Level-independent and level-dependent *QBD*s and their processes of departure

David Green
Teletraffic Research Centre
Department of Applied Mathematics
The University of Adelaide
South Australia
5005.
dgreen@maths.adelaide.edu.au

Abstract

The topic of *Quasi Birth and Death* (*QBD*) processes is best introduced by considering the simplicity of the *birth and death* (*BD*) process. By extending the possibilities for both the *birth* process and the *death* process, we develop the structure and theory for *QBD*s. The development includes both the homogeneous and in-homogeneous cases for *BD* and *QBD* processes. The departure process from the *MAP/PH/1* queue, treated as a *QBD* is then considered, as such a process may constitute the arrival process in another *QBD*. It appears that a finite *MAP* description for the output process of a stationary non-trivial *MAP/PH/1* queue does not exist. A family of *MAP* approximations to the departure process are proposed and an analysis of these approximations reveals some very special properties. These approximations are also shown to be applicable to *MAP/PH/1* feedback queues which extends their usefulness into the realm of the in-homogeneous *QBD*.

Contents

1 Birth and death processes

Imagine a population of individuals. A *birth* increases the population by one, and a *death* decreases it by one. Let X_t represent the total number of individuals in the population at time t. (X_t) is called a homogeneous *birth and death process* if (X_t) is a Markov process with a state transition diagram of the form depicted in Figure 1.

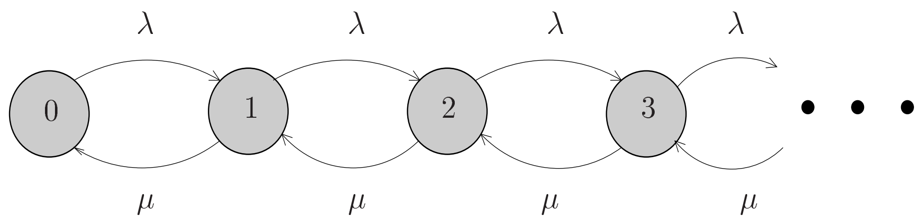


Figure 1: *State diagram of a birth and death process.*

The transition rate matrix Q for a homogeneous *birth* and *death* process with constant rates λ and μ respectively has the following tri-diagonal form

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \ldots \\ \mu & -(\lambda+\mu) & \lambda & 0 & \ldots \\ 0 & \mu & -(\lambda+\mu) & \lambda & \\ \vdots & & & \ddots & \end{bmatrix}. \tag{1}$$

Note that the above matrix is skip free in both the upwards and downwards direction for the number in the population (only one birth or one death can occur at a single time). The birth and death rates may also be state dependent, that is they may be of the form $\{\lambda_i\}$ and $\{\mu_i\}$. Many Markov processes in applications are of this kind and the simplicity of the process allows a detailed analysis.

In the homogeneous case depicted in Figure 1 and equation (1), $\lambda_i \equiv \lambda$ and $\mu_i \equiv \mu$ for all states i. Then, the matrix Q is also the transition rate matrix for an *M/M/1* queue, with arrival rate λ and service rate μ.

The transition rate matrix Q of a Markov process with state-dependent birth and death rates (a level-dependent *BD* process) is given by

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \ldots \\ \mu_1 & -\lambda_1-\mu_1 & \lambda_1 & 0 & \ldots \\ 0 & \mu_2 & -\lambda_2-\mu_2 & \lambda_2 & \ldots \\ \vdots & \vdots & \vdots & & \end{bmatrix}. \tag{2}$$

Before generalising to the QBD, we will briefly introduce some of the building blocks for the service time distributions and arrival processes that will be used. This introduction although not thorough, gives references to where the reader might find a more thorough treatment.

2 Phase-type (PH-type) distributions

An excellent treatment of the PH-type distributions is given in [37]. See also [31]. We will consider the continuous time version here, the discrete time versions are covered in the afore-mentioned references.

Consider a Markov chain with $n+1$ states, initial probability vector $(\boldsymbol{\alpha}, \alpha_{n+1})$ and transition rate matrix

$$Q = \begin{bmatrix} S & \boldsymbol{S}^0 \\ \mathbf{0} & 0 \end{bmatrix}, \tag{3}$$

where S is a non-singular $n \times n$ matrix with

$$\begin{aligned} S_{ii} &< 0, \\ S_{ij} &\geq 0 \quad \text{for all } i \neq j \text{ and} \\ \boldsymbol{S}^0 &\geq 0, \end{aligned}$$

where $\boldsymbol{S}^0$ is an n-vector of rates such that

$$S\boldsymbol{e} + \boldsymbol{S}^0 = \mathbf{0}.$$

The initial state is chosen according to an initial probability vector $(\boldsymbol{\alpha}, \alpha_{n+1})$ and as can be seen, all states are transient except the absorbing state. Any continuous distribution on $[0, \infty)$ which can be obtained as the distribution of time until absorption in a continuous time finite space Markov chain which has a single absorbing state into which absorption is certain, is said to be of phase-type. The probability the process has been absorbed by time t (i.e. its sojourn in states $(1, \ldots, n) \leq t$) is given by

$$\boldsymbol{F(t)} = 1 - \boldsymbol{\alpha} e^{\boldsymbol{S}t}\boldsymbol{e} \text{ for all } t \geq 0,$$

which is the probability distribution of time until absorption into state $n+1$. This is classified as a PH-type distribution with representation $(\boldsymbol{\alpha}, S)$.

The PH-type distributions are dense in the field of all distributions and encapsulate such as the negative-exponential distribution, Hyper-exponential distribution, Erlang distribution and mixtures thereof.

3 Markovian Arrival Processes (MAP)s

We will begin the discussion for the MAP by construction and extension of the Poisson process as in [45]. The MAP is actually a special case of the Batch Markovian Arrival Process or $BMAP$, a class of point processes that has since been shown to be the equivalent of the Versatile Markovian point process of Neuts [36]. See also Latouche and Ramaswami [31].

Consider a continuous-time Markov process, with a single revisited state.

- Sojourn times are negative exponentially distributed with parameter λ.
- When a sojourn expires, another immediately begins as the single state is immediately revisited.

If we associate an arrival with each transition in the above Markov process, we can construct a point process which is Poisson.

We can then generalise the Poisson process by adding auxiliary states or phases to the Markov process, and by associating arrivals in the point process with particular transitions in the new underlying Markov process.

The resultant process

- Allows other than negative exponentially distributed inter-arrival times.
- Preserves the Markov property for state transitions.
- Has a convenient matrix form representation.

It is convenient to consider such processes in terms of two filtration matrices D_0 and D_1 which filter those parts of the Markov process which correspond to non-arrival transitions and arrival transitions respectively. In the Poisson arrival case we would have the following representation

$$D_0 = [-\lambda], \quad D_1 = [\lambda]$$

Here every transition corresponds to an arrival.

Let's now consider a two state process and associate with probability $p_{i,i}(1)$ an arrival at the end of a sojourn in state i for $i \in \{1, 2\}$, which immediately returns the process to state i. With probability $p_{i,j}(0) = 1 -$

$p_{i,i}(1)$, there is no arrival but a transition to the state j where $j \neq i$. Hence we may write

$$D_0 = \begin{bmatrix} -\lambda_1 & p_{1,2}(0)\lambda_1 \\ p_{2,1}(0)\lambda_2 & -\lambda_2 \end{bmatrix}, \quad D_1 = \begin{bmatrix} p_{1,1}(1)\lambda_1 & 0 \\ 0 & p_{2,2}(1)\lambda_2 \end{bmatrix}.$$

This is a special case of the *MAP* known as the Markov Modulated Poisson Process or *MMPP*. The *MMPP* has been successfully used to model such things as the error environment for data transmission channels [32]. It is a natural extension to allow transitions between states to correspond to arrivals as well, which generalises to the *MAP* and in the two state case has the following representation

$$D_0 = \begin{bmatrix} -\lambda_1 & p_{1,2}(0)\lambda_1 \\ p_{2,1}(0)\lambda_2 & -\lambda_2 \end{bmatrix}, \quad D_1 = \begin{bmatrix} p_{1,1}(1)\lambda_1 & p_{1,2}(0)\lambda_1 \\ p_{2,1}(0)\lambda_2 & p_{2,2}(1)\lambda_2 \end{bmatrix},$$

where $\sum_j p_{i,j}(.) = 1$ for each $i \in \{1, 2\}$. The *MAP* can have many more states associated with it and has been found to be extremely versatile. Another special case or sub class of the *MAP* which is worthy of mention is the Phase renewal (*PH*) process. *PH*-type distributions can be used to model arrival processes by using them as inter-event times in a renewal process. That is, for the distribution described in Section 2, when an absorbing transition occurs, we instantaneously restart the process according to the initial probability vector $\boldsymbol{\alpha}$, so that the process consisting of absorption epochs becomes a renewal process. This can be viewed in the following diagram

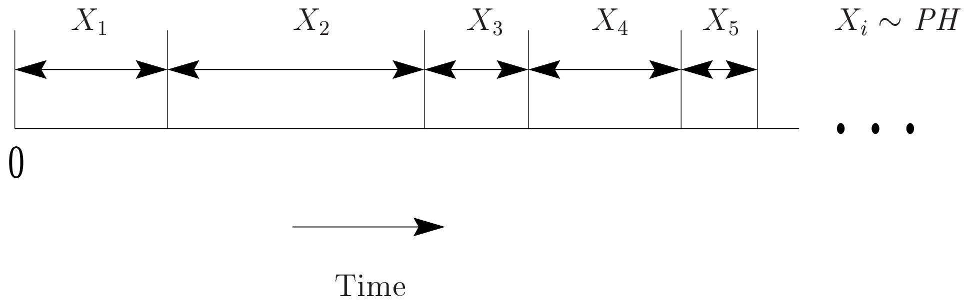


Figure 2: *PH-renewal process.*

The resultant *PH*-renewal process represented by $(\boldsymbol{\alpha}, S)$ has the following *MAP* representation

$$D_0 = S \text{ and } D_1 = \boldsymbol{S}^0\boldsymbol{\alpha}.$$

In order to display the capability of the *MAP* it is useful to observe some realisations of point processes as in Figure 3, where time runs left to right.

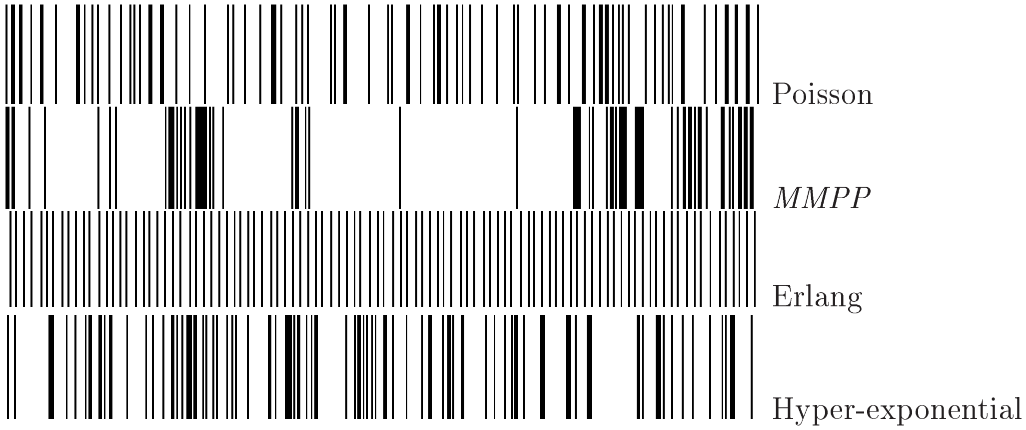


Figure 3: *Point processes of arrivals*

Each vertical line in each the four traces in Figure 3 represents an arrival event. There is an appearance of thicker lines, which correspond to arrival events which are very close together and avoid clear resolution in the figure. In each of the traces there are 100 arrival events. The processes marked Erlang and Hyper-exponential are renewal processes constructed using the Erlang and Hyper-exponential distributions respectively.

The Poisson trace has no apparent form, it is random, whereas the *MMPP* trace is obviously very bursty.

The Erlang trace is of a high order Erlang, which shows a very regular inter-arrival time being almost deterministic.

The Hyper-exponential trace also shows a bursty nature but not to the same extent as the *MMPP*. The latter can be parameterised so that it has a very bursty nature, like that of the *MMPP*. The main difference between the two is that for the renewal process, the lengths of the burst can be shown to have a geometric distribution whereas the *MMPP* does not.

The actual distribution function for the given Hyper-exponential trace is

$$F(t) = \frac{10}{11}(1 - e^{-10.0t}) + \frac{1}{11}(1 - e^{-0.1t})$$

Here, the process enters an inter-arrival time sojourn governed by a long inter-arrival time distribution with probability $\frac{1}{11}$ or it is governed by a short inter-arrival time distribution with probability $\frac{10}{11}$. So the probability that the burst has length n is given by

$$p(n) = \left(\frac{1}{11}\right)\left(\frac{10}{11}\right)^n.$$

4 *QBD* examples

The *QBD* like the simple *BD* process is also skip free in both the upwards and downwards direction for what is known as the level (often the number in the population). The difference being that for a *QBD*, the distribution of time between births and the time between the birth and death of an individual is allowed to take on more general form rather than just negative-exponential.

We will now consider an *M/M/1* queue in a random environment [31], and build up a *QBD* description for the system. Here the arrival rate may vary according to an underlying Markov chain. The service rate for this example will remain constant at rate μ and the arrival rate changes only as depicted in Figure 4, which for instance may mimic customers arriving to a bank teller at various times of the day.

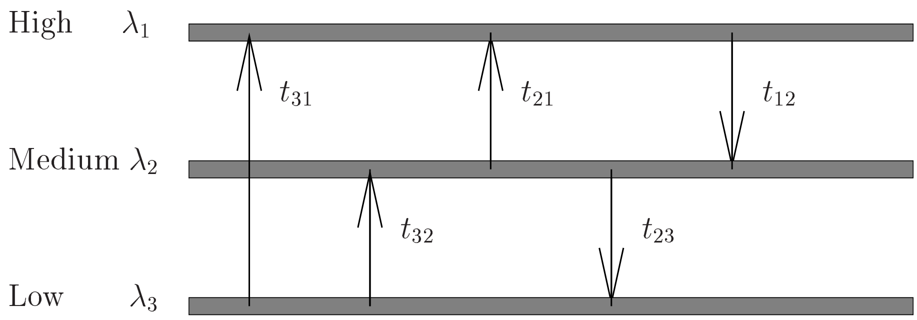


Figure 4: *Transition rates for the underlying Markov process.*

If we consider these changes in arrival rate as phase changes and the changes in customer numbers as levels, we can construct a Q matrix with a block design. First we consider the state transition diagram for this process in Figure 5.

We see that a transition is either between levels or between phases, so we partition the Q matrix into blocks of square dimension equal to the number of phases as in equation (4). In this block partitioned Q matrix, the level 0 corresponds to the first row of blocks etc. Note that a transition which moves into a new level remains in the same phase at the corresponding new level.

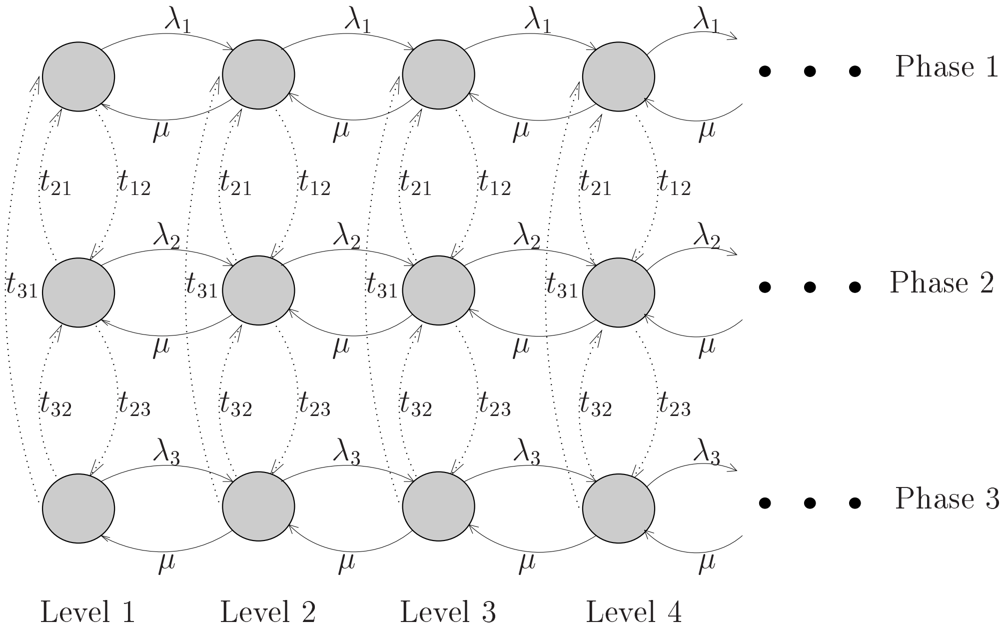


Figure 5: *An M/M/1 queue in a random environment.*

$$Q = \begin{bmatrix} \begin{bmatrix} \alpha & t_{12} & 0 \\ t_{21} & \beta & t_{23} \\ t_{31} & t_{32} & \gamma \end{bmatrix} & \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} \mu & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \mu \end{bmatrix} & \begin{bmatrix} \delta & t_{12} & 0 \\ t_{21} & \epsilon & t_{23} \\ t_{31} & t_{32} & \eta \end{bmatrix} & \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} & \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} \mu & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \mu \end{bmatrix} & \begin{bmatrix} \delta & t_{12} & 0 \\ t_{21} & \epsilon & t_{23} \\ t_{31} & t_{32} & \eta \end{bmatrix} & \\ \vdots & & & \ddots \end{bmatrix}, \quad (4)$$

where

$$
\begin{aligned}
\alpha &= -(t_{12} + \lambda_1), \\
\beta &= -(t_{21} + \lambda_2 + t_{23}), \\
\gamma &= -(t_{31} + t_{32} + \lambda_3), \\
\delta &= \alpha - \mu, \\
\epsilon &= \beta - \mu \quad \text{and} \\
\eta &= \gamma - \mu.
\end{aligned}
$$

This Q matrix is also the Q matrix for an *MMPP/M/1* queue, where the arrival (birth) process is the Markov Modulated Poisson Process. It is a simple extension to allow arrivals to coincide with phase changes as well, which is the case for many *MAP/M/1* queues (recall Section 3).

Another example: Consider an *M/M/1/*$(c-1)$ queue where arrivals finding the buffer full, go away and retry after an exponentially distributed length of time with parameter α [31]. We can represent this diagrammatically as in Figure 6.

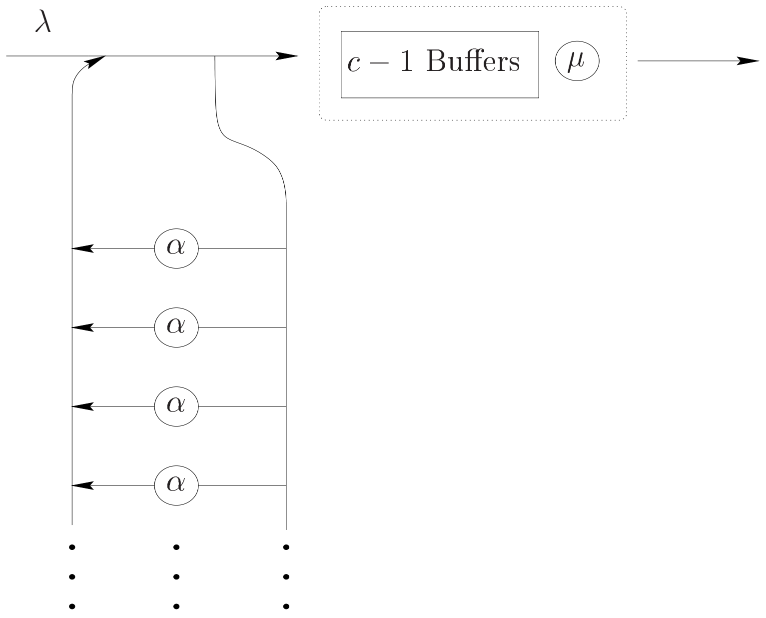


Figure 6: *A retrial scheme*

If we let N_1 be the number of customers in the queue and N_2 be the number of customers in retrial, we can establish a two dimensional Markov process description using a block Q matrix as follows. We first chose N_1 to be the phase, which in this case will give us a finite block size. If we had chosen N_2 to be the phase, this would not have been the case. In all treatment of the level-independent QBD we will consider the block matrices to be of finite size. Ramaswami has considered the case of a doubly infinite structure.

The following transitions and rates as given in (5) apply, where $(n,\ i)$ gives the (level, phase).

$$\begin{array}{lcll}
(n,i) & \to & (n,i+1) & \text{at rate } \lambda \text{ if } 0 \leq i \leq c-1, \\
(n,i) & \to & (n,i-1) & \text{at rate } \mu \text{ if } 1 \leq i \leq c, \\
(n,i) & \to & (n+1,i) & \text{at rate } \lambda \text{ if } i \geq c, \\
(n,i) & \to & (n-1,i) & \text{at rate } n\alpha \text{ if } n \geq 1, \quad 0 \leq i \leq c-1.
\end{array} \tag{5}$$

From these transition rates we may construct the following Q matrix in equation (6).

$$Q = \begin{bmatrix}
B_1 & A_{00} & \mathbf{0} & \cdots & & \\
A_{20} & A_{10} & A_{01} & \mathbf{0} & \cdots & \\
\mathbf{0} & A_{21} & A_{11} & A_{02} & \mathbf{0} & \cdots \\
\vdots & \mathbf{0} & A_{22} & A_{12} & A_{03} & \ddots \\
\vdots & \vdots & & \ddots & \ddots & \ddots
\end{bmatrix}, \tag{6}$$

where the $(c+1) \times (c+1)$ matrices

$$B_1 = \begin{bmatrix}
-\lambda & \lambda & 0 & \cdots & 0 \\
\mu & -(\lambda+\mu) & \lambda & & \vdots \\
0 & \mu & -(\lambda+\mu) & \ddots & 0 \\
\vdots & & \ddots & \ddots & \lambda \\
0 & \cdots & 0 & \mu & -(\lambda+\mu)
\end{bmatrix}, \quad A_{00} = \begin{bmatrix}
0 & & \cdots & & 0 \\
 & & & & \vdots \\
\vdots & & \ddots & & \vdots \\
 & & & 0 & 0 \\
0 & \cdots & \cdots & 0 & \lambda
\end{bmatrix},$$

$$A_{20} = \begin{bmatrix}
0 & \alpha & 0 & \cdots & 0 \\
 & 0 & \alpha & & \vdots \\
\vdots & & \ddots & \ddots & 0 \\
 & & & 0 & \alpha \\
0 & \cdots & \cdots & 0 & 0
\end{bmatrix}, \quad A_{0i} = A_{00} \text{ and } A_{2i} = (i+1)A_{20} \text{ for all } i \geq 1,$$

and for $\varpi = (\lambda + \alpha + \mu)$, we have

$$A_{10} = \begin{bmatrix} -(\lambda+\alpha) & \lambda & 0 & \cdots & & 0 \\ \mu & -\varpi & \lambda & & & \\ 0 & \mu & -\varpi & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \\ & & & & -\varpi & \lambda \\ 0 & & & & \mu & -(\lambda+\mu) \end{bmatrix},$$

and for $i \geq 1$ with $\vartheta = (\lambda + (i+1)\alpha + \mu)$, we have

$$A_{1i} = \begin{bmatrix} -\vartheta & \lambda & 0 & \cdots & & 0 \\ \mu & -\vartheta & \lambda & & & \\ 0 & \mu & -\vartheta & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \\ & & & & -\vartheta & \lambda \\ 0 & & & & \mu & -(\lambda+\mu) \end{bmatrix}.$$

This Q matrix is an example of a level-dependent QBD because as the level increases, the rate of retrial increases due to the number in retrial. This is reflected in the entries of the matrices $A_{2,i}$ and $A_{1,i}$. In full generality, the level-dependent QBD may also have super-diagonal matrices $A_{0,i}$ that have level-dependent entries. This is not the case here.

5 The equilibrium distribution for a level-independent QBD

This section is based on a tutorial given by Professor Guy Latouche during his visit to the TRC in 1996 and the book by Latouche and Ramaswami [31]. The development ensues under the assumption that the processes under scrutiny are irreducible.

Consider a Markov chain with a finite state space S. If we partition the state space into two sets T and D, that is

$$|T| + |D| = |S| < \infty,$$

such that there are transitions within and between these partitions as depicted in Figure 7.

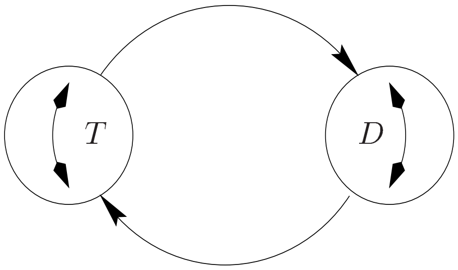


Figure 7: *Partitioned state space*

We have a probability transition matrix P such that

$$\boldsymbol{\pi} P = \boldsymbol{\pi}, \qquad \text{with } \boldsymbol{\pi}\boldsymbol{e} = 1,$$

where $\boldsymbol{e}$ is a column of ones of the appropriate dimension. The probability transition matrix P can be written as

$$P = \begin{array}{cc} \left[\begin{array}{cc} P_T & P_{TD} \\ P_{DT} & P_D \end{array}\right] & \begin{array}{l} \}T \\ \}D \end{array} \\ \begin{array}{cc} \underbrace{\qquad}_{T} & \underbrace{\qquad}_{D} \end{array} & \end{array},$$

with $\boldsymbol{\pi} = (\boldsymbol{\pi}_T,\ \boldsymbol{\pi}_D)$, so that we have the following.

$$\begin{aligned} \boldsymbol{\pi}_T P_T + \boldsymbol{\pi}_D P_{DT} &= \boldsymbol{\pi}_T \quad \text{and} && (7) \\ \boldsymbol{\pi}_T P_{TD} + \boldsymbol{\pi}_D P_D &= \boldsymbol{\pi}_D. && (8) \end{aligned}$$

Now, equation (8) implies

$$\boldsymbol{\pi}_D (I - P_D) = \boldsymbol{\pi}_T P_{TD}, \tag{9}$$

where I is the identity matrix of the appropriate dimension. Combining equations (9) and (7) we see that

$$\boldsymbol{\pi}_T \left(P_T + P_{TD} (I - P_D)^{-1} P_{DT}\right) = \boldsymbol{\pi}_T. \tag{10}$$

A probabilistic interpretation allows an infinite state space and assuming positive recurrence we define

$$\begin{aligned} N_D &= (I - P_D)^{-1} \\ &\equiv \sum_{k \geq 0} P_D^k \\ &= E\Big[\text{number of visits to } D\Big], \end{aligned}$$

with for $i, j \in D \notin T$

$$(N_D)_{i,j} = E\Big[\text{number of visits to } j \in D \text{ before visiting } T \mid i \in D\Big],$$

Then from equation (9) we have that the distribution $\boldsymbol{\pi}_D$ for the partition D is given in terms of the distribution for the partition T as follows

$$\begin{aligned} \boldsymbol{\pi}_D &= \boldsymbol{\pi}_T P_{TD} (I - P_D)^{-1} \\ &= \boldsymbol{\pi}_T P_{TD} N_D, \end{aligned} \tag{11}$$

where $P_{TD} (I - P_D)^{-1}$ is the expected time spent in D between two visits to T. We now define for $i \in D$ and $j \in T$

$$[G_{DT}]_{i,j} = \left[(I - P_D)^{-1} P_{DT}\right]_{i,j} \tag{12}$$

$$= \left[\sum_{k \geq 0} P_D^k P_{DT}\right]_{i,j}, \tag{13}$$

which is the probability of starting in state $i \in D$ and after an arbitrary number of transitions, moving to state $j \in T$. We now use this definition and rewrite equation (10) as

$$\boldsymbol{\pi}_T (P_T + P_{TD} G_{DT}) = \boldsymbol{\pi}_T,$$

where we note that the matrix $(P_T + P_{TD} G_{DT})$ is the transition matrix of the censored Markov chain which keeps track of the observed transitions in T, but not the unobserved transitions in D (see Figure 8). Schematically

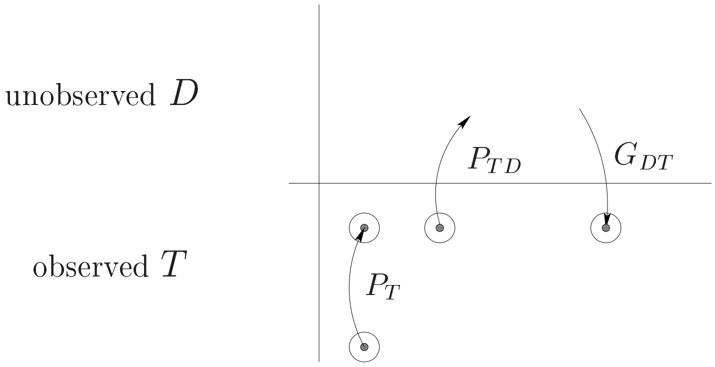


Figure 8: *Schematic of censored chain*

We now consider the following probability transition matrix P for a level-independent QBD given by

$$
\begin{array}{cc}
 & \hspace{6em}(n+1) \\
(n) & \left[\begin{array}{ccccc|ccccc}
B_1 & A_0 & 0 & \cdots & 0 & 0 & \cdots & & & \\
A_2 & A_1 & A_0 & & \vdots & \vdots & \ddots & & & \\
0 & A_2 & A_1 & \ddots & 0 & \vdots & & & & \\
\vdots & & \ddots & \ddots & A_0 & 0 & & & & \\
0 & \cdots & 0 & A_2 & A_1 & A_0 & 0 & \cdots & & \\
\hline
0 & \cdots & \cdots & 0 & A_2 & A_1 & A_0 & 0 & \cdots & \\
\vdots & \ddots & & \vdots & 0 & A_2 & A_1 & A_0 & & \\
 & & & & \vdots & 0 & A_2 & A_1 & A_0 & \\
 & & & & & \vdots & & \ddots & \ddots & \ddots
\end{array}\right]
\end{array}. \tag{14}
$$

Using equations (11) and the form given in (14), we may write

$$
(\boldsymbol{\pi}_{n+1}, \boldsymbol{\pi}_{n+2}, \ldots) \;=\; (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)\, P_{TD} N_D,
$$

where

$$
N_D \;=\; \left[\begin{array}{ccccc}
N_{11} & N_{12} & \ldots & N_{1n} & \ldots \\
N_{21} & & & & \\
\vdots & & & &
\end{array}\right]
$$

and

$$P_{TD}N_D \;=\; \begin{bmatrix} 0 & & \cdots & & \cdots \\ \vdots & & & & \\ 0 & & & & \\ A_0N_{11} & A_0N_{12} & A_0N_{13} & \cdots & \end{bmatrix}.$$

Hence we have that

$$\boldsymbol{\pi}_{n+1} \;=\; \boldsymbol{\pi}_n A_0 N_{11},$$

which is true for all $n \geq 0$ so that

$$\begin{aligned} \boldsymbol{\pi}_1 &= \boldsymbol{\pi}_0 A_0 N_{11} \\ \boldsymbol{\pi}_2 &= \boldsymbol{\pi}_1 A_0 N_{11} = \boldsymbol{\pi}_0 (A_0N_{11})(A_0N_{11}) \\ \vdots\quad & \vdots \qquad \vdots \quad\;\; \vdots \qquad\qquad \vdots \qquad . \end{aligned}$$

Let $R = A_0N_{11}$, so that we may write

$$\boldsymbol{\pi}_n \;=\; \boldsymbol{\pi}_0 R^n.$$

Now recalling the definition of the matrix N_D, that is

$$\begin{aligned} N_D &\equiv \sum_{k\geq 0} P_D^k \\ &= (I - P_D)^{-1}, \end{aligned}$$

we see that it is the minimal non-negative solution to

$$(I - P_D)\, N_D \;=\; I.$$

So $R = A_0N_{11}$ where

$$N_{11} \;=\; E\left[\begin{array}{c} \text{number of visits to level 1} \\ \text{before level 0 is reached} \end{array} \middle| \begin{array}{c} \text{start in} \\ \text{level 1} \end{array} \right].$$

For the matrix $\hat{P}$ in Figure 9, define $\displaystyle\sum_{n=0}^{\infty} (\hat{P})^n = (I - \hat{P})^{-1}$.

Calculating R or N:

$$\begin{aligned} \text{Let } \theta &= \inf_{n\geq 1}\{n : X_n \in l(1)\}, &&\text{(level 1)} \\ \tau &= \inf_{n\geq 1}\{n : X_n \in l(0)\}, &&\text{(level 0)}\ , \end{aligned}$$

$$
\begin{bmatrix}
B_1 & A_0 & 0 & \dots & \\
A_2 & A_1 & A_0 & 0 & \dots \\
0 & A_2 & A_1 & A_0 & \\
\vdots & & \ddots & \ddots & \ddots
\end{bmatrix}
\qquad \hat{P}
$$

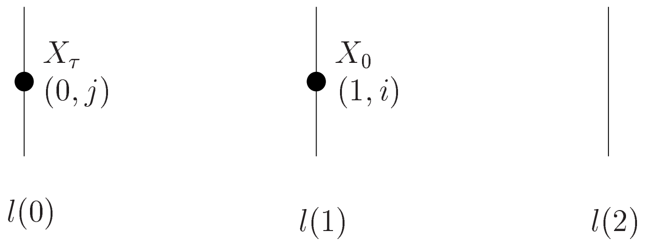
Figure 9:

$$
U_{ij} \quad = \quad P\Big[\theta < \infty,\ \theta < \tau,\ X_\theta = (1,j) \mid X_0 = (1,i)\Big]
$$

$$
\text{and } G_{kj} \quad = \quad P\Big[\tau < \infty,\ X_\tau = (0,j) \mid X_0 = (1,k)\Big],
$$

$$
\text{so that } U_{ij} \quad = \quad (A_1)_{ij} + \sum_k (A_0)_{ik}(G)_{kj} \quad \text{(by the symmetry of } P\text{)}\ .
$$

$$
\text{That is, } U \quad = \quad A_1 + A_0 G. \tag{15}
$$



Also we have

$$
\begin{aligned}
G_{ij} \quad &= \quad (A_2)_{ij} + \sum_k \Big[(A_1)_{ik} + \sum_l (A_0)_{il}(G)_{lk}\Big](G)_{kj} \\
&= \quad (A_2)_{ij} + \sum_k [U_{ik}]\,(G)_{kj} \\
\Rightarrow G \quad &= \quad A_2 + UG \\
\Rightarrow G \quad &= \quad (I-U)^{-1}A_2,
\end{aligned}
$$

which leads to the following set of equations

$$
\begin{aligned}
G &= (I-U)^{-1}A_2, && (16)\\
U &= A_1 + A_0 G && (17)\\
R &= A_0(I-U)^{-1} && (18)
\end{aligned}
$$

Using the above equations 16, 17 and 18, we may we may establish a method for finding the matrices G and U.

Algorithm 5.1 *If we let*

$$
\begin{aligned}
G_0 &= 0 \quad \textit{and then let}\\
U_0 &= A_1 + A_0 G_0,
\end{aligned}
$$

and then set

$$
\begin{aligned}
U_n &= A_1 + A_0 G_n \quad \textit{and}\\
G_{n+1} &= (I-U_n)^{-1}A_2.
\end{aligned}
$$

we have a simple linear iterative method for finding the matrices G and U. Now, G is stochastic and U is sub-stochastic by their definitions, that is

$$
\begin{aligned}
G &: \textit{start in level } 1 \textit{ and eventually go to level } 0,\\
U &: \textit{start in level } 1 \textit{ and eventually go back to level } 1 \textit{ in taboo of level } 0,
\end{aligned}
$$

so that we may stop the iterations when

$$|G\boldsymbol{e} - \boldsymbol{e}| < \epsilon\, \boldsymbol{e}$$

for some ϵ.

If we define

$$
\begin{aligned}
G_k &= P\left[\begin{array}{c}\text{process starts in level 1 and goes}\\ \text{to level 0 without reaching level } k+1\end{array}\right] \quad \text{and}\\
U_k &= P\left[\begin{array}{c}\text{process starts in level 1 and goes back}\\ \text{to level 1 in taboo of levels 0 and } k+1\end{array}\right],
\end{aligned}
$$

then the iterates take on some meaning if we begin with $G_0 = 0$. This may be seen in the following figures (Figures 10, 11 and 12) and is a result of the structure of the transition matrix, which is repetitive after level 1.

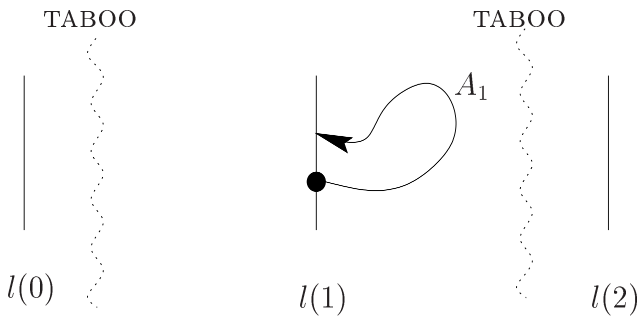


Figure 10: $U_1 \ = \ A_1$

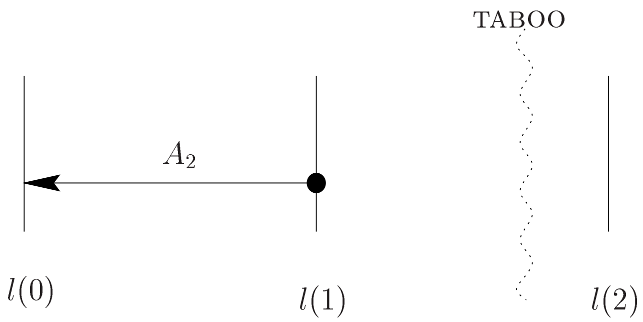


Figure 11: $G_1 \ = \ A_2$

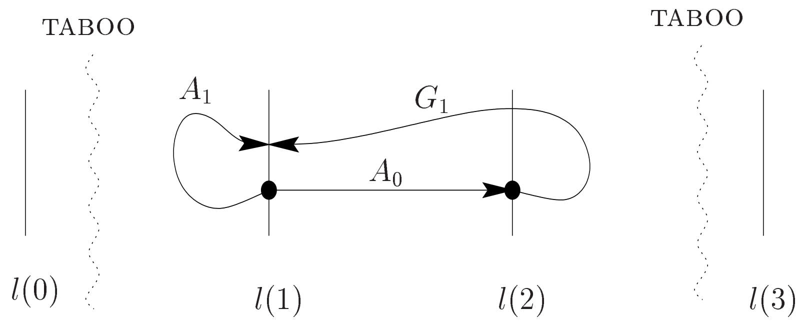


Figure 12: $U_2 \ = \ A_1 \ + \ A_0 \ G_1$

$$\begin{aligned}
G_2 &= (I - U_2)^{-1} A_2 = A_2 + U_2 A_2 + U_2 U_2 A_2 + \cdots \\
U_3 &= A_1 + A_0 G_2 \\
G_3 &= (I - U_3)^{-1} A_2 = A_2 + U_3 A_2 + U_3 U_3 A_2 + \cdots \\
\vdots \quad &\vdots \qquad \vdots
\end{aligned}$$

Therefore we see that as $k \to \infty$ we have

$$\begin{aligned} G_k &\to \lim_{k\to\infty} (I - U_k)^{-1} A_2 = (I-U)^{-1} A_2 \\ U_k &\to \lim_{k\to\infty} A_1 + A_0 G_{k-1} = A_1 + A_0 G \end{aligned}$$

Convergence:

$$\begin{array}{ccccccccccc} G_0 & = & 0 & \le & G_1 & = & A_2 & \le & G & = & (I-U)^{-1}A_2 \\ U_0 & = & A_1 & \le & U_1 & = & A_1 + A_0 A_2 & \le & U & = & A_1 + A_0 G \\ G_1 & = & A_2 & \le & G_2 & = & (I-U_1)^{-1}A_2 & \le & G & = & (I-U)^{-1}A_2 \\ U_1 & = & A_1 + A_0 A_2 & \le & U_2 & = & A_1 + A_0 G_1 & \le & U & = & A_1 + A_0 G \\ \vdots & = & \vdots & \le & \vdots & = & \vdots & \le & \vdots & = & \vdots \end{array},$$

therefore both iterates are monotonic increasing and bounded $\Rightarrow$ convergent. The convergence is quite intuitive, using the interpretation of the matrices U_k and G_k given in equations 19. The convergence rate can also be shown to be $o\{\rho\}$, where

$$\begin{aligned} \rho &= \eta\,\zeta, \quad \text{with} \\ \eta &= \text{maximal eigenvalue of } R \text{ and} \\ \zeta &= \text{maximal eigenvalue of } G. \end{aligned}$$

In this case, G is stochastic since the level-independent QBD is positive recurrent so that $\zeta \equiv 1$ and $\rho = \eta$.

5.1 The matrix quadratic equations

Formally we can state the method of analysis of the level-independent QBD as follows. Neuts showed in (1981) [37] and (1989) [40] how level-independent QBDs can be analysed. We attempt to derive the minimal non-negative solutions G and R to the matrix quadratic equations

$$\hat{G} = A_2 + A_1 \hat{G} + A_0 \hat{G}^2 \tag{19}$$

and

$$\hat{R} = A_0 + \hat{R} A_1 + \hat{R}^2 A_2. \tag{20}$$

The $(i,j)^{th}$ entry of the matrix G is the probability that the process will first enter level 0 in phase j given that it starts in phase i of level 1.

G is a sub-stochastic matrix and the level-independent QBD is recurrent if and only if it is stochastic.

The (i,j)th entry of the matrix R is the expected number of visits to phase j of level 1 before first return to level 0 conditional on the process starting in phase i of level 0.

In general, R has spectral radius which is less than or equal to one, and the level-independent QBD is positive recurrent if and only if the spectral radius of R is less than one.

Many performance measures for level-independent QBDs can be derived from a knowledge of G and R. The most well-known of these is the matrix-geometric form of the stationary distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)$ of a positive recurrent level-independent QBD given by

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 R^n. \tag{21}$$

5.2 Ergodicity

Thus far we have assumed positive recurrence without giving any conditions under which this can be assumed. If we let $\boldsymbol{\psi}$ be the solution to

$$\boldsymbol{\psi} \quad = \quad \boldsymbol{\psi}[A_0 + A_1 + A_2],$$

such that $\boldsymbol{\psi}\boldsymbol{e} = 1$. Then the level-independent QBD is positive recurrent if

$$\boldsymbol{\psi} A_0 \boldsymbol{e} \quad < \quad \boldsymbol{\psi} A_2 \boldsymbol{e},$$

null recurrent if

$$\boldsymbol{\psi} A_0 \boldsymbol{e} \quad = \quad \boldsymbol{\psi} A_2 \boldsymbol{e},$$

and transient if

$$\boldsymbol{\psi} A_0 \boldsymbol{e} \quad > \quad \boldsymbol{\psi} A_2 \boldsymbol{e},$$

5.3 Algorithms and complexity

Each iteration of the Algorithm 5.1 can be shown to take approximately

$$14M^3/3$$

floating point operations.

Remark 5.1 *The number N of iterations required is related to the physics of the level-independent QBD. Recall that when the level-independent QBD is recurrent we can use $|G\boldsymbol{e} - \boldsymbol{e}| < \epsilon\, \boldsymbol{e}$ as a stopping criterion. In this case N is given by the $1-\epsilon$ quantile of the distribution of the maximum level reached during a sojourn at level 1 or above. It is asymptotically*

$$\frac{\log \epsilon}{\log \eta},$$

where η is the maximum eigenvalue of R.

Algorithm 5.2 (Improved Algorithm)

In the recurrent case, we are looking for a stochastic solution G. Therefore it might make sense to start our recursion with a stochastic matrix. That is we take $G_0 = I$, which is sometimes quicker to converge, although it loses the nice probabilistic interpretation of the iterates. It can be shown that the matrices G_i for $i \geq 0$ are all stochastic, so we should stop the iterations when

$$\max |G_k - G_{k-1}| < \epsilon$$

for some ϵ.

Meini [33] showed that asymptotically this algorithm requires

$$\frac{\log \epsilon}{\log \eta_2}$$

iterations where η_2 is the second largest eigenvalue of R.

Latouche and Ramaswami [29] provided examples where the first algorithm can take around 800 iterations to achieve 10 digit accuracy, while the second algorithm requires only around 60 iterations to achieve the same accuracy.

For a transient level-independent QBD, the minimal G is sub-stochastic and not stochastic. The original algorithm converges to the minimal G_{MIN}. However the improved algorithm still converges to a stochastic G_{STOCH}, which implies there is a stochastic solution even when the level-independent QBD is transient.

The stochastic solution has applications. For example, it can be used to construct a matrix-geometric invariant measure for a transient level-independent QBD (Latouche, Pearce and Taylor [28]).

It follows from Gail, Hantler and Taylor [20] that, under an appropriate irreducibility assumption,

$$G_{STOCH} = G_{MIN} + [I - G_{MIN}]\, \boldsymbol{e}\, \boldsymbol{g}$$

where $\boldsymbol{g}$ is the maximal left eigenvector of G_{MIN}.

The original and improved algorithms are identical apart from the initial points $G_0 = 0$ and $G_0 = I$. In general we could take $G_0 = B$ where B is sub-stochastic. Latouche and Taylor (1999) have recently shown that, for a transient level-independent QBD such an algorithm converges to G_{STOCH} if and only if B is stochastic; otherwise it will converge to G_{MIN}.

Algorithm 5.3 (Logarithmic Reduction Algorithm)

A more efficient quadratically convergent algorithm known as the Logarithmic Reduction Algorithm has been devised by Latouche and Ramaswami [30]. For G, this works by evaluating the closed-form expression

$$G = \sum_{\ell=0}^{\infty} \left[\prod_{i=0}^{\ell-1} U^i \right] D^{\ell}, \tag{22}$$

where the matrices U^{ℓ} and D^{ℓ} satisfy the recursion

$$U^{\ell+1} = \left[I - U^{\ell}D^{\ell} - D^{\ell}U^{\ell}\right]^{-1} \left[U^{\ell}\right]^2$$

and

$$D^{\ell+1} = \left[I - U^{\ell}D^{\ell} - D^{\ell}U^{\ell}\right]^{-1} \left[D^{\ell}\right]^2,$$

with $U^0 = (I - A_1)^{-1}A_0$ and $D^0 = (I - A_1)^{-1}A_2$.

Let $G(n)$ be the matrix that is produced when the sum in (22) is truncated after n terms. Then the (i, j)th entry of $G(n)$ is the probability that, starting in state $(1, i)$, the process first visits level 0 in state $(0, j)$ under taboo of levels $2^n + 1$ and above.

Recall that each iteration of the first algorithm discussed above required

$$14M^3/3$$

floating point operations. With the Logarithmic-Reduction Algorithm each iteration is more complex. It uses

$$50M^3/3$$

floating point operations. However the number of iterations required to reach a given accuracy is substantially less. In the positive recurrent case, the number of iterations required for the simpler algorithms discussed to converge to an accuracy ϵ was of the order

$$\frac{\log \epsilon}{\log \eta}$$

where η is the largest or second largest eigenvalue of the matrix R. For the Logarithmic-Reduction Algorithm, the number of iterations required is of the order

$$\frac{\log \log \epsilon - \log \log \eta_1}{\log 2}$$

where η_1 is the largest eigenvalue of the matrix R. In some experiments with "well-behaved" positive recurrent level-independent QBDs, Latouche and Ramaswami [30] concluded that the total time taken by the Logarithmic-Reduction Algorithm was substantially less than that used by the recursion when $G(0) = 0$, but of the same order as that taken by the algorithm with $G(0) = I$. However, the Logarithmic-Reduction Algorithm also works efficiently with positive recurrent processes which are not "well-behaved", and with null-recurrent and transient processes.

5.4 Continuous time

The same theory may be applied to continuous time level-independent QBDs by considering the embedded process at epochs of a change of level (see [30]). That is, given the transition rate matrix

$$Q = \begin{bmatrix} B_1^* & A_0^* & 0 & \cdots & & \\ A_2^* & A_1^* & A_0^* & 0 & \cdots & \\ 0 & A_2^* & A_1^* & A_0^* & 0 & \cdots \\ \vdots & 0 & A_2^* & A_1^* & A_0^* & \ddots \\ \vdots & \vdots & & \ddots & \ddots & \ddots \end{bmatrix},$$

the Markov chain observed at epochs of a change of level has the structure as shown in (14), where

$$\begin{aligned} B_1 &= A_1 = 0 \\ A_0 &= (-A_1^*)^{-1} A_0^* \\ A_2 &= (-A_1^*)^{-1} A_2^*. \end{aligned}$$

The continuous time QBD and the embedded Markov chain have the same matrix G. Hence for the continuous time QBD, equations (16) and (17) still hold, but the R matrix has a different form given by

$$R = A_0^*(-U)^{-1}. \tag{23}$$

An implementation of the Logarithmic Reduction Algorithm in the continuous time case can therefore be stated as follows.

Let $C_0 = (-A_1^*)^{-1}A_0^*$,

$C_2 = (-A_1^*)^{-1}A_2^*$,

$X = C_2$,

$Y = C_0$,

$H = YC_2$.

While $\left(\max(\boldsymbol{e} - X\boldsymbol{e}) \leq \epsilon\right)$

$J = (I - C_0C_2 - C_2C_0)^{-1}$,
$C_0 = JC_0C_0$,
$C_2 = JC_2C_2$,
$H = YC_2$,
$X = X + H$,
$Y = YC_0$.

End

Then we have

$G = X$,

$U = A_1^* + A_0^*G$,

$R = A_0^*(-U)^{-1}$.

The level-independent QBD has found wide usage in modelling such things as high speed communications systems, mobile cellular systems, database systems and multiprogramming systems (see for example [14], [15], [26], [35], [38], [39], [54] and [55]). The level-dependent QBDs are a generalisation of the level-independent QBD and are considered in the next section. The use of the level-dependent QBD therefore enables the modelling of a much wider class of problems.

Ergodicity in continuous time:

In the continuous time level-independent QBD, recurrence and transience can be ascertained in a completely analogous way to the discrete time case by letting $\boldsymbol{\psi}$ be the solution to

$$\boldsymbol{\psi}\left(A_0^* + A_1^* + A_2^*\right) = \mathbf{0},$$

such that $\boldsymbol{\psi}\ \boldsymbol{e} = 1$. Then, the following drift condition can be shown to correspond to positive-recurrence

$$\boldsymbol{\psi} A_0^* \boldsymbol{e} < \boldsymbol{\psi} A_2^* \boldsymbol{e}.$$

Null-recurrence corresponds to

$$\boldsymbol{\psi} A_0^* \boldsymbol{e} = \boldsymbol{\psi} A_2^* \boldsymbol{e},$$

and transience corresponds to

$$\boldsymbol{\psi} A_0^* \boldsymbol{e} > \boldsymbol{\psi} A_2^* \boldsymbol{e}.$$

6 The equilibrium distribution for a level-dependent QBD

Recall the example given by equation (6), which obviously has the form of a level-dependent QBD. The level-dependent QBD form is not always immediately obvious, as is demonstrated by the following example.

Consider now a *PH/M/1/1* retrial queue where arrivals finding the server busy, go away and retry after an exponentially distributed length of time with parameter γ. Let m be the number of phases of the arrival process given by $(\boldsymbol{\alpha}, S)$. The customers are persistent, in that they do not leave the system until they have received service. We can represent this diagrammatically as in Figure 13.

Arranging the states in lexicographical order, and defining

$$D_{n,j} = \begin{cases} S - I_m\mu & \text{if } j = 1 \\ S - nI_m\gamma & \text{if } j = 0 \end{cases},$$

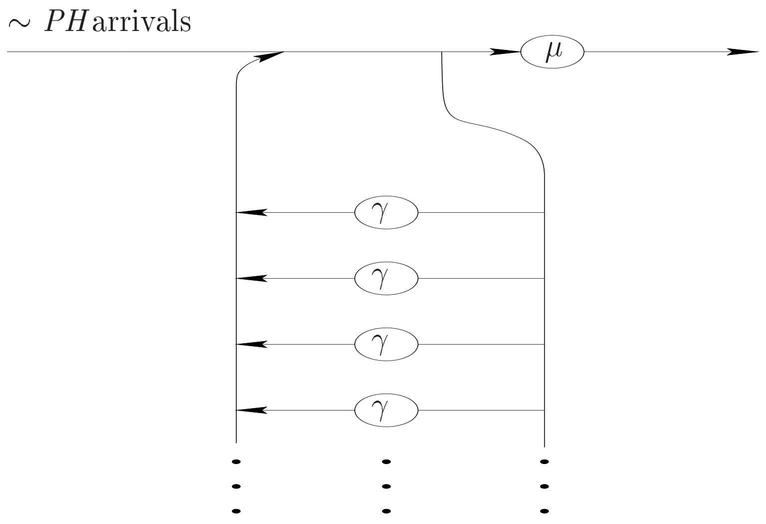


Figure 13: *A retrial scheme*

we can construct a state transition diagram for the retrial scheme as in Figure 14.

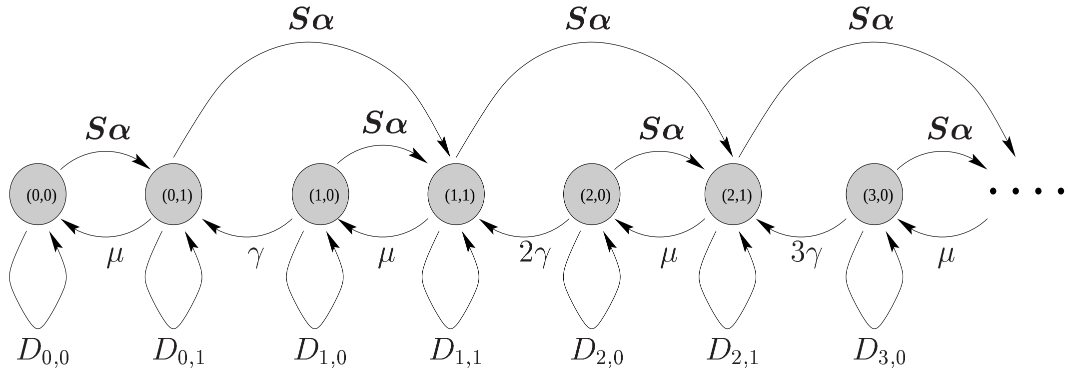


Figure 14: *State transitions*

Here the n and j subscripts on the $D_{n,j}$ matrices represent the number of customers in retrial and the server occupancy respectively.

It is then a simple step to construct the following transition rate matrix.

$$Q = \begin{bmatrix} D_{0,0} & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ I_m\mu & D_{0,1} & 0 & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & 0 & 0 & \\ 0 & I_m\gamma & D_{1,0} & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & 0 & 0 & \\ 0 & 0 & I_m\mu & D_{1,1} & 0 & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & \\ 0 & 0 & 0 & 2I_m\gamma & D_{2,0} & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & \\ 0 & 0 & 0 & 0 & I_m\mu & D_{2,1} & 0 & \boldsymbol{S}^0\boldsymbol{\alpha} & \\ 0 & 0 & 0 & 0 & 0 & 3I_m\gamma & D_{3,0} & \boldsymbol{S}^0\boldsymbol{\alpha} & \\ 0 & 0 & 0 & 0 & 0 & 0 & I_m\mu & D_{3,1} & \\ \vdots & & & & & & & & \ddots \end{bmatrix}, \tag{24}$$

The state transition diagram and block matrix form are not skip free in the upward direction and hence does not obviously fall into either of the level-independent *QBD* or level-dependent *QBD* categories.

However, consider Figure 15 where we regroup the states so that the transitions between "levels" become skip free. Formally, we let n be the

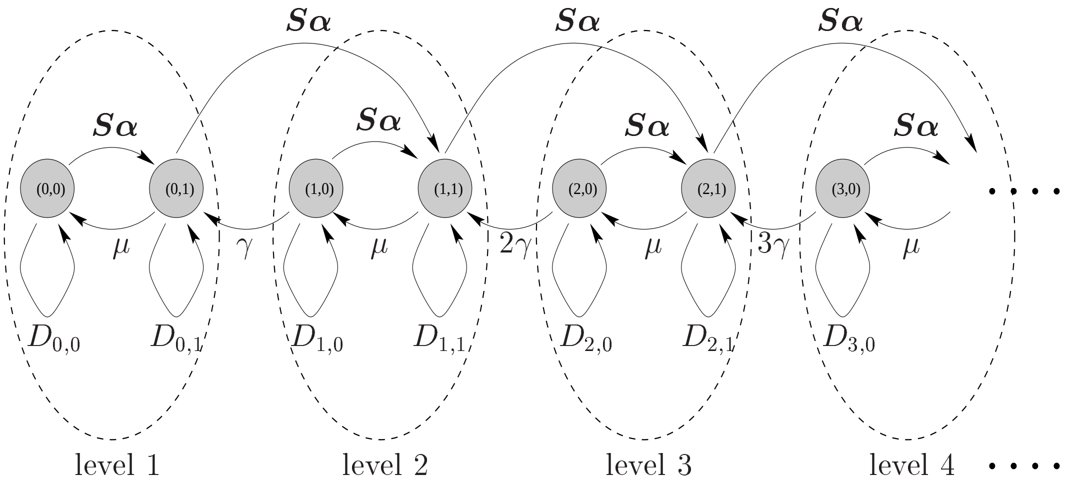


Figure 15: *State transitions*

number of customers in retrial, i be the phase of the *PH*-type arrival process and j represent the server occupancy. Then we let the number of customers in retrial correspond to the level and let the description within each level include information about the arrival phase and the server occupancy as (i, j). Then we can re-block the rate matrix Q in the form of a level-dependent *QBD* as in equation (25).

$$Q = \left[\begin{array}{cc|cc|cc|cc|c}
D_{0,0} & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
I_m\mu & D_{0,1} & 0 & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & 0 & 0 & \\
\hline
0 & I_m\gamma & D_{1,0} & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & 0 & 0 & \\
0 & 0 & I_m\mu & D_{1,1} & 0 & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & \\
\hline
0 & 0 & 0 & 2I_m\gamma & D_{2,0} & \boldsymbol{S}^0\boldsymbol{\alpha} & 0 & 0 & \\
0 & 0 & 0 & 0 & I_m\mu & D_{2,1} & 0 & \boldsymbol{S}^0\boldsymbol{\alpha} & \\
\hline
0 & 0 & 0 & 0 & 0 & 3I_m\gamma & D_{3,0} & \boldsymbol{S}^0\boldsymbol{\alpha} & \\
0 & 0 & 0 & 0 & 0 & 0 & I_m\mu & D_{3,1} & \\
\hline
\vdots & & & & & & & & \ddots
\end{array}\right], \qquad (25)$$

Note here that this re-blocked form is skip free in both the upward and downward directions, with an also obvious level-dependency. The fact that this Markov chain can be considered as a level-dependent *QBD* is not surprising, as any process which has a finite jump size in either direction can be similarly re-blocked to form either a level-independent *QBD* or level-dependent *QBD* structure. The phase and level descriptions may become less appealing but after finding the associated equilibrium distribution, will provide full details of the system. In far greater generality, Ramaswami [46] has in fact shown that the general *M/G/1* and *GI/M/1* paradigms for which matrix-analytic methods apply can be analysed through an appropriately defined level-independent *QBD* or level-dependent *QBD*.

A level-dependent *QBD* cannot be considered in the same way as a level-independent *QBD*, since the level-independent *QBD* has a nice homogeneous structure above level 0. However, the level-dependent *QBD* has been addressed by Ramaswami [45] and an algorithm for numerical calculation of the steady state distribution has both been given and demonstrated by Bright and Taylor in [10] and [11]. Their algorithm for the level-dependent *QBD* is based on the Logarithmic Reduction Algorithm of Latouche and Ramaswami.

We have seen already that the level-dependent *QBD* can be directly used to model queues in a random environment, re-trial queues with multiple servers, re-trial queues with *PH*-type arrival processes and can be easily extended to model infinite server queues in a random environment.

The general block partitioned form of the rate matrix for a level-dependent

QBD is given in equation (26)

$$Q = \begin{bmatrix} A_1^{(0)} & A_0^{(0)} & \mathbf{0} & \cdots & \cdots \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & \mathbf{0} & \cdots \\ \mathbf{0} & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & \\ \vdots & \mathbf{0} & A_2^{(3)} & A_1^{(3)} & \ddots \\ \vdots & \vdots & & \ddots & \ddots \end{bmatrix}, \tag{26}$$

where the matrices $A_2^{(k)}, A_1^{(k)}$ and $A_0^{(k)}$ for $k \geq 0$ are of order $m_k \times m_{k-1}$, $m_k \times m_k$ and $m_k \times m_{k+1}$ respectively, with $A_1^{(0)}$ of order $m_0 \times m_0$. This allows for a full generalisation of a level dependent number of phases.

Remark 6.1 Finite state space: *Thus far we have only considered the level-independent QBD and level-dependent QBD with an infinite state space. The equilibrium distribution for finite state space level-independent QBDs do not generally have a simple analytic form. The M/PH/1/K queue is one of the exceptions and is demonstrated in [37]. Various treatments of the finite level-independent QBD are given in for example [9], [21], [24], [53], and discussed in [31]. Finite state space level-dependent QBDs have been considered in [15] and [14] for evaluation of network overload control and mobile cellular systems. The Folding Algorithm of Ye and Li [53] has also found application in finite state space level-dependent QBDs, where there is a piece-wise level-independence structure.*

Assuming positive recurrence, Ramaswami [45] showed that the equilibrium distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)$ for the level-dependent QBD with rate matrix given by equation (26) satisfies the following relationship.

$$\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n R_n, \quad \text{for } n \geq 0, \tag{27}$$

which may be re-written as

$$\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_0 \prod_{k=o}^{n} R_k, \quad \text{for } n \geq 0, \tag{28}$$

where the matrices R_n for $n \geq 0$ have the following physical interpretation. In discrete time the matrix R contains information about expected number of visits, whereas in continuous time it contains information about expected sojourns. In fact $(R_n)_{ij}$ is the expected sojourn in state j of level $n+1$ per unit of sojourn in state i of level n before returning to level n given that the process started in state i of level n [47]. This family of matrices $\{R_n, n \geq 0\}$ are the minimal non-negative solutions to the system of equations

$$A_0^{(n)} + R_n A_1^{(n+1)} + R_n R_{n+1} A_2^{(n+2)} = 0, \text{ for } n \geq 0. \tag{29}$$

Note that $\boldsymbol{\pi}_0$ is the solution to

$$\boldsymbol{\pi}_0 \left(A_1^{(0)} + R_0 A_2^{(1)} \right) = 0, \tag{30}$$

subject to

$$\boldsymbol{\pi}_0 \sum_{n=0}^{\infty} \left(\prod_{k=o}^{n-1} R_k \right) \boldsymbol{e} = 1.$$

In order to find the equilibrium distribution of the level-dependent QBD, we need to know the matrices $\{R_n, n \geq 0\}$. As we no longer have a homogeneous form above level 0, this becomes a difficult task as the matrices $\{R_n, n \geq 0\}$ may generally only be numerically computed. A general method would be therefore to construct a truncated version of the level-dependent QBD and calculate an approximation to the equilibrium distribution using the truncated state space. The selection of a truncation level M should be such that the amount of probability which is lost above level M is minimal. Neuts and Rao use such a method in [42], to compute the equilibrium distribution for a multi-server retrial queue. They also proposed another more efficient method for approximating the equilibrium distribution for the particular class of model they considered. The augmented truncation method has the following finite rate matrix for a given truncation level M.

$$Q = \begin{bmatrix} A_1^{(0)} & A_0^{(0)} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & \mathbf{0} & & \vdots \\ \mathbf{0} & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & \ddots & \vdots \\ \vdots & & A_2^{(3)} & A_1^{(3)} & \ddots & \mathbf{0} \\ \vdots & & & \ddots & \ddots & A_0^{(M-1)} \\ \mathbf{0} & \dots & \dots & \mathbf{0} & A_2^{(M)} & \hat{A}_1^{(M)} \end{bmatrix}, \tag{31}$$

with the matrix $\hat{A}_1^{(M)}$ given by either

$$\hat{A}_1^{(M)} = A_1^{(M)} + A_0^{(M)}, \tag{32}$$

or

$$\hat{A}_1^{(M)} = A_1^{(M)} + diag(A_0^{(M)} \boldsymbol{e}). \tag{33}$$

In the first case (32), the transition rate from phase i at level M to phase j at level $M+1$ is re-directed to phase j of level M (This corresponds to a phase change only). In the second case (33), the transition rate from phase i at level M to phase j at level $M+1$ is set to zero (There is no level or phase change). This renormalisation at the truncation level can be done in a variety of ways, but as is shown in Gibson and Seneta [22], not all approximations to the equilibrium distribution actually tend to the true stationary vector as the number of states tends to infinity.

In the "exact" truncation used by Bright and Taylor in [11], the matrix $\hat{A}_1^{(M)}$ is given by

$$\hat{A}_1^{(M)} = A_1^{(M)} + R_M A_2^{(M+1)}. \tag{34}$$

The term "exact" truncation refers to results in Ramaswami [45], where the generator matrix for transitions above level M is shown to be the matrix $A_1^{(M)}$ of equation (34). This process then has the interpretation of being the original level-dependent QBD on the infinite state space observed only at time points where it is in levels less than or equal to M. Hence the equilibrium distribution $\boldsymbol{\pi}(M)$ for the process described by equations (31) and (34) has the following interpretation. The probability $(\boldsymbol{\pi}_k(M))_j$ is the stationary probability that the infinite process is in state (k,j), conditional on the process being in the set of states $\{(i,j)|0 \leq i \leq M, 1 \leq j \leq m_i\}$. That is, the invariant measure is exactly that of the original process, with $\boldsymbol{\pi}(M)$ being the normalisation over the truncated state space. Moreover, we have element wise that

$$\boldsymbol{\pi}_k \;<\; \boldsymbol{\pi}_k(M) \text{ for all } M \geq 0 \text{ and } \boldsymbol{\pi}_k = \lim_{M\to\infty} \boldsymbol{\pi}_k(M), \text{ for all } k \geq 0. \tag{35}$$

Therefore by choosing M large enough, we have $\boldsymbol{\pi}_k \approx \boldsymbol{\pi}_k(M)$ for $0 \leq k \leq M$.

Bright and Taylor [11] advocate the direct computation of the rate matrix R_M for some truncation index M and then recursively calculating R_{M-1}, R_{M-2}, etc. The following closed form expression for R_k given in [11] is the continuous time analogue of Theorem 2.2 given by Ramaswami and Taylor in [47]. If the level-dependent QBD is positive recurrent, then

$$R_k \;=\; \sum_{\ell=0}^{\infty} U_k^{\ell} \prod_{i=0}^{\ell-1} D_{k+2^{\ell-1}}^{\ell-i-1} \text{ for all } k \geq 0, \tag{36}$$

where U_k^{ℓ} and D_k^{ℓ} are $m_k \times m_{k+2^\ell}$ and $m_k \times m_{k-2^\ell}$ matrices respectively defined recursively by

$$\begin{aligned} U_k^0 &= A_0^{(k)}(-A_1^{(k+1)})^{-1}, & (37)\\ D_k^0 &= A_2^{(k)}(-A_1^{(k-1)})^{-1}, & (38)\\ U_k^{\ell+1} &= U_k^\ell U_{k+2^\ell}^\ell \left[I - U_{k+2^{\ell+1}}^\ell D_{k+3.2^\ell}^\ell - D_{k+2^{\ell+1}}^\ell U_{k+2^\ell}^\ell\right]^{-1}, & (39)\\ D_k^{\ell+1} &= D_k^\ell D_{k-2^\ell}^\ell \left[I - U_{k-2^{\ell+1}}^\ell D_{k-2^\ell}^\ell - D_{k-2^{\ell+1}}^\ell U_{k-3.2^\ell}^\ell\right]^{-1}. & (40) \end{aligned}$$

In order to calculate the matrix R_k, the infinite sum in (36) must be truncated (note that this truncation is not the same as mentioned above). If we define $R_k(n)$ to be the sum of the first $n+1$ terms in (36), then we truncate the sum when

$$\max\Big(R_k(L) - R_k(L-1)\Big) < \epsilon$$

for some ϵ and set $R_k = R_k(L)$. Bright and Taylor point out that this does not generally guarantee that $\max(R_k(n) - R_k(n-1)) < \epsilon$ holds for all $n > L$, but is the case for most processes of potential interest.

Latouche and Ramaswami favour the calculation of the matrix G_{M+1} and provide a closed form and recursion for calculation of the same in [31]. The matrix R_k can be shown to be given by

$$R_k = A_0^{(k-1)}(I - A_1^{(k)} - A_0^{(k)}G_{k+1})^{-1}.$$

Then a closed form for G_k is given by

$$G_k = \sum_{\ell=0}^{\infty}\left[\prod_{i=0}^{\ell-1} U_{k-1+2^i}^i\right] D_{k-1+2^\ell}^\ell, \tag{41}$$

with the matrices U_k^ℓ and D_k^ℓ defined recursively by

$$\begin{aligned} U_k^0 &= (I - A_1^{(k)})^{-1}A_0^{(k)}, & (42)\\ D_k^0 &= (I - A_1^{(k)})^{-1}A_2^{(k)}, & (43)\\ U_k^{\ell+1} &= \left[I - U_k^\ell D_{k+2^\ell}^\ell - D_k^\ell U_{k-2^\ell}^\ell\right]^{-1} U_k^\ell U_{k+2^\ell}^\ell, & (44)\\ D_k^{\ell+1} &= \left[I - U_k^\ell D_{k+2^\ell}^\ell - D_k^\ell U_{k-2^\ell}^\ell\right]^{-1} D_k^\ell D_{k-2^\ell}^\ell. & (45) \end{aligned}$$

For sufficiently large L, we can approximate G_k by

$$G_k = \sum_{\ell=0}^{L}\left[\prod_{i=0}^{\ell-1} U_{k-1+2^i}^i\right] D_{k-1+2^\ell}^\ell. \tag{46}$$

The method suggested by Latouche and Ramaswami for the positive recurrent case has the nicer stopping condition of

$$|G_k(L)\boldsymbol{e} - \boldsymbol{e}| < \epsilon\, \boldsymbol{e},$$

so that the stochastic matrix $G_k \approx G_k(L)$. The truncation of the infinite level-dependent QBD has the following alternative "exact" form which can be exploited using the matrix G_{M+1}.

$$Q = \begin{bmatrix} A_1^{(0)} & A_0^{(0)} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & \mathbf{0} & & \vdots \\ \mathbf{0} & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & \ddots & \vdots \\ \vdots & & A_2^{(3)} & A_1^{(3)} & \ddots & \mathbf{0} \\ \vdots & & & \ddots & \ddots & A_0^{(M-1)} \\ \mathbf{0} & \dots & \dots & \mathbf{0} & A_2^{(M)} & A_1^{(M)} + A_0^{(M)} G_{M+1} \end{bmatrix}, \tag{47}$$

It is then a simple matter to use any one of the available methods for calculating the equilibrium distribution for the resultant finite state space. The calculated equilibrium distribution is then just a renormalisation of the actual stationary distribution of the infinite level-dependent QBD over the state space $0 \leq k \leq M$.

6.1 Recurrence and Truncation level

It is important to consider whether the level-dependent QBD is positive recurrent or not. If it is, the distribution $\boldsymbol{\pi}(M)$ calculated above is an upper bound for the equilibrium distribution for the infinite level-dependent QBD. Then choosing M large enough will mean that for each k, $\boldsymbol{\pi}_k(M)$ will be very close to $\boldsymbol{\pi}_k$ for the infinite level-dependent QBD.

The question of stability has been looked at in for example [31], where recurrence is expressed in terms of the vector $\boldsymbol{\pi}_0$ or matrix $G^{(1)}$. For example, if the matrix $G^{(1)}$ is stochastic, then it can be proven that the level-dependent QBD is recurrent. However, for a sub-stochastic matrix $G^{(1)}$ which has been numerically computed and where $G^{(1)}\boldsymbol{e}$ is close to $\boldsymbol{e}$, it may be impossible to determine it to be sub-stochastic and hence to ascertain the non-recurrent nature of the level-dependent QBD. Bright and Taylor stated in [11] that to their knowledge, no general method for determining positive recurrence for a level-dependent QBD on an infinite state space existed and should therefore be addressed on an individual basis. To the authors knowledge, this is still the case.

Algorithm 6.1 *A general method for selection of an appropriate truncation M and an algorithm for the calculation of the invariant measure is given in [11]. Their basic general method is as follows*

1. *Set* $M_{old} = 0$
2. *Choose* $M_{new} > M_{old}$, *set* $p = 1$
3. **While** $(p \geq \epsilon)$

    Calculate the first 5 terms of $R_{M_{new}-1}$ *in (36)*

    If $\left(\max \left(R(4)^{(M_{new}-1)} - R(3)^{(M_{new}-1)} \right) < \epsilon \right)$

    Recursively calculate $R_{M_{new}-2}, R_{M_{new}-3}, \ldots$

    Solve for $\boldsymbol{\pi}_0(M_{new})$ *in (30) using* $\boldsymbol{\pi}_0(M_{new})\boldsymbol{e} = 1$

    Calculate invariant measure $\boldsymbol{\pi}(M_{new})$ *using (28) and re-normalise*

    Set $p = \boldsymbol{\pi}_{M_{new}}(M_{new})\boldsymbol{e}$

    Else

    $M_{old} = M_{new}$

    Goto 2.

    End If

    End While
4. *Set* $M = M_{new}$

The above method will generally find an appropriate truncation level M for level-dependent QBDs of practical interest, although as pointed out in [11], a level-dependent QBD can easily be constructed where this will not be the case. For instance, the process can be transient because of transition rates beyond M. However this is unlikely to happen in realistic applications.

The level-dependent QBD under investigation should therefore be carefully considered as there are sometimes specific characteristics which facilitate ease of selection of a truncation level M and the determination of positive recurrence. For example, in the case where we have $A_2^{(k)} = A_2$, $A_1^{(k)} = A_1$ and $A_0^{(k)} = A_0$ for all $k \geq \widehat{M}$, the level-dependent QBD can be considered as a level-independent QBD with a large number of boundary states in the levels $0 \leq k \leq \widehat{M}$. Such processes have also been considered in Neuts [37] where the same approach as Bright and Taylor to find $\boldsymbol{\pi}_k$ for all $k > \widehat{M}$ is computed using

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_{\widehat{M}} \left(R_{\widehat{M}} \right)^{k-\widehat{M}}.$$

For the levels $0 \le k \le \widehat{M}$, Neuts solved

$$\begin{aligned}
\boldsymbol{\pi}_0 A_1^{(0)} + \boldsymbol{\pi}_1 A_2^{(1)} &= \mathbf{0}, \\
\boldsymbol{\pi}_i A_0^{(i)} + \boldsymbol{\pi}_{i+1} A_1^{(i+1)} + \boldsymbol{\pi}_{i+2} A_2^{(i+2)} &= \mathbf{0}, \text{ for } 0 \le i \le \widehat{M} - 2, \\
\boldsymbol{\pi}_{\widehat{M}-1} A_0^{(\widehat{M}-1)} + \boldsymbol{\pi}_{\widehat{M}} A_1^{(\widehat{M})} + \boldsymbol{\pi}_{\widehat{M}} R_{\widehat{M}} A_2^{(\widehat{M})} &= \mathbf{0}, \\
\sum_{i=0}^{\widehat{M}-1} \boldsymbol{\pi}_i \boldsymbol{e} + \boldsymbol{\pi}_{\widehat{M}} (I - R_{\widehat{M}})^{-1} \boldsymbol{e} &= 1.
\end{aligned}$$

Note that the matrix analytic structure is not displayed by this solution as it is by the method of Bright and Taylor.

Positive recurrence in this instance has been shown by Neuts to correspond to the drift condition

$$\boldsymbol{\psi} A_0 \boldsymbol{e} < \boldsymbol{\psi} A_2 \boldsymbol{e},$$

where $\boldsymbol{\psi}$ is the solution to $\boldsymbol{\psi}\,(A_0 + A_1 + A_2) = \mathbf{0}$ such that $\boldsymbol{\psi}\ \boldsymbol{e} = 1$.

For a class of level-dependent *QBD*s demonstrated in [10], it is possible to construct another level-dependent *QBD* that stochastically dominates the original. This can reveal a simple drift condition which guarantees positive recurrence for the original level-dependent *QBD* and makes the selection of a truncation level M much simpler. For instance, they demonstrate a sub-class of these where the marginal distribution over the levels of the dominating process is equal to the equilibrium distribution of a standard *BD* process. It basically is such that if

Condition 6.1 $\sum_j \left(A_2^{(k)}\right)_{i,j} > 0$ *for all* i *and* k*,*

then we can construct a birth and death process which stochastically dominates the level-dependent *QBD*. The rate matrix $\overline{Q}$ for the dominating *BD* process is constructed as follows (see [10] for proof).

$$
\begin{aligned}
(\overline{Q})_{0,1} &= 0 \\
(\overline{Q})_{i,i+1} &= \max\left(A_0^{(i-1)}\boldsymbol{e}\right) \text{ for all } i \geq 0 \\
(\overline{Q})_{1,0} &= 0 \\
(\overline{Q})_{i,i-1} &= \min\left(A_2^{(i-1)}\boldsymbol{e}\right) \text{ for all } i \geq 0 \\
(\overline{Q})_{i,j} &= 0 \text{ otherwise}
\end{aligned}
$$

Then we find M so that the BD process has negligible mass for $k > M$. That is, we find M such that

$$\sum_{i=M}^{\infty} \frac{(\rho)^i e^{-\rho}}{i!} < \epsilon$$

for some ϵ, where

$$\rho = \frac{\max\left(A_0^{(i-1)}\boldsymbol{e}\right)}{\min\left(A_2^{(i-1)}\boldsymbol{e}\right)} .$$

The same will also hold for the level-dependent QBD.

6.2 Numerical examples

Consider an $M/M/\infty$ queue in a random environment, where both the service rate and arrival rate are dependent on the state (phase) of the underlying environment. This is similar to the *M/M/1* queue in a random environment as outlined in Section 4, except that the service rate is now dependent on the state (phase) of the underlying environment and level-dependence is also introduced by having an infinite number of servers. The resultant continuous time level-dependent QBD form for this system has the rate matrix as given in (26), where the matrices $A_2^{(i)}$, $A_1^{(i)}$ and $A_0^{(i)}$ are given by

$$
\begin{aligned}
A_2^{(i)} &= i\mathcal{S}, \\
A_1^{(i)} &= Q^* - \Lambda - i\mathcal{S}, \\
A_0^{(i)} &= \Lambda,
\end{aligned}
$$

Here Q^* is the $m \times m$ generator matrix corresponding to the underlying random environment (modulating Markov process) which determines the arrival rate and service rate given respectively by the diagonal matrices of rates Λ and $\mathcal{S}$. That is

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_m \end{bmatrix} \quad \text{and} \quad \mathcal{S} = \begin{bmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \mu_m \end{bmatrix}$$

This example was used in [10] with a two state matrix Q^* given by

$$Q^* \quad = \quad \begin{bmatrix} -8 & 8 \\ 5 & -5 \end{bmatrix}$$

$\mu_1 = 2$, $\mu_2 = 1$, $\lambda_1 = 40$ and $\lambda_2 = 5$. We reproduce their results here to illustrate some points about their level-dependent QBD techniques. In this case we see that Condition 6.1 is satisfied and that a dominating BD process can therefore be constructed to determine a truncation level M. Using a value $\epsilon = 10^{-10}$ yields a value of $M = 88$, which was subsequently verified as adequate by calculating the equilibrium distribution again using $M = 188$. The difference in the two results was less than 10^{-14} in all cases. The marginal probabilities of being in states $0 \leq k \leq M$ are given in Table 1.

They then chose to set $\mu = 0$ so that Condition 6.1 no longer applied in order to demonstrate the general algorithm 6.1. This corresponds to the situation where the servers are all idle when the underlying Markov modulating process given by Q^* is in state 2. Assuming that $M = 88$ would be at least required for this case, this was used as an initial guess and found to be sufficient. The level-dependent QBD was not shown to be positive-recurrent, but the results shown in Table 2 tend to indicate that this is the case.

Bright and Taylor finally considered the $PH/M/\infty$ queue with the same mixture of two ten stage Erlangs used by Latouche and Ramaswami in [29]. The distribution function had the form

$$0.05841196\ E_{10}(1.168239125)\ +\ 0.94158804\ E_{10}(18.83176088)$$

The service distribution was selected to have a value of $\mu = 0.5$, and hence Condition 6.1 applies in this case and the dominating BD process yields a value of $M = 84$. The marginal probabilities are presented in Table 3.

In all three of the examples given in [10], the values of M are larger than necessary, so that the methods for determination of the truncation level all

k	ℓ_k	k	ℓ_k
0	2.43033×10^{-5}	18	4.36024×10^{-2}
1	2.12737×10^{-4}	19	3.34413×10^{-2}
2	9.61911×10^{-4}	20	2.47627×10^{-2}
3	2.99837×10^{-3}	21	1.77270×10^{-2}
4	7.25313×10^{-3}	22	1.22830×10^{-2}
5	1.45280×10^{-2}	23	8.24654×10^{-3}
6	2.50976×10^{-2}	24	5.36972×10^{-3}
7	3.84463×10^{-2}	25	3.39415×10^{-3}
8	5.32737×10^{-2}	26	2.08432×10^{-3}
9	6.77654×10^{-2}	27	1.24447×10^{-3}
10	8.00201×10^{-2}	28	7.22940×10^{-4}
11	8.84792×10^{-2}	29	4.08889×10^{-4}
12	9.22315×10^{-2}	30	2.25306×10^{-4}
13	9.11276×10^{-2}	31	1.21022×10^{-4}
14	8.57099×10^{-2}	32	6.34068×10^{-5}
15	7.70106×10^{-2}	33	3.24203×10^{-5}
16	6.62927×10^{-2}	34	1.61860×10^{-5}
17	5.48054×10^{-2}	>34	$< 10^{-5}$

Table 1:

over estimate M. For example, in the case of the *PH/M/1* queue we have $M = 84$, yet the table of marginal probabilities indicate that the probability of being above level 18 is less than 10^{-14}.

Latouche and Ramaswami [30] have observed that it is possible to construct level-independent *QBD* processes where an excessive number of iterations are required for their algorithm. This is also relevant to the level-dependent *QBD* case. The possibility of overflow or underflow errors also becomes a potential problem when recursively calculating the matrices R_k. Similar problems were observed by Gaver, Jacobs and Latouche [21] in their computations. Bright and Taylor state that this is more likely when either $A_0^{(i)}$ or $A_2^{(i)}$ are very large or small compared to the other matrices in Q, where they suggest a numerical scaling technique may be appropriate.

k	ℓ_k	k	ℓ_k
0 to 5	$< 10^{-5}$	30	4.07063×10^{-2}
6	1.54870×10^{-5}	31	3.34230×10^{-2}
7	4.99438×10^{-5}	32	2.67954×10^{-2}
8	1.41353×10^{-4}	33	2.10015×10^{-2}
9	3.56725×10^{-4}	34	1.61113×10^{-2}
10	8.12863×10^{-4}	35	1.21114×10^{-2}
11	1.68958×10^{-3}	36	8.93141×10^{-3}
12	3.23061×10^{-3}	37	6.46767×10^{-3}
13	5.72303×10^{-3}	38	4.60375×10^{-3}
14	9.45034×10^{-3}	39	3.22418×10^{-3}
15	1.46231×10^{-2}	40	2.22364×10^{-3}
16	2.13018×10^{-2}	41	1.51153×10^{-3}
17	2.93325×10^{-2}	42	1.01352×10^{-3}
18	3.83202×10^{-2}	43	6.70883×10^{-4}
19	4.76513×10^{-2}	44	4.38713×10^{-4}
20	5.65699×10^{-2}	45	2.83620×10^{-4}
21	6.42890×10^{-2}	46	1.81386×10^{-4}
22	7.01145×10^{-2}	47	1.14829×10^{-4}
23	7.35521×10^{-2}	48	7.20016×10^{-5}
24	7.43740×10^{-2}	49	4.47413×10^{-5}
25	7.26348×10^{-2}	50	2.75666×10^{-5}
26	6.86384×10^{-2}	51	1.68492×10^{-5}
27	6.28697×10^{-2}	52	1.02210×10^{-5}
28	5.59082×10^{-2}	>53	$< 10^{-5}$
29	4.83435×10^{-2}		

Table 2:

k	ℓ_k	k	ℓ_k
0	2.94726×10^{-1}	10	4.00557×10^{-5}
1	1.68517×10^{-1}	11	4.09573×10^{-6}
2	1.57537×10^{-1}	12	3.30731×10^{-7}
3	1.55108×10^{-1}	13	2.12891×10^{-8}
4	1.19684×10^{-2}	14	1.10142×10^{-9}
5	6.70135×10^{-2}	15	4.61353×10^{-11}
6	2.71629×10^{-3}	16	1.57479×10^{-12}
7	8.09367×10^{-3}	17	4.40641×10^{-14}
8	1.80460×10^{-3}	> 18	$< 10^{-14}$
9	3.06062×10^{-4}		

Table 3:

7 Departure processes

The departure process of a queue is important in the analysis of networks of queues, as it may be the arrival process to another queue in the network. There have been many papers [12], [13], [17], [18], [34], [49], dealing with the output process of a single queue. In 1956, Burke [12] and later Reich [48], independently proved the 1955 claim by Morse [34] that “A little thought will convince one that the efflux from a single-channel, fed by Poisson arrivals, must be Poisson with the same rate as the arrivals,...”. This result simplified much of the analysis for tandems that had already been undertaken, for example Jackson [25].

For a *MAP/PH/1* queue with a finite buffer, there exists an exact finite *MAP* description of the output process which has dimension given by the product of the maximum buffer size, the dimension of the matrix descriptors of the input *MAP* and the *PH*-server. This exact representation suffers from the “curse of dimensionality”, especially when considered as a tool in the analysis of a network. Approximate techniques which reduce the size of these representations therefore become necessary. An extensive list of references for methods of analysis of various tandem queues with a finite intermediate buffer is given in [40].

In the infinite buffer case, with which we will be primarily concerned, no such exact finite description has been found. In fact, it appears that there does not exist a finite *MAP* description for the output process of a stationary *MAP/PH/1* queue, in which the *MAP* is not a Poisson process. The possibility of a finite description was addressed in Olivier and Walrand [43], who claimed that the output from an *MMPP/M/1* queue can not have a *MAP* description. Although their proof was shown to be flawed by the authors in [6], we believe the result to be true. In a related paper [5] we addressed the question of when a *MAP* is just a complicated description of a Poisson process, in which case, the output of the *MAP/M/1* queue must also be a Poisson process.

We propose a family of *MAP* approximations to the departure process of the *MAP/PH/1* queue. To check the viability of these approximations, they are used as input to another single-server queue, and the second queue length distributions compared with their exact counterparts, calculated using matrix-analytic techniques. Other techniques to approximate point processes have been given in, for example, [1], [2], [3], [8], [19], [50], [51] and [52]. Where possible, the approximations given here are compared to the results of these publications.

The structures of the defining processes are exploited in the approximations used and as a direct result of this, all of the approximations yield the exact output process for the trivial situation of Poisson arrivals to a negative-exponential first server.

7.1 A Family of Approximations

Let D_0 and $D_1 \geq 0$ be the $m \times m$ matrix descriptors of a MAP and (β, S) be a PH-type service distribution, where the matrix S is $n \times n$. Define $\boldsymbol{S}^0 = -S\boldsymbol{e}$, where $\boldsymbol{e}$ is a column of ones of the appropriate dimension. Then consider a QBD with this MAP arrival process and PH-type service service distribution. This QBD has the conservative rate matrix

$$Q = \begin{bmatrix} B_1 & B_0 & & & \\ B_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & A_2 & A_1 & A_0 \\ & & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{48}$$

where

$$\begin{aligned} B_0 &= D_1 \otimes \boldsymbol{\beta}, \\ B_1 &= D_0, \\ B_2 &= I_m \otimes \boldsymbol{S}^0, \\ A_0 &= D_1 \otimes I_n, \\ A_1 &= I_m \otimes S + D_0 \otimes I_n, \\ A_2 &= I_m \otimes \boldsymbol{S}^0 \boldsymbol{\beta} \end{aligned}$$

and $\otimes$ is the Kronecker product. Assume it defines an irreducible, regular Markov chain. Therefore this Markov chain has at most one stationary distribution $\boldsymbol{\Psi}$ such that $\boldsymbol{\Psi} Q = 0$. This stationary distribution is given by the following modified matrix geometric form

$$\boldsymbol{\Psi} = \boldsymbol{\pi}_0 [I, R_0, R_0 R, R_0 R^2, \ldots], \tag{49}$$

where the $mn \times mn$ matrix R is the minimal non-negative solution to the matrix quadratic equation

$$R^2 A_2 + R A_1 + A_0 = 0. \tag{50}$$

The $m \times mn$ matrix R_0 is given by

$$R_0 = -B_0 \left(A_1 + RA_2\right)^{-1} \tag{51}$$

and the vector $\boldsymbol{\pi}_0$ is the unique positive solution to the system of equations

$$\boldsymbol{\pi}_0(B_1 + R_0 B_2) = \mathbf{0} \quad \text{and} \quad \boldsymbol{\pi}_0 \boldsymbol{e} + \boldsymbol{\pi}_0 R_0 (I - R)^{-1} \boldsymbol{e} = 1. \tag{52}$$

We note that the inverse $\left(A_1 + RA_2\right)^{-1}$ always exists (see, for example, equation (16) of [27]).

Partition (48), so that $Q = Q_0 + Q_1$

$$\text{with } Q_0 = \begin{bmatrix} B_1 & B_0 & & & & \\ & A_1 & A_0 & & & \\ & & A_1 & A_0 & & \\ & & & A_1 & A_0 & \\ & & & & \ddots & \ddots \end{bmatrix} \text{ and } Q_1 = \begin{bmatrix} 0 & & & & & \\ B_2 & 0 & & & & \\ & A_2 & 0 & & & \\ & & A_2 & 0 & & \\ & & & \ddots & \ddots & \end{bmatrix}. (53)$$

The observed transitions recorded by Q_1 are departure transitions and the observed process is the departure process. However, this is not a *MAP* because there are infinitely many states.

One obvious method to obtain a *MAP* approximation for the departure process is to truncate the matrices Q_0 and Q_1 in (53) at a "sufficiently" large level ℓ and use these as the *MAP* descriptors. A major problem with this method is that as the traffic intensity increases, the value of ℓ which gives an accurate approximation grows quickly with ℓ, making the size of the matrix descriptors extremely large.

We also need to think about how to adjust the transition rates at level ℓ in order that the generator remains conservative. Two common ways to do this are to add $\text{diag}(A_0\boldsymbol{e})$ to A_1 or to add A_0 to A_1, both at level ℓ. The first of these effectively assumes that transitions to level $\ell+1$ do not happen, while the second allows these transitions to occur but assumes the process remains at level ℓ. However, it was pointed out in Theorem 1 on page 362 of Bright and Taylor [10] that the best way to do this is to add RA_2 to A_1 at level ℓ. This takes into account the influence of the structure of the *QBD* beyond level ℓ on the invariant measure at level ℓ and below. In fact, the stationary distribution of this truncation is a renormalised version of the stationary distribution of the whole process.

Another way to obtain a *MAP* approximation for the departure process can be obtained by thinking physically about what is happening in the *QBD*.

During the busy period, the departure process from the queue is just the PH-renewal process $(\boldsymbol{\beta}, S)$. It is only when the queue becomes empty that this simplicity is lost. However, if we knew the phase of the arrival process when the busy period finishes, we could model the time until the next busy period. To characterise the departure process exactly, we also need to know the length of the busy period.

Ramaswami in [44] has studied the busy period of queues with a matrix-geometric steady state probability vector. He gives expressions for the Laplace-Stieltjes transform for the distribution function of the busy period which are awkward, even in the case of the simple *M/M/1* queue. The actual distribution function of the busy period of an *M/M/1* queue is given explicitly in equation 2.33 of Cohen [16] as an integral of a mixture of Bessel functions. Neither of these expressions lead easily to a *PH*-type description of the busy period. Thus, we approximate the busy period with a *PH*-type random variable in order to get a MAP description of the departure process. Below we construct a family of approximations indexed by a parameter k. The k^{th} approximation assumes that

1. the phase of the arrival process when the QBD moves from level k to level $k-1$ is given its correct marginal distribution, and

2. the number of services during a sojourn at level k and above is geometrically distributed with the mean chosen such that the sojourn at level k and above has the correct mean.

Thus the $k = 1$ approximation assumes that

1. the phase of the arrival process when a busy period ends has the correct marginal distribution, and

2. the number of services during a busy period is geometrically distributed with the mean chosen such that the busy period has the correct mean.

Physically the k^{th} approximation amalgamates levels k and above into a super level $\bar{k}$ and approximates the distribution of the sojourn in level $\bar{k}$ by a geometric mixture of convolutions of *PH*-type distributions, and the phase on return to level $k-1$ by its correct marginal distribution. What is lost in this approximation is the exact distribution of the sojourn at level $\bar{k}$ and correlations between the return phases and sojourn times.

The distribution of the QBD at level $k-1$, conditional on a departure having just occurred, can be calculated from the stationary distribution of the QBD by (see [37])

$$\boldsymbol{x}_{k-1} = \begin{cases} \boldsymbol{\pi}_0 R_0 B_2(\boldsymbol{\nu} D_1 \boldsymbol{e})^{-1} & \text{for } k = 1, \\ \boldsymbol{\pi}_0 R_0 R^{k-1} A_2(\boldsymbol{\nu} D_1 \boldsymbol{e})^{-1} & \text{for } k > 1, \end{cases} \tag{54}$$

where we recall that $\boldsymbol{\nu}$ is the unique stationary distribution of the MAP arrival process. In our approximation, the probability of return to level $k-1$ after each service in the super-level $\bar{k}$ is

$$\frac{\boldsymbol{x}_{k-1}\boldsymbol{e}}{(\sum_{j=k-1}^{\infty} \boldsymbol{x}_j \boldsymbol{e})},$$

and the distribution of the return phase given that a return occurs is given by

$$\frac{\boldsymbol{x}_{k-1}}{\boldsymbol{x}_{k-1}\boldsymbol{e}}.$$

Thus, the unconditioned distribution of return phase at level $k-1$ is given by

$$\boldsymbol{y}_{k-1} = \frac{\boldsymbol{x}_{k-1}}{(\sum_{j=k-1}^{\infty} \boldsymbol{x}_j \boldsymbol{e}}). \tag{55}$$

The k^{th} MAP approximation for the departure process of the $MAP/PH/1$ queue is given by the $\Big(m+n+(k-1)mn\Big) \times \Big(m+n+(k-1)mn\Big)$ matrices

$$Q_0(k) = \begin{bmatrix} B_1 & B_0 & & & & \\ & A_1 & A_0 & & & \\ & & A_1 & \ddots & & \\ & & & \ddots & A_0 & \\ & & & & A_1 & E_0 \\ & & & & & E_1 \end{bmatrix}, Q_1(k) = \begin{bmatrix} 0 & & & & & \\ B_2 & 0 & & & & \\ & A_2 & \ddots & & & \\ & & \ddots & 0 & & \\ & & & A_2 & 0 & \\ & & & & E_2 & E_3 \end{bmatrix}. \tag{56}$$

Here the sub-matrices are as defined previously with

$$\begin{aligned} E_0 &= D_1 \boldsymbol{e} \otimes I_n, \\ E_1 &= S, \\ E_2 &= \boldsymbol{S}^0 \boldsymbol{y}_{k-1}, \\ E_3 &= (1 - \boldsymbol{y}_{k-1}\boldsymbol{e})\boldsymbol{S}^0 \boldsymbol{\beta}. \end{aligned}$$

The stationary distribution of this MAP approximation can be shown to be given by

$$\boldsymbol{\nu}(k) = \boldsymbol{\pi}_0[I, R_0, R_0 R, \ldots, R_0 R^{k-2}, R_0 R^{k-1}(I_{mn} - R)^{-1}(\boldsymbol{e}_m \otimes I_n)]. \tag{57}$$

It seems intuitive that neither of the approximations discussed above should be too drastic. The test of this however, will come with the numerical results that we discuss in Section 7.3.

Special case:

For the special case of $k = 1$, the MAP approximation to the departure process of the *MAP/PH/1* queue reduces to a PH-renewal process with the exact stationary inter-departure time distribution. This approximation precisely captures the distribution of the inter-departure times but ignores any correlation structures between these times.

Bitran and Dasu [8] show that a *MAP/PH/1* queue has a stationary inter-departure time distribution given by the PH distribution $(\boldsymbol{\alpha}, Q_0(1))$ where

$$\boldsymbol{\alpha} = (\boldsymbol{x}_0, (1 - \boldsymbol{x}_0\boldsymbol{e})\boldsymbol{\beta}), \quad Q_0(1) = \begin{pmatrix} D_0 & D_1\boldsymbol{e}\boldsymbol{\beta} \\ 0 & S \end{pmatrix}, \tag{58}$$

and $\boldsymbol{x}_0$ is the distribution of phases of the arrival process immediately after a departure that leaves the queue empty, calculated using (54). The stationary distribution of this MAP approximation can be shown to be given by

$$\boldsymbol{\nu}(1) = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_0 R_0 (I_{mn} - R)^{-1}(\boldsymbol{e}_m \otimes I_n)].$$

This $(m+n) \times (m+n)$ MAP approximation is satisfactory in a great many situations, as is shown by the results in Section 7.3. For the case of Poisson arrivals of rate $\lambda > 0$ to a negative exponential server of rate $\mu > 0$, such that $\lambda < \mu$, the representation is exact as can be seen by inserting

$$D_0 = -\lambda, D_1 = \lambda, S = -\mu, \boldsymbol{\beta} = 1,$$

so that

$$\boldsymbol{\alpha} = \left(\frac{\mu - \lambda}{\mu}, \frac{\lambda}{\mu}\right), \quad Q_0(1) = \begin{pmatrix} -\lambda & \lambda \\ 0 & -\mu \end{pmatrix}.$$

Note that in this case, because the arrivals are Poisson and service is negative exponential, $\boldsymbol{\pi}_0 \equiv \boldsymbol{x}_0$. The vector $\boldsymbol{\beta} = 1$, so that from (58) the vector $\boldsymbol{\alpha}$ is identical to the stationary distribution $\boldsymbol{\nu}(1)$. The stationary probability vector $\boldsymbol{\nu}(1)$ of this process is also the left eigenvector of the matrix $Q_(0)$ corresponding to eigenvalue λ, which by Corollary 2.10 of [5] implies that the departure process under stationary conditions is Poisson of rate λ.

These MAP approximations to the departure process give excellent accuracy, as can be seen from the results given in Section 7.3. Increasing the value of k clearly increases the accuracy of the approximation, but also the

Figure 16: Tandem queueing model.

complexity of the *MAP*. This then requires a trade off be made by the modeller between the degree of accuracy required and the time taken to do the analysis.

7.2 The Tandem Queueing Models

To measure the accuracy of the approximations of the departure process of the *MAP/PH/1* queue, we used the approximations as input to another single-server queue and the stationary distribution of the queue length of the second server compared to that of a tandem queueing system as in Figure 16.

The tandem queueing system of two *FIFO* servers in series can be represented by a *QBD* as outlined in [7], by setting the number in the first and second queues to be the level and part of the phase description respectively. Under this regime, the size of one of the queues must be truncated at a "sufficiently large" value that will not affect the calculation of the stationary distribution of the second queue length. We refer to the results calculated for this *QBD* model as "exact" throughout this paper.

For comparison of numerical results, we calculated the probability distributions of the stationary second queue length so that any queue length probability being less than 10^{-14} was considered as 0. This distribution was then used to calculate the mean and variance for the stationary queue length.

Six different arrival processes, as given in Appendix A, were used to feed the tandem queueing system, including three *PH*-renewal processes and three correlated *MAP*s. The first *PH*-type server was one of the four given in Appendix B and the second *PH*-type server was one of the four given in Appendix C. The service distributions were scaled appropriately to give the correct traffic intensity at each queue. Each of the two FIFO servers had an infinite buffer with the traffic intensity at each queue being one of $(0.1, 0.25, 0.5, 0.75)$.

All permutations of the above were investigated, except the situation where the fifth arrival process was fed to the first queue with traffic intensities 0.5 and 0.75 and with traffic intensity 0.75 at the second queue. The "exact" results were not able to be evaluated for these traffic intensities because of the very bursty nature of the arrival process. To achieve a truncation for the *QBD* models according to the "sufficiently large" regime proved too large. Our approximation however, still delivers a result. Overall there were 1504 different queueing tandems to check the approximation methods via the moments of the second queue.

In Appendix A, the squared coefficient of variation $\frac{\sigma^2}{\mu^2}$ is given for each of the arrival processes. For those processes which have a non-zero lag correlation structure, an indication of the level of this structure is given by the first two serial correlations c_1, c_2, calculated by (see [41])

$$c_i = \frac{\frac{2}{\boldsymbol{\nu} D_1 \boldsymbol{e}} \boldsymbol{\nu}[(-D_0)^{-1} D_1]^i (-D_0)^{-1} \boldsymbol{e} - \frac{1}{(\boldsymbol{\nu} D_1 \boldsymbol{e})^2}}{\frac{2}{\boldsymbol{\nu} D_1 \boldsymbol{e}} \boldsymbol{\nu}(-D_0)^{-1} \boldsymbol{e} - \frac{1}{(\boldsymbol{\nu} D_1 \boldsymbol{e})^2}} \text{ for } i \geq 1, \tag{59}$$

where $\boldsymbol{\nu}$ is the stationary probability vector of the *MAP*. Note that all *PH*-renewal arrival processes are un-correlated by their very nature.

7.3 The results

The results are compiled for the approximations with four variants for the aggregated state approximation. These are

- AGG1: $k = 1$, exact inter-departure time
- AGG2: $k = 2$
- AGG3: $k = 3$
- AGG-3: $k = \max(j : p(j) > 10^{-3})$,

where $p(j)$ is the marginal probability of there being j customers in the first queue of the tandem. Overall results for the 1504 different queueing tandems are broken up into 16 sets which specify the traffic intensities for both queues, then subsequent smaller groupings are given which tend to show the particular applicability of each variant.

The sets $\{a, b, \ldots, k\}$ and $\{m, n, o\}$ given in Table 1, each comprise 96 different combinations of the 6 arrival processes, 4 first server and 4 second server processes for the values of traffic intensity at each queue specified by ρ_1 and ρ_2. Sets (l, p) both comprise 80 different combinations of the arrival processes (all except for the fifth arrival process), 4 first server and 4 second server processes for the values of traffic intensity at each queue specified by ρ_1 and ρ_2.

The second group of 16 sets $\{a^*, b^*, \ldots, p^*\}$ given in Table 2 are subsets of the sets $\{a, b, \ldots, p\}$ given in Table 1. They are comprised of 48 different combinations of the 3 un-correlated PH-renewal arrival processes given in Appendix A, 4 first server and 4 second server processes for the values of traffic intensity at each queue specified by ρ_1 and ρ_2.

The results given in Tables 1 and 2 are the average over each set of the absolute value of the relative percentage error, given by

$$\frac{100\ |\text{exact- approximation}|}{\text{exact}},$$

for the first and second central moments of the queue length of the stationary second queue.

Table 1.

Set	*Traffic intensities*		*Averaged absolute relative percentage error for the 1^{st} and 2^{nd} central moments of the stationary queue length of the second queue*								
			AGG1		AGG2		AGG3		AGG-3		
	ρ_1	ρ_2	E[μ]	E[σ^2]	E[μ]	E[σ^2]	E[μ]	E[σ^2]	E[μ]	E[σ^2]	k
a	0.1	0.1	0.3501	1.1388	0.0154	0.0650	0.0018	0.0090	0.0000	0.0002	3.6
b	0.1	0.25	2.6681	6.3460	0.1127	0.3817	0.0097	0.0484	0.0003	0.0021	3.6
c	0.1	0.5	5.6243	10.6426	0.1986	0.5438	0.0103	0.0440	0.0003	0.0017	3.6
d	0.1	0.75	6.3970	11.2535	0.2196	0.5240	0.0114	0.0352	0.0004	0.0014	3.6
e	0.25	0.1	0.2119	0.7201	0.0100	0.0414	0.0011	0.0052	0.0000	0.0000	6.0
f	0.25	0.25	2.2317	5.8062	0.2839	1.0148	0.0766	0.3366	0.0001	0.0009	6.0
g	0.25	0.5	6.1140	12.4001	0.8809	2.7517	0.2898	1.1332	0.0036	0.0211	6.0
h	0.25	0.75	7.4304	13.9464	1.1038	2.9392	0.3747	1.1412	0.0084	0.0428	6.0
i	0.5	0.1	0.0706	0.2601	0.0038	0.0166	0.0004	0.0018	0.0000	0.0000	12.3
j	0.5	0.25	0.6097	1.9343	0.0666	0.2604	0.0128	0.0579	0.0000	0.0000	12.3
k	0.5	0.5	4.6695	11.3070	1.4906	4.6422	0.8400	2.8534	0.0033	0.0206	12.3
l	0.5	0.75	2.7466	8.5090	1.2067	4.2924	0.5655	2.2551	0.0027	0.0189	8.9
m	0.75	0.1	0.0234	0.0887	0.0014	0.0061	0.0001	0.0007	0.0000	0.0000	26.3
n	0.75	0.25	0.1761	0.6198	0.0202	0.0839	0.0037	0.0170	0.0000	0.0000	26.3
o	0.75	0.5	1.3195	3.9419	0.3408	1.2225	0.1401	0.5490	0.0000	0.0000	26.3
p	0.75	0.75	2.8511	9.0210	1.6456	5.7547	0.9628	3.7110	0.0005	0.0040	18.6
Ave. abs. relative percentage error			2.7184	6.1210	0.4750	1.5338	0.2063	0.7624	0.0012	0.0071	

The results show a marked difference in the accuracy for both the mean queue length and its variance between the AGG1 approximation and the AGG2 approximation, with still further improvement for the AGG3 approximation. The AGG-3 approximation is not a fixed level aggregation but is dynamically decided by a specific first queue length probability truncation, making some of the MAP approximations smaller or larger than those of the AGG3 approximation. With this type of parameterisation, excellent results can be gained as is seen by Tables 1 and 2, but the MAP approximations become large as ρ_1 becomes large. The average value of k is given for the AGG-3 approximation in the extreme right hand column.

Table 2.

Set	*Traffic intensities*		*Averaged absolute relative percentage error for the* 1^{st} *and* 2^{nd} *central moments of the queue length of the second queue*								
			AGG1		AGG2		AGG3		AGG-3		
	ρ_1	ρ_2	E[μ]	E[σ^2]	E[μ]	E[σ^2]	E[μ]	E[σ^2]	E[μ]	E[σ^2]	k
a*	0.1	0.1	0.0172	0.0715	0.0013	0.0064	0.0001	0.0006	0.0000	0.0000	3.3
b*	0.1	0.25	0.0850	0.3544	0.0059	0.0311	0.0006	0.0039	0.0000	0.0003	3.3
c*	0.1	0.5	0.2074	0.7835	0.0148	0.0694	0.0017	0.0092	0.0001	0.0007	3.3
d*	0.1	0.75	0.3126	0.9396	0.0213	0.0751	0.0024	0.0095	0.0001	0.0006	3.3
e*	0.25	0.1	0.0339	0.1357	0.0030	0.0142	0.0004	0.0020	0.0000	0.0000	4.9
f*	0.25	0.25	0.1733	0.6955	0.0231	0.1112	0.0041	0.0223	0.0000	0.0001	4.9
g*	0.25	0.5	0.6910	2.5918	0.1175	0.5370	0.0264	0.1406	0.0002	0.0019	4.9
h*	0.25	0.75	1.3269	4.1176	0.2637	0.9543	0.0641	0.2633	0.0008	0.0044	4.9
i*	0.5	0.1	0.0354	0.1384	0.0030	0.0138	0.0004	0.0018	0.0000	0.0000	8.5
j*	0.5	0.25	0.1670	0.6470	0.0278	0.1267	0.0062	0.0315	0.0000	0.0000	8.5
k*	0.5	0.5	1.0137	3.7074	0.2724	1.1610	0.0851	0.4099	0.0001	0.0008	8.5
l*	0.5	0.75	3.0406	9.7222	1.1981	4.3764	0.5236	2.1395	0.0024	0.0168	8.5
m*	0.75	0.1	0.0183	0.0709	0.0014	0.0064	0.0002	0.0008	0.0000	0.0000	18.0
n*	0.75	0.25	0.1121	0.4237	0.0175	0.0760	0.0036	0.0173	0.0000	0.0000	18.0
o*	0.75	0.5	0.6099	2.1485	0.1757	0.7126	0.0595	0.2666	0.0000	0.0000	18.0
p*	0.75	0.75	3.0753	9.9312	1.6636	5.9560	0.9312	3.6546	0.0004	0.0032	18.0
Ave. abs. relative percentage error			0.6825	2.2799	0.2381	0.8892	0.1069	0.4358	0.0003	0.0018	

7.4 Comparison to other work

Whitt [52] used the work of Shimshak [50] on expected waiting times of queues in series, to compare his work on approximating departure processes. Shimshak used simulation results for the total expected waiting time in two single server queues in series. The Fraker, Page and Marchal approximations are approximations devised by Shimshak using previous approximations. The *M/M/1*, *M/G/1*, *QNA* and modified *QNA* approximations are described in Section 5 of [52].

We produce results for Shimshak's experiments *I*, *III* and *IV* using the methods that we have proposed. These experiments involved three different renewal arrival processes used to feed two queues in series. The arrival processes were a Poisson process, a two state hyper-exponential process with squared coefficient of variation 4 and an Erlang renewal process of order 10. The service distributions were either negative-exponential (E_1) or Erlang of order 10 (E_{10}), with the traffic intensities either 0.6 or 0.8.

Experiment *I* has a Poisson arrival process of rate one with the first server fixed as E_{10}. In experiments *III* and *IV*, the arrival processes of rate one are the two state hyper-exponential and Erlang (E_{10}) process respectively with the first queue traffic intensity set to 0.8 for both cases. Shimshak's experiment *IV* has the same traffic intensities and service time distribution as his experiment *III*. The traffic intensity and service time distributions for Shimshak's experiments *I* and *III* are given in Tables 3 and 4.

Table 3.

Shimshak's experiment I				
System	*Queue 1*		*Queue 2*	
number	ρ_1	distr.	ρ_2	distr.
1	0.6	E_{10}	0.6	E_1
2	0.8	E_{10}	0.6	E_1
3	0.6	E_{10}	0.6	E_{10}
4	0.8	E_{10}	0.6	E_{10}
5	0.6	E_{10}	0.8	E_1
6	0.8	E_{10}	0.8	E_1
7	0.6	E_{10}	0.8	E_{10}
8	0.8	E_{10}	0.8	E_{10}

Table 4.

Shimshak's experiment III				
System	*Queue 1*		*Queue 2*	
number	ρ_1	distr.	ρ_1	distr.
1	0.8	E_1	0.6	E_1
2	0.8	E_{10}	0.6	E_1
3	0.8	E_1	0.6	E_{10}
4	0.8	E_{10}	0.6	E_{10}
5	0.8	E_1	0.8	E_1
6	0.8	E_{10}	0.8	E_1
7	0.8	E_1	0.8	E_{10}
8	0.8	E_{10}	0.8	E_{10}

The results for Shimshak's experiments *I*, *III* and *IV* are given respectively in Tables 5, 6 and 7. The simulation estimates of the total expected waiting time at the first and second queues were obtained by Shimshak and have a 95% confidence interval given in brackets below each estimate. The tables were produced in the same form as that of [52] for easy comparison. The total expected waiting times at both queues as calculated by each of the approximation methods are given, together with the relative error (in brackets).

Table 5.

Shimshak's experiment I, with Poisson arrival stream										
		Approximations								
Sys. no.	*Sim. est.*	M/M/1	M/G/1	Fraker	Page	Marchal	QNA (st. int.)	Mod. QNA	AGG1	AGG2
1	1.20	1.80	1.40	1.19	1.20	1.18	1.25	1.30	1.11	1.15
	(±0.09)	(0.50)	(0.17)	(-0.01)	(0.00)	(-0.02)	(0.04)	(0.08)	(-0.08)	(-0.04)
2	2.27	4.10	2.66	2.30	2.31	2.28	2.38	2.42	2.23	2.25
	(±0.23)	(0.81)	(0.17)	(0.01)	(0.02)	(0.01)	(0.05)	(0.07)	(-0.02)	(-0.01)
3	0.78	1.80	0.99	0.77	0.84	0.84	0.84	0.89	0.66	0.68
	(±0.06)	(1.31)	(0.27)	(-0.02)	(0.08)	(0.07)	(0.08)	(0.14)	(-0.15)	(-0.13)
4	1.83	4.10	2.26	1.90	1.99	1.98	1.98	2.01	1.83	1.83
	(±0.22)	(1.24)	(0.23)	(0.04)	(0.09)	(0.08)	(0.08)	(0.10)	(0.00)	(0.00)
5	3.41	4.10	3.70	3.07	3.10	3.06	3.21	3.27	2.98	3.18
	(±0.43)	(0.20)	(0.09)	(-0.10)	(-0.09)	(-0.10)	(-0.06)	(-0.04)	(-0.13)	(-0.07)
6	4.33	4.33	6.40	4.96	3.85	3.70	3.84	4.07	3.78	3.91
	(±0.60)	(0.48)	(0.15)	(-0.14)	(-0.10)	(-0.11)	(-0.06)	(-0.05)	(-0.13)	(-0.10)
7	1.93	4.10	2.26	1.60	1.73	1.72	1.77	1.83	1.48	1.64
	(±0.27)	(1.12)	(0.17)	(-0.17)	(-0.10)	(-0.11)	(-0.08)	(-0.05)	(-0.23)	(-0.15)
8	2.48	6.40	3.52	2.43	2.58	2.57	2.63	2.68	2.27	2.33
	(±0.29)	(1.58)	(0.42)	(-0.02)	(0.04)	(0.04)	(0.06)	(0.08)	(-0.08)	(-0.06)
Average relative error		0.91	0.21	-0.05	-0.01	-0.02	0.01	0.04	-0.10	-0.07
Average absolute relative error		0.91	0.21	0.06	0.07	0.07	0.06	0.08	0.10	0.07

Table 6.

Shimshak's experiment III, with hyper-exponential arrival stream										
		Approximations								
Sys. no.	*Sim. est.*	M/M/1	M/G/1	Fraker	Page	Marchal	QNA (st. int.)	Mod. QNA	AGG1	AGG2
1	9.08	4.10	4.10	10.30	–	10.39	9.39	9.56	9.02	9.08
	(±1.38)	(-.055)	(-0.55)	(0.13)		(0.14)	(0.03)	(0.05)	(0.01)	(0.00)
2	6.49	4.10	6.49	–	–	7.91	7.72	7.94	6.78	6.79
	(±0.73)	(-0.37)	(-0.51)			(0.22)	(0.19)	(0.22)	(0.04)	(0.05)
3	8.55	4.10	3.18	10.17	–	9.86	8.98	9.16	8.53	8.53
	(±1.20)	(-0.52)	(-0.63)	(0.19)		(0.15)	(0.05)	(0.07)	(0.00)	(0.00)
4	6.01	4.10	2.26	–	–	7.43	7.31	7.54	6.25	6.25
	(±0.73)	(-0.32)	(-0.62)			(0.24)	(0.22)	(0.25)	(0.04)	(0.04)
5	12.31	6.40	6.40	13.50	–	13.54	12.92	13.09	12.38	12.55
	(±2.26)	(-0.48)	(-0.48)	(0.10)		(0.10)	(0.05)	(0.06)	(0.01)	(0.02)
6	9.64	6.40	4.96	–	–	10.78	10.67	10.88	9.30	9.44
	(±1.33)	(-0.34)	(-0.49)			(0.12)	(0.11)	(0.13)	(-0.04)	(-0.02)
7	11.13	6.40	4.96	12.55	–	11.90	11.49	11.65	10.68	10.79
	(±1.37)	(-0.42)	(-0.55)	(-0.13)		(0.07)	(0.03)	(0.05)	(-0.04)	(-0.03)
8	7.40	6.40	3.52	–	–	9.21	9.23	9.44	7.17	7.25
	(±0.95)	(-0.14)	(-0.52)			(0.24)	(0.25)	(0.28)	(-0.03)	(-0.02)
Average relative error		-0.39	-0.54	0.14		0.16	0.12	0.14	0.00	0.00
Average absolute relative error		0.39	0.54	0.14		0.16	0.12	0.14	0.03	0.02

Table 7.

Shimshak's experiment IV, with Erlang (E_{10}) arrival stream										
		Approximations								
Sys. no.	*Sim. est.*	M/M/1	M/G/1	Fraker	Page	Marchal	QNA (st. int.)	Mod. QNA	AGG1	AGG2
1	2.30	4.10	4.10	2.24	2.30	2.21	2.29	2.25	2.30	2.24
	(±0.19)	(0.78)	(0.78)	(-0.03)	(0.00)	(-0.04)	(0.00)	(-0.02)	(0.00)	(-0.03)
2	0.59	4.10	3.18	0.58	0.65	0.59	0.56	0.56	0.58	0.57
	(±0.04)	(5.95)	(4.39)	(-0.02)	(0.10)	(0.00)	(-0.05)	(-0.05)	(-0.02)	(-0.03)
3	1.95	4.10	3.18	1.81	1.92	1.84	1.89	1.85	1.92	1.89
	(±0.45)	(1.10)	(0.63)	(-0.07)	(-0.02)	(-0.06)	(-0.03)	(-0.05)	(-0.02)	(-0.03)
4	0.25	4.10	2.26	0.27	0.38	0.35	0.20	0.20	0.25	0.25
	(±0.02)	(15.40)	(8.04)	(0.08)	(0.52)	(0.40)	(-0.20)	(-0.20)	(0.00)	(0.00)
5	3.84	6.40	6.40	4.27	4.30	4.26	4.21	4.17	4.31	4.02
	(±0.33)	(0.67)	(0.67)	(0.11)	(0.12)	(0.11)	(0.10)	(0.09)	(0.12)	(0.05)
6	1.82	6.40	4.96	1.77	1.85	1.75	1.85	1.85	1.83	1.77
	(±0.19)	(2.52)	(1.73)	(-0.03)	(0.02)	(-0.04)	(0.02)	(0.02)	(0.01)	(-0.03)
7	2.68	6.40	4.96	2.79	2.92	2.88	2.77	2.73	2.90	2.68
	(±0.52)	(1.39)	(0.84)	(0.04)	(0.09)	(0.07)	(0.03)	(0.02)	(0.08)	(0.00)
8	0.46	6.40	3.52	0.50	0.61	0.58	0.43	0.43	0.49	0.45
	(±0.02)	(12.91)	(6.65)	(0.09)	(0.33)	(0.26)	(-0.06)	(-0.06)	(0.07)	(-0.02)
Average relative error		5.09	2.97	0.02	0.15	0.09	-0.02	-0.03	0.03	-0.01
Average absolute relative error		5.09	2.97	0.06	0.15	0.12	0.06	0.06	0.04	0.02

These relative errors, were calculated by taking the difference of the approximation to the simulation estimate and dividing by the simulation estimate. Following the form of the tables in [52], we have given the average relative and average absolute relative errors for each approximation method across the eight runs.

Note that the two columns labelled QNA are in general superior to all the columns to the left. Therefore we shall confine our comparison to those marked QNA. It is interesting to note that our approximations do not show their best results with a Poisson arrival stream, but this is not the case with the QNA methods. When the arrival stream is Hyper-exponential, it appears that our methods strongly outperform the QNA methods as well as the others. When the arrival stream is Erlang, our method again outperforms the QNA methods with the method labelled AGG2 the better. When the arrival process is Poisson, it is hard to separate any of the methods and so the simplest method should be chosen. It would be interesting to understand why both of our methods perform so much worse with a Poisson arrival stream.

8 Correlation structures

It can be shown (see [23]) that the stationary inter-event time distribution for each of our approximations is in fact the same as the stationary inter-departure time distribution for the *MAP/PH/1* queue. This result implies that both the mean and variance of the stationary inter-event time distribution for each of our approximations is in fact the same as the mean and variance of the stationary inter-departure time distribution for the *MAP/PH/1* queue. The correlation structure between successive inter-event times must therefore be all that separates our approximations from the actual departure process. One measure of this correlation structure is given by the lag-correlation coefficients. Our k^{th} approximation captures the first $k-1$ lag-correlation coefficients exactly for the stationary inter-departure times of the *MAP/PH/1* queue.

We will first show the form of these lag-correlation coefficients for the *MAP/PH/1* queue and then show the direct equivalence of the first $k-1$ lag-correlation coefficients to those of our k^{th} approximation.

8.1 Correlation structure of *MAP/PH/1* departure processes

By considering the filtration of the *MAP/PH/1* queue given by the matrices Q_0 and Q_1, we will establish the equation for the lag-correlation coefficients of the stationary inter-departure times. The matrix $Q = Q_0 + Q_1$ has finite, negative diagonal entries and finite, non-negative off diagonal entries. The row sums are all equal to zero, so that Q is a stable and in fact uniform conservative generator matrix. The matrix Q_1 also has some positive entries at every level greater than zero (as departures are possible from every occupied level). It can also be shown (see [23]) that the matrix $-Q_0^{-1}$ exists and has a simple form.

Let W_1 and W_r be the random variables representing the first and r^{th} inter-departure time respectively, immediately after an arbitrary departure under the stationary regime. The r^{th} lag-correlation coefficient, c_r, is given by

$$c_r = \frac{Cov(W_1, W_r)}{\Big(Var(W_1)Var(W_r)\Big)^{\frac{1}{2}}} = \frac{E[W_1 W_r] - E[W_1]E[W_r]}{\Big(Var(W_1)Var(W_r)\Big)^{\frac{1}{2}}} \tag{60}$$

Under the stationary regime, the inter-departure times for the *MAP/PH/1* queue have the same expected mean and variance as the stationary inter-event times for our family of approximations. Hence to prove the equivalence of any lag-correlation coefficients, we need only prove equivalence of the joint expectation between the 1^{st} and r^{th} inter-departure times of the stationary *MAP/PH/1* queue and the 1^{st} and r^{th} inter-event times of the stationary versions of our approximations.

We will construct an expression for the joint expectation between the 1^{st} and r^{th} inter-departure times of the stationary *MAP/PH/1* queue using a probabilistic argument. This will involve the embedded Markov chain immediately after departure epochs. Another approach is given in [23] using Laplace-Stieltjes transforms and matrix resolvents as defined in Chapter 1.3 of Anderson [4].

The expectation $E[W_1 W_r]$ clearly involves two dependent random variables. However, these random variables are conditionally independent given the state j_1 at the end of W_1 and the state i_{r-1} at the beginning of W_r. Thus we can write

$$\begin{aligned} E[W_1 W_r] &= \sum_{j_1, i_{r-1}} E[W_1 W_r | j_1, i_{r-1}] P(j_1, i_{r-1}) \\ &= \sum_{j_1, i_{r-1}} E[W_1 | j_1] P(j_1, i_{r-1}) E[W_r | i_{r-1}] \\ &= \sum_{j_1, i_{r-1}} E[W_1 | j_1] P(j_1) P(i_{r-1} | j_1) E[W_r | i_{r-1}]. \end{aligned} \tag{61}$$

This is illustrated in Figure 17.

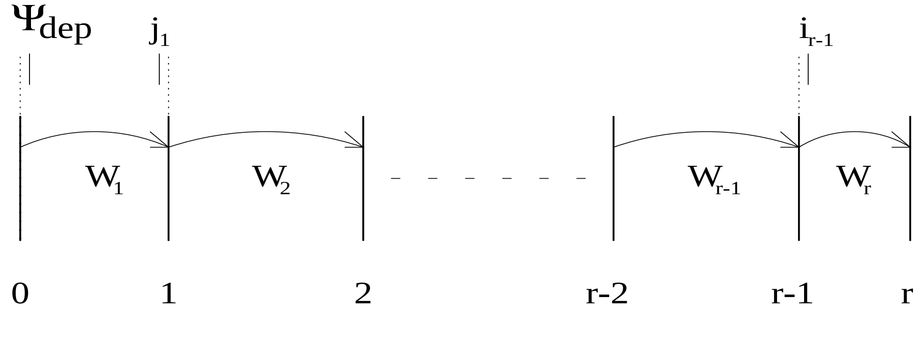


Figure 17: Stationary inter-departure intervals.

Recall that in our filtration of the *MAP/PH/1* queue, the matrix Q_0 governs transitions which are un-observed and the matrix Q_1 governs those transitions which are observed (the departure process). We will construct individual expressions for each of the terms $P(i_{r-1}|j_1)$, $E[W_r|i_{r-1}]$, $E[W_1|j_1]$ and $P(j_1)$.

The matrix $-Q_0^{-1}Q_1$ is the probability transition matrix of the embedded Markov chain immediately after departure epochs of the *MAP/PH/1* queue. Given that the process is in state j_1 immediately before the observed transition which ends W_1, the probability that it is in state i_1 immediately after this transition is given by

$$\xi_{j_1,i_1} = \begin{cases} \dfrac{[Q_1]_{j_1,i_1}}{\sum\limits_{\ell}[Q_1]_{j_1,\ell}} & \text{if } \sum\limits_{\ell}[Q_1]_{j_1,\ell} > 0, \\ \\ 0 & \text{otherwise.} \end{cases}$$

Let Ξ be the matrix of terms $\xi_{j,i}$. This then allows us to use the probability transition matrix embedded immediately after departures to find $P(i_{r-1}|j_1)$. Hence if the process is in state j_1 immediately before the first observed transition, then the probability of being in state i_{r-1} immediately after the $r-1^{st}$ observed transition is given by

$$P(i_{r-1}|j_1) = \left[\Xi\Big(-Q_0^{-1}Q_1\Big)^{r-2}\right]_{(j_1,i_{r-1})}. \tag{62}$$

Given that the process is in state i_{r-1} immediately after the $r-1^{st}$ observed transition, the expected sojourn time before the r^{th} observed transition is given by

$$E[W_r|i_{r-1}] = \left[-Q_0^{-1}\boldsymbol{e}\right]_{i_{r-1}}. \tag{63}$$

In order to evaluate $P(j_1)$ and $E[W_1|j_1]$, we will need to use the time reverse process. If we let $\Phi = diag(\boldsymbol{\Psi})$, then the time reverse matrix corresponding to Q_0 is given by

$$Q_0^R \;=\; \Phi^{-1}Q_0^T\Phi, \tag{64}$$

where the superscript T indicates the transpose. The j_1^{th} term of the row vector

$$\left[-\left(Q_0^R\right)^{-1}\boldsymbol{e}\right]^T$$

contains the expected sojourn time before an observed transition in the reverse time process, given that it starts in state j_1. Thus

$$\begin{aligned}
E[W_1|j_1] &= \left[- \left(Q_0^R\right)^{-1} \boldsymbol{e} \right]_{j_1}^T \\
&= \left[\left(-\Phi^{-1} Q_0^T \Phi\right)^{-1} \boldsymbol{e} \right]_{j_1}^T \\
&= \left[- \Phi^{-1} \left(Q_0^{-1}\right)^T \Phi \boldsymbol{e} \right]_{j_1}^T \\
&= \left[- \Phi^{-1} \left(Q_0^{-1}\right)^T \boldsymbol{\Psi}^T \right]_{j_1}^T \\
&= \left[- \boldsymbol{\Psi} Q_0^{-1} \left(\Phi^{-1}\right) \right]_{j_1}.
\end{aligned}$$

This expected sojourn time corresponds to the first inter-departure time W_1, given that the process ends in j_1. Similarly $P(j_1)$ is given by the j_1^{th} term of the distribution embedded immediately after arrivals in the reverse time process, that is, $\left[\boldsymbol{\Psi}_{dep}^R\right]_{j_1}$. This distribution is given by

$$\boldsymbol{\Psi}_{dep}^R = \frac{\boldsymbol{\Psi} Q_1^R}{\boldsymbol{\Psi} Q_1^R \boldsymbol{e}} = \frac{\boldsymbol{\Psi}\left(\Phi^{-1} Q_1^T \Phi\right)}{\boldsymbol{\Psi}\left(\Phi^{-1} Q_1^T \Phi\right) \boldsymbol{e}} = \frac{\boldsymbol{e}^T Q_1^T \Phi}{\boldsymbol{e}^T Q_1^T \boldsymbol{\Psi}^T} = \frac{\left(\Phi Q_1 \boldsymbol{e}\right)^T}{\boldsymbol{\Psi} Q_1 \boldsymbol{e}},$$

hence

$$P(j_1) = \frac{\left[\Phi Q_1 \boldsymbol{e}\right]_{j_1}}{\boldsymbol{\Psi} Q_1 \boldsymbol{e}}. \tag{65}$$

Therefore, we have

$$P(j_1)P(i_{r-1}|j_1) = \frac{\left[\Phi Q_1 \boldsymbol{e}\right]_{j_1}}{\boldsymbol{\Psi} Q_1 \boldsymbol{e}} \left[\Xi \Bigg(- Q_0^{-1} Q_1 \Bigg)^{r-2} \right]_{(j_1, i_{r-1})},$$

which by the definition of Ξ gives us

$$P(j_1)P(i_{r-1}|j_1) = \frac{1}{\boldsymbol{\Psi} Q_1 \boldsymbol{e}} \left[\Phi Q_1 \Bigg(- Q_0^{-1} Q_1 \Bigg)^{r-2} \right]_{(j_1, i_{r-1})}.$$

Hence we may write the joint expectation $E[W_1 W_r]$ as

$$\sum_{j_1, i_{r-1}} E[W_1|j_1]P(j_1)P(i_{r-1}|j_1)E[W_{r-1}|i_{r-1}]$$

$$= \sum_{j_1, i_{r-1}} \left[-\boldsymbol{\Psi} Q_0^{-1} \left(\Phi^{-1}\right) \right]_{j_1} \frac{1}{\boldsymbol{\Psi} Q_1 \boldsymbol{e}} \left[\Phi Q_1 \Big(-Q_0^{-1} Q_1 \Big)^{r-2} \right]_{(j_1, i_{r-1})} \left[-Q_0^{-1} \boldsymbol{e} \right]_{i_{r-1}}$$

$$= \frac{\boldsymbol{\Psi} \Big(-Q_0^{-1} Q_1 \Big)^{r-1} (-Q_0^{-1} \boldsymbol{e})}{\boldsymbol{\Psi} Q_1 \boldsymbol{e}}. \tag{66}$$

These same probabilistic arguments may be used to form a similar expression for the joint expectation $E[W_1(k)W_r(k)]$ for the k^{th} *MAP* approximation. We note here that the departure rate from the *MAP/PH/1* queue is identical to the event rate of each of our approximations, so that

$$\boldsymbol{\Psi} Q_1 \boldsymbol{e} = \boldsymbol{\nu}(k) Q_1(k) \boldsymbol{e}.$$

Hence to show equivalence between the first $k-1$ lag-correlation coefficients for the *MAP/PH/1* queue and our k^{th} approximation, we need only prove for all $k \geq 2$ and $g \in \{1, \ldots, k-1\}$ that

$$\boldsymbol{\Psi} \Big(-Q_0^{-1} Q_1 \Big)^{g} (-Q_0^{-1} \boldsymbol{e}) = \boldsymbol{\nu}(k) \Big(-Q_0(k)^{-1} Q_1(k) \Big)^{g} \Big(-Q_0(k)^{-1} \boldsymbol{e} \Big). \tag{67}$$

8.2 The captured lag-correlation structure

Theorem 8.1 *The r^{th} lag-correlation coefficient $c_r(k)$ of the stationary inter-event times for the k^{th} MAP approximation to the departure process of the MAP/PH/1 queue is identical to the r^{th} lag-correlation coefficient c_r of the stationary inter-departure times for the MAP/PH/1 queue, for each $r \in \{2, 3, \ldots k\}$, for all $k \geq 2$.*

Proof:

The proof will be given using the similarities between the form of the matrices on each side of (67). It can be shown (see [23]) that

$$-Q_0^{-1} = \begin{pmatrix} X(0) & X_1 & X_2 & X_3 & \cdots \\ 0 & Y_0 & Y_1 & Y_2 & \cdots \\ & 0 & Y_0 & Y_1 & \\ \vdots & & 0 & Y_0 & \ddots \\ & & & 0 & \ddots \\ & & & & \ddots \end{pmatrix}, \tag{68}$$

and for all $k \geq 2$

$$-Q_0(k)^{-1} = \begin{pmatrix} X(0) & X_1 & X_2 & \cdots & X_{k-1} & Z(k) \\ 0 & Y_0 & Y_1 & \cdots & Y_{k-2} & Z_{k-1} \\ & 0 & Y_0 & & \vdots & \vdots \\ \vdots & & 0 & \ddots & Y_1 & Z_2 \\ & & & \ddots & Y_0 & Z_1 \\ 0 & & \cdots & & 0 & Z(0) \end{pmatrix}, \tag{69}$$

where

$$\begin{aligned} X(0) &= -D_0^{-1}, \\ X_i &= D_0^{-1}(D_1 \otimes \boldsymbol{\beta})\Big(-(D_0 \oplus S)^{-1}(D_1 \otimes I_n)\Big)^{i-1}(D_0 \oplus S)^{-1}, \text{ for } i \in \{1, 2, \ldots, k-1\}, \\ Y_i &= -\Big(-(D_0 \oplus S)^{-1}(D_1 \otimes I_n)\Big)^{i}(D_0 \oplus S)^{-1}, \text{ for } i \in \{0, 1, \ldots, k-2\}, \\ Z(0) &= -S^{-1}, \\ Z_i &= -\Big(-(D_0 \oplus S)^{-1}(D_1 \otimes I_n)\Big)^{i}(\boldsymbol{e}_m \otimes I_n)S^{-1} \text{ for } i \in \{1, 2, \ldots, k-1\} \text{ and} \\ Z(k) &= D_0^{-1}(D_1 \otimes \boldsymbol{\beta})\Big(-(D_0 \oplus S)^{-1}(D_1 \otimes I_n)\Big)^{k-1}(\boldsymbol{e}_m \otimes I_n)S^{-1}. \end{aligned} \tag{70}$$

Using the form of Q_1 and $Q_1(k)$ given in (53) and (56) respectively, it follows that we may write

$$-Q_1 Q_0^{-1}\boldsymbol{e} = \begin{pmatrix} 0 \\ \boldsymbol{\epsilon} \\ \boldsymbol{\theta} \\ \boldsymbol{\theta} \\ \vdots \end{pmatrix},$$

and

$$-Q_1(k)Q_0(k)^{-1}\boldsymbol{e} = \begin{pmatrix} \mathbf{0} & \\ \boldsymbol{\epsilon} & \\ \left.\begin{matrix}\boldsymbol{\theta} \\ \vdots \\ \boldsymbol{\theta}\end{matrix}\right\} & \begin{matrix} k-2 \\ \text{repeats}\end{matrix} \\ \boldsymbol{\sigma} & \end{pmatrix}, \text{ for all } k \geq 2,$$

where (see [23])

$$\begin{aligned} \boldsymbol{\sigma} &= (\boldsymbol{S}^0\boldsymbol{\beta})(-S)^{-1}\boldsymbol{e}_n, \\ \boldsymbol{\theta} &= (\boldsymbol{e}_m \otimes I_n)\boldsymbol{\sigma} \text{ and} \\ \boldsymbol{\epsilon} &= -(I_m \otimes (-\boldsymbol{S}^0))D_0^{-1}\boldsymbol{e}_m + \boldsymbol{\theta}. \end{aligned}$$

We may then construct

$$Q_0^{-1}Q_1Q_0^{-1}\boldsymbol{e} = \begin{pmatrix} X_1\boldsymbol{\epsilon} + \sum_{i=2}^{\infty} X_i\boldsymbol{\theta} \\ Y_0\boldsymbol{\epsilon} + \sum_{i=1}^{\infty} Y_i\boldsymbol{\theta} \\ \sum_{i=0}^{\infty} Y_i\theta \\ \sum_{i=0}^{\infty} Y_i\theta \\ \vdots \end{pmatrix},$$

$$Q_0(2)^{-1}Q_1(2)Q_0(2)^{-1}\boldsymbol{e} = \begin{pmatrix} X_1\boldsymbol{\epsilon} + Z(2)\boldsymbol{\sigma} \\ Y_0\boldsymbol{\epsilon} + Z_1\boldsymbol{\sigma} \\ Z(0)\boldsymbol{\sigma} \end{pmatrix}$$

and for all $k \geq 3$

$$Q_0(k)^{-1}Q_1(k)Q_0(k)^{-1}\boldsymbol{e} = \begin{pmatrix} X_1\boldsymbol{\epsilon} + \sum_{i=2}^{k-1} X_i\boldsymbol{\theta} + Z(k)\boldsymbol{\sigma} \\ Y_0\boldsymbol{\epsilon} + \sum_{i=1}^{k-2} Y_i\boldsymbol{\theta} + Z_{k-1}\boldsymbol{\sigma} \\ \sum_{i=0}^{k-3} Y_i\boldsymbol{\theta} + Z_{k-2}\boldsymbol{\sigma} \\ \vdots \\ \sum_{i=0}^{1} Y_i\boldsymbol{\theta} + Z_2\boldsymbol{\sigma} \\ Y_0\boldsymbol{\theta} + Z_1\boldsymbol{\sigma} \\ Z(0)\boldsymbol{\sigma} \end{pmatrix}.$$

Using the definitions given in (70), these columns may be reduced (see [23]) to the following

$$Q_0^{-1}Q_1Q_0^{-1}\boldsymbol{e} = \begin{pmatrix} X_1\boldsymbol{\epsilon} + Z(2)\boldsymbol{\sigma} \\ Y_0\boldsymbol{\epsilon} + Z_1\boldsymbol{\sigma} \\ (\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)\boldsymbol{e}_{mn} \\ (\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)\boldsymbol{e}_{mn} \\ \vdots \end{pmatrix}, \tag{71}$$

and for all $k \geq 2$

$$Q_0(k)^{-1}Q_1(k)Q_0(k)^{-1}\boldsymbol{e} = \begin{pmatrix} X_1\boldsymbol{\epsilon} + Z(2)\boldsymbol{\sigma} \\ Y_0\boldsymbol{\epsilon} + Z_1\boldsymbol{\sigma} \\ \left.\begin{matrix} \left(\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n\right)\boldsymbol{e}_{mn} \\ \vdots \\ \left(\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n\right)\boldsymbol{e}_{mn} \end{matrix}\right\} \quad k-2 \text{ repeats} \\ \left(\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n\right)\boldsymbol{e}_n \end{pmatrix}, \tag{72}$$

where we emphasise the fact that $\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n$ is a scalar.

From the definitions of $Q_1, Q_1(k), Q_0^{-1}$ and $Q_0(k)^{-1}$ given in (53), (56), (68) and (69), it may be seen that

$$-Q_0^{-1}Q_1 = \begin{pmatrix} X_1B_2 & X_2A_2 & X_3A_2 & X_4A_2 & \cdots \\ Y_0B_2 & Y_1A_2 & Y_2A_2 & Y_3A_2 & \cdots \\ 0 & Y_0A_2 & Y_1A_2 & Y_2A_2 & \\ 0 & 0 & Y_0A_2 & Y_1A_2 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{73}$$

and

$$-Q_0(k)^{-1}Q_1(k) = \begin{pmatrix} X_1B_2 & X_2A_2 & X_3A_2 & \cdots & X_{k-1}A_2 & Z(k)E_2 & Z(k)E_3 \\ Y_0B_2 & Y_1A_2 & Y_2A_2 & \cdots & Y_{k-2}A_2 & Z_{k-1}E_2 & Z_{k-1}E_3 \\ 0 & Y_0A_2 & Y_1A_2 & & Y_{k-3}A_2 & Z_{k-2}E_2 & Z_{k-2}E_3 \\ & 0 & Y_0A_2 & \ddots & & \vdots & \vdots \\ \vdots & & \ddots & \ddots & Y_1A_2 & Z_2E_2 & Z_2E_3 \\ & & & \ddots & Y_0A_2 & Z_1E_2 & Z_1E_3 \\ 0 & & \cdots & & 0 & Z(0)E_2 & Z(0)E_3 \end{pmatrix}. \tag{74}$$

We will use the fact that (73) and (74) are the probability transition matrices of the embedded Markov chains immediately after observed epochs of the *MAP/PH/1* queue departure process and of the k^{th} *MAP* approximation to the departure process respectively.

Using (71), (72), (73) and (74), we may write for all $g \geq 1$ that

$$\left(-Q_0^{-1}Q_1\right)^g \left(-Q_0^{-1}\boldsymbol{e}\right) \;=\; \begin{pmatrix} \mathbb{C}_{g,0} \\ \mathbb{C}_{g,1} \\ \vdots \\ \mathbb{C}_{g,g} \\ (\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)\boldsymbol{e}_{mn} \\ \vdots \end{pmatrix}, \qquad (75)$$

and for $k \geq 2$ with $1 \leq g < k$, that

$$\Big(-Q_0(k)^{-1}Q_1(k)\Big)^g \left(-Q_0(k)^{-1}\boldsymbol{e}\right) = \begin{pmatrix} \mathbb{C}_{g,0} \\ \mathbb{C}_{g,1} \\ \vdots \\ \mathbb{C}_{g,g} \\ \left.\begin{matrix} (\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)\boldsymbol{e}_{mn} \\ \vdots \\ (\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)\boldsymbol{e}_{mn} \end{matrix}\right\} \; k-g-1 \text{ repeats} \\ (\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)\boldsymbol{e}_{n} \end{pmatrix}, (76)$$

where for a given g, the terms $\mathbb{C}_{g,i}$, for each $i \in \{0, \ldots, g\}$, are common to all *MAP* approximations with $g < k$ and the *MAP/PH/1* queue. This can be readily seen from the fact that for all $k \geq 2$, the first $k-1$ entries of the first k rows of the matrix $-Q_0(k)^{-1}Q_1(k)$ are identical to those of the matrix $-Q_0^{-1}Q_1$, and both are stochastic.

Using (49), (57), (75) and (76), all that remains to be proved for

$$\boldsymbol{\Psi}\Big(-Q_0^{-1}Q_1\Big)^g(-Q_0^{-1}\boldsymbol{e}) \;=\; \boldsymbol{\nu}(k)\Big(-Q_0(k)^{-1}Q_1(k)\Big)^g(-Q_0(k)^{-1}\boldsymbol{e}), \text{ for all } k > g \geq 1,$$

to hold is that for all $k \geq 2$

$$\boldsymbol{\pi}_k(I_{mn}-R)^{-1}(\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)\boldsymbol{e}_{mn} \;=\; \boldsymbol{\pi}_k(I_{mn}-R)^{-1}(\boldsymbol{e}_m \otimes I_n)(\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)\boldsymbol{e}_n.$$

This is trivial since $(\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e}_n)$ is just a scalar. ■

8.3 Some Numerical examples

A tandem queueing model (see Figure 16) is used to demonstrate the approximations and to observe some interesting facets of the importance of capturing the lag-correlations. The departure process from the first queue is approximated by a *MAP* which is used to feed the second queue. The average second queue length is used to compare the approximations to the "real solution" calculated using large k. Two arrival processes are used for comparison, namely a Poisson process and a positively correlated Markov modulated Poisson process (MMPP). Note that the service distributions are scaled appropriately to achieve the required traffic intensities.

8.3.1 The arrival processes and service distributions.

1.
   - Negative-exponential arrivals of rate 1 or
   - Markov Modulated Poisson process (*MMPP*) arrivals

$$D_0 = \begin{pmatrix} -10.0 & 1.0 \\ 0.4 & -0.8 \end{pmatrix}, D_1 = \begin{pmatrix} 9.0 & 0.0 \\ 0.0 & 0.4 \end{pmatrix}.$$

2. Hyper-exponential first server

$$S = \begin{pmatrix} -6 & 0 \\ 0 & -12 \end{pmatrix}, \boldsymbol{\beta} = (0.2, 0.8).$$

3. Erlang second server E_2.

The plots in figures 19, 20, 21 and 22 of Appendix D give the calculated mean second queue length against the number of captured lag-correlations for Poisson arrivals for given ρ_1 and a value of $\rho_2 = 0.75$. The plots in figures 23, 24, 25 and 26 of Appendix D give the calculated mean second queue length against the number of captured lag-correlations for MMPP arrivals for given ρ_1 and a value of $\rho_2 = 0.5$. The maximum number j of captured lag-correlations used to calculate the "true" mean second queue length $l(j)$ is such that

$$\left| \frac{l(j) - l(j-1)}{l(j-1)} \right| < 10^{-6}. \tag{77}$$

It is interesting to note that the number of lag-correlations for given service distributions needed to satisfy (77) appears to depend on the arrival process and the value of ρ_1 and ρ_2 in a non trivial manner.

All of the approximations used for the tandem with Poisson arrivals are less than 1% in error in the calculated value of the mean to the "true" mean second queue length.

In the case of MMPP arrivals, when $\rho_1 = 0.75$ or 0.9, all of the approximations used are less than 1% in error in the calculated value of the mean to the "true" mean second queue length. If $\rho_1 = 0.25$, it is necessary to capture three lag-correlation coefficients and for $\rho_1 = 0.5$ we need seven lag-correlation coefficients to achieve this same degree of accuracy. For $\rho_1 = 0.5$, our $k = 2$ approximation (1 captured lag-correlation coefficient) yields the worst result, being approximately 6.4% in error to the "true" mean second queue length.

9 Approximating the departure process from a *MAP/PH/1* feedback queue

Let D_0 and $D_1 \geq 0$ be the $m \times m$ matrix descriptors of the *MAP* and $(\boldsymbol{\beta}, S)$ be the *PH*-type service distribution, where the matrix S is $n \times n$. The probability of feedback may be level dependent and has a description given by some probability $f(i)$, where i is the number in the system. For $i > 0$, a customer who completes service leaving $i-1$ in the queue will immediately rejoin the queue with probability $f(i)$ or with probability $(1 - f(i))$ leaves the system. Define $\boldsymbol{S}^0 = -S\boldsymbol{e}$. The *MAP/PH/1* level dependent feedback queue then has the following quasi birth and death process description given by

$$Q = \begin{bmatrix} B_1 & B_0 & & & & \\ (1-f(1))B_2 & A_1 + f(1)A_2 & A_0 & & & \\ & (1-f(2))A_2 & A_1 + f(2)A_2 & A_0 & & \\ & & (1-f(3))A_2 & A_1 + f(3)A_2 & A_0 & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}, \quad (78)$$

where

$$\begin{aligned} B_0 &= D_1 \otimes \boldsymbol{\beta}, && (79)\\ B_1 &= D_0, && (80)\\ B_2 &= I_m \otimes \boldsymbol{S}^0, && (81)\\ A_0 &= D_1 \otimes I_n, && (82)\\ A_1 &= D_0 \oplus S, && (83)\\ A_2 &= I_m \otimes \boldsymbol{S}^0\boldsymbol{\beta}. && (84) \end{aligned}$$

Here, $\otimes$ and $\oplus$ are the Kronecker product and sum respectively. Assume Q defines an irreducible, regular Markov chain, so that it has at most one stationary distribution $\mathbf{\Psi}$ such that $\mathbf{\Psi}Q = 0$. If it exists, this stationary distribution is given by the following matrix form

$$\mathbf{\Psi} = \boldsymbol{\pi}_0[I, R_0, R_0R_1, R_0R_1R_2, \ldots], \tag{85}$$

where for $i \geq 1$, the $mn \times mn$ matrices R_i and the $m \times mn$ matrix R_0 are the minimal non-negative solutions to the system of equations

$$\begin{aligned} R_iR_{i+1}\Big(1 - f(i+2)\Big)A_2 + R_i\Big(A_1 + (1 - f(i+1))A_2\Big) + A_0 = 0 \\ B_0 + R_0\Big(A_1 + f(1)A_2 + R_1A_2(1 - f(2))\Big) = 0 \end{aligned} \tag{86}$$

and the vector $\boldsymbol{\pi}_0$ is the unique positive solution to the system of equations

$$\boldsymbol{\pi}_0(B_1 + R_0B_2) = \mathbf{0} \quad \text{and} \quad \boldsymbol{\pi}_0 \sum_{j=0}^{\infty} \prod_{i=0}^{j-1} R_i\boldsymbol{e} = 1. \tag{87}$$

Let's now consider the departure process from the *MAP/PH/1* feedback queue, which can be observed using the following filtration matrices Q_0^* and Q_1^*, where $Q = Q_0^* + Q_1^*$.

$$Q_0^* = \begin{bmatrix} B_1 & B_0 & & & & \\ & A_1 & A_0 & & & \\ & & A_1 & A_0 & & \\ & & & A_1 & A_0 & \\ & & & & \ddots & \ddots \end{bmatrix} \quad \text{and}$$

$$Q_1^* = \begin{bmatrix} 0 & & & & \\ B_2(1 - f(1)) & f(1)A_2 & & & \\ & A_2(1 - f(2)) & f(2)A_2 & & \\ & & A_2(1 - f(3)) & f(3)A_2 & \\ & & & \ddots & \ddots \end{bmatrix}, \tag{88}$$

where the A_i and B_i have the same interpretation as those given in equation (84). Note that a measure $f(i)$ of the "departures" at level i is fed back to level i. These "departures" do not leave the system, but immediately return to service. This effectively maintains the number i of customers present at the server and associated queue. As we are concerned with those departures which actually leave the system, we partition this particular Q matrix into the following filtration matrices Q_0 and Q_1, so as to capture the actual process of departures leaving the system (hereafter referred to as the departure process).

$$Q_0 = \begin{bmatrix} B_1 & B_0 & & & & \\ & A_1 + f(1)A_2 & A_0 & & & \\ & & A_1 + f(2)A_2 & A_0 & & \\ & & & A_1 + f(3)A_2 & A_0 & \\ & & & & \ddots & \ddots \end{bmatrix} \text{ and}$$

$$Q_1 = \begin{bmatrix} 0 & & & & & \\ B_2(1 - f(1)) & 0 & & & & \\ & A_2(1 - f(2)) & 0 & & & \\ & & A_2(1 - f(3)) & 0 & & \\ & & & \ddots & \ddots & \end{bmatrix}, \quad (89)$$

The observed transitions recorded by Q_1 are departure transitions and the observed process is the departure process.

The resultant filtration matrices in (89) when $f(i) = \rho$ for all $i \geq 1$ are in fact representative of another *MAP/PH/1* queue without feedback, where the arrival process is the same but with a modified service time distribution.

Similarly to before, we construct a family of approximations indexed by a parameter k, where the accuracy of the approximation increases with the value of the parameter k. Intuitively the stationary rate of departures from a *MAP/PH/1* feedback queue must be equivalent to the stationary rate of arrivals given by $\boldsymbol{\nu} D_1 \boldsymbol{e}$, where $\boldsymbol{\nu}$ is the stationary probability vector of the *MAP* satisfying $\boldsymbol{\nu}(D_0 + D_1) = 0$. We will use this result in the construction of the distribution of the *QBD* at level $k-1$, conditional on a departure having just occurred.

The distribution of the *QBD* given in (89) at level $k-1$, conditional on a departure having just occurred, can be calculated from its stationary distribution by

$$\boldsymbol{x}_{k-1} = \begin{cases} \boldsymbol{\pi}_0 R_0 (1 - f(1)) B_2 (\boldsymbol{\nu} D_1 \boldsymbol{e})^{-1} & \text{for } k = 1 \\ \\ \boldsymbol{\pi}_0 \prod_{i=0}^{k-1} R_i (1 - f(k-1)) A_2 (\boldsymbol{\nu} D_1 \boldsymbol{e})^{-1} & \text{for } k > 1. \end{cases} \quad (90)$$

The k^{th} *MAP* approximation to the departure process of the *MAP/PH/1* feedback queue is given by the $\Big(m+n+(k-1)mn\Big) \times \Big(m+n+(k-1)mn\Big)$

matrices

$$Q_0 = \begin{bmatrix} B_1 & B_0 & & & & & \\ & A_1 + f(1)A_2 & A_0 & & & & \\ & & A_1 + f(2)A_2 & \ddots & & & \\ & & & \ddots & A_0 & & \\ & & & & A_1 + f(k-1)A_2 & E_0 & \\ & & & & & E_1 & \end{bmatrix} \text{ and}$$

$$Q_1 = \begin{bmatrix} 0 & & & & & & \\ B_2(1-f(1)) & 0 & & & & & \\ & A_2(1-f(2)) & \ddots & & & & \\ & & \ddots & 0 & & & \\ & & & A_2(1-f(k-1)) & 0 & & \\ & & & & E_2 & E_3 & \end{bmatrix}. \quad (91)$$

Here the sub-matrices are as defined previously with

$$\begin{aligned} E_0 &= D_1 \boldsymbol{e} \otimes I_n, \\ E_1 &= S + \gamma_k \boldsymbol{S}^0 \boldsymbol{\beta}, \\ E_2 &= (1-\gamma_k) \boldsymbol{S}^0 \boldsymbol{y}_{k-1}, \\ E_3 &= (1 - \boldsymbol{y}_{k-1}\boldsymbol{e})(1-\gamma_k)\boldsymbol{S}^0 \boldsymbol{\beta}, \end{aligned}$$

where γ_k is a scaling factor for the service time distribution which reflects the level dependent feedback mechanism in operation above the super level k. When $f(i) = \rho \in [0, 1)$ for all $i \geq 1$, we have a level-independent *QBD* with $\gamma_k = \rho$ for all $k \geq 1$. When $f(i)$ is not constant, we have a level-dependent *QBD* and the scaling factor γ_k may be found by an iteration method which establishes the stationary distribution of this *MAP* approximation to be

$$\boldsymbol{\nu}(k) = \boldsymbol{\pi}_0 \left[I, R_0, R_0 R_1, \ldots, \prod_{i=0}^{k-2} R_i, \left(\sum_{j=k-1}^{\infty} \prod_{i=0}^{j} R_i \right) (\boldsymbol{e}_m \otimes I_n) \right]. \quad (92)$$

Special case:

For the special case of $k = 1$, the *MAP* approximation to the departure process of the *MAP/PH/1* feedback queue reduces to a *PH*-renewal process. This approximation closely captures the distribution of the inter-departure times but ignores any correlation structures between these times. When $f(i) = \rho \in [0, 1)$ for all $i \geq 1$, the approximation exactly captures the distribution of the inter-departure times. For $k=1$, the MAP approximation has a PH-distribution $(\boldsymbol{\alpha}, Q_0(1))$, where $Q_0(1)$ is given by

$$Q_0(1) = \begin{pmatrix} D_0 & D_1 \boldsymbol{e}\boldsymbol{\beta} \\ 0 & \gamma_1 \boldsymbol{S}^0 \boldsymbol{\beta} + S \end{pmatrix}.$$

and

$$\boldsymbol{\alpha} = (\boldsymbol{x}_0, (1 - \boldsymbol{x}_0 \boldsymbol{e})\boldsymbol{\beta})\,. \tag{93}$$

Here, $\boldsymbol{x}_0$ is the distribution of phases of the arrival process immediately after a departure that leaves the queue empty, calculated using (90). The stationary distribution of this MAP approximation is given by

$$\boldsymbol{\nu}(1) = \left[\boldsymbol{\pi}_0, \boldsymbol{\pi}_0 \left(\sum_{j=0}^{\infty} \prod_{i=0}^{j} R_i\right)(\boldsymbol{e}_m \otimes I_n)\right].$$

This approximation closely captures the distribution of the inter-departure times but ignores any correlation structures between these times.

9.1 Numerical results

In keeping with the previous numerical results for the departure approximations, we will use a tandem queueing model as in Figure 18.

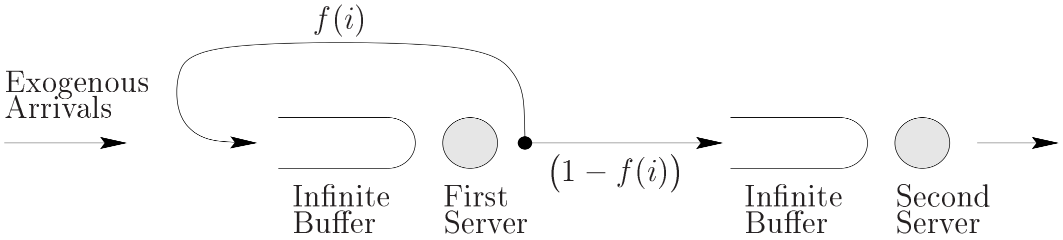


Figure 18: Tandem queueing model with feedback.

9.1.1 A constant feedback probability.

In the case where $f(i) = \rho \in [0,1)$ for all $i \geq 1$, $\gamma_k = \rho$ for all $k \geq 1$ and we have effectively already demonstrated the accuracy of the approximations. This is because the resultant filtration matrices given in (89) form another level-independent QBD with the same arrival process, but with a modified PH-type service distribution.

In demonstrating the accuracy of the approximations however, it became clear that Poisson and or negative-exponential assumptions are inappropriate in many instances. Therefore, we present our approximations for the case where $f(i) = \rho \in [0, 1)$ with a the two fold purpose. Let's consider a tandem network of queues, where we fix the feedback rate and traffic intensities at each server. In Table 4, the "exact" first two central moments for the second queue length are presented for a Poisson arrival stream against an Erlang renewal arrival process and a bursty *MMPP*. We also compare the actual results against our $k = 10$ approximation for these tandems in Table 5.

The moments for the Poisson arrivals with the H_2 server can be considered as the results for a Poisson approximation to both the Erlang renewal and *MMPP* arrival processes. This Poisson approximation ($\frac{\sigma^2}{\mu^2} = 1$) over-estimates the stationary second queue length for the Erlang renewal arrival process ($\frac{\sigma^2}{\mu^2} = 0.5$) and greatly under-estimates for the *MMPP* ($\frac{\sigma^2}{\mu^2} = 4.9721$). Hence the case for using processes other than the Poisson process is obviated.

Note that in the above network, a Poisson approximation to an Erlang renewal arrival process, progressively gets worse as the order k of the Erlang renewal process gets larger. In fact it is relatively easy to prove that as k increases, E_k becomes more regular to the point where the inter-arrival times become deterministic as $k \to \infty$.

Feedback Probability	*Traffic intensities*		*Two moments of the stationary second queue length for different arrival processes to the tandem (first server also given)*							
			Erlang H_2 server		Poisson H_2 server		MMPP H_2 server		Poisson neg-exp server	
ρ	η_1	η_2	E[μ]	E[σ^2]	E[μ]	E[σ^2]	E[μ]	E[σ^2]	E[μ]	E[σ^2]
0.25	0.25	0.25	0.302	0.266	0.336	0.369	0.593	1.719	0.333	0.361
0.25	0.5	0.25	0.317	0.312	0.338	0.378	0.434	0.763	0.333	0.361
0.25	0.25	0.5	0.839	0.897	1.006	1.52	3.244	26.326	1.000	1.500
0.25	0.5	0.5	0.898	1.086	1.02	1.580	2.122	10.446	1.000	1.500
0.25	0.25	0.75	2.368	5.415	3.008	9.777	12.631	232.060	3.000	9.750
0.25	0.5	0.75	2.494	5.882	3.041	9.959	10.764	189.796	3.000	9.750
0.5	0.25	0.25	0.301	0.265	0.335	0.367	0.590	1.702	0.333	0.361
0.5	0.5	0.25	0.317	0.31	0.337	0.374	0.432	0.753	0.333	0.361
0.5	0.25	0.5	0.837	0.892	1.004	1.512	3.243	26.314	1.000	1.500
0.5	0.5	0.5	0.893	1.064	1.014	1.554	2.107	10.291	1.000	1.500
0.5	0.25	0.75	2.364	5.404	3.005	9.766	12.632	232.081	3.000	9.750
0.5	0.5	0.75	2.479	5.808	3.026	9.873	10.751	189.614	3.000	9.750
0.75	0.25	0.25	0.301	0.263	0.334	0.364	0.587	1.682	0.333	0.361
0.75	0.5	0.25	0.315	0.305	0.336	0.369	0.429	0.738	0.333	0.361
0.75	0.25	0.5	0.835	0.886	1.002	1.506	3.242	26.308	1.000	1.500
0.75	0.5	0.5	0.887	1.043	1.007	1.527	2.091	10.125	1.000	1.500
0.75	0.25	0.75	2.361	5.396	3.002	9.757	12.633	232.109	3.000	9.750
0.75	0.5	0.75	2.464	5.747	3.012	9.803	10.74	189.466	3.000	9.750

Table 4: Poisson and negative-exponential assumptions comparison.

Feedback Probability	Traffic intensities		*Two moments of the stationary second queue length for different arrival processes to the tandem*							
			Erlang				*MMPP*			
			exact		$k=10$		*exact*		$k=10$	
ρ	η_1	η_2	$\mathrm{E}[\mu]$	$\mathrm{E}[\sigma^2]$	$\mathrm{E}[\mu]$	$\mathrm{E}[\sigma^2]$	$\mathrm{E}[\mu]$	$\mathrm{E}[\sigma^2]$	$\mathrm{E}[\mu]$	$\mathrm{E}[\sigma^2]$
0.25	0.25	0.25	0.302	0.266	0.302	0.266	0.593	1.719	0.593	1.719
0.25	0.5	0.25	0.317	0.312	0.317	0.312	0.434	0.763	0.434	0.763
0.25	0.25	0.5	0.839	0.897	0.839	0.897	3.244	26.326	3.242	26.236
0.25	0.5	0.5	0.898	1.086	0.898	1.086	2.122	10.446	2.104	10.035
0.25	0.25	0.75	2.368	5.415	2.368	5.415	12.631	232.060	12.620	231.179
0.25	0.5	0.75	2.494	5.882	2.495	5.883	10.764	189.796	10.238	156.993
0.5	0.25	0.25	0.301	0.265	0.301	0.265	0.590	1.702	0.590	1.701
0.5	0.5	0.25	0.317	0.31	0.317	0.309	0.432	0.753	0.432	0.753
0.5	0.25	0.5	0.837	0.892	0.837	0.892	3.243	26.314	3.241	26.222
0.5	0.5	0.5	0.893	1.064	0.893	1.064	2.107	10.291	2.089	9.875
0.5	0.25	0.75	2.364	5.404	2.364	5.404	12.632	232.081	12.621	231.212
0.5	0.5	0.75	2.479	5.808	2.479	5.808	10.751	189.614	10.218	156.403
0.75	0.25	0.25	0.301	0.263	0.301	0.263	0.587	1.682	0.587	1.682
0.75	0.5	0.25	0.315	0.305	0.315	0.305	0.429	0.738	0.429	0.738
0.75	0.25	0.5	0.835	0.886	0.835	0.887	3.242	26.308	3.240	26.217
0.75	0.5	0.5	0.887	1.043	0.887	1.043	2.091	10.125	2.073	9.704
0.75	0.25	0.75	2.361	5.396	2.361	5.396	12.633	232.109	12.622	231.269
0.75	0.5	0.75	2.464	5.747	2.464	5.747	10.740	189.466	10.200	155.930

Table 5: The $k=10$ approximation comparison.

9.1.2 A geometrically decaying feedback probability.

Recall that when $f(i)$ is not constant, we have a level-dependent *QBD* and the scaling factor γ_k must be found by some method. The proposed method is a simple iteration, which establishes the stationary distribution of this *MAP* approximation to be identical to the level-dependent *QBD* in all levels up to level k, with the remaining mass being concentrated at level k. That is we look for

$$\boldsymbol{\nu}(k) = \boldsymbol{\pi}_0 \left[I, R_0, R_0R_1, \ldots, \prod_{i=0}^{k-2} R_i, \left(\sum_{j=k-1}^{\infty} \prod_{i=0}^{j} R_i \right) (\boldsymbol{e}_m \otimes I_n) \right].$$

In the scenario where we have a geometrically decaying feedback parameter, we have that $f(i) = \rho^i$ for all $i \geq 1$. This regime seems to be sensible from the perspective that a customer wanting to rejoin a queue and finding a large number already present, is less likely to rejoin it.

No results yet, work in progress

References

[1] S. Albin. On Poisson Approximations for Superposition Arrival Processes in Queues. *Management Science*, 28(2):126–137, 1982.

[2] S. Albin. Approximating a Point Process by a Renewal Process,II : Superposition Arrival Processes to Queues. *Operations Research*, 32(5):1133–1162, 1984.

[3] S. Albin and Sheng-Roan Kai. Approximation for the Departure Process of a Queue in a Network. *Naval Research Logistics Quarterly*, 33(1):129–143, 1986.

[4] W.J. Anderson. *Continuous-time Markov Chains.* Springer Series in Statistics : Probability and its Applications. Springer-Verlag, New York, Berlin, Heidelberg, 1991.

[5] N.G. Bean and D.A. Green. When is a *MAP* Poisson? *Mathematical and Computer Modelling (Special Issue)*, 31:10–12:31–46, 2000.

[6] N.G. Bean, D.A. Green, and P.G. Taylor. The Output Process of an *MMPP/M/1* queue. *Journal of Applied Probability*, 35(4):998–1002, 1998.

[7] N.G. Bean, J. Li, and P.G. Taylor. Some Asymptotic Properties of Two-stage Tandem Networks of *PH/PH/1* Queues. In *Advances in Matrix Analytic Methods for Stochastic Models - Proceedings of the 2nd International Conference on Matrix Analytic Methods*, pages 171–193. Notable Publications Inc., NJ, 1998.

[8] G. Bitran and S. Dasu. Analysis of the $\sum Ph_i/Ph/1$ Queue. *Operations Research*, 42(1):158–174, 1994.

[9] J.Y. Le Boudec. An Efficient Solution Method for Markov Models of ATM Links with Loss Priorities. *IEEE Journal of Selected Areas in Communications*, 9:408–417, 1991.

[10] L.W. Bright and P.G. Taylor. Calculating the equilibrium distribution in level-dependent quasi-birth-and-death processes. *Communications in Statistics*, 11(2):497–526, 1995.

[11] L.W. Bright and P.G. Taylor. *Lecture notes in Pure and Applied Maths*, volume 183, chapter Equilibrium distributions for level-dependent quasi-birth-and-death processes, pages 359–375. Dekker, New York, 1997.

[12] P.J. Burke. The Output of a Queueing System. *Operations Research*, 4:699–704, 1956.

[13] P.J. Burke. The Output Process of a Stationary *M/M/s* Queueing System. *Annals of Mathematical Statistics*, 39(4):1144–1152, 1968.

[14] Sung-Ho Choi and K. Sohraby. Analysis of a Mobile Cellular Systems with Hand-off Priority and Hysteresis Control. In *IEEE INFOCOM'2000*, pages 217–224, 2000.

[15] Sung-Ho Choi, K. Sohraby, and B. Kim. General *QBD* Processes with Applications to Overload Control. In *IEEE INFOCOM'2000*, pages 165–172, 2000.

[16] J.W. Cohen. *The Single Server Queue*. Applied Mathematics and Mechanics. North-Holland, Amsterdam, London, 1969.

[17] D.J. Daley. The Correlation Structure of the Output Process of Some Single Server Queueing Systems. *The Annals of Mathematical Statistics*, 39:1007–1019, 1968.

[18] R.L. Disney and T.L. Vlacht. The Departure Process from the *GI/G/1* Queue. *Journal of Applied Probability*, 6:704–707, 1969.

[19] J.R. Fraker. *Approximate Techniques for the Analysis of Tandem Queueing Systems*. PhD thesis, Department of Industrial Engineering, Clemson University, 1971.

[20] H.R. Gail, S.L. Hantler, and B.A. Taylor. Solutions of the basic matrix equation for *M/G/1* and *G/M/1* type Markov chains. *Communications in Statistics : Stochastic Models*, 10(1):1–43, 1994.

[21] D.P. Gaver, P.A. Jacobs, and G. Latouche. Finite Birth-and-Death Models in a Randomly Changing Environments. *Advances in Applied Probability*, 16:715–731, 1984.

[22] S. Gibson and E. Seneta. Augmented Truncation of Infinite Stochastic Matrices. *Journal of Applied Probability*, 24:600–608, 1987.

[23] D.A. Green. *Departure Processes from MAP/PH/1 Queues*. PhD thesis, Department of Applied Mathematics, The University of Adelaide, 1998.

[24] B. Hajek. Birth-and-Death Processes on the Integers with Phases and General Boundaries. *Journal of Applied Probability*, 19:488–499, 1982.

[25] R.R.P. Jackson. Queueing Systems with Phase Type Service. *Operational Research Quarterly*, 5(1):109–120, 1954.

[26] G. Latouche. Algorithmic Analysis of a multiprogramming-multiprocessor computer system. *Journal of the Association for Computing Machinery*, 28:662–679, 1981.

[27] G. Latouche. A Note on Two Matrices Occurring in the Solution of Quasi-birth-and-death Processes. *Communications in Statistics : Stochastic Models*, 3(2):251–257, 1987.

[28] G. Latouche, C.E.M. Pearce, and P.G. Taylor. Invariant measures for quasi-birth-and-death processes. *Communications in Statistics : Stochastic Models*, 14(1-2):443–460, 1998.

[29] G. Latouche and V. Ramaswami. An Experimental Evaluation of the Matrix-Geometric Method for the *GI/PH/1* Queue. *Communications in Statistics : Stochastic Models*, 5:629–667, 1989.

[30] G. Latouche and V. Ramaswami. A Logarithmic Reduction Algorithm for Quasi-Birth-Death-Processes. *Journal of Applied Probability*, 30:650–674, 1993.

[31] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modelling.* ASA-SIAM Series on Statistics and Applied Probability, 1999.

[32] D.M. Lucantoni. New Results on the Single Server Queue with a Batch Markovian Arrival Process. *Communications in Statistics : Stochastic Models*, 7:1–46, 1991.

[33] B. Meini. New convergence results on functional iteration techniques for the numerical solution of *M/G/1* type Markov chains. *Numerische Mathematik*, 78(1):39–58, 1997.

[34] P.M. Morse. Stochastic Properties of Waiting Lines. *Journal of the Operations Research Society of America*, 3:255–261, 1955.

[35] R.D. Nelson. A performance evaluation of a general parallel processing model. *Performance Evaluation Review*, 18:13–26, 1990.

[36] M.F. Neuts. A versatile Markovian point process. *Journal of Applied Probability*, 16:764–769, 1979.

[37] M.F. Neuts. *Matrix-geometric Solutions in Stochastic Models : An Algorithmic Approach*. The John Hopkins University Press, Baltimore, 1981.

[38] M.F. Neuts. Further results on the *M/M/1* queue with randomly varying rates. *Opsearch*, 15:158–168, 1987.

[39] M.F. Neuts. The *M/M/1* queue with randomly varying arrival and service rates. *Opsearch*, 15:139–157, 1987.

[40] M.F. Neuts. *Structured stochastic matrices of the M/G/1 type and their applications*. Dekker, New York, 1989.

[41] M.F. Neuts. *Algorithmic Probability: A collection of problems*. Chapman and Hall, London, 1995.

[42] M.F. Neuts and B.M. Rao. Numerical Investigation of a Multi-server Retrial Model. *Queueing Systems*, 7:169–190, 1990.

[43] C. Olivier and J. Walrand. On the existence of finite-dimensional filters for Markov-modulated traffic. *Journal of Applied Probability*, 31:515–525, 1994.

[44] V. Ramaswami. The busy period of queues which have a matrix-geometric steady state probability vector. *Opsearch*, 19:238–261, 1982.

[45] V. Ramaswami. Matrix-Analytic Methods: A Tutorial Overview with some Extensions and New Results. In *Matrix Analytic Methods in Stochastic Models (Flint, MI)*, pages 261–296. Dekker, New York, 1997.

[46] V. Ramaswami. The Generality of Quasi Birth-and-Death Processes. In *Advances in Matrix Analytic Methods for Stochastic Models - Proceedings of the 2nd International Conference on Matrix Analytic Methods*, pages 93–113. Notable Publications Inc., NJ, 1998.

[47] V. Ramaswami and P.G. Taylor. Some Properties of the Rate Operators in Level Dependent Quasi-Birth and Death Processes with a Countable Number of Phases . *Communications in Statistics: Stochastic Models*, 12(1):143–164, 1996.

[48] E. Reich. Waiting times when queues are in tandem. *Annals of Mathematical Statistics*, 28:768–773, 1957.

[49] E. Reich. *Departure processes (with discussion) in Congestion Theory*. University of North Carolina Press, Chapel Hill, 1965. editors W.L.Smith and W.E.Wilkinson.

[50] D.G. Shimshak. A comparison of waiting time approximations in series queueing systems. *Naval Research Logistics Quarterly*, 26:499–509, 1979.

[51] W. Whitt. Approximating a Point Process by a Renewal Process, I : Two Basic Methods. *Operations Research*, 30(1):125–147, 1982.

[52] W. Whitt. Approximations for Departure processes and queues in series. *Naval Research Logistics Quarterly*, 31:499–521, 1984.

[53] J. Ye and S.Q. Li. Folding Algorithm: A computational method for finite QBD processes with level-dependent transitions. *IEEE Transactions on Communications*, 42:625–639, 1994.

[54] M. Zuckerman. Applications of Matrix-Geometric solutions for queueing performance evaluation of a hybrid switching system. *Journal of the Australian Mathematical Society Series B, Applied mathematics*, 31:219–239, 1989.

[55] M. Zuckerman and P. Kirton. Applications of Matrix-Geometric solutions to the analysis of the bursty data queue in a B-ISDN switching system. In *Proceedings GLOBECOM'88*, pages 1635–1639, 1988.

A The arrival processes

1. Poisson of rate 1, $\dfrac{\sigma^2}{\mu^2} = 1.0000$

2.
$$D_0 = \begin{pmatrix} -3 & 2 \\ 4 & -6 \end{pmatrix}, D_1 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix},$$
$$\frac{\sigma^2}{\mu^2} = 1.0969, c_1 = -0.0044, c_2 = 0.0004.$$

3. Erlang (E_2) of rate 1, $\dfrac{\sigma^2}{\mu^2} = 0.5000.$

4. Hyper-exponential
$$D_0 = \begin{pmatrix} -2 & 0 \\ 0 & -1 \end{pmatrix}, \boldsymbol{\alpha} = (0.75, 0.25), \frac{\sigma^2}{\mu^2} = 1.2400.$$

5. Markov Modulated Poisson process *MMPP*
$$D_0 = \begin{pmatrix} -10.0 & 1.0 \\ 0.4 & -0.8 \end{pmatrix}, D_1 = \begin{pmatrix} 9.0 & 0.0 \\ 0.0 & 0.4 \end{pmatrix},$$
$$\frac{\sigma^2}{\mu^2} = 4.9721, c_1 = 0.1892, c_2 = 0.0896.$$

6.
$$D_0 = \begin{pmatrix} -4 & 1 \\ 1 & -6 \end{pmatrix}, D_1 = \begin{pmatrix} 1 & 2 \\ 1 & 4 \end{pmatrix},$$
$$\frac{\sigma^2}{\mu^2} = 1.1183, c_1 = 0.0046, c_2 = 0.0004.$$

B The first server

a Negative-exponential, $\dfrac{\sigma^2}{\mu^2} = 1.0000$.

b Erlang E_2, $\dfrac{\sigma^2}{\mu^2} = 0.5000$.

c Hyper-exponential

$$S = \begin{pmatrix} -6 & 0 \\ 0 & -12 \end{pmatrix}, \boldsymbol{\beta} = (0.2, 0.8), \frac{\sigma^2}{\mu^2} = 1.2222.$$

d Mixed-Erlang

$$\boldsymbol{\beta} = (0.6, 0.4), \frac{\sigma^2}{\mu^2} = 0.7188.$$

C The second server

i Negative-exponential, $\dfrac{\sigma^2}{\mu^2} = 1.0000$.

ii Hyper-exponential

$$S = \begin{pmatrix} -15 & 0 \\ 0 & -8.75 \end{pmatrix}, \boldsymbol{\beta} = (0.3, 0.7), \frac{\sigma^2}{\mu^2} = 1.0952.$$

iii Erlang E_2, $\dfrac{\sigma^2}{\mu^2} = 0.5000$.

iv Mixed-Erlang

$$\boldsymbol{\beta} = (0.25, 0.75), \frac{\sigma^2}{\mu^2} = 0.9200.$$

D Correlation plots

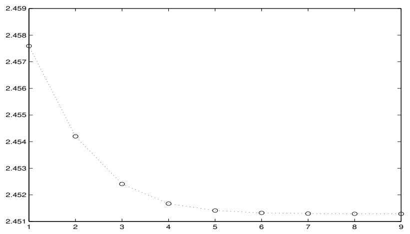


Figure 19: Poisson $\rho_1 = 0.25$.

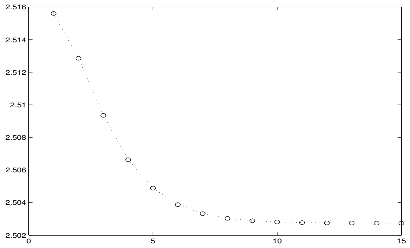


Figure 20: Poisson $\rho_1 = 0.5$.

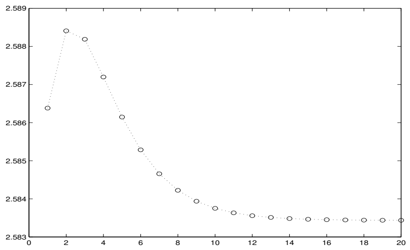


Figure 21: Poisson $\rho_1 = 0.75$.

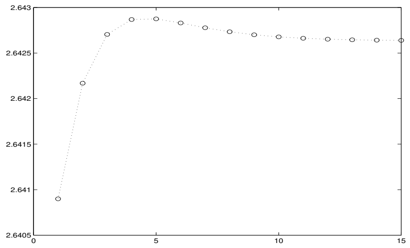


Figure 22: Poisson $\rho_1 = 0.95$.

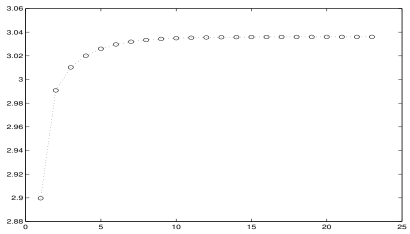


Figure 23: MMPP $\rho_1 = 0.25$.

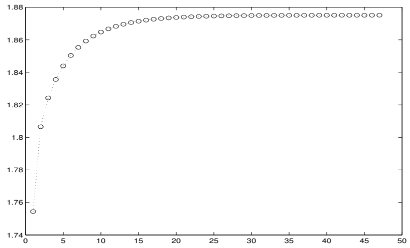


Figure 24: MMPP $\rho_1 = 0.5$.

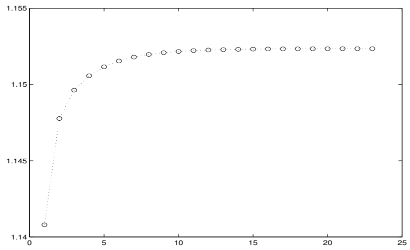


Figure 25: MMPP $\rho_1 = 0.75$.

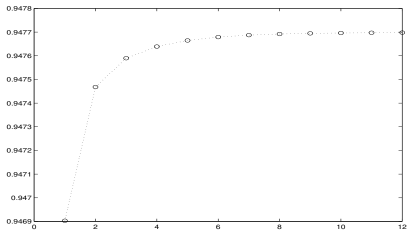


Figure 26: MMPP $\rho_1 = 0.95$.