2021

Queueing Models with MAP Arrivals Useful in Service Sectors

Srinivas R. Chakravarthy

Queueing Models and
Service Management
Vol. 4, No. 1, page 55-84, 2021

QMSM
© PU 2018

Queueing Models with MAP Arrivals Useful in Service Sectors

Srinivas R. Chakravarthy

Departments of Industrial and Manufacturing Engineering and Mathematics
Kettering University
Flint, MI-48504, USA
(Received May 2020 ; accepted November 2020)

Abstract: Queueing models have found applications in many fields, notably in service sectors. In this paper, we study queueing models that have significant applications in service sectors. We look at multi-server systems with *MAP* arrivals. We assume phase type services for single server systems and exponential services when dealing with multi-server systems. All arriving customers finding no idle server will not wait in the system to receive services but rather leave their information in a registry list. These customers will be reached out on a first-come-first-served basis (*FCFS*) by an idle server soon after completing its current service. The reach out time is assumed to be exponential and at the end of this time, with a certain probability the reached out customer is available for service; with complementary probability the customer is not reachable due to various reasons including the customer not picking up the call from the service system to receive a service. In the case when the reach out is unsuccessful, the server will remain idle should there be no customer in the registry list. However, if there is at least one customer in the registry, then the server will start another reach out. The classical approach using matrix-analytic methods is employed and discuss a few illustrative examples that bring out the qualitative nature of the models in steady-state. When dealing with *MAP*/*G*/*c* queues we resort to simulation and present a few examples. Some concluding remarks including a few extensions to the models studied here are presented.

Keywords: Algorithmic probability, call-back, Markovian arrival process, phase type distributions, queues, simulation.

1. Introduction

Queueing models have been (and still are being) used in many areas such as production and manufacturing, telecommunications, and service sectors. With the advent of modern technology and a stiff competition in service sectors to capture the market share, the service providers try to assure their customers of efficient and timely services. However, the service providers do not have unlimited resources, a common scene in our day-to-day activities, to provide an immediate access to all their arriving customers. Retrial queues (see, e.g., [3, 4, 16, 27]) were introduced to model the situations where customers who cannot receive service immediately upon their arrivals but who are willing try again to capture a free server. One major drawback in using retrial queues including any variation, such as balking, reneging among others, introduced later on, is its applicability in certain service sectors. For example,

*Corresponding author
Email : schakrav@kettering.edu

in this modern era, many service providers have the ability to assure customers who cannot access a server upon their requests, that they will be called back as soon as a free server is available and that their position in the queue will not change because of that option. That is, the customers who cannot access a server immediately will be reached out by the service provider on a $FCFS$ basis soon after a server becomes free. Towards this end the customers leave behind their information in what we call henceforth as *registry list* for the service provider to reach out to them at a later point in time.

It should be pointed out the first study of queueing models wherein a server, upon completion of a service, will search for a customer to offer a service was done in [23]. Such models are very useful in the context of call center applications. However, the models studied here do not fall under the category of searching for customers but rather reaching out to those customers who entered the system but only to find the servers to be busy, and left the system to be called back.

In real-life applications, not all customers are able to answer the service provider's call when their turn comes. For example, the customers maybe away from their phones or maybe on an another call, and so might miss the service calls. In such cases, the service provider has to remove this customer from the registry list of customers needing service, and move on to the next customer either on the list or wait for a new arrival. This motivated us to study the queueing models in this paper. We believe, to the best of our knowledge, only a few papers dealing with such models have been published in the queueing literature [1, 2, 12, 15].

While the authors in [1, 2] focus on studying $M/M/N$-type queues with two types of customers (one of which is call-back customers) asymptotically, the models studied in [12, 15] look at two types of customers, one of which is call-back type, and the analysis is done in steady-state. In [15], $MAP/M/N$-type queueing system is used to study "real" and "virtual or call-back" customers such that "real" customers have a finite buffer whereas the "virtual" customers have an infinite buffer. The model studied in [12] generalizes the one in [15] in a more general context, namely, modeling the customers' impatience as well as using phase type services. However, the nature of a strict "call-back" following the $FCFS$ basis rule is lost in their models. This is due to the fact that in [12, 15], the virtual customers are offered service only after all the "real" customers are cleared. Thus, it is possible for the new arrivals to become "real" (and get a non-preemptive priority service) while the earlier arrivals who became virtual are still waiting for service. This fact was not discussed numerically as well as analytically in [12, 15]. Thus, our additional motivation for our paper is to bring out the salient features when the $FCFS$ rule is strictly applied as it should be a significant one due to the service centers pretty much promising the customers that the call-back option will not affect their position in the queue. In this sense, our paper not only complements with that of [12] to help practitioners to adapt the models in the context of call-back in a more general context, but also to provide qualitative insight into such models.

It is worth to point out here that the models studied here do not fall under the umbrella of retrial queues as neither the customers nor the servers retry. Only the servers try to reach out to the customers in the registry, and a failure to capture a customer results in the removal of that customer from the list. Further, the customers who left their information in the registry

do not attempt to call the service provider again.

We model the arrivals using a Markovian arrival process (MAP), a special case (namely, single arrivals) of versatile Markovian point process ($VMPP$) as originally introduced by Neuts in 1970s. Since its introduction and since with major notational simplifications of $VMPP$ as $BMAP$ in 1990 (see, [18]), this versatile process is widely used in stochastic modeling [5, 6, 7, 9, 18, 19, 21, 24, 25, 26].

Recall that MAP is a powerful point process in stochastic modeling due to its capability to model a variety of scenarios that occur naturally in practice. A MAP is described by two parameter (square) matrices, say, E_0 and E_1, of dimension, say, m. The (irreducible) generator, say, E, given by $E = E_0 + E_1$, governs transitions of the MAP. Let $\boldsymbol{\eta}$ be such that

$$\boldsymbol{\eta} E = \mathbf{0}, \;\; \boldsymbol{\eta} \boldsymbol{e} = 1, \tag{1}$$

where $\boldsymbol{e}$ stands for a column vector of 1's of dimension m. The arrival rate is then given by $\lambda = \boldsymbol{\eta} E_1 \boldsymbol{e}$.

In addition to the notation $\boldsymbol{e}$, we introduce a few others needed in subsequent sections. $\boldsymbol{e}_i$ is a unit column vector with 1 in the i^{th} position and 0 elsewhere; I an identity matrix; The notation $'$ is for the transpose of a matrix or a vector. Also, we will be using the Kronecker product and Kronecker sum of matrices (see, e.g., [13, 20, 28] for more details). These operations are denoted, respectively, by $\otimes$ and $\oplus$.

Thus, the objective of this paper is to analyze these queueing models useful in service sector, and bring out the qualitative behavior of the models studied with regard to the arrival processes, the service times, and other parameters. The rest of the paper is organized as follows. In Section 2, we present and analyze $MAP/PH/1$-type of our model, and in Section 3, we look at $MAP/M/c$-type model and carry out the requisite analysis. Selected key system performance measures needed for discussion are given in Section 4. In Section 5, we discuss a few examples to highlight some qualitative nature of the models under study. A simulation approach in the context of $MAP/G/c$ queues along with a few (simulated) examples are discussed in Section 6. Finally, in Section 7, some concluding remarks are outlined.

2. $MAP/PH/1$-Model

In this section, we look at $MAP/PH/1$-type queueing model in the context of call-back services. Specifically the model assumptions are as follows. Customers arrive according to a MAP with representation matrices (E_0, E_1) of dimension m.

There is a single server with the service times following a phase type ($PH-$) distribution with representation $(\boldsymbol{\beta}, S)$ of dimension n. Recall that a $PH-$ distribution is obtained as the time until absorption in a finite state (irreducible) Markov chain with generator $S + \boldsymbol{S^0}\boldsymbol{\beta}$, where $S\boldsymbol{e} + \boldsymbol{S^0} = \mathbf{0}$. The service rate is given by $\mu = [\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}]^{-1}$ (see, e.g., [22]).

An arriving customer finding the server idle will get into service; otherwise, the customer will not be waiting in the system (so as to not waste their time and tend to other

activities) but their contact information will be in a registry list of infinite capacity in order for the service provider to reach out to those customers at a later point in time. Upon completion of a service, the server will either become idle (if the registry list is empty) or will try to "reach" the next customer in the list to provide a service. The reaching out duration is random and lasts for an exponential amount of time with parameter θ. At the end of a reaching out time, either the customer is taken into service with probability $p, 0 \leq p < 1$, or with probability $q = 1 - p$, the customer is unavailable for service. The unavailable customer is dropped from the list of customers waiting for service by the system and no further contact will be made to offer a service to such a customer. At that instant, the server becomes idle and will follow the process of either remaining idle or reach out for the next customer depending on whether the registry list is empty or positive.

Thus, at any given time, the server is in one of the following states: (a) idle (due to the system being empty); or (b) busy serving a customer; or (c) busy reaching out to a customer.

2.1. Steady-state analysis

Suppose that we define the following random variables at time t.

- $J_1(t)$ = number of customers in the system (in service plus the ones not reached)
- $J_2(t)$ = phase of the service process
- $J_3(t)$ = phase of the MAP arrival process
- $J_4(t)$ = status of the server (idle/busy/reach)

Verify that the stochastic process $\{(J_1(t), J_2(t), J_3(t), J_4(t))\}$ has the state space given by

$$\Omega = \{(0,k) : 1 \leq k \leq m\} \bigcup \{(i,j,k,r) : 1 \leq j \leq n,\ 1 \leq k \leq m,\ r = 1,2,\ i \geq 1\}.$$

The server busy serving a customer is denoted by $r = 1$ and when the server is busy reaching out to a call-back customer is denoted by $r = 2$. Defining

$$B_0 = \boldsymbol{e}_1' \otimes \boldsymbol{\beta} \otimes E_1, \quad B_1 = \begin{bmatrix} \boldsymbol{S^0} \otimes I \\ q\theta \otimes I \end{bmatrix}, \tag{2}$$

$$F_0 = I \otimes E_1, \quad F_1 = \begin{bmatrix} S \oplus E_0 & 0 \\ \theta p \boldsymbol{\beta} \otimes I & E_0 - \theta I \end{bmatrix}, \quad F_2 = \boldsymbol{e}_2' \otimes B_1, \tag{3}$$

verify that the generator, Q, of the $LIQBD$ is of the form

$$Q = \begin{pmatrix} E_0 & B_0 & 0 & 0 & 0 & \cdots \\ B_1 & F_1 & F_0 & 0 & 0 & \cdots \\ 0 & F_2 & F_1 & F_0 & 0 & \cdots \\ 0 & 0 & F_2 & F_1 & F_0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \cdots \end{pmatrix}. \tag{4}$$

Suppose that $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_0, \boldsymbol{\varphi}_1)$ denotes the steady-state probability vector of $F = F_0 + F_1 + F_2$. The vector $\boldsymbol{\varphi}$ satisfying $\boldsymbol{\varphi} F = \mathbf{0}$ and $\boldsymbol{\varphi}\boldsymbol{e} = 1$ can be solved using the equations:

$$\begin{aligned} \boldsymbol{\varphi}_0(S \oplus E) + \theta p \boldsymbol{\varphi}_1(\boldsymbol{\beta} \otimes I) &= \mathbf{0}, \\ \boldsymbol{\varphi}_0(\boldsymbol{S^0} \otimes I) + \boldsymbol{\varphi}_1(E - \theta p I) &= \mathbf{0}, \\ \boldsymbol{\varphi}\boldsymbol{e} &= 1. \end{aligned} \tag{5}$$

The following, intuitively obvious, internal accuracy checks can easily be proved.

$$\begin{aligned} \boldsymbol{\varphi}_0(\boldsymbol{e} \otimes I) + \boldsymbol{\varphi}_1 &= \boldsymbol{\eta}, \\ \boldsymbol{\varphi}_0(I \otimes \boldsymbol{e}) &= d\mu\boldsymbol{\beta}(-S)^{-1}, \end{aligned} \tag{6}$$

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

$$\boldsymbol{\varphi}_0(\boldsymbol{S^0} \otimes \boldsymbol{e}) = \theta p \boldsymbol{\varphi}_1 \boldsymbol{e}. \tag{7}$$

Now using equations (5-7), it is easy to verify that $\boldsymbol{\varphi}_0\boldsymbol{e} = d$, and hence we have

$$d = \frac{\theta p}{\mu + \theta p}. \tag{8}$$

From the classical stability condition for $LIQBD$ process (see, e.g., [22]), namely, $\boldsymbol{\varphi} F_0 \boldsymbol{e} < \boldsymbol{\varphi} F_2 \boldsymbol{e}$, we immediately get

$$\lambda < \frac{\theta\mu}{\mu + \theta p}. \tag{9}$$

For use in sequel, we define the traffic load of the system as

$$\rho_q = \frac{\lambda(\mu + \theta p)}{\theta\mu}. \tag{10}$$

Under the stability condition given in (9), we can determine the steady-state probability vector of Q. Towards this end, suppose we define $\boldsymbol{z} = (\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots)$ with $\boldsymbol{z}_i,\ i \geq 1$, further partitioned as $\boldsymbol{z}_i = (\boldsymbol{u}_i,\ \boldsymbol{v}_i),\ i \geq 1$. Note that $\boldsymbol{z}_0$, which is of dimension m gives the steady-state probability vector of the system being empty and the phase of the MAP in various phases; $\boldsymbol{u}_i$, of dimension mn, gives the steady-state probability vector that the system has i customers (one in service and $i-1$ in the registry) with the service and the arrival processes in various states; $\boldsymbol{v}_i$, of dimension m, gives the steady-state probability vector that the server is trying to reach the (earliest) customer among the i that are in "reach" category, and the arrival process is in one of m phases.

The steady-state vector $\boldsymbol{z}$ is obtained by solving $\boldsymbol{z}Q = \mathbf{0}$ and $\boldsymbol{z}\boldsymbol{e} = 1$. Exploiting the special structure of the generator Q, we solve for $\boldsymbol{z}$ as follows.

$$\begin{aligned} \boldsymbol{z}_0 F_0 + \boldsymbol{u}_1(\boldsymbol{S^0} \otimes I) + q\theta\boldsymbol{v}_1 &= \mathbf{0}, \\ \boldsymbol{z}_0(\boldsymbol{\beta} \otimes E_1) + \boldsymbol{u}_1(S \oplus F_0) + \theta p \boldsymbol{v}_1(\boldsymbol{\beta} \otimes I) &= \mathbf{0}, \\ \boldsymbol{u}_1[R_1(\boldsymbol{S^0} \otimes I) + q\theta R_2] + \boldsymbol{v}_1[F_0 - \theta I + R_3(\boldsymbol{S^0} \otimes I) + q\theta R_4] &= \mathbf{0}, \\ \boldsymbol{z}_i = \boldsymbol{z}_1 R^{i-1},\ i \geq 1, & \end{aligned} \tag{11}$$

subject to the normalizing condition

$$\boldsymbol{z}_0 + \boldsymbol{z}_1(I - R)^{-1}\boldsymbol{e} = 1, \tag{12}$$

and where R, the minimal nonnegative solution to the matrix-quadratic equation: $R^2F_2 + RF_1 + F_0 = 0$, is partitioned as

$$R = \begin{pmatrix} R_1 & R_2 \\ R_3 & R_4 \end{pmatrix}. \tag{13}$$

The matrix R, of dimension $m(n+1)$, can be computed either by using the well-known logarithmic reduction algorithm (see, e.g., [17]) when the dimensions are of reasonable size or by using (block) Gauss-Seidel method (see, e.g., [29]) taking advantage of the sparsity of the coefficient matrices appearing in the matrix-quadratic equation. The details are omitted. Suppose we denote by ρ, the traffic intensity of the classical $MAP/PH/1$ queue. That is, , $\rho = \frac{\lambda}{\mu}$. One item that is of interest is to find the value of θ for which $\rho_q = \rho$. Setting $\rho_q = \rho$ and using equation (10), it is easy to see that

$$\theta = \frac{\mu}{1-p}. \tag{14}$$

Obviously, the above equation implies that as $p \to 1$, θ increases to ∞. Note that θ is undefined when $p = 1$ indicating that if every customer in the registry is available at the time of the server reaching out to receive a service, our model will never outperform the corresponding classical $MAP/PH/1$ queue. This is not surprising since the customers who do not enter into service upon their arrivals incur an additional reach out time as part of their services (from the system's point of view). We use the term "outperform" in the context of the system having less customers (on the average) waiting in the system as compared to the corresponding classical queue. It should be pointed out that "outperform" here does not necessarily mean the system is serving more customers as there are some customers who leave the system without getting served due to their choice. Further, the customers, who cannot enter into service immediately upon their arrivals to the system will probably tend to their other activities while waiting for the service provider to reach out to them. Thus, these customers will not negatively view (or rate) the service provider, and may possibly result in giving a higher rating irrespective of whether they receive a service or not. When $p < 1$, one can choose θ for which $\theta > \frac{\mu}{1-p}$, so that the mean number of customers waiting in the system will be less than the corresponding classical $MAP/PH/1$ queue.

By taking $p = 0$ and $\theta = \mu$, under the assumptions of exponential service times, it is easy to verify that our current model behaves like the classical $MAP/M/1$ queue. That is, the measures such as the mean number of customers in the system and the probability that the server is idle (and hence the probability that the server is busy) of the current model are identical to the corresponding classical $MAP/M/1$ model. An intuitive explanation is as follows. When $p = 0$, the server after reaching out to a customer will become idle and repeat the same process of reaching out until there are no more customers waiting to be reached

out. Every reach out results in spending an exponential amount of time with parameter θ which is taken to be μ. Thus, a reaching out duration is similar to a service duration in the corresponding classical queue.

3. $MAP/M/c$-Model

In this section we relax the single server system to a multi-server one but restrict the service times to be exponential. We now define the following random variables at time t.

- $\hat{J}_1(t)$ = number of customers in the system (in service plus the ones not reached)
- $\hat{J}_2(t)$ = number of servers in "reach" state trying to access customers (note that the number of busy servers is given by $min\{\hat{J}_1(t), c\} - \hat{J}_2(t)$).
- $\hat{J}_3(t)$ = phase of the MAP arrival process

Verify that the stochastic process $\{(\hat{J}_1(t), \hat{J}_2(t), \hat{J}_3(t))\}$ has the state space given by

$$\hat{\Omega} = \{(0,k) : 1 \leq k \leq m\} \bigcup \{(i,j,k) : 1 \leq j \leq min\{i,c\},\ 1 \leq k \leq m,\ i \geq 1\}.$$

Defining

$$\hat{B}_{1,0} = \begin{bmatrix} \mu I \\ q\theta I \end{bmatrix},\ \hat{B}_{i,i+1} = \begin{bmatrix} E_1 & 0 \cdots & 0 & 0 \\ 0 & E_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & 0 & 0 \\ 0 & 0 & \cdots & E_1 & 0 \end{bmatrix},\ 1 \leq i \leq c-1, \tag{15}$$

$$\hat{B}_i = \begin{bmatrix} E_0 - i\mu I & & & & & \\ \theta p I & E_0 - (i-1)\mu I & & & & \\ & 2\theta p I & E_0 - (i-2)\mu I & & & \\ & & \ddots & \ddots & & \\ & & & (i-1)\theta q I & \mu I & \\ & & & & i\theta q I & \end{bmatrix},\ 1 \leq i \leq c-1, \tag{16}$$

$$\hat{B}_{i,i-1} = \begin{bmatrix} i\mu I & & & & & \\ \theta q I & (i-1)\mu I & & & & \\ & 2\theta q I & (i-2)\mu I & & & \\ & & \ddots & \ddots & & \\ & & & (i-1)\theta q I & \mu I & \\ & & & & i\theta q I & \end{bmatrix},\ 1 \leq i \leq c, \tag{17}$$

$$\hat{F}_0 = I \otimes E_1,\ \hat{F}_1 = \begin{bmatrix} E_0 - c\mu I & & & & & & \\ \theta p I & E_0 - (c-1)\mu I & & & & & \\ & 2\theta p I & E_0 - (c-2)\mu I & & & & \\ & & \ddots & \ddots & & & \\ & & & (c-1)\theta q I & \mu I & & \\ & & & & c\theta q I & & \end{bmatrix}, \tag{18}$$

$$\hat{F}_2 = \begin{bmatrix} c\mu I & & & & & \\ \theta q I & (c-1)\mu I & & & & \\ & 2\theta q I & (c-2)\mu I & & & \\ & & \ddots & \ddots & & \\ & & & (c-1)\theta q I & \mu I & \\ & & & & c\theta q I & \end{bmatrix}. \tag{19}$$

Verify that the generator, $\hat{Q}$, of the $LIQBD$ is of the form

$$\hat{Q} = \begin{pmatrix} E_0 & \boldsymbol{e}_1' \otimes E_1 & & & & & & & \\ \hat{B}_{1,0} & \hat{B}_1 & \hat{B}_{1,2} & & & & & & \\ & \hat{B}_{2,1} & \hat{B}_2 & \hat{B}_{2,3} & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & & \hat{B}_{c,c-1} & \hat{F}_1 & \hat{F}_0 & & \\ & & & & & \hat{F}_2 & \hat{F}_1 & \hat{F}_0 & \\ & & & & & & \hat{F}_2 & \hat{F}_1 & \hat{F}_0 \\ & & & & & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{20}$$

Suppose that $\hat{\boldsymbol{\varphi}} = (\hat{\boldsymbol{\varphi}}_0, \cdots, \hat{\boldsymbol{\varphi}}_c)$ denotes the steady-state probability vector of $\hat{F} = \hat{F}_0 + \hat{F}_1 + \hat{F}_2$. The vector $\hat{\boldsymbol{\varphi}}$ satisfying $\hat{\boldsymbol{\varphi}}\hat{F} = \mathbf{0}$ and $\hat{\boldsymbol{\varphi}}\boldsymbol{e} = 1$, can be solved using the equations:

$$\begin{gathered} \hat{\boldsymbol{\varphi}}_0(E - c\mu I) + \theta p \hat{\boldsymbol{\varphi}}_1 = \mathbf{0}, \\ (c-i+1)\mu \hat{\boldsymbol{\varphi}}_{i-1} + \hat{\boldsymbol{\varphi}}_1(E - [(c-i)\mu + i\theta p]I) + (i+1)\theta p \hat{\boldsymbol{\varphi}}_{i+1} = \mathbf{0},\ 1 \le i \le c-1, \\ \mu \hat{\boldsymbol{\varphi}}_{c-1} + \hat{\boldsymbol{\varphi}}_c(E - c\theta p I) = \mathbf{0}, \\ \hat{\boldsymbol{\varphi}}\boldsymbol{e} = 1. \end{gathered} \tag{21}$$

It is easy to verify the following results using the equations in (21).

$$(c-i)\mu \hat{\boldsymbol{\varphi}}_i \boldsymbol{e} = (i+1)\theta p \hat{\boldsymbol{\varphi}}_{i+1}\boldsymbol{e},\ 0 \le i \le c-1, \tag{22}$$

$$\mu \sum_{i=0}^{c} (c-i)\hat{\boldsymbol{\varphi}}_i \boldsymbol{e} = \theta p \sum_{i=0}^{c} i \hat{\boldsymbol{\varphi}}_i \boldsymbol{e}, \tag{23}$$

$$\sum_{i=0}^{c} \hat{\boldsymbol{\varphi}}_i = \boldsymbol{\eta}, \tag{24}$$

where $\boldsymbol{\eta}$ is as given in (1). Through some simple algebraic manipulations it is easy to verify that

$$\hat{\boldsymbol{\varphi}}_i \boldsymbol{e} = \binom{c}{i} \left(\frac{\mu}{\mu+\theta p}\right)^i \left(\frac{\theta p}{\mu+\theta p}\right)^{c-i}, \ 0 \le i \le c. \tag{25}$$

That is, $\hat{\boldsymbol{\varphi}}_i \boldsymbol{e}$ is given by the binomial probabilities with parameters c and $\frac{\mu}{\mu+\theta p}$. The classical stability condition for the $LIQBD$ process with generator given in (20), namely, $\hat{\boldsymbol{\varphi}} \hat{F}_0 \boldsymbol{e} < \hat{\boldsymbol{\varphi}} \hat{F}_2 \boldsymbol{e}$, yields, $\lambda < \theta \sum_{i=0}^{c} i \hat{\boldsymbol{\varphi}}_i \boldsymbol{e}$. Now, using the fact that $\sum_{i=0}^{c} i \hat{\boldsymbol{\varphi}}_i \boldsymbol{e} = c \frac{\mu}{\mu+\theta p}$, we see that this queueing model is stable if and only if

$$\lambda < \frac{c\theta\mu}{\mu+\theta p}. \tag{26}$$

Suppose that we define the traffic intensity for our current model in the context of $MAP/M/c$ model as

$$\hat{\rho}_q = \frac{\lambda(\mu+\theta p)}{c\theta\mu}. \tag{27}$$

The steady-state vector $\hat{\boldsymbol{z}}$, partitioned into vectors of smaller dimension as, $\hat{\boldsymbol{z}} = (\hat{\boldsymbol{z}}_0, \hat{\boldsymbol{z}}_1, \cdots)$ with $\hat{\boldsymbol{z}}_i = (\hat{\boldsymbol{z}}_{i,0}, \hat{\boldsymbol{z}}_{i,1}, \cdots, \hat{\boldsymbol{z}}_{i,min\{i,c\}})$, $i \ge 1$, is obtained by solving $\hat{\boldsymbol{z}}\hat{Q} = \boldsymbol{0}$ and $\hat{\boldsymbol{z}}\boldsymbol{e} = 1$. Exploiting the special structure of the generator $\hat{Q}$, we solve for $\hat{\boldsymbol{z}}$ as follows.

$$\begin{aligned}
\hat{\boldsymbol{z}}_0 E_0 + \mu \hat{\boldsymbol{z}}_{1,0} + q\theta \hat{\boldsymbol{z}}_{1,1} &= \boldsymbol{0}, \\
\hat{\boldsymbol{z}}_0 E_1 + \hat{\boldsymbol{z}}_{1,0}(E_0 - \mu I) + \theta p \hat{\boldsymbol{z}}_{1,1} + 2\mu \hat{\boldsymbol{z}}_{2,0} + q\theta \hat{\boldsymbol{z}}_{2,1} &= \boldsymbol{0}, \\
\hat{\boldsymbol{z}}_{1,1}(E_0 - \mu I) + \mu \hat{\boldsymbol{z}}_{2,1} + 2q\theta \hat{\boldsymbol{z}}_{2,2} &= \boldsymbol{0}, \\
\hat{\boldsymbol{z}}_{i-1} \hat{B}_{i-1,i} + \hat{\boldsymbol{z}}_i \hat{B}_i + \hat{\boldsymbol{z}}_{i+1} \hat{B}_{i+1,i} &= \boldsymbol{0}, \ 2 \le i \le c-1, \\
\hat{\boldsymbol{z}}_{c-1} \hat{B}_{c-1,c} + \hat{\boldsymbol{z}}_c [A_1 + \hat{R} A_2] &= \boldsymbol{0},
\end{aligned} \tag{28}$$

subject to the normalizing condition

$$\sum_{i=0}^{c-1} \hat{\boldsymbol{z}}_i \boldsymbol{e} + \hat{\boldsymbol{z}}_c (I - \hat{R})^{-1} \boldsymbol{e} = 1, \tag{29}$$

and where $\hat{R}$ is the minimal nonnegative solution to the matrix-quadratic equation: $\hat{R}^2 \hat{F}_2 + \hat{R}\hat{F}_1 + \hat{F}_0 = 0$.

Here the matrix R, which is of dimension $m(c+1)$, can be computed along the lines pointed out earlier. The details are omitted.

Like in the $MAP/PH/1$ case (see Section 2), the value of θ for which $\hat{\rho}_q = \rho$ for the current model, has the same expression as given in (14). Hence, the discussion on θ and p holds good here too. Since the service times are exponential here, in the case when $p = 0$ and $\theta = \mu$, our model behaves like the classical $MAP/M/c$ with regard to the mean number in the system and the probability of the server being idle (and hence the utilization of the resource).

4. System Performance Measures

In this section we will list four system performance measures: (a) P(idle) = P(the system is idle); (b) P(busy) = P(the server is busy serving a customer); (c) P(reach) = P(the server is busy reaching out to a customer); and (d) μ_{NQ} = mean number of customers in the registry, for our qualitative study of both the models. The relevant expressions for these measures for both the models are displayed in Table 1 below. It is important to note that P(busy) refers to the utilization of the server in serving the customers and does not include reaching out probabilities. A similar interpretation holds good for P(reach).

5. Illustrative Numerical Examples

In this section, we will discuss a few qualitative numerical examples. Towards this end, we split this section into many to discuss $MAP/PH/1$-type and $MAP/M/c$-type models. In all the cases, we use the following nineteen $MAPs$. The ones corresponding positive and negative correlated inter-arrival times are generated using the techniques first spelled out in [7] and further elaborated with more details in [10].

5.1. Types of arrival processes for the numerical examples

For arrival processes, we consider three processes representing renewal arrivals, eight negatively correlated arrivals, and eight positively correlated arrivals. In all our examples, we fix, without any loss of generality, the arrival rate to be one. That is, we take $\lambda = 1$. Before, we proceed further, we need the following notation.

$$\boldsymbol{\zeta_{m-1}} = (1, 0, \cdots, 0), \;\; U_{m-1} = \begin{pmatrix} -\lambda_1 & \lambda_1 & & \\ & -\lambda_1 & \lambda_1 & \\ & & \ddots & \\ & & & -\lambda_1 \end{pmatrix}, \;\; \boldsymbol{U^0_{m-1}} = -U_{m-1}\boldsymbol{e},$$

$$E_0 = E_0(\lambda_1, \lambda_2, m) = \begin{pmatrix} U_{m-1} & 0 \\ 0 & -\lambda_2 \end{pmatrix},$$

$$E_1 = E_1(\lambda_1, \lambda_2, r_1, r_2, m) = \begin{pmatrix} r_1 \boldsymbol{U^0_{m-1}\zeta_{m-1}} & (1-r_1)\boldsymbol{U^0_{m-1}} \\ (1-r_2)\lambda_2 \boldsymbol{\zeta_{m-1}} & r_2\lambda_2 \end{pmatrix},$$

where $0 < r_i < 1, i = 1, 2..$ Note that the representation $(\boldsymbol{\zeta}_{m-1}, U_{m-1})$ of dimension $m-1$ stands for an Erlang distribution of order $m-1$ (see e.g., [22]). The representation $(E_0(\lambda_1, \lambda_2, m), E_1(\lambda_1, \lambda_2, p_1, p_2, m))$ of dimension m can be used to generate a variety of correlated arrival MAP processes. For more details, we refer the reader to [10].

Renewal processes (RP):

aE**:** This is Erlang of order 5 with a rate of 5 in each stage.

aX**:** This is exponential with parameter 1.

Table 1. System Performance Measures for $MAP/PH/1$ and $MAP/M/c$.

$Measure$	$MAP/PH/1$	$MAP/M/c$
$P(idle)$	$\boldsymbol{z_0 e}$	$\hat{\boldsymbol{z}}_{\mathbf{0}}\boldsymbol{e}$
$P(busy)$ [Utilization in serving]	$\boldsymbol{z_1}(I-R)^{-1}\begin{pmatrix}\boldsymbol{e}\\ \mathbf{0}\end{pmatrix}$	$\frac{1}{c}\left[\sum_{i=1}^{c-1}\sum_{j=0}^{i}(i-j)\hat{\boldsymbol{z}}_{ij}\boldsymbol{e}+\sum_{i=c}^{\infty}\sum_{j=0}^{c}(c-j)\hat{\boldsymbol{z}}_{ij}\boldsymbol{e}\right]$
$P(reach)$ [Utilization in reaching]	$\boldsymbol{z_1}(I-R)^{-1}\begin{pmatrix}\mathbf{0}\\ \boldsymbol{e}\end{pmatrix}$	$\frac{1}{c}\left[\sum_{i=1}^{c-1}\sum_{j=0}^{i}j\hat{\boldsymbol{z}}_{ij}\boldsymbol{e}+\sum_{i=c}^{\infty}\sum_{j=0}^{c}j\hat{\boldsymbol{z}}_{ij}\boldsymbol{e}\right]$
μ_{NQ}	$\boldsymbol{z_1}R(I-R)^{-2}\boldsymbol{e}$	$\sum_{i=1}^{c-1}i\hat{\boldsymbol{z}}_i\boldsymbol{e}+c\hat{\boldsymbol{z}}_c(I-\hat{R})^{-1}\boldsymbol{e}+\hat{\boldsymbol{z}}_c\hat{R}(I-\hat{R})^{-2}\boldsymbol{e}$ $-c[P(idle)+P(reach)]$

aH**:** This is hyperexponential having mixing probabilities as $(0.5, 0.3, 0.15, 0.05)$ with the corresponding rates given by $(68.5, 6.85, 0.685, 0.0685)$.

Negatively correlated processes (NC**):**

Ni**:** This is a MAP of order $i+2,\ 1 \le i \le 8$, with representation given by

$$(E_0(1.25+0.5(i-1), 2+0.5i, i+2), E_1(1.25+0.5(i-1), 2+0.5i, 0.01, 0.01, i+2).$$

Positively correlated processes (PC**):**

Pi**:** This is a MAP of order $i+2,\ 1 \le i \le 8$,) with representation given by

$$(E_0(1.25+0.5(i-1), 2+0.5i, i+2), E_1(1.25+0.5(i-1), 2+0.5i, 0.99, 0.99, i+2).$$

Note that we will normalize the $MAPs$ so that they all will have a mean of 1. The standard deviations (SD) and one lag correlation coefficients (ρ_a of the inter-arrival times of these $MAPs$ are displayed in Tables 2-4 below.

5.2. Examples dealing with $MAP/PH/1$*-type model*

In this section we will present a few representative examples dealing with MAP arrivals and $PH-$type service times. Towards this end, we identify the three service times used in these examples.

Table 2. σ_a and ρ_a of the renewal process $MAPs$.

$Measure$	aE	aX	aH
σ_a	0.4472	1.0000	4.5787
ρ_a	0	0	0

Table 3. σ_a and ρ_a of the negatively correlated $MAPs$.

$Measure$	$N1$	$N2$	$N3$	$N4$	$N5$	$N6$	$N7$	$N8$
σ_a	1.0392	1.0202	1.0123	1.0082	1.00590	1.00443	1.0035	1.0028
ρ_a	-0.3267	-0.4804	-0.5786	-0.6454	-0.6935	-0.7296	-0.7577	-0.7802

Table 4. σ_a and ρ_a of the positively correlated $MAPs$.

$Measure$	$P1$	$P2$	$P3$	$P4$	$P5$	$P6$	$P7$	$P8$
σ_a	1.0392	1.0202	1.0123	1.0082	1.00590	1.00443	1.0035	1.0028
ρ_a	0.3267	0.4804	0.5786	0.6454	0.6935	0.7296	0.7577	0.7802

5.2.1. Types of $PH-$ services for the numerical examples

sE**:** This is Erlang of order 4 with a rate of 4 in each stage. The rate in each stage is taken to be 4μ so that the service mean will be $\frac{1}{\mu}$.

sX**:** This is exponential with parameter μ.

sH**:** This is hyperexponential having mixing probabilities as $(0.6, 0.25, 0.10, 0.05)$ with the corresponding rates given by $\mu(63.1, 6.31, 0.631, 0.0631)$.

Note that the above $PH-$distributions all have a mean of $\frac{1}{\mu}$ and are qualitatively different covering a wide variety of scenarios in practice.

5.2.2. Example 1:

In this example, we compare our model to the corresponding classical $MAP/PH/1$ queueing model by fixing $\theta = \frac{\mu}{1-p}$, and varying p from 0 to 0.99. Recall that λ is fixed to be 1. As pointed out earlier, a comparison of the two models makes sense and also being fair only by taking $\rho_q = \rho$. The value of μ is therefore obtained as $\mu = \frac{\lambda}{\rho}$. We consider four values for $\rho = 0.50, 0.90, 0.95, 0.99$ and look at various combinations of arrival processes and service times. Due to lack of space, we display the graphs of the measure, the ratio of the mean number in queue. That is, we look at $\frac{\mu_{NQ}}{\mu_{CNQ}}$, where μ_{CNQ} is the mean number of customers in the queue of the corresponding classical $MAP/PH/1$ model. In Figures 1 and 2, we display the selected but representative graphs of this ratio, respectively, for renewal and for (selected) correlated arrival processes.

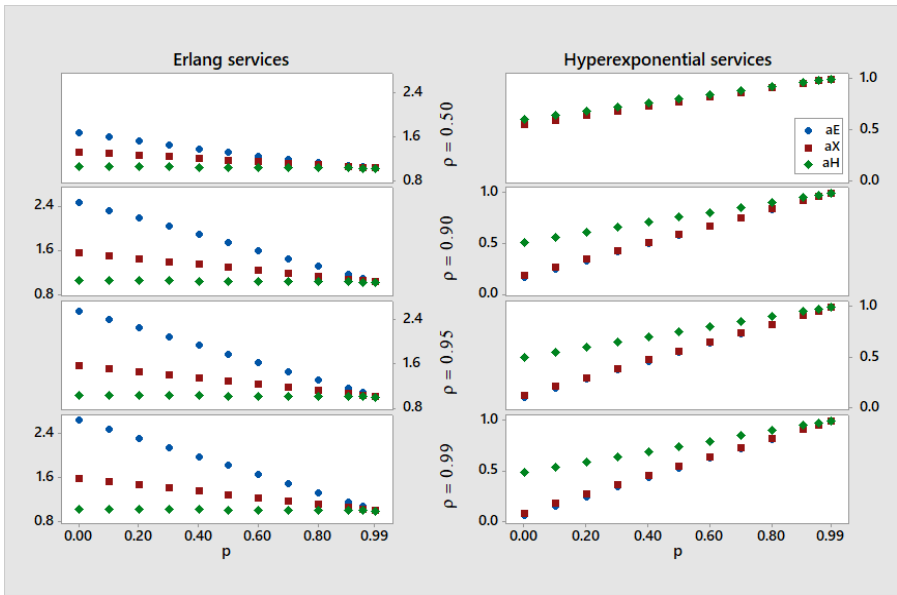

Figure 1. Ratio of mean number in queue for renewal arrivals under various scenarios.

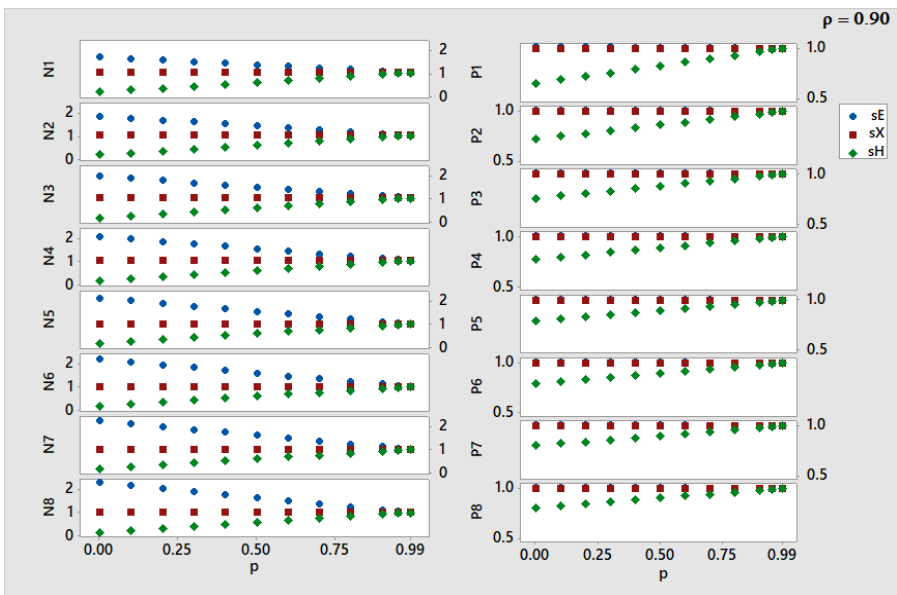

Figure 2. Ratio of mean number in queue for correlated arrivals under various scenarios.

From these figures, we notice some interesting trends and observations, and are listed below.

1. In the case of renewal arrivals, we notice that

   (a) when the service times are Erlang (i.e., sE), the ratio decreases as p increases. However, the rate of decrease is smallest in the case of hyperexponential arrivals, namely, aH. We also observe that the sensitivity to the type of arrival process

reduces when p is close to 1. Further, the higher the variability in the arrival process, the smaller the values of the ratio is. The above observations appear to be true for the four values of ρ considered. Another interesting point to register is that the hyperexponential arrivals (i.e., aH) appear to have the ratio to be consistently less than 1.

(b) when the services are hyperexponential (i.e., sH), the trend is exactly the opposite of Erlang service case. Here, the ratio increases as p is increased. Note also how the ratios widen (between Erlang and hyperexponential arrivals) as ρ is increased from 0.5 to 0.99. Also, as p increases, the "gap" (the difference in the ratios between hyperexponential and Erlang arrivals) decreases to a point whether they are converge to 1. Further, the smaller the variability in the arrival process, the smaller the values of the ratio. This is direct opposites of what we noticed for Erlang services. The above observations appear to be true for the four values of ρ considered. An important point is that the ratios are all less than 1 and approach 1 as p is increased.

2. In the case of correlated arrivals, we observe that

(a) as p is increased, Erlang services provide a decreasing trend, whereas hyperexponential services provide a increasing trend. The decreasing trend is such that the ratio decreases from a value greater than 1 to 1; where as the increasing trend is from a value less than 1 to 1.

(b) while in the negatively correlated arrivals one sees the sensitivity of the ratios to the type of service time distribution whether the coefficient of variation is < 1 or = 1 or > 1, especially for $p \leq 0.8$, we can only say that the sensitivity appears to be whether the coefficient of variation is ≤ 1 or > 1 for the case of positively correlated arrivals. That is, for positively correlated arrivals, the ratio behaves similarly for Erlang and Exponential services.

5.2.3. Example 2:

The purpose of this example is to first search for the minimum value, μ^*, of μ, for which the mean number in the queue (or registry) of the model under study is close to the mean number in the queue of the corresponding classical queue. That is, find μ^* such that $\mu_{NQ} \simeq \mu_{CNQ}$, for all $\mu \geq \mu^*$. As pointed out earlier, θ depends on p as well as on μ if we have to compare our model to the corresponding classical queue. Thus, once we identify μ^* for a given ρ and p, we take θ as $\theta = \frac{\mu^*}{\rho\mu^* - p}$. In obtaining μ^*, we start with a value for μ, say, $\mu = 1$, and use a multiplier (d) which is greater than 1. The steps involved are as follows.

Step 0: $\mu \leftarrow 1$
Step 1: $\mu \leftarrow d\mu$, $\theta \leftarrow \frac{\mu}{\rho\mu - p}$.
Step 2: If $\hat{\rho}_q < 1$, go to Step 3. If not, go to Step 1.

Step 3: Compute μ_{NQ} and go to Step 4.
Step 4: If $|\mu_{NQ} - \mu_{CNQ}| < \epsilon$, then $\mu^* = \mu$; otherwise, $\mu \leftarrow d\mu$, and repeat Step 3.

Note: (a) The multiplier, d, is chosen such that μ^* can be found as close to the theoretical minimum as one needs. Noting that as μ increases μ_{NQ} decreases to μ_{CNQ} provided $p < 1$ and hence such a μ^* is guaranteed to exist. Further, it is obvious that d should be greater than 1; otherwise μ will decrease and therefore will not be able to identify μ^*.
(b) The quantity ϵ is a small positive number and here we chose this as follows.

$$\epsilon = \begin{cases} 10^{-1}, \ if \ \mu_{CNQ} \leq 1, \\ 10^{-3}, \ otherwise. \end{cases} \tag{30}$$

For this example, we set $\lambda = 1$, vary p from 0.1 to 0.99, consider two values for ρ, namely, $\rho = 0.50, 0.99$, and consider various combinations of arrival processes and service times. In searching for μ^* we set $d = 1.05$. Note that any value for d as long as it is greater than 1 will suffice but the degree of closeness to the theoretical value of μ^* depends on the choice of d. A value closer to 1 will result in more iterations. First, we display the values of μ^* in Figure 3.

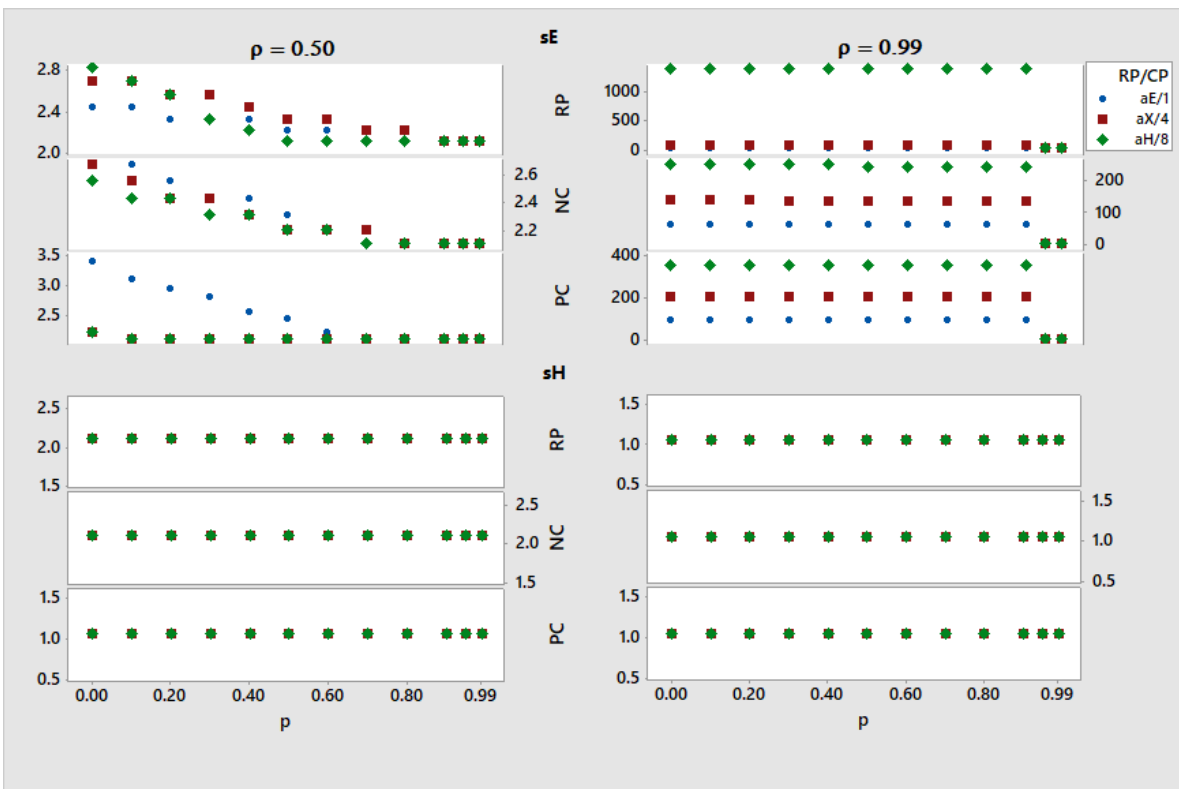

Figure 3. Minimum service rates (μ^*) under various scenarios.

Before we highlight some key findings, a few clarification on the legend in this figure. In the text body of the legend, aE, aX, aH are for RP while $1, 4, 8$ refers to $N1, N4, N8$

and $P1, P4, P8$, respectively, for NC and PC segments. A quick at this figure shows that for

1. Erlang services (sE)
   (a) one needs a larger μ^* as ρ is increased from 0.5 to 0.99;
   (b) the sensitivity to the type of arrival processes is seen in the case of small ρ for all $p \leq 0.90$;
   (c) the insensitivity to the type of arrival processes as well as to p is seen in the case of large ρ;
2. Hyperexponential services (sH)
   (a) we notice that μ^* decreases as ρ is increased from 0.5 to 0.99. This is exactly the opposite of what we noticed under sE services indicating that a higher variability in the services somehow neutralizes the need for a higher service rate. Note that in the case of PC arrivals, μ^* appears to be insensitive to ρ.
   (b) we notice the insensitivity to the type of arrivals and to p. This pattern is again quite different to the one seen under sE services.

In Figures 4, and 5, respectively, we display the graphs of the two measures, $P(busy)$ and $P(reach)$, under various scenarios.

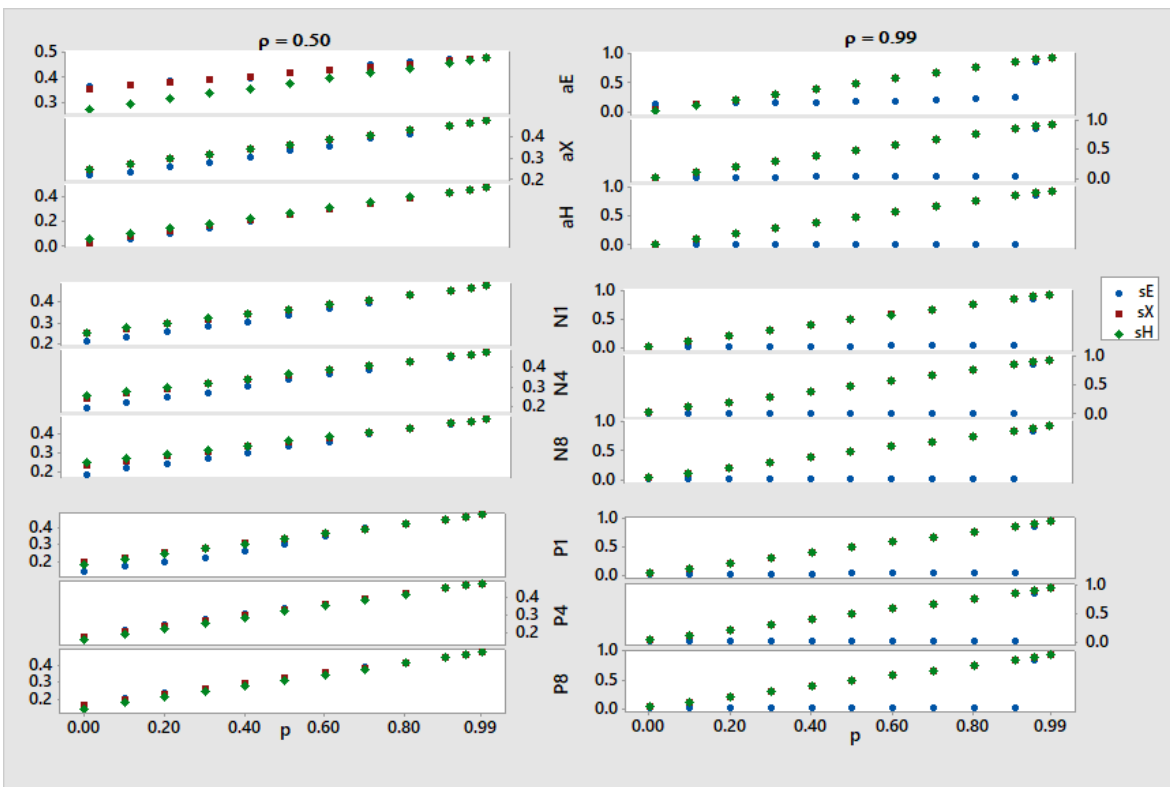

Figure 4. P(server is busy serving) under various scenarios.

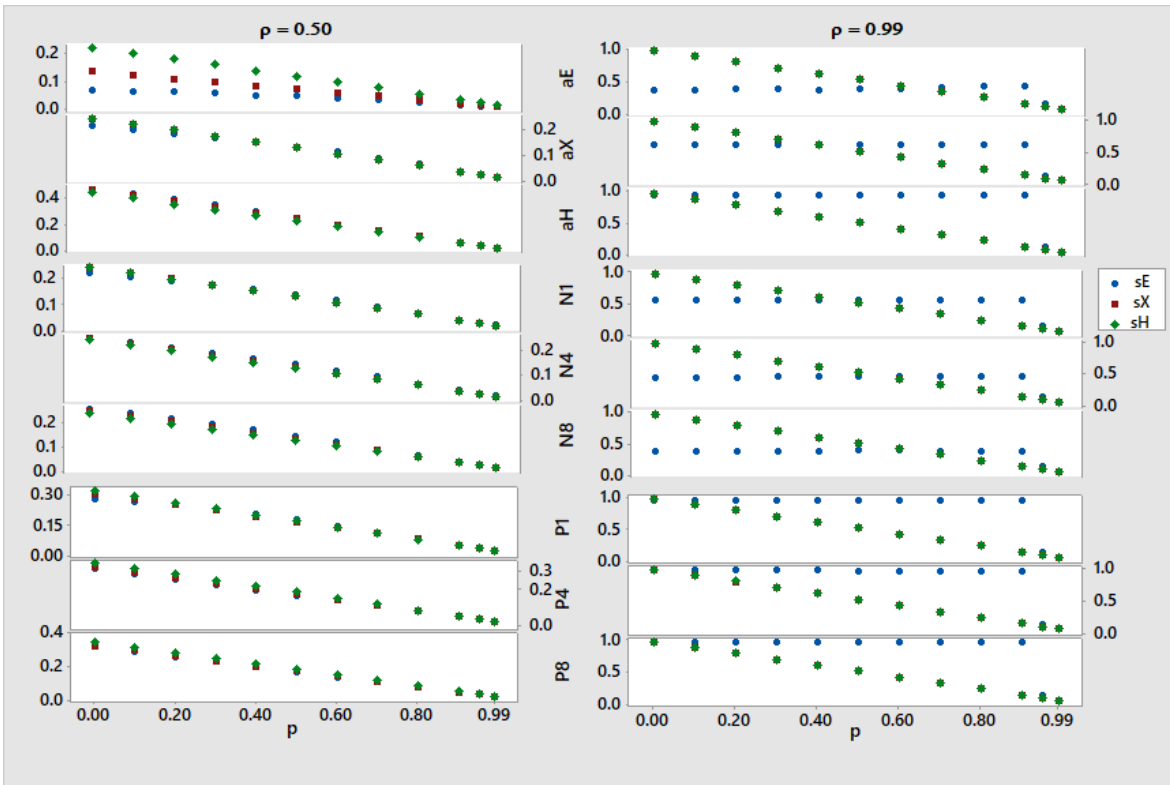

Figure 5. P(server is busy reaching out) under various scenarios.

A brief summary of observations based on these two figures is given below. First, we want to mention that these measures are calculated for the parameters set earlier for this example and for $\mu = \mu^*$. Thus, a common platform for comparison purposes is due to the fact that both these models pretty much have the same mean number in the queue.

1. With respect to $P(busy)$, we notice that

   (a) as p increases, a non-decreasing (an increasing for most scenarios) trend is seen under all scenarios.

   (b) the insensitivity to the type of services is seen for $p = 0.9$ through $p = 0.99$.

   (c) in the case of aH arrivals, we notice a high insensitivity to the type of services considered for low traffic intensity. However, for a high traffic intensity, the sensitivity to the type of services considered can be noticed.

   (d) for a high traffic intensity, sE and sH services yield different results in that sH services produce a larger value as compared with sE services. Further, the sensitivity to all values of p is seen for sE only.

2. With respect to $P(reach)$, we observe that

(a) in the case of low traffic intensity, the sensitivity to the type of services is seen only in the case of aE; for all other arrivals including CP, there appears to be no sensitivity to the type of services.

(b) in the case of high traffic intensity, there appears to be an interesting pattern. Except for PC case, where sE services appear to have a higher value as compared to sH services, there appears to be a cut-off point, say, p^*, such that for all $p < p^*$, sH services appear to have a higher value for the measure as compared to sE services and for other values of p, sE services have a higher value compared to sH services. The value of p^* depends on the type of arrivals considered.

5.3. Examples dealing with $MAP/M/c$-type model

In this section, we will discuss a few representative examples dealing with the same type of MAP arrivals considered in Section 5.2.

5.3.1. Example 3:

This example is similar to Example 1, in that we are comparing our model here to the corresponding classical $MAP/M/c$ under various scenarios using the mean number of customers in the queue. We use the same parameter values for λ, μ, and θ. In addition to these, we consider two values for $\rho = 0.5, 0.99$ and four values for $c = 1, 2, 4, 8$. In Figure 6, we display the ratio of the mean number of customers in queue, but for $c = 4$ and $c = 8$ cases. This is due to the fact that the ratios happen to be 1 for $c = 1$ and $c = 2$.

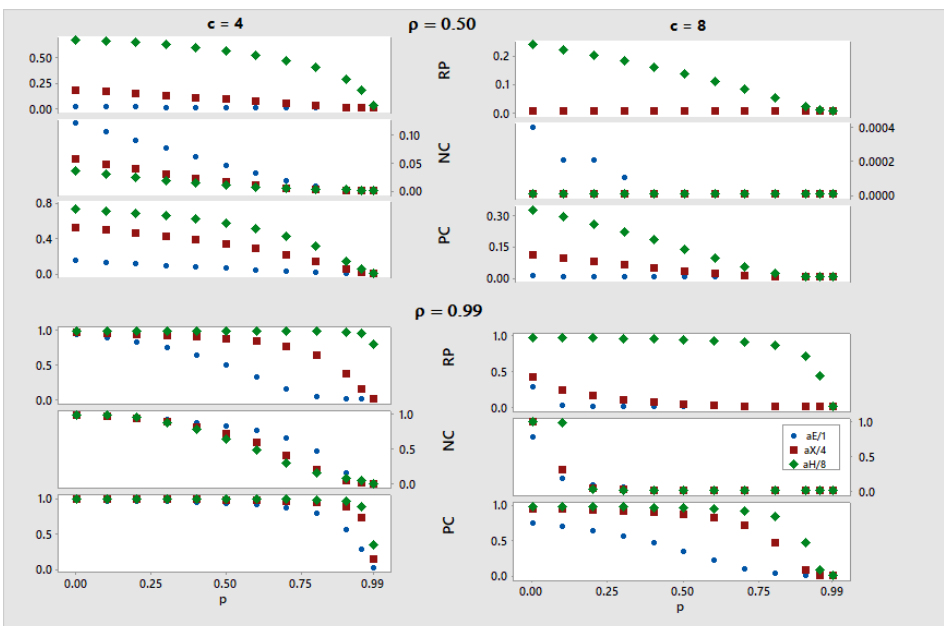

Figure 6. Ratio of mean number in queue under various scenarios.

It is clear from this figure that the ratios are such that

1. both arrival processes, aH (which has the highest variability among all others) and $P8$ (which has the highest positive correlation) have higher ratios compared to all other arrivals. This seems to be the case across all values of p. This indicates the roles of variability and (positive) correlation.

2. when comparing aH and $P8$ arrivals, we notice that the graphs of the ratios (as functions of p) cross each other. Initially, $P8$ arrival has a higher ratio up to a certain point, say, p^*, and then aH arrival has a higher ratio. This is the case for both $\rho = 0.50, 0.99$ and for $c = 4$. However, for $c = 8$, we see the ratios for these arrivals are such that for p up to a certain point there is no significant difference, and beyond this point aH has a higher value. This is the case for both ρ values.

3. a non-increasing (for most scenarios it is a decreasing) trend is seen under all scenarios as p is increased. It should be pointed out that as p is increased the values of θ increase and the rate of increase grows exponentially as p is close to 1 (recall that $\theta = \frac{\mu}{1-p}$).

4. for $\rho = 0.50$ the ratio is less than 1 under all scenarios. This can be intuitively explained as follows. When p is small, there is a high probability that the service facility will not have any customers to serve when reaching out to them, and further the stability condition is such that $\theta > \mu$, for all $p > 0$. [Note that when $p = 0$, the ratio will always be less than or equal to one depending on, respectively, whether $\theta > \mu$ or $\theta = \mu$]. When p is large but less than 1, there is still a positive probability that some customers will not be available (and hence not served) when reaching out to them, and furthermore, θ will have to very large to meet the stability condition resulting in the system having less customers (on the average) waiting to be served.

5.3.2. Example 4:

This example is similar to Example 2, in that we are seeking the minimum service rate, μ^*, so that the mean number in queue for the model under study is very close to the corresponding classical $MAP/M/c$ queue. The same criterion as mentioned in that example in getting μ^*. Similar to the previous examples, we fix $\lambda = 1$ and look at two values for $\rho = 0.50, 0.99$. Since we are dealing with a multi-server queueing system, the parameter c is taken as $c = 1, 2, 4$, and 8. In Figure 7, we display μ^* under various scenarios.

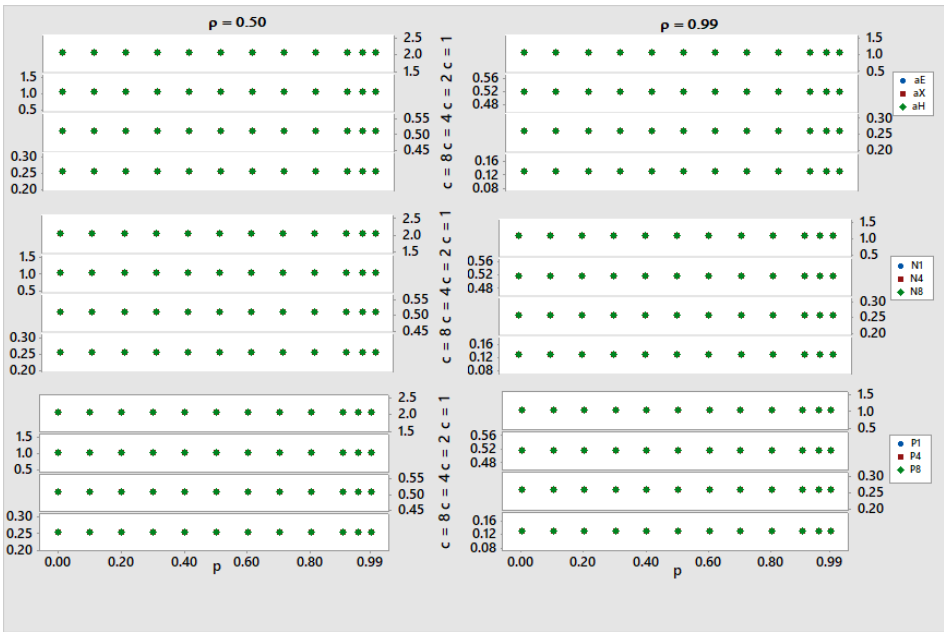

Figure 7. Minimum service rates under various scenarios.

It is clear from this figure that μ^* is insensitive to p under all scenarios. However, the values of μ^* depend on the number of servers as well as the type of arrival processes.

Next, we display $P(busy)$ for RP, NC, and PC, respectively, in Figures 8, 9, and 10.

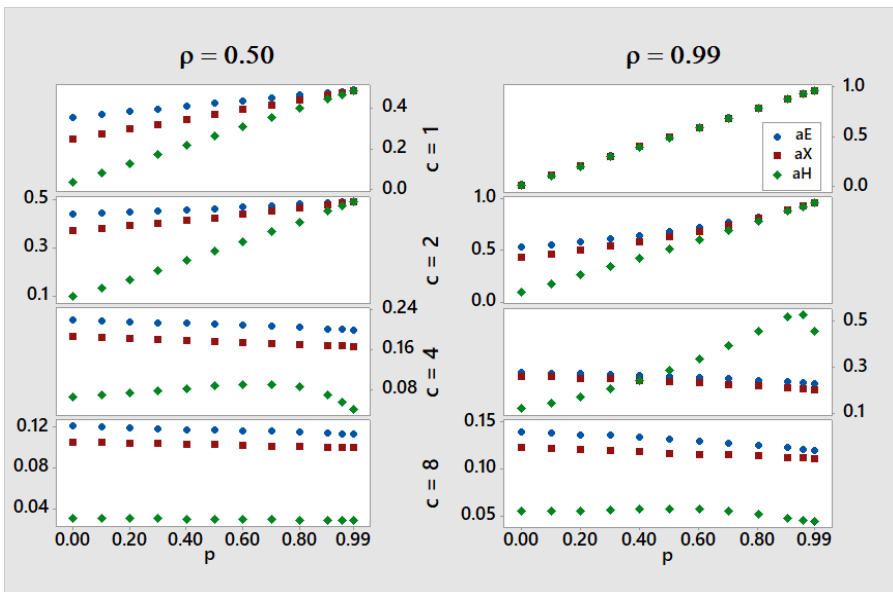

Figure 8. P(server is busy serving) under various scenarios for renewal arrivals.

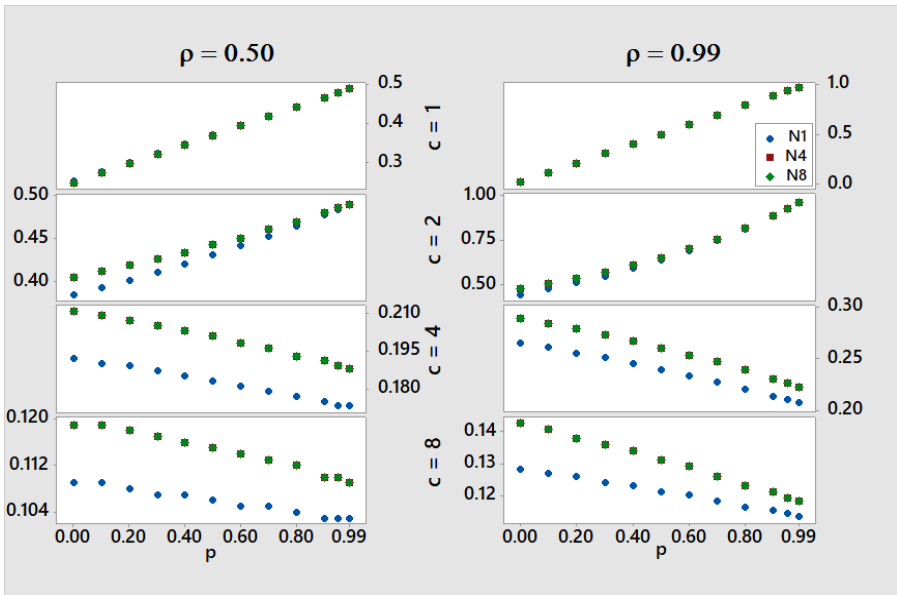

Figure 9. P(server is busy serving) under various scenarios for negatively correlated arrivals.

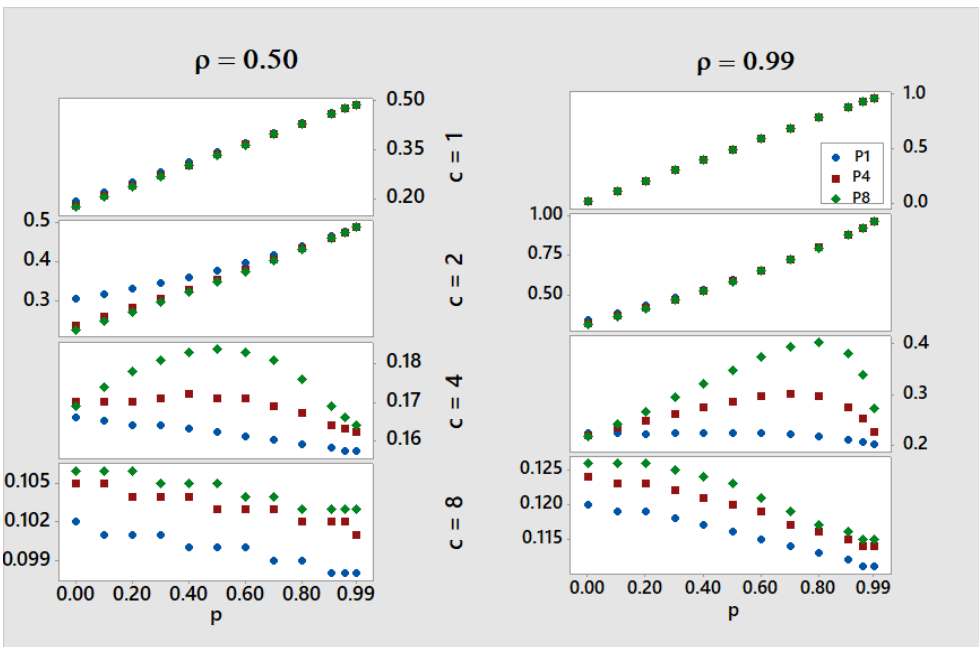

Figure 10. P(server is busy serving) under various scenarios for positively correlated arrivals.

Looking at these figures carefully, we notice some interesting observations on the measure, $P(busy)$, and are summarized below into three groups.

1. For renewal arrivals, the measure is such that

(a) the behavior is predominantly quite distinct as we go from $c = 1$ to $c = 8$. In the case of $\rho = 0.50$ we notice an increasing trend as p is increased for all three arrivals for both $c = 1$ and $c = 2$. Also, notice the insensitivity to the type of arrivals when p is close to 0.9 and higher. when $c = 1$. In the case when $c = 4$ we see a decreasing trend in p for both aE and aX arrivals, while for aH an increasing trend initially and then a decreasing trend as p is increased. The insensitivity to p is seen in all three arrivals when $c = 8$. This indicates that one needs a reasonably large c to see the insensitivity to p.

(b) for the case when $\rho = 0.99$, like in item (a) above, we do see a distinct behavior as c is increased. While for $c = 1$ and $c = 2$, we see an increasing trend as p is increased, the insensitivity to the type of arrivals is seen only in the case of $c = 1$. The behavior of this measure, for other cases, exhibit an interesting trend. For both aE and aX arrivals, a decreasing trend as p increases is seen; however, for aH arrivals, we notice an initial increasing and then a decreasing trend for both $c = 4$ and $c = 8$. The value of p at which the peak occurs depends on the value of c.

2. For negatively correlated arrivals, the measure is such that

(a) the behavior is predominantly quite distinct as we go from $c = 1$ to $c = 8$. In the case of $\rho = 0.50$ we notice an increasing trend as p is increased for all three arrivals, namely, $N1, N4,$ and $N8$, for both $c = 1$ and $c = 2$. Also, notice the insensitivity to the type of arrivals when $c = 1$. However, for $c = 4$ and $c = 8$, we see a non-increasing trend in p for all arrivals. An interesting thing is we notice the measures differ significantly between $N1$ and $N8$ arrivals as c is increased. However, $N4$ arrivals appear to have values of this measure almost coincide with those of $N8$.

(b) for the case when $\rho = 0.99$, like in item (a) above, we do see a distinct behavior as c is increased. While for $c = 1$ and $c = 2$, we see an increasing trend as p is increased as well as the insensitivity to the type of arrivals, the behavior in the case of $c = 4$ and $c = 8$ is exactly the opposite, namely, a decreasing trend in p, as well as the sensitivity to $N1$ and $N8$ (or $N4$) arrivals.

3. For positively correlated arrivals, the measure is such that

(a) the behavior is similar to the negatively correlated arrivals when $c = 1$ and $c = 2$ for both $\rho = 0.50$ and $\rho = 0.99$.

(b) for both cases $\rho = 0.5$ and $\rho = 0.99$, we notice similar patterns for $c = 4$ and $c = 8$. In the case when $c = 4$, we see initially an increasing trend and then a decreasing trend. In the case of $c = 8$, we notice a non-increasing trend. Further we see the sensitivity to the type of correlated processes under these combinations.

Finally, we display $P(reach)$ for RP, NC, and PC, respectively, in Figures 11, 12, and 13. As is to be expected this measure either stays constant or decreases as p is increased. An intuitive reasoning for a decrease in this measure as a function of p is as more and more customers are available at the time of the servers reaching out, the more likely the servers will start to become busy serving and hence less probability to be in "reach" mode. However, the interesting thing is the behavior of this measure varies from renewal arrivals to correlated arrivals. Specific key points are as follows.

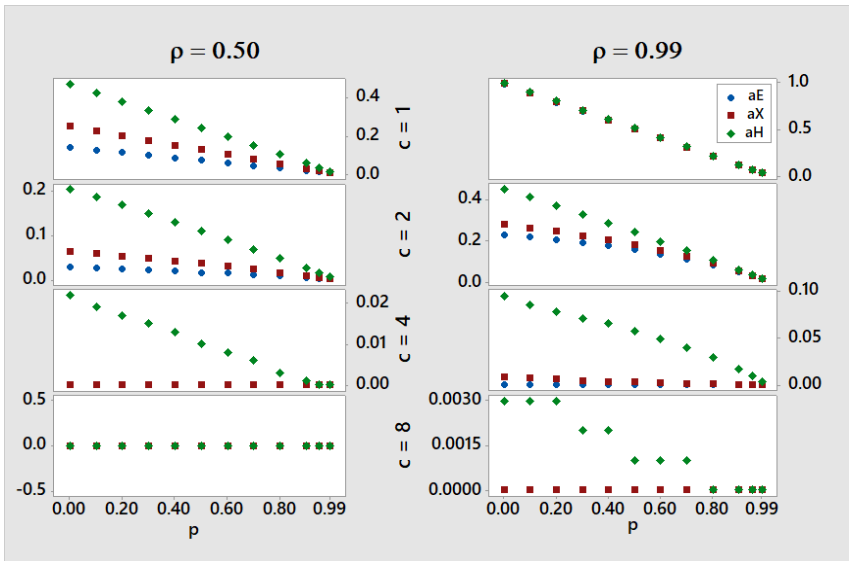

Figure 11. P(server is busy reaching) under various scenarios for renewal arrivals.

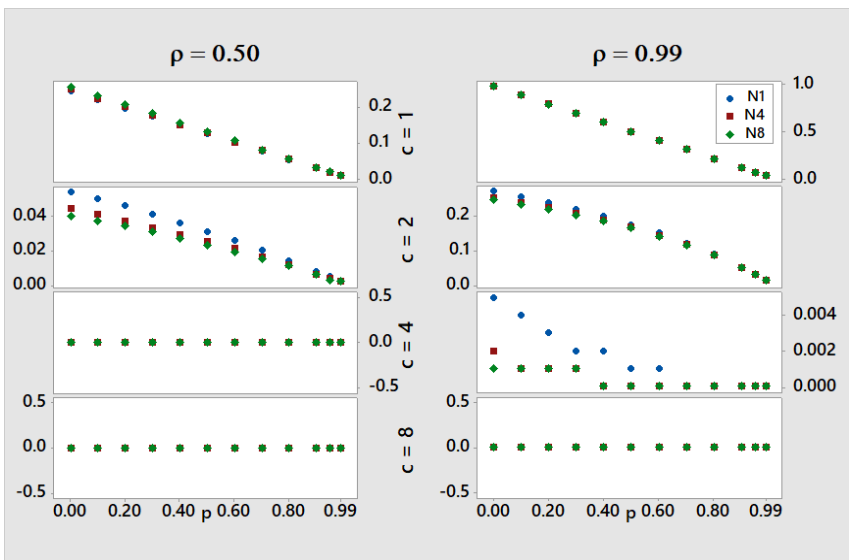

Figure 12. P(server is busy reaching) under various scenarios for negatively correlated arrivals.

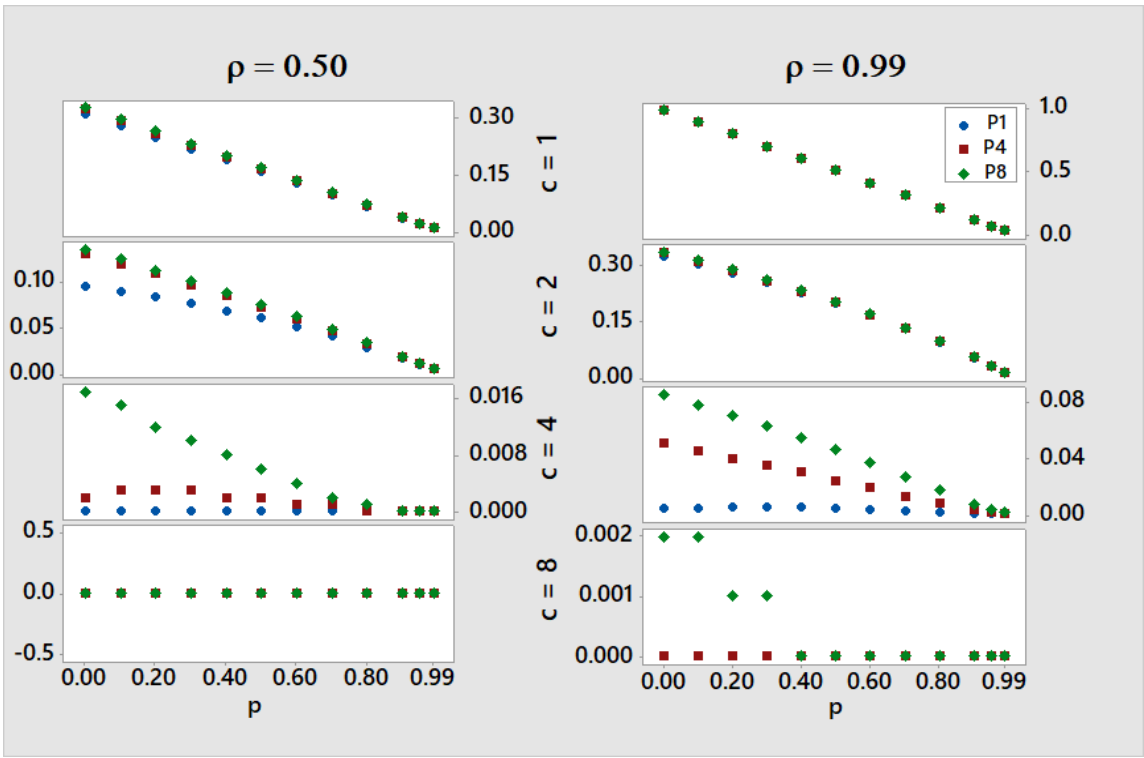

Figure 13. P(server is busy reaching) under various scenarios for positively correlated arrivals.

1. aH arrivals appear to have a higher rate of decrease as compared to the other renewal arrivals under many scenarios. Only when $c = 8$ and for $\rho = 0.50$, we notice an insensitivity to p for all three renewal arrivals. Further, for aE and aX arrivals, this measure is zero when c is increased to 4 and above.

2. When dealing with correlated arrivals, as the 1-lag correlation increases (from negative to positive values, i.e., going from N8 to P8) we notice a non-decreasing trend in $P(reach)$ under most scenarios. This indicates the role of correlation.

A few takeaways from the above set of examples for the models under study are as follows.

First, note that it is important from the customers' points of view that they get their services as quickly as they can upon their arrivals. However, from the service providers' points of view, resources are scarce and should be used very productively and hence to balance from both the customers and servers points of view, it is better to let the customers know, when they cannot be served immediately upon their arrivals, that they will be contacted soon after a free server is available, and that the customers' position in the queue (if there was actually a queue) will not change. This way the customers can tend to other errands without having to wait in non-value added activities. If a customer is not available at the time of the reaching out by the service provider, the customer will not feel that the quality of service is bad due to its own unavailability as opposed to the fault of the service provider.

The set of examples in the context of $MAP/PH/1$ as well as that of $MAP/M/c$ show that

1. even a small fraction of customers not available to receive services at the time of the service provider reaching out results in a better performance (with respect to the backlog in terms of the mean number of customers waiting in the queue).

2. the trends in the (ratio) of the mean number of customers as a function of p are opposite to each others when dealing with Erlang services and Hyperexponential services. While we have seen a large increase in the mean number in the queue when going from Erlang to hyperexponential services, this is the first time we are seeing trends that are exact opposites to each other.

3. from both customers' and service provider's points of view having a "reach" out approach to provide services have satisfactory results in that customers benefit from tending to their other activities without having to worry about loosing their positions in the queue; the service provider benefits from the fact that the customers have less non-value added times in their waiting times even though the service provider has to incur some additional time in the form of reaching out to the customers. It is worth mentioning that a similar observation has been reported in [1, 2].

6. Simulation Approach

In this section, we will briefly discuss a few examples out of many simulated in the context of $MAP/G/c$-type call-back queues. The servers are assumed to be homogeneous. Towards this end, we used ARENA, a powerful and most commonly used simulation software in academia and industry (see, e.g., [11, 14]). We look at both constant and Weibull services. For Weibull, we looked at the 2-parameter family which has the probability density function given by

$$F(t) = \begin{cases} 1 - e^{-\sqrt{2\mu t}}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

The simulated models were validated against the analytical models first. Then, we ran a number of examples under a variety of scenarios for 250,000 units of time using five replicates. A pictorial description of the simulation model in ARENA is displayed in Figure 14.

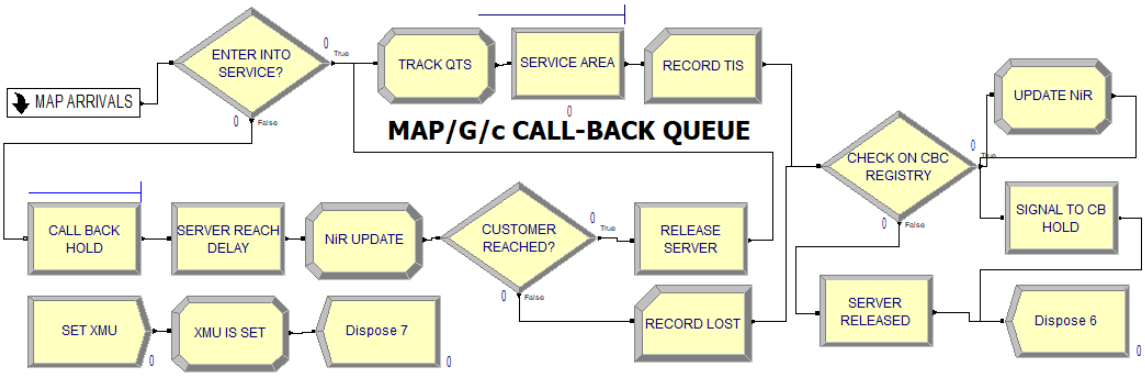

Figure 14. ARENA simulation model for $MAP/G/c$ queues.

For the examples summarized below, we fix $\lambda = 1, \mu = \frac{1.5}{c}, \theta = 5$, and consider various scenarios by varying (i) $c = 1, 2, 4, 8$, (ii) $p = 0, 0.50, 0.95$, (iii) the arrival process to be $aE, aH, N1, P1$, and (iv) the service times to be constant and Weibull. In Figure 15, we display the two measures, the probability that the server is busy serving a customer and the probability that the server is busy reaching out, under a variety of scenarios. It is worth pointing out that identifier used in the radar chart. The first letter is to identify the type of arrival process, the second is for the service times, and third one is for the value of p. Thus, "$N\ C\ 0.5$" in the figure indicates the plotted value is for $N1$ arrivals having a constant service time and the value of p is set at 0.5; similarly "$P\ W\ 0$" is for $P1$ arrivals having Weibull services and the value of p is 0. The reach out time is assumed to be constant with a mean given by $\frac{1}{\theta}$.

Looking at this figure corresponding to the two measures, we notice (a) as c is increased, the probability that the server is busy serving customers increases. This might seem to be counter-intuitive at first; however, on noticing that having multiple servers helps the system to accept customers at their arrival points as opposed to capturing them through reach out durations. This is the case for all four arrival processes and for the two services considered; (b) as p is increased, the probability that the server is busy serving shows a non-decreasing trend under all scenarios. This is to be expected; (c) as c is increased, the probability that the server is busy reaching out shows a non-increasing trend. This is to be expected and also appears to be the case under all scenarios; (d) as p is increased, the probability of the server reaching out shows a non-decreasing trend. This behavior was observed in our earlier examples; and (e) the sensitivity to the type of services (constant vs Weibull) used.

A key takeaway from the above simulated example is that the results appear to hold good for more general services like constant and heavy tailed distribution. Further, this also gives confidence in the simulation models which can be explored in more detail covering a wide variety of arrival, service, and reach out distributions.

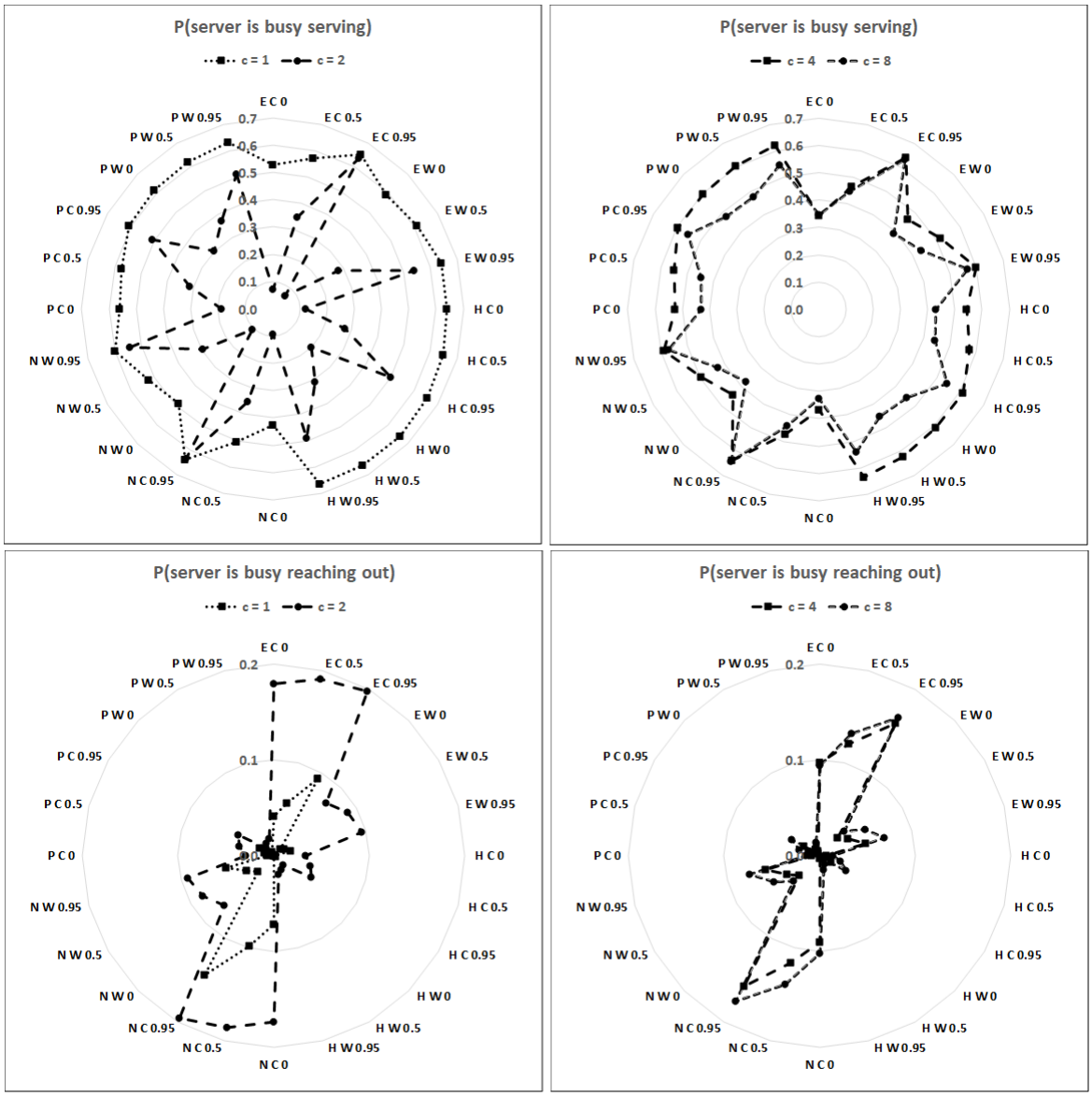

Figure 15. P(server is busy serving) and P(server is busy reaching out) under various scenarios.

7. Concluding Remarks

In this paper, queueing models in the context of $MAP/PH/1$ and $MAP/M/c$ useful in service industries are studied analytically, and in the case of $MAP/G/c$, we resorted to simulation. Illustrative numerical examples are provided to study the behavior of various performance measures. The models studied in this paper can be generalized in several ways. First, we can relax the assumption of striking out the customers who do not answer the service calls at the first instant by allowing a finite number of attempts to reach such customers. Secondly, an exhaustive simulation approach similar to the ones studied in [8, 10] can be ap-

plied under a variety of scenarios including batch arrivals. Thirdly, having a common timer for all registered customers or individual timer for each registered customer to model the reach out mechanism would be interesting to study. Finally, one can introduce vacationing of the server(s) as well as failures and repairs of the server(s). These are currently being investigated and the results will be reported elsewhere.

Acknowledgments: The comments from the anonymous referees that improved the presentation of the paper are greatly appreciated by the author.

References

[1] Armony, M., & Maglaras, C. (2004). On Customer Contact Centers with a Call-Back Option: Customer Decisions, Routing Rules, and System Design. *Operations Research*, 52, 271-292.

[2] Armony, M., & Maglaras, C. (2004). Contact Centers with a Call-Back Option and Real-Time Delay Information. *Operations Research*, 52, 527-545.

[3] Artalejo, J. R. (1999). Accessible Bibliography on Retrial Queues. *Mathematical and Computer Modelling*, 30, 1-6.

[4] Artalejo, J. R., & Gomez-Corral, A. (2008). Retrial Queueing Systems: A Computational Approach. Springer-Verlag, Berlin, Heidelberg.

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

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

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

[8] Chakravarthy, S. R. (2013). Analysis of $MAP/PH/c$ retrial queue with phase type retrials - Simulation approach. In: A. Dudin et al. (Eds.): BWWQT 2013, CCIS 356, 37-49.

[9] Chakravarthy, S. R. (2015). Matrix-Analytic Queueing Models, Chapter 8 in "An Introduction to Queueing Theory", U. Narayan Bhat, Second Edition, Birkhauser, Springer Science + Business Media, New York.

[10] Chakravarthy, S. R. (2020). Queueing Models in Services - Analytical and Simulation Approach. To appear in Advanced Trends in Queueing Theory. In: Prof.Vladimir Anisimov and Prof.Nikolaos Limnios. (Eds.): Series of books "Mathematics and Statistics", Sciences, ISTE & J. Wiley, London.

[11] Dias, L. M. S., Vierira, A. A. C., Pereira, G. A. B., & Oliveira, J. A. (2016). Discrete Simulation Software Ranking - A Top List Of The Worldwide Most Popular And Used Tools. Proceedings of the 2016 Winter Simulation Conference, Eds: T. M. K. Roeder, et al., 1060-1071.

[12] Dudin, A. N., Kim, C., Dudina, O., & Dudin, S. (2016). Multi-server queueing system with a generalized phase-type service time distribution as a model of call center with a call-back option. *Annals of Operations Research*, 239, 401-428.

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

[14] Kelton, W. D., Sadowski, R. P., & Swets, N. B. (2010). Simulation with ARENA, Fifth ed., McGraw-Hill, New York.

[15] Kim, C., Dudin, A. N., Dudina, O., & Dudin, S.(2012). Queueing System MAP/M/N as a Model of Call Center with Call-Back Option. In K. Al-Begain, et.al. (Eds.): ASMTA 2012, LNCS 7314, 1-15.

[16] Kim, J., Kim, B. (2016). A survey of retrial queueing systems. *Annals of Operations Research*, 247, 3-36.

[17] Latouche, G., & Ramaswami, V. (1999). Introduction to matrix analytic methods in stochastic modeling. SIAM.

[18] Lucantoni, D. M., 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.

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

[20] Marcus, M., & Minc, H. (1961). A Survey of Matrix Theory and Matrix Inequalities. Allyn and Bacon, Boston, MA.

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

[22] Neuts, M. F. (1981). Matrix-geometric solutions in stochastic models: An algorithmic approach. The Johns Hopkins University Press, Baltimore, MD. [1994 version is Dover Edition].

[23] Neuts, M. F., & Ramalhoto, M. F. (1984). A service model in which the server is required to search for customers. *Journal of Applied Probability*, 21, 157-166.

[24] Neuts, M. F. (1989). Structured stochastic matrices of $M/G/1$ type and their applications. Marcel Dekker, Inc., New York.

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

[26] Neuts, M. F. (1995). Algorithmic Probability: A collection of problems. Chapman and Hall, NY.

[27] Phung-Duc, T. (2017). Retrial queueing models: A survey on theory and applications. In: T. Dohi,et.al.(eds.): Stochastic Operations Research in Business and Industry, Singapore: World Scientific, [Online]. Available: http://infoshako.sk.tsukuba.ac.jp/ tuan/papers/ Tuan_chapter_ver3.pdf

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

[29] Stewart, W. J. (1994). Introduction to the Numerical Solution of Markov Chains. Princeton, NJ: Princeton University Press.