Loss}$ is as given in (35).

### *4.3. Case* 3*: M/PH/*1*-model with* $N = 1, S = 1$

In this case, we assume that the arrivals occur according to a Poisson process and the service times are of phase type. Noting that for this case,

$$A_0 = \lambda I, A_1 = T - (\lambda+\theta)I, A_2 = \tilde{A}_2 = 0, E = \boldsymbol{T}^0,$$

verify that $R$ has an explicit expression given by $R = \lambda[(\lambda+\theta)I - T]^{-1}$ and the steady-state probability vector, $\boldsymbol{x}$, is given explicitly as follows.

$$x^* = \frac{x(0)}{\delta_1}\boldsymbol{\beta}R(I-R)^{-1}\boldsymbol{T}^0, \hat{x}^* = \frac{\theta x(0)}{\delta_2}[I + \boldsymbol{\beta}R(I-R)^{-1}\boldsymbol{e}], \boldsymbol{x}(i) = x(0)\boldsymbol{\beta}R^i, i \geq 1,$$

$$\mu_W = \frac{x(0)}{1-P_{Loss}}\boldsymbol{\beta}(I-R)^{-1}(\theta I - T)^{-1}\boldsymbol{e}, \mu_{W_q} = \frac{x(0)}{1-P_{Loss}}\boldsymbol{\beta}R\,(I-R)^{-1}(\theta I - T)^{-1}\boldsymbol{e},$$

where $P_{Loss}$ is as given in (35) and $x(0)$ is the normalizing constant.

### *4.4. Case* 4*: MAP/PH/*1*-model with* $\theta = 0$

In this case, we assume that θ = 0, which implies that there are no catastrophic events. We assume that $S < \infty$ so that the queueing model in this special is always stable. The only way the system can be in down state is due to zero cash. Thus, in this special case of our model, we study *MAP*/*PH*/1-type queueing-inventory model in which customers (or demands) including those waiting in the queue are lost whenever the cash (or stocks) level becomes zero, and the customers who are admitted (due to availability of inventory) may be served with one or more inventory based on a probabilistic rule. Note that in this case $\delta_2$ doesn't play a role. One can easily verify that the steady-state equations given in (7) through (11) along with the normalizing equation (12) hold good here by taking $\widehat{\boldsymbol{x}}^*$**,** and θ to be zero. However, there are no other simplifications or explicit expressions available for this special case.

### *4.5. Case 5: MAP /PH/1-model with S = ∞*

In this case, we assume that there is no limit on S and hence the system can be down only through catastrophic events. Thus, the model reduces to *MAP /PH*/1 catastrophic models with recovery time that is exponentially distributed with parameter $\delta_2$. Unless $\theta = 0$, the queue is always stable. This becomes a special case of the model considered in [8]. However, for the sake of completeness, we will briefly summarize the results below with the notations set forth in this paper. First note that due to unlimited value for *S*, all admitted customers leaving the system with a service will see no shortage of cash. Thus, there is no need to keep track of the cash level. Secondly, the parameters *N* and $\delta_1$ play no role. Finally, the state space of Section 3 will be reduced significantly for the current case as shown below.

$$\widehat{\Omega} = \{\underline{*}\} \cup \{\underline{\boldsymbol{i}}, i \geq 0\},$$

where the set of states and their definitions are as follows:

- The set of states, $\underline{*} = \{k, 1 \leq k \leq m\}$, of dimension m corresponds to the system being down due to catastrophic event, and the arrival process is in one of *m* phases.
- While the set of states, $\underline{\mathbf{0}} = \{(0, k), 1 \leq k \leq m\}$, of dimension m corresponds to the system being in idle state with the arrival process in one of *m* phases, the set of states, $\underline{\boldsymbol{i}} = \{(i, r, k), 1 \leq r \leq n, 1 \leq k \leq m\}$, of dimension *mn* corresponds to the system being busy with the arrival process in one of *m* phases, the service process in one of *n* phases.

Denoting $\hat{Q}$ to be the generator of the CTMC governing the system

$$\hat{Q} = \begin{bmatrix} D-\delta_2 I & \delta_2 I & & & & & \\ \theta I & D_0 - \theta I & \boldsymbol{\beta} \otimes D_1 & & & & \\ \theta(\boldsymbol{e} \otimes I) & \boldsymbol{T^0} \otimes I & (T \oplus D_0) - \theta I & I \otimes D_1 & & & \\ \theta(\boldsymbol{e} \otimes I) & & \boldsymbol{T^0\beta} \otimes I & (T \oplus D_0) - \theta I & I \otimes D_1 & & \\ \theta(\boldsymbol{e} \otimes I) & & & \boldsymbol{T^0\beta} \otimes I & (T \oplus D_0) - \theta I & I \otimes D_1 & \\ \vdots & & & \ddots & \ddots & \ddots & \end{bmatrix}. \quad (51)$$

Suppose that $\boldsymbol{z}$, partitioned as $\boldsymbol{z} = (\boldsymbol{z}^*, \boldsymbol{z}_0, \boldsymbol{z}_1, \ldots)$, is the steady-state probability vector of $\hat{Q}$. That is, $\boldsymbol{z}$ satisfies $\boldsymbol{z}\hat{Q} = \mathbf{0}$, $\boldsymbol{ze} = 1$. Then we have the following theorem.

**Theorem 8.** *In the case of MAP/PH/1−type model for ATM system with unlimited value for S, the steady-state vector* $\boldsymbol{z}$ *is obtained as follows.*

$$\begin{aligned} \boldsymbol{z}^* &= \theta[\boldsymbol{z}_0 + \boldsymbol{z}_1(I - \hat{R})^{-1}(\boldsymbol{e} \otimes I](\delta_2 I - D)^{-1}, \\ \boldsymbol{z}_0 &= [\delta_2 \boldsymbol{z}^* + \boldsymbol{z}_1(\boldsymbol{T^0} \otimes I)](\theta I - D_0)^{-1}, \\ \boldsymbol{z}_1 &= \boldsymbol{z}_0(\boldsymbol{\beta} \otimes D_1)\left[-(A_1 + \hat{R}A_2)^{-1}\right], \\ \boldsymbol{z}_i &= \boldsymbol{z}_1 \hat{R}^{i-1}, i \geq 1, \end{aligned} \quad (52)$$

*where* $\hat{R}$ *satisfies*

$$\hat{R}^2(\boldsymbol{T^0\beta} \otimes I) + \hat{R}[(T \oplus D_0) - \theta I] + I \otimes D_1 = 0, \quad (53)$$

*and the normalizing condition is given by*

$$\boldsymbol{z}^*\boldsymbol{e} + \boldsymbol{z}_0\boldsymbol{e} + \boldsymbol{z}_1(I - \hat{R})^{-1}\boldsymbol{e} = 1, \quad (54)$$

*and the probability,* $P_{Down}$*, that the system is down at an arbitrary time is given by*

$$P_{Down} = \boldsymbol{z}^*\boldsymbol{e} = \frac{\theta}{\theta + \delta_2}. \quad (55)$$

**Proof.** First note that (52) and (53) follow immediately due to the generator given in (51) is of *GI/M*/1−type and hence one can apply matrix-analytic methods (see e.g., [20]). To get the result given in (55), we rewrite the first equation in (52) as

$$\boldsymbol{z}^*(\delta_2 I - D) = \theta\left[\boldsymbol{z}_0 + \boldsymbol{z}_1\left(I - \hat{R}\right)^{-1}(\boldsymbol{e} \otimes I)\right]. \quad (56)$$

Now post-multiplying (56) by $\boldsymbol{e}$ and using the normalizing condition given in (54) we get the stated result in (55).

### *4.6. Case 6: MAP /PH/1-model with θ = 0 and S = ∞*

In this final case, when we take θ = 0 and $S = \infty$, the model reduces to the classical *MAP/PH/*1 queue. Note that in this case the queueing model is stable if and only if λ < μ.

## 5. Illustrative Examples

In this section we discuss the qualitative aspects of the queueing model useful in service industries under study through illustrative numerical examples. In order to verify the correctness and the accuracy of the FORTRAN code written for the qualitative study of the model, we used (a) the results of Theorem 3, Theorem 4, the Little's result; (b) the explicit results available for some special cases; and (c) results obtained for the Poisson arrivals in its simple form and in the form involving eigenvalue and eigenvector [23] which the general algorithm doesn't distinguish but the numerical results are identical.

For our illustrative examples we consider five arrival processes and three service time distributions. These five *MAP*s and three PH-representations are as follows.

1. *ErA:* This *MAP* corresponds to Erlang of order 2 with parameter 2λ for inter-arrival times.
2. *ExA:* This *MAP* corresponds to exponential distribution with parameter λ for inter-arrival times.
3. *HeA:* This *MAP* corresponds to hyperexponential distribution with mixing probabilities 0.9 and 0.1, and their rates are, respectively, 1.9λ and 0.19λ.
4. *MnA:* This corresponds to a *MAP* with negative correlation between two successive inter-arrival times and the representation matrices are given by

$$D_0 = \lambda\begin{pmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{pmatrix}, D_1 = \lambda\begin{pmatrix} 0 & 0 & 0 \\ 0.01002 & 0 & 0.9922 \\ 223.4925 & 0 & 2.2575 \end{pmatrix}$$

5. *MpA*: This corresponds to a *MAP* with positive correlation between two successive inter-arrival times and the representation matrices are given by

$$D_0 = \lambda\begin{pmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{pmatrix}, D_1 = \lambda\begin{pmatrix} 0 & 0 & 0 \\ 0.9922 & 0 & 0.01002 \\ 2.2575 & 0 & 223.4925 \end{pmatrix}$$

Note that the ratio of the standard deviation of the inter-arrival times of these five arrival processes with respect to *ErA* are, respectively, 1, 1.41421, 3.17451, 1.99336, and 1.99336. Also, it can be verified that *MnA* has a negative correlation of -0.4889 and *MpA* has a positive correlation of 0.4889. Thus, the above *MAP* processes are all qualitatively different. These will be normalized so as to have a specific value for λ.

For the service times we consider the following three PH −distributions. These distributions will be normalized so as to arrive at a desired value for μ.

A. *ErS:* This is an Erlang of order 2 with parameter given by 2μ.
B. *ExS:* This is an exponential distribution with rate μ.
C. *HeS:* We take this to be an hyperexponential distribution with mixing probabilities 0.9 and 0.1, and their rates are, respectively, 1.9μ and 0.19μ.

It can easily be verified that these three PH−distributions are qualitative different in that the ratio of the standard deviation of *ExS* and *HeS* to *ErS* are, respectively, 1.41421 and 3.17451. By looking at these we ought to be able consider various scenarios for the service times.

**Example 1:** In this example we fix $N = 4$, μ = 1.0, $p_1 = p_2 = p_3 = p_4 = 0.25$, $S = KN$, θ = 0.1, $\delta_1 = 1.0$, $\delta_2 = 2.0$, and consider three values for the arrival rate: λ = 0.1, 0.8, 0.95 and vary $K$ from 1 to 50. That is,

= 2.0, and consider three values for the arrival rate: λ = 0.1, 0.8, 0.95 and vary *K* from 1 to 50. That is, we vary *S* from 4 to 200 in increments of 4.

In Figure 1 we display the plots of selected system performance measures under different scenarios. Since some of the plots are similar in shapes, for example, the ones dealing with *ExS* are similar to *ErS* or the ones with λ = 0.80 and λ = 0.95, and also due to space limit, we display only those as representative ones. To point out the significance of the variation in the service times, we also display the ratios (*HeS* to *ErS*) of the values of the measures in Figure 1 (the right most set of plots) for selected measures. The observations summarized for the first four measures (see below) are based on the figures displayed. For the other selected measures the figures are not displayed due to space consideration.

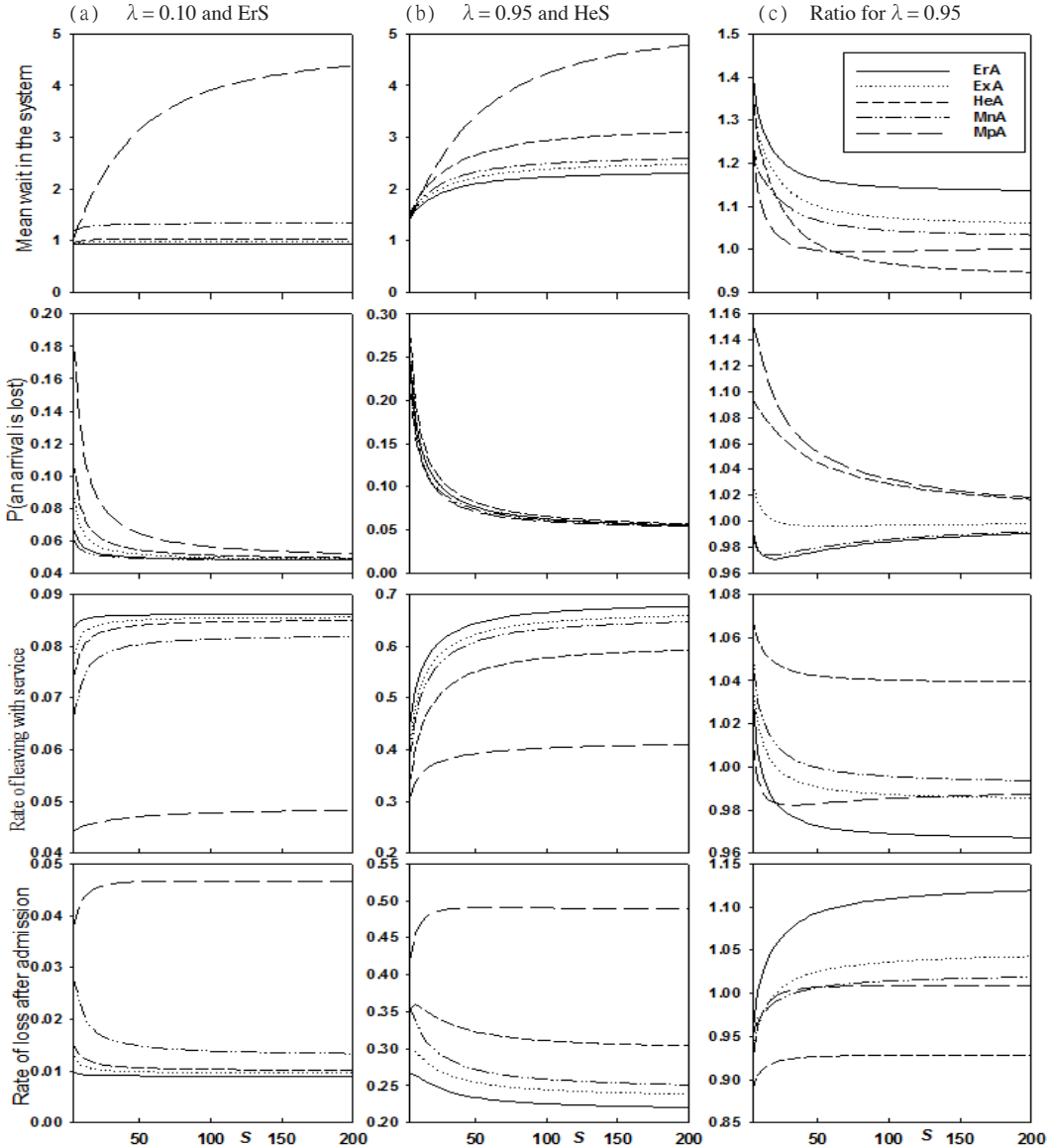


Figure 1a. Plots of selected measures under different scenarios for Erlang services.
Figure 1b. Plots of selected measures under different scenarios for Hyperexponential services.
Figure 1c. Plots of selected ratios under different scenarios.

*1. The mean waiting time in the system* ($\mu_W$)*:*

- This measure appears to be sensitive to the type of service times as well as the type of arrival process. The sensitivity is more pronounced as λ becomes larger. Further, the level of sensitivity with respect to the arrival process as well as service times goes down as S is increased. This appears to be the case for all values λ.
- When compared to all other arrival processes, this measure appears to be significantly higher for *MpA* indicating the key role played by correlation, especially positive one.
- Unlike in the classical *MAP /PH*/1 queue this measure does not appear to grow large as λ is increased. This is the case even for *MpA* arrivals. This can be explained intuitively as follows. When λ is increased the system will be in down state (and hence all customers will be removed from the system) due to cash shortage, and hence the mean waiting time in the system is not allowed to grow beyond a certain point. We verified this to be true even for λ = 0.999.

*2. The probability that an arriving customer is lost* ($P_{Loss}$)*:*

- This measure giving the probability that the system is down at an arrival epoch appears to be sensitive to the type of arrival and service processes; however, the degree of sensitivity goes down as *S* is increased.
- We notice an interesting observation with regard to this measure. While for *ErA* and *MnA* arrivals this measure appears to be smaller when services are *HeS* as compared to *ErS,* we notice that for the other three arrival processes, the reverse appears to be true. That is, for all other three arrivals, *HeS* appears to yield a larger value compared to *ErS*.
- For all scenarios, we see that this measure appears to approach a limiting value as *S* increases (when all other parameters are fixed). This can be explained intuitively as follows. First note that an arriving is customer is lost if and only if the system is found to be in down state at that instant. When *S* is large, the system will be in down state mainly due to catastrophic events and the limiting value corresponds to that of the *MAP /PH*/1 queueing model with catastrophic events with an exponential recovering time for the system. Note from Theorem 8 that this limiting value is given by $\frac{\theta}{\theta + \delta_2}$.

*3. The rate of admitted customers leaving with a service* ($R_{ADLS}$)*:*

- First observe that this measure is highest for *ErA* and lowest for *M pA* arrivals. As *S* is increased, this measure appears to increase at a steady rate for all *MAP* arrivals.
- The sensitivity of this measure with regard to the arrivals and the service times can be seen for all scenarios.

*4. The rate of admitted customers lost* ($R_{ADCL}$)*:*

- First note that this measure is highest for *MpA* arrivals and is lowest for *ErA* arrivals (across all values of *S*) as compared to the other arrival processes. Also, in *MpA* case this measure appears to show a non-decreasing trend as *S* is increased; however for other *MAP*s, this we see a non-increasing trend in *S*. This trend is seen for small as well as for large values of λ.
- It is interesting to observe (see the right most plots in Figure 1) that for *ErA*, this measure appears to be higher for *HeS* as compared to *ErS* for most values of S; however, for *HeA* arrivals it is exactly the opposite of this behavior.

*5. The probability that the system is down due to cash shortage* ($P_{DNCZ}$):

- This measure approaches zero as *S* is increased but at a slower rate.
- The sensitivity of this measure to the type of arrival process is more apparent for large values of λ.
- In all cases, the *MpA* arrivals appear to have the smallest value for this measure as compared to the other arrival processes, while *ErA* has the highest value.

*6. The probability that the system is down due to catastrophic event* ($P_{DNCE}$):

- While this measure appears to be almost the same across all five *MAP* s for very small and for large values of *S*, we notice some small deviations among various *MAP* s for medium values of *S*. This can be intuitively explained as follows. For very small values of *S*, the system will be in down state mostly due to cash shortage and hence $P_{DNCE}$ will be very small. However, for large

values of $S$, the system will be down mostly due to catastrophic events, which are independent of the arrival process. For moderate values of $S$, the role of variation and correlation (if any) in the inter-arrival times of the five *MAP*s determines whether the system will be in down state through cash shortage or catastrophic events.

*7. The probability that the system is down* ($P_{Down}$):

- As is to be expected, this measure appears to decrease as $S$ is increased for all scenarios. This measure is more sensitive to the service times in the case of *HeA* arrivals even for small values of $S$ and for $\lambda$ reasonably large. However, for large S and for other scenarios, the measure appears to be insensitive to the type of services.
- For all scenarios, we see that this measure appears to approach a limiting value as $S$ increases (when all other parameters are fixed). The intuitive explanation for this is the same as the one given for the measure $P_{Loss}$ since an arriving is customer is lost if and only if the system is found to be down at an arrival instant.
- It is interesting to observe that the degree of sensitivity to the variation in the service times is more for *MnA* arrivals as compared to that of *MpA*, especially when $\lambda$ is large.

*8. The mean cash level* ($\mu_{Cash}$):

- This measure, for all scenarios, appears to increase linearly as $S$ is increased. This is the case for low as well as high traffic intensities. This might seem to be counter-intuitive since one would expect the cash level to decrease as $\lambda$ is increased due to an increase in the number of customers arriving to the system. A possible intuitive explanation for this phenomenon (i.e., linearity) is for large $S$, the system will be down mostly through catastrophic events which will remove only the customers present in the system and hence the cash will not deplete much; on the other hand, for small $S$ more replenishments will occur due to the system in down state mostly through cash shortage.

The measures, the rate of customers leaving the system with a service and the mean waiting in the system, are important among others for the management. In the next example, we will specifically focus on these two measures as functions of $\theta$ and $\delta_2$ by fixing all other parameters.

**Example 2:** Here we fix $N = 3$, $\lambda = 0.8$, $\mu = 1.0$, $p_1 = p_2 = p_3 = 1/3$, $S = 225$, $\delta_1 = 1.0$, and vary $\theta$ from 0.01 to 0.1, and vary $\delta_2$ from 1.0 to 2.0.

In Figures 2 and 3, respectively, we display the two measures, $\mu_W$ and $R_{ADLS}$. Looking at these two figures, we register the following key observations.

- While we see a significant change in $\mu_W$ when going from *ErS* to *HeS* for both *ErA* and *HeA* arrivals for the ranges of ($\theta$, $\delta_2$) considered, we do not see such a significant difference in the case of *MpA* arrivals. The rate of change, as a function of $\theta$, in this measure is significant, whereas the rate of change appears to be insignificant when $\delta_2$ is varied.
- Similar to the case of $\mu_W$, we notice a significant change in $R_{ADLS}$ when going from *ErS* to *HeS* for *ErA* arrivals for the ranges of ($\theta$, $\delta_2$) considered. While for *HeA* arrivals, we see somewhat moderate change, there is not a significant change for the case of *MpA* arrivals. Again, the rate of change, as a function of $\theta$, in this measure is significant, whereas the rate of change appears to be insignificant when $\delta_2$ is varied.

Based on the above two illustrative examples and the facts that the management (i) may not be able to control the variation or the sources from where the customers arrive leading to possibly higher variation in the inter-arrival times as well as possible correlation between two successive arrivals; and (ii) should be able to control the services, we can recommend the following.

- Whenever the inter-arrival times of the customers have less variation and are independent of each other, the type of services offered play an important role. That is, a higher variation in the service times yield a lower value for $\mu_W$ and a higher value for $R_{ADLS}$. This is a desirable situation.
- Whenever the inter-arrival times of the customers have more variation and or have (positive) correlation, the type of services offered appears to play no role. Also, in such cases $\mu_W$ has a higher value and a lower value for $R_{ADLS}$. This is not desirable one but as mentioned earlier the

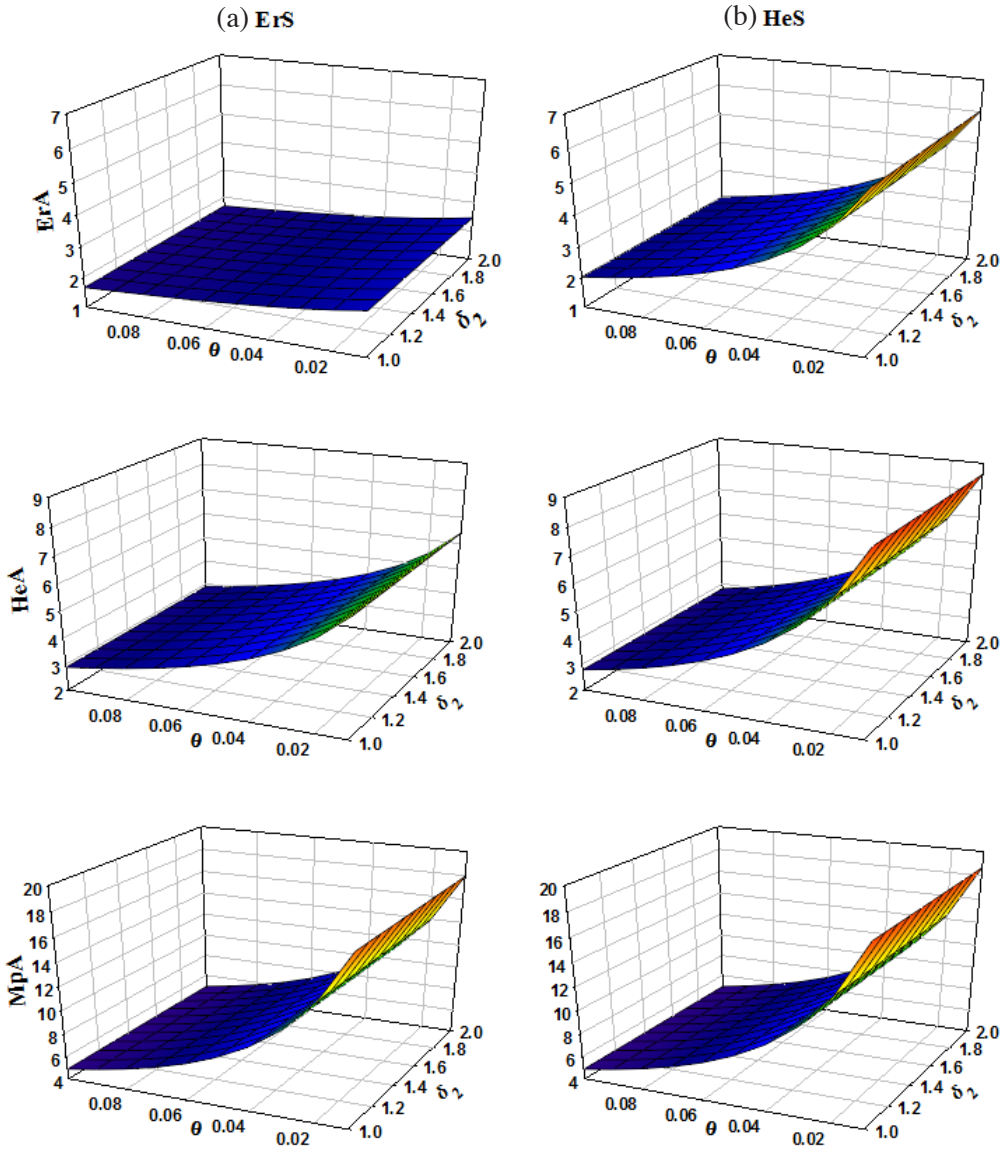


Figure 2a. Mean waiting time in the system as a function of $\theta$ and $\delta_2$ for Erlang services.
Figure 2b. Mean waiting time in the system as a function of $\theta$ and $\delta_2$ for Hyperexponential services.

management may not have much of a choice in controlling. However, by increasing the rate of fixing the system upon an external shock, the management may be able to reduce the $\mu_W$ as well as increase $R_{ADLS}$ but this will also increase the cost on a per unit basis.

## 6. Concluding Remarks and Future Research Work

In this paper we proposed a queueing model useful in automatic teller machine applications. Assuming *ATM* is subject to being down due to (a) failures from catastrophic events and (b) shortage in cash level), repairs and replenishment, and with *MAP* arrivals and phase type services, we analyzed the model with the help of matrix-analytic methods. A few illustrative examples were presented. It was pointed out that if the inter-arrival times of the customers have less variation and are independent of

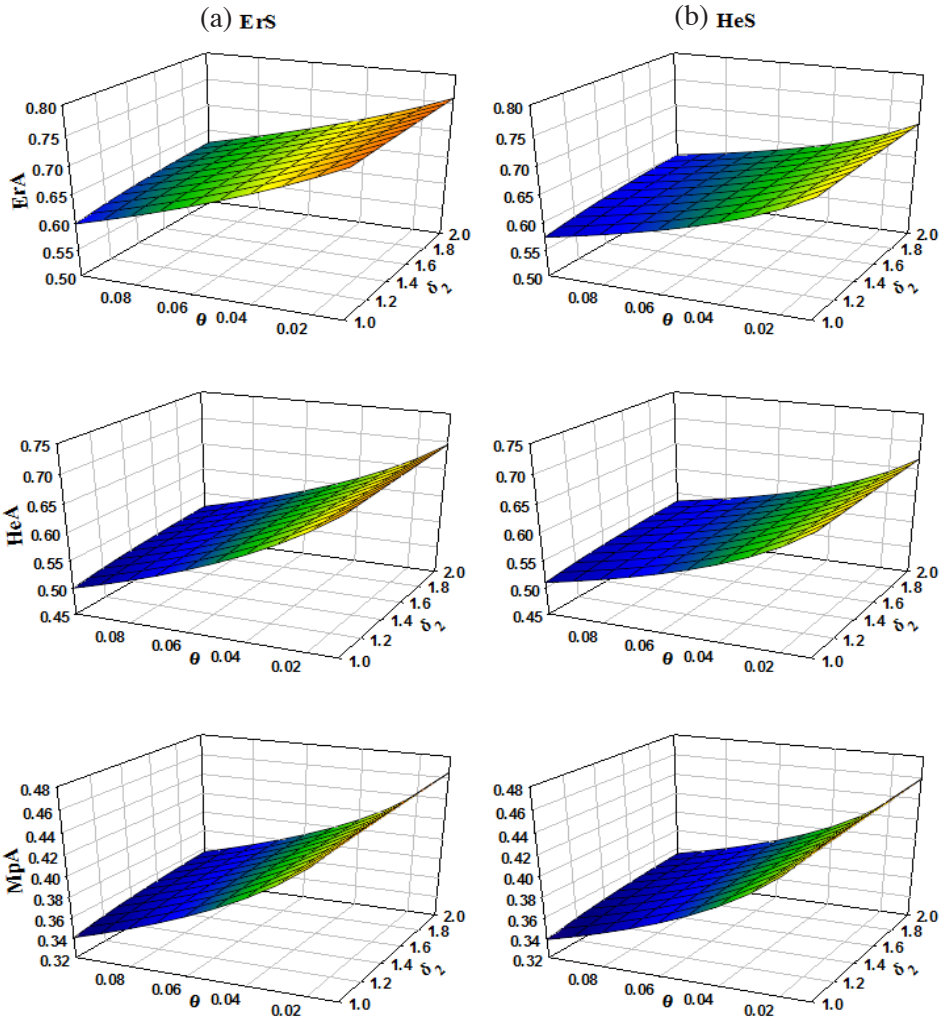


Figure 3a. Rate of customers leaving with service as a function of $\theta$ and $\delta_2$ for Erlang services.
Figure 3b. Rate of customers leaving with service as a function of $\theta$ and $\delta_2$ for Hyperexponential services.

each other, then having less variation in the service time will help to have a (a) higher value for the rate of customers leaving with service; and (b) lower value for the mean waiting time in the system. The methodology of the current paper can be employed to address a few variations of the current model as future work. First, we can relax the restriction of $s = 0$ in the $(s, S)$-type replenishment that we assumed for the cash. That is, we can send a request for cash replenishment when the cash level hits $s$ or below as opposed to waiting for the cash level to deplete to zero. Secondly, since customers using *ATM* opt for obtaining a receipt of their transactions before leaving the area, we can model this using a probability. Note that if we assume the time to print a receipt is exponential or of phase type, then it is very easy to modify the existing phase type distribution of order $n$ with a suitable of dimension higher than $n$. This is due to the fact that phase type distributions are closed under finite convolutions and mixtures. Thirdly, there are some locations where more than one *ATM* machine is present with a common queue. However, each machine has its own cash reserve but can have same source of catastrophic events (due to being in the same location). Thus, our model can be extended to include more than one *ATM* machine with its own cash reserve and a common source of catastrophic events, and all of these *ATM*s will have one common queue to house the incoming customers. All customers are lost if and only if all *ATM*s are down.

## References

[1] Artalejo, J. R., Gomez-Correl, A., & He, Q. M. (2010). Markovian arrivals in stochastic mdelling: a survey and some new results. *SORT*, 34(2), 101-144.

[2] Bedman, N. (2013). Service quality in automated teller machines: an empirical investigation. *Managing Service Quality: An International Journal*, 23(1), 62-89.

[3] Chakravarthy, S. R. (2001). The batch Markovian arrival process: A review and future work. *Advances in Probability Theory and Stochastic Processes*. Eds., A. Krishnamoorthy et al. Notable Publications Inc., NJ, 21-39.

[4] Chakravarthy, S. R. (2010). Markovian arrival processes. Wiley Encyclopedia of Operations Research and Management Science. Published Online: 15 JUN 2010.

[5] Chakravarthy, S. R., Maity, A., & Gupta, U. C. (2015). Modeling and Analysis of Bulk Service Queues with an Inventory under "(s, S)" Policy. *Annals of Operations Research*, Published Online: 31 October 2015; DOI 10.1007/s10479-015-2041-z.

[6] Chakravarthy, S. R. (2017). A catastrophic queueing model with delayed action. *Applied Mathematical Modeling*, 46, 631-649.

[7] Choi, K. H., & Yoon, B. K. (2016). A survey on the queueing inventory systems with phase- type service distributions. QTNA '16, December 13-15, 2016, Wellington, New Zealand.

[8] Dudin, A. N., & Karolik, A. V. (2001). BMAP /SM/1 queue with Markovian input of disasters and non-instantaneous recovery. *Performance Evaluation*, 45, 19-32.

[9] Graham, A. (1981). Kronecker Products and Matrix Calculus with Applications. Ellis Horwood, Chichester, UK.

[10] He, Qi-Ming. (2014). Fundamentals of Matrix-Analytic Methods. Springer, New York.

[11] Krishnamoorthy, A., Lakshmy, B., & Manikandan, R. (2011). A survey on inventory models with positive service time. *Opsearch*, 48(2), 153169.

[12] Krishnamoorthy, A., Manikandan, R., & Shajin, D. (2015). Analysis of a Multiserver Queueing-Inventory System. *Advances in Operations Research*, Volume 2015, Article ID 747328, 16 pages.

[13] Krishnamoorthy, A., Manikandan, R., & Lakshmy, B. (2016). A revisit to queueing-inventory system with positive service time. *Annals of Operations Research*, 233, 221-236.

[14] Krishnamoorthy, A., Shajin, D., & Lakshmy, B. (2016). GI /M/1 type queueing-inventory systems with postponed work, reservation, cancellation and common life time. *Indian Journal of Pure Applied Mathematics*, 47(2), 357-388.

[15] Latouche G., & Ramaswami V. (1999). Introduction to Matrix Analytic Methods in Stochastic Modeling. SIAM.

[16] Lucantoni D., Meier-Hellstern, K. S., & Neuts, M. F. (1990). A single-server queue with server vacations and a class of nonrenewal arrival processes. *Advances in Applied Probability*, 22, 676-705.

[17] Lucantoni, D. M. (1991). New results on the single server queue with a batch Markovian arrival process. *Stochastic Models*, 7, 1-46.

[18] Marcus, M., & Minc, H. (1964). A Survey of Matrix Theory and Matrix Inequalities. Allyn and Bacon, Boston, MA.

[19] Neuts, M. F. (1979). A versatile Markovian point process. *Journal of Applied Probability*, 16, 764-779.

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

[21] Neuts, M. F. (1989). Structured Stochastic Matrices of M/G/1 Type and Their Applications. Marcel Dekker, NY.

[22] Neuts, M. F. (1992). Models based on the Markovian arrival process. *IEICE Transactions on Communications*, E75B, 1255-1265.

[23] Neuts, M. F. (1995). Algorithmic Probability: A Collection of Problems. Chapman and Hall, NY.

[24] Saffari, M., Haji, R., & Hassanzadeh, F. (2011). A queueing system with inventory and mixed exponentially distributed lead times. *International Journal of Advanced Manufacturing Technology*, 53, 1231-1237.

[25] Saffari, M., Asmussen, S., & Haji, R. (2013). The M/M/1 queue with inventory, lost sale, and general lead times. *Queueing Systems*, 75, 65-77.

[26] Steeb, W.-H., & Hardy, Y. (2011). Matrix Calculus and Kronecker Product. World Scientific Publishing, Singapore.

[27] Venkataraman, A., Sibia, A., Pandeya, J., Kumar, K., & Alex, R. (2016). Predicting ATM Failure from Logs: A Machine Learning Approach. Paper #189, Business Analytics and Intelligence: A Compendium, Eds., M. Mathirajan and U. Dinesh Kumar. Proceedings of the 4th International Conference on Business Analytics and Intelligence, Bangalore, India, December 19-21, 2016.

[28] Wang, F., Bhagat, A., & Chang, T. (2016). Analysis of priority multi-server retrial queueing inventory systems with MAP arrivals and exponential services. OPSEARCH, Published online: 16 June