Modelling dependence between interarrival and service times with Markovian arrival processes

M.B. Combé

Department of Operations Research, Statistics, and System Theory

Report BS-R9412 April 1994

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

Modelling Dependence between Interarrival and Service Times with Markovian Arrival Processes

M.B. Combé
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

This paper discusses dependence between the interarrival and service time of a customer. Such a dependence structure naturally arises in queueing systems, for example it can be the result of collection or reservation mechanisms for messages in communication networks. The present paper shows how this type of dependence can be constructed using Markov Modulated Arrival Processes; in particular the framework of the Matrix geometric method is applied by using the Batch Markovian Arrival Process. Earlier work on dependence between interarrival and service times of customers is extended in several directions.

AMS Subject Classification (1991): 60K25, 90B22
Keywords & Phrases: dependence, batch Markovian arrival process
Note: The author was supported by NFI

1 INTRODUCTION

In early queueing models the characteristics of the arrival and service time distributions are constant over time. Hence those queueing models have limited ability to adequately model the typical traffic characteristics of modern communication networks, such as bursty arrival processes and time varying arrival rates.

In the last few decades, analysis concentrating on the typical properties of queueing systems in communication networks has led to the development of the theory of the Markov Modulated Queueing System (MMQS). In an MMQS a Markov process describes the time varying behaviour of some of the arrival and service processes. The state of this Markov process contains information about the parameters of the current arrival process and the service time distributions of customers in the system, or it may contain information about current system characteristics, like the speed of the server. For example, the arrival rate of customers generated by an On/Off source may well be described by a Markov process, some states representing On phases of the source, the other states representing the Off phases.

In general, the time varying behaviour of the arrival process and service characteristics in a queueing system is reflected by a number of dependence structures in the sequences of interarrival and service times of customers. In Fendick et al.[4] three types of dependence are mentioned: between consecutive interarrival times, between consecutive service times, and between the interarrival and service time of a customer.

Report BS-R9412
ISSN 0924-0659
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

In the past most attention has been paid to describing dependence between consecutive interarrival times and between consecutive service times of customers; dependence between interarrival and service times has been less extensively treated. The purpose of this paper is to show how an MMQS can be used to model queueing systems with the latter dependence structure. In particular we explore the power of the BMAP (Batch Markovian Arrival Process) for modelling and analysing queueing systems with dependence between the interarrival process and the service times. The main reason for using the BMAP is that the well analysed $BMAP/G/1$ queue, the single server queue with the BMAP as arrival process, allows convenient numerical procedures for most performance measures of interest.

Correlation between the interarrival and service time of a customer naturally arises in queueing systems. Kleinrock[5] discusses the traffic characteristics of message flows in a message switching communication network. In the corresponding queueing network the service times of messages at each queue are proportional to the message length. For example, in a tandem configuration successive service times of the same message are correlated, hereby creating a positive correlation between interarrival and service times at the second queue. A second example arises in bridge queues between communication networks. Every now and then, the messages in a network that are destined for other networks are collected and delivered at a bridge queue. In this situation the time between two consecutive arrivals of batches of collected messages at a bridge queue and the number of messages collected is (positively) correlated. This type of correlation structure has been studied by Borst et al.[2, 3] and Bisdikian et al.[1]. For a survey of literature on correlated interarrival and service times we refer to [3].

Overview of the paper.
In section 2 we describe how an MMQS can be used to model queueing systems with correlation between the interarrival and service time of a customer. Section 3 illustrates the modelling technique with a number of examples. Section 4 numerically shows the potential of the modelling technique to obtain performance measures and insights for queueing systems with dependence between interarrival and service times. In section 5 we consider queueing models in which various dependence structures occur simultaneously.

2 Modelling dependence between interarrival and service times with MMQS

In this section we concentrate on modelling correlation between interarrival and service times of customers with the use of the Batch Markovian Arrival Process (BMAP). First we describe the characteristics of the BMAP.

2.1 The Batch Markovian Arrival Process.

In the BMAP, the interarrival and service times of customers are directed by a continuous time Markov process $\{\mathbf{J}(t), t \geq 0\}$ on a finite state space E. A transition from a state i to a state j may induce the arrival of a *batch customer*, the size of the batch depending on i and j. Let the sojourn time in state $i \in E$ be exponentially distributed with parameter $\lambda_i > 0$, and given that a transition takes place, let p_{ij} be the probability of a transition from i to a state $j \in E\backslash\{i\}$, $\sum_{j \in E\backslash\{i\}} p_{ij} = 1$. Defining $p_{ii} = -1$, then the generator of the Markov process

$\{\mathbf{J}(t), t \geq 0\}$ is given by the matrix $D = (p_{ij}\lambda_i)$. Next, conditional on a transition from a state i to a state j, the probability of a batch arrival of size k is denoted by q_{ij}^k, $k = 0, 1, \ldots$, where q_{ij}^0 may be interpreted as the probability of having no arrival or of the arrival of an empty batch. With this notation, $p_{ij}q_{ij}^k\lambda_i$ is the transition rate from state i to state j inducing a batch arrival of size k. The BMAP is completely characterized by the sequence of matrices $D_k = \left(p_{ij}q_{ij}^k\lambda_i\right)$, $k = 0, 1, \ldots$, $\sum_k D_k = D$. The BMAP is a special variant of the general Markov modulated arrival process; in the latter case the *service time* distribution depends on the transitions in $\{\mathbf{J}(t), t \geq 0\}$.

The $BMAP/G/1$ queue is defined as the single server queue with the BMAP describing the arrival process of batch customers, where batches are served in FCFS order, and the service times of single customers are independent and identically distributed with a general distribution function $H(\cdot)$. The $BMAP/G/1$ has been studied in Lucantoni[6], in which one can find results for most of the performance measures of interest, such as the number of customers, waiting times, and the busy period.

2.2 Modelling dependence between interarrival and service times.

A key observation of the paper is that dependence between interarrival and service times can be modeled by viewing $\{\mathbf{J}(t), t \geq 0\}$ as a two-dimensional Markov process, $\{(\mathbf{J}_1(t), \mathbf{J}_2(t)), t \geq 0\}$. $\mathbf{J}_1$ generates the arrivals of batches, and the state of $\mathbf{J}_1$ contains information about the remaining interarrival time; if the state of $\mathbf{J}_1$ is j_1, this might for example imply that the remaining interarrival time consists of j_1 exponential phases. The component $\mathbf{J}_2$ contains information about the batch size distribution; for example, the state of $\mathbf{J}_2$ might stand for the number of customers in a batch (here and in the remainder of the paper we use the convenient notation $\mathbf{J}_i$ for $\{\mathbf{J}_i(t), t \geq 0\}$, i=1,2).

The key idea is to let the interarrival time, the time in $\mathbf{J}_1$ between two batch generating epochs, be the time parameter in $\mathbf{J}_2$. So, as the time between two arrivals goes by, the batch size distribution is described by the evolution of $\mathbf{J}_2$. The state of $\mathbf{J}_2$ just before a batch arrival contains information about the batch size.

A typical example of this mechanism is given by the Compound Poisson Process (CPP), which can be viewed as a special batch arrival process. The interarrival time of a batch is exponentially distributed, the size of a batch is the number of customers generated by a second Poisson process during that interval. In this example, $\mathbf{J}_1$ is a one state Markov process, describing the exponentially distributed interarrival times of batch customers, $\mathbf{J}_2$ is a Markov chain with state space $\{0, 1, \ldots\}$, describing the current size of the batch, i.e., the number of customers that have arrived in the second Poisson process since the last batch arrival.

In the next section we show the potential of this construction by elaborating on a queueing model with the CPP as the arrival process.

Consider the following model. At a bus stop, customers arrive according to a Poisson process with rate λ. According to a second Poisson process with rate γ, a bus visits this bus stop, collects the waiting customers, and delivers the customers as a batch at a single server service facility. With this procedure, the number of customers in a batch is positively correlated with

the interarrival time of batches at the service facility. The queueing model arising from this collecting or reservation mechanism, an $M/G/1$ queue with a typical dependence between interarrival and service times, has been studied in Borst et al.[3]. In [3] results are presented for various performance measures of interest, such as the waiting time, the sojourn time, and the number of customers.

3 Exploring the model

In this section we use the BMAP for modelling a number of variants (and generalisations) of the collector model that we described in the previous section.

• *Variant 1: Finite bus capacity.*

In Borst et al.[3] the number of customers in the bus can be arbitrarily large. However, the method used in [3] to analyse this case does not seem to be applicable for the natural variant where the number of customers in the bus can not exceed M. With the BMAP, there are two ways of modelling this restriction.

-First approach. If the number of customers at the bus stop has reached M, all future arrivals are rejected until the bus has collected the M customers at the stop. This is an example of the most pure form of the construction; the Markov process $\mathbf{J}_1$ only describes the phase of the interarrival process, the Markov process $\mathbf{J}_2$ only describes the number of customers in the bus, and the interaction of the two chains is limited to the resetting of $\mathbf{J}_2$ to the state corresponding to an empty bus stop, at the moment at which $\mathbf{J}_1$ generates a batch arrival. The Markov chain of $\mathbf{J}_1$ has a single state because the interarrival time is exponentially distributed, the Markov chain of $\mathbf{J}_2$ has state space $E = \{0, 1, \ldots, M\}$, each state representing the number of customers at the bus stop. In this second Markov chain M is an absorbing state.

Note that the arrival process of the busses is still a Poisson process.

In figure 3.1 the transition rate diagram for $(\mathbf{J}_1, \mathbf{J}_2)$ is presented for the case $M = 3$. Figure 3.2 shows the generator matrix D and the matrices $D_0, \ldots, D_3$ it decomposes into. In figure 3.1, the nodes are numbered $0, \ldots, 3$, representing the states of $\mathbf{J}_2$.

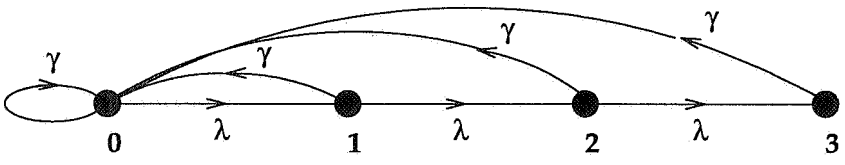

FIGURE 3.1. Transition rate diagram for the collector model with finite bus capacity and blocking of customers.

$$D = \begin{bmatrix} -\lambda & \lambda & 0 & 0 \\ \gamma & -(\lambda+\gamma) & \lambda & 0 \\ \gamma & 0 & -(\lambda+\gamma) & \lambda \\ \gamma & 0 & 0 & -\gamma \end{bmatrix}, D_0 = \begin{bmatrix} -\lambda & \lambda & 0 & 0 \\ 0 & -(\lambda+\gamma) & \lambda & 0 \\ 0 & 0 & -(\lambda+\gamma) & \lambda \\ 0 & 0 & 0 & -\gamma \end{bmatrix},$$

$$D_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, D_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, D_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 \end{bmatrix}.$$

Figure 3.2: Generator matrices for the collector model with finite bus capacity and blocking of customers.

For the $BMAP/G/1$ queue with such generator matrices and a general distribution function $H(\cdot)$ for the service time of a single customer, analytical results and numerical evaluation procedures are presented in Lucantoni[6] for many performance measures of interest, such as the waiting time, the queue length and the busy period. For $M \to \infty$ the $BMAP/G/1$ queue reduces to the model studied in Borst et al.[3]. As a particular result of this, we are able to approximate higher moments of the busy period in the latter model as accurately as we wish. In [3] we were only able to obtain the first moment of the busy period.

-Second approach. Here we let $\mathbf{J}_1$ generate batch arrivals, but we also have the bus visit the bus stop when the number of customers at the bus stop reaches M. Hence every time $\mathbf{J}_2$ is in state $M-1$, the next transition, bus or customer, generates a batch arrival. For this model, for the case $M = 3$, figures 3.3 and 3.4 show the transition rate diagram for $(\mathbf{J}_1, \mathbf{J}_2)$ and the matrices $D_0, D_1, \ldots, D_M$ respectively.

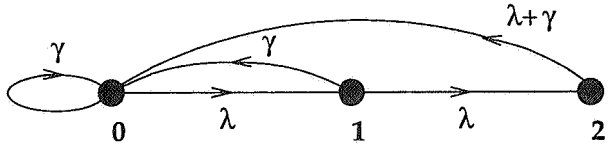

Figure 3.3: Transition rate diagram for the collector model with finite bus capacity and no blocking of customers.

$$D = \begin{bmatrix} -\lambda & \lambda & 0 \\ \gamma & -(\lambda+\gamma) & \lambda \\ \lambda+\gamma & 0 & -(\lambda+\gamma) \end{bmatrix}, D_0 = \begin{bmatrix} -\lambda & \lambda & 0 \\ 0 & -(\lambda+\gamma) & \lambda \\ 0 & 0 & -(\lambda+\gamma) \end{bmatrix},$$

$$D_1 = \begin{bmatrix} 0 & 0 & 0 \\ \gamma & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \gamma & 0 & 0 \end{bmatrix}, D_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \lambda & 0 & 0 \end{bmatrix}.$$

Figure 3.4: Generator matrices for the collector model with finite bus capacity and no blocking of customers.

REMARK 3.1 In the first example, the batch size of an arrival is equal to the state of $\mathbf{J}_2$ at the moment of an arrival. In the second example, a batch arrival may be caused by the arrival of a bus or by the arrival of the $M-th$ customer at the bus stop. Hence, when a batch arrival occurs while $\mathbf{J}_2$ is in state $M-1$, the batch size is with probability $\frac{\gamma}{\lambda+\gamma}$ equal to $M-1$, and with probability $\frac{\lambda}{\lambda+\gamma}$ equal to M. In general, the state of $\mathbf{J}_2$ may describe a batch size *distribution*, rather than just being the number of customers in a batch. □

• *Variant 2: General arrival processes.*

In the original collector model of Borst et al.[3] the arrival process of customers at the bus stop and the arrival process of busses are both Poisson. For more general arrival processes it seems hard to extend the approach of [3] to obtain analytical results. However, when the interarrival times of busses and customers are of semi-Markov type, the BMAP modelling can provide approximations for the case of infinite bus capacity, and exact results for the case of finite bus capacity. Here we remark that most results on the $BMAP/G/1$ theoretically seem to extend for the case of a countable underlying Markov chain. However, numerical methods available are mainly developed for the case of a finite underlying Markov chain.

As an example of more general arrival processes we present the case in which the interarrival times of busses are Erlang-2 distributed, the customers arrive according to a Poisson process at the bus stop, and customers are rejected once the bus is full. Figure 3.5 shows the corresponding transition rate diagram for the case $M = 3$. The nodes $(j_1, j_2) \in \{0,1\} \times \{0,1,2,3\}$ in figure 3.5 represent the state of $(\mathbf{J}_1, \mathbf{J}_2)$, $\mathbf{J}_1$ being the state of the bus interarrival process, $\mathbf{J}_2$ representing the number of customers at the bus stop.

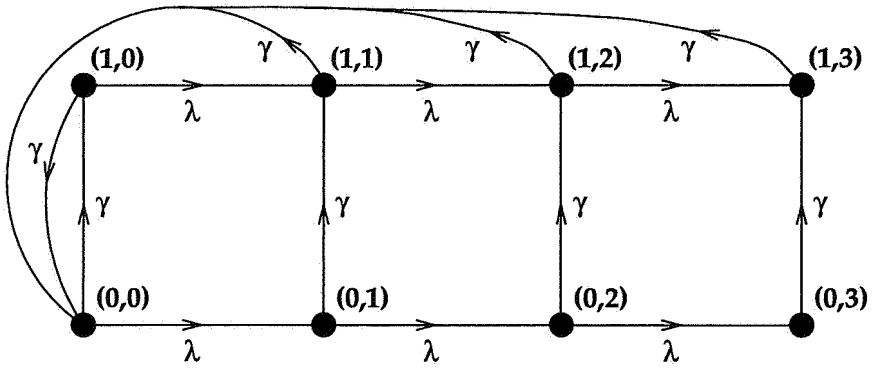

Figure 3.5: Transition rate diagram for the collector model with Erlang-2 distributed interarrival times of batches.

• *Variant 3: Other types of correlation.*

The above mentioned variants of the model with customer collection all consider a positive correlation between the interarrival and service time of a batch customer. More specific, the batch size distribution is stochastically non-decreasing as a function of the interarrival time. In particular, the Markov chain in the first example (cf. figure 3.1) is related to a birth process. An obvious extension is to construct a negative correlation between interarrival and service times, for example by letting the batch size distribution be stochastically non-increasing as a function of the interarrival time. A second extension is to model customer

behaviour at the bus stop; customers may for example leave the bus stop after a while when the bus is 'late'. This queueing model with impatient customers can be adequately modeled with Markov processes $\mathbf{J}_1$ and $\mathbf{J}_2$.

4 NUMERICAL RESULTS

In this section we numerically illustrate the possibilities of our model. We show how infinite collector models might adequately be approximated by a $BMAP/G/1$ queue. We pay attention to the effect of the maximum batch size and we also investigate the influence of the distribution of the intercollecting interval on performance measures of the queueing model. The performance measures considered are the waiting time and the busy period length.

We have restricted ourself to a few examples, because numerical exploration of $BMAP/G/1$ procedures is not the main purpose of this paper. Lucantoni[6] presents both for batches as well as for individual customers many results and numerical procedures.

4.1 Approximating infinite bus capacity.

While discussing variant 2 we remarked that most numerical methods for the $BMAP/G/1$ queue are mainly for the case of a finite underlying Markov chain. In table 4.1 we examine various performance measures as a function of the dimension of this Markov chain for the case of Poisson arrivals of customers and exponentially distributed intercollecting intervals (variant 1). For $M \to \infty$ exact results are adopted from Borst et al.[3]. Unforced collect and forced collect respectively are the first and second approach of variant 1. In the example, the arrival rate λ of customers at the bus stop is 1, the mean service time of a customer is 0.5, and the intercollecting distribution has parameter 1. For this collector model we distinguish between two types of busy periods: busy periods $\mathbf{B}$ that also include busy periods of length 0, i.e. busy periods started by empty busses, and busy periods $\mathbf{B}'$, for which only strictly positive busy periods are taken into account.

We notice that for moderate M the values of the mean waiting times EW and mean busy period length EB' are already reasonably close to these values for $M = \infty$. The difference for M small is explained by the fact that for unforced collect not all customers arriving at the bus stop will receive service; when $\mathbf{J}_2 = M$, newly arriving customers are rejected. In connection with remark 3.1 we note that this might be avoided by having a batch size *distribution* rather than fixing the batch size to M for $\mathbf{J}_2 = M$.

	Exponential service				Deterministic service			
	unforced collect		forced collect		unforced collect		forced collect	
M	EW	EB'	EW	EB'	EW	EB'	EW	EB'
5	0.594280	1.507388	0.630328	1.527809	0.398027	1.454060	0.417376	1.441434
6	0.639372	1.554330	0.657011	1.562283	0.436232	1.490878	0.445759	1.486589
7	0.666197	1.578507	0.674647	1.581514	0.459451	1.511221	0.464061	1.509814
8	0.681677	1.590791	0.685685	1.591904	0.473076	1.522049	0.475293	1.521601
9	0.690428	1.596987	0.692327	1.597393	0.480881	1.527681	0.481952	1.527540
10	0.695301	1.600099	0.696207	1.600245	0.485275	1.530565	0.485796	1.530522
20	0.701151	1.603123	0.701182	1.603130	0.490637	1.533424	0.701182	1.533416
∞	0.701155	1.603122	0.701155	1.603122	0.490648	1.533418	0.490648	1.533418

Table 4.1: Mean waiting times and mean busy period lengths as a function of the bus capacity M.

4.2 The distribution of the collecting interval.

Figure 4.1 illustrates variant 2. Mean busy period lengths and mean waiting times are presented as functions of the coefficient of variation of Erlang distributed intercollection times. The mean collecting interval is 1, the number of phases in the Erlang intercollection distribution ranges from 1 to 7, and we also examined **EW**, **EB**, and **EB′** for a deterministically distributed collecting interval. For the last two instances we performed a simulation experiment. Customers arrive according to Poisson processes with rate 1, the mean service time of a customer is 0.5.

Figure 4.1 illustrates the possibility of the BMAP framework to obtain insight in queueing systems with dependence between interarrival and service times, in particular when this dependence is the result of a collection/reservation mechanism. The first observation in figure 4.1 is that both mean waiting times and busy period lengths are decreasing as the intercollection interval becomes 'more deterministic'. Secondly, the performance measures are almost linear functions of the coefficient of variation of the intercollection interval distribution. Finally, not shown here, we observed that the coefficient of variation of the busy period remains almost constant.

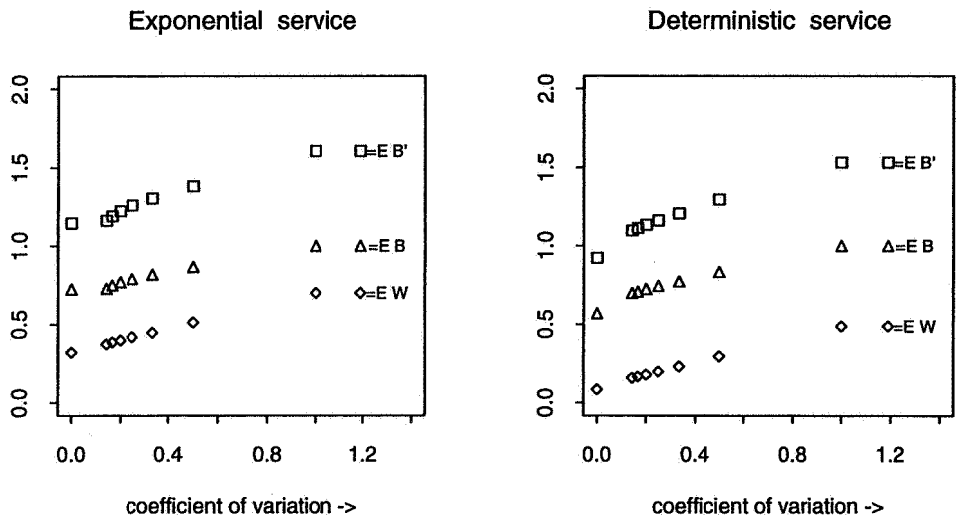

Figure 4.1: The effect of the intercollection interval distribution on mean waiting times and mean busy period lengths.

5 Integrating various dependence structures.

In general one considers three types of dependence in queueing systems: between consecutive interarrival times, between consecutive service times, and between the interarrival and service times of customers. Fendick et al.[4] state that in packet queues with multiple classes of traffic and variable packet lengths all three types of dependence occur *simultaneously*. The BMAP allows us to model this phenomenon. To illustrate this, we present an example in which the arrival process of busses is a two-state Markov Modulated Poisson Process (MMPP), the

batch size corresponds to a collecting procedure, and consecutive batch sizes are positively correlated. The example features all three types of dependence.

The dependence between two consecutive batch sizes can be modeled in several ways; for example by letting the arrival process of customers at the bus stop be a Markov Modulated Arrival Process, or by assuming that when the number of customers at the bus stop equals the maximum capacity M of the bus, newly arriving customers will wait for the next bus. The latter can be modeled as follows; at the moments $\mathbf{J}_1$ generates a bus arrival, also allow transitions in $\mathbf{J}_2$ to states other than 0 (this state representing the empty batch).

For our example we choose to model the dependency of consecutive service times by a two-state MMPP. The two-state MMPP's for the arrivals of busses and customers are as follows: while the underlying Markov process is in state i, the arrival rate for the arrival process of busses (of customers at the bus stop) is γ_i (λ_i), and the sojourn time in state i is exponentially distributed with parameter η_i (ν_i), $i = 1, 2$. For this model it is convenient to describe the evolution of the batch size distribution by a two dimensional Markov chain $\mathbf{J}_2 = (\mathbf{J}_2^1, \mathbf{J}_2^2)$, $\mathbf{J}_2^1$ representing the state of the MMPP for the individual customers, and the state of $\mathbf{J}_2^2$ representing the number of customers at the bus stop.

In figure 5.1 the transition rate diagram is presented for the case $M = 3$.

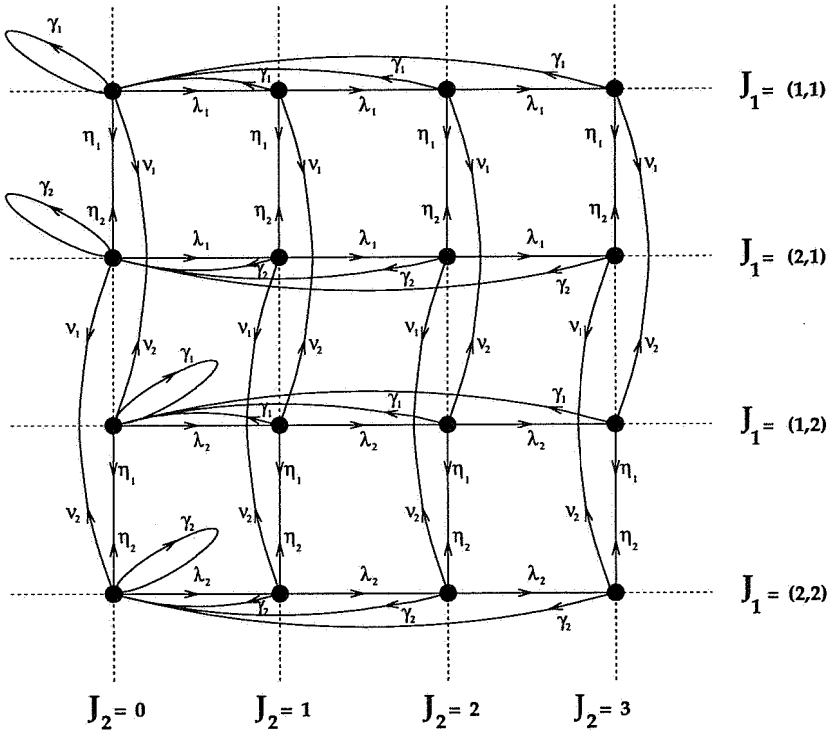

Figure 5.1: transition rate diagram for the collector model. Busses and customers arrive according to two-state Markov Modulated Poisson Processes.

REFERENCES

[1] Bisdikian, C., Lew, J.S., Tantawi, A.N. (1992). The Generalized $D^{[X]}/D/1$ queue and its application in the analysis of bridged high speed token-ring networks. IBM Res. Rep., RC 18387.

[2] Borst, S.C., Boxma, O.J., Combé, M.B. (1992). Collection of Customers: a Correlated M/G/1 Queue. *Perf. Eval. Review* **20**, 47-59.

[3] Borst, S.C., Boxma, O.J., Combé, M.B. (1993). An $M/G/1$ queue with customer collection. *Commun. Statist.-Stochastic Models* **9**, 341-371.

[4] Fendick, K.W., Saksena, V.R., Whitt, W. (1989). Dependence in packet queues. *IEEE Trans. Comm.* **37**, 1173-1183.

[5] Kleinrock, L. (1964). *Communication Nets.* (McGraw-Hill, New York).

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