Models of Production Lines as Quasi-Birth-Death Processes

MEHMET MURAT FADILOGLU, *Bilkent University*

SENCER YERALAN, *Rigel Corporation*

Industrial Engineering Department
Bilkent University
Bilkent, Ankara 06533
Turkey
mmurat@bilkent.edu.tr

Abstract: The aim of this work is to illustrate the suitability of Quasi-Birth-Death Processes (QBDs) for stochastic modeling of production lines. With this end in mind, first, an introduction to QBDs is made, so that the reader who may not be acquainted with this aspect of stochastic modeling may be introduced to the basics of the topic. Then, a formal definition of QBD is given and the QBDs are contrasted with the traditional birth-death processes. Later, examples of QBD models pertaining to production lines are presented.   The rational of this exposition is to show how QBDs present themselves within the context of production lines and to show the kind of work that needs to be performed to fully specify the corresponding QBD. By compiling the aforementioned models, the strength of QBDs in modeling production lines is demonstrated.

1 Introduction to Quasi-Birth-Death Processes:

The study of Quasi-Birth-Death Processes (QBDs) goes back to R. V. Evans [1] and to V. Wallace [2]. In fact, Wallace coined the term "Quasi-Birth-Death Process". Yet, the most comprehensive study is by Neuts [3]. Since all the following research refers to Neuts for the basics of the field, it is appropriate to give his definition:

A QBD is a Markov process on the state space $E = \{(i, j), i \geq 0, 1 \leq j \leq m\}$,

with the infinitesimal generator $\tilde{Q}$, given by

$$
\tilde{Q} = \begin{bmatrix}
B_0 & C & & & & & \\
A_0 & B & C & & & & \\
& A & B & C & & & \\
& & A & B & C & . & . & . \\
& & & A & B & . & . & . \\
& & & & . & . & \\
& & & & . & . & \\
& & & & . & . &
\end{bmatrix}
\tag{1}
$$

where $B_0 \underline{1}^T + C \underline{1}^T = A_0 \underline{1}^T + B \underline{1}^T + C \underline{1}^T = A \underline{1}^T + B \underline{1}^T + C \underline{1}^T = \underline{0}$.

The QBD, as defined, is a generalization of the birth-death process. Birth-death processes are stochastic processes of Markov type where each state can be associated with an element of a counting set, which effectively means that the states can be ordered in a linear fashion. Once the correct ordering is established, transitions should occur only to neighboring states: that means given that the process is in state n, it can have a transition only to state n+1 or state n-1. And the transition times are exponentially distributed which allows their classification under the general category of Markov processes. As with any Markov process, a state transition diagram (Figure 1) can be used to represent it.
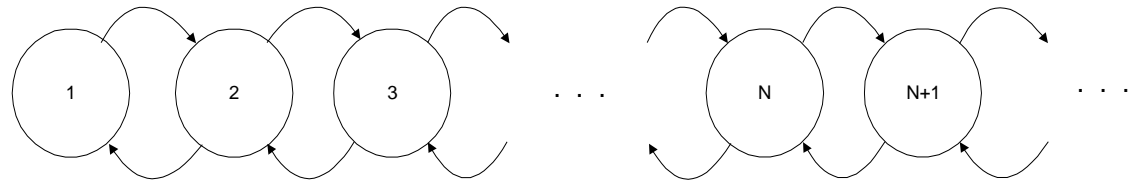

Figure 1 Basic birth-death process model

In the case of the QBD, instead of one state corresponding to an element of the counting set, there is a group of states corresponding to it. Thus, from a given group of states, the process can have transitions to the group of states corresponding to the previous element of the counting set or to the group of states corresponding to the next element of the counting set. This kind of Markov process yields a state transition diagram of the form shown in Figure 2.
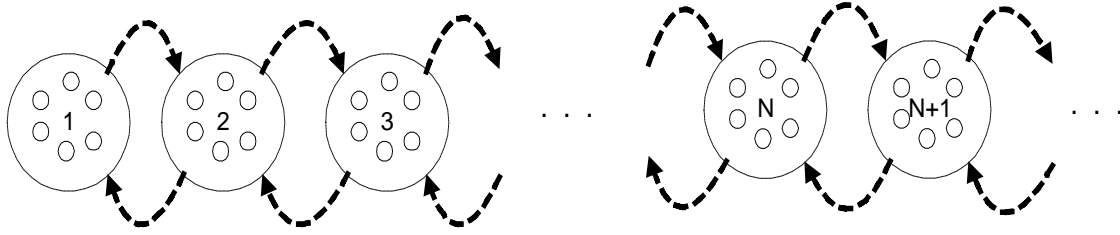
Figure 2   Representation of a QBD (The dashed lines represent a collection of transitions originating from a state in one group and terminate in a state in another group.)

According to Neuts' definition, a QBD is defined by a generator matrix that has a block nested tridiagonal structure in which the same blocks are repeated infinitely many times except for the initial states. For this to occur, one would need to have the same number of states corresponding to each counting set element. Moreover, the same kind of transitions with the same transition rates should be present in between these states. Thus all literature that adopts the definition of Neuts really relates to state-homogeneous quasi-birth-death processes.

Within the framework defined for QBDs, it is most appropriate to define the state space as the Kroenicker product of two sets, one of which is the counting set. Thus, the other set is used to distinguish all possible things that can happen at a given counting set level. The same things should be possible at all counting set levels and the way the process goes from one of them to another should be independent of the level.

The QBDs that Neuts investigates differ in one way from the ones that are the subject of this study. The counting set is infinite, in other words the counting set is the set of natural numbers for Neuts' case. A finite counting set is used here, since no production system is likely to have infinite buffers. This means that instead of having only one boundary there are two boundaries for these QBDs.

A formal definition for the kind of QBD used here is:

A   QBD   is   a   Markov   process   on   the   state   space $E = \{(i,j), 0 \leq i \leq M, 1 \leq j \leq n\}$, with the transition rate matrix –or the infinitesimal generator-- R, given by

$$
R = \begin{bmatrix}
B_0 & C & & & & & \\
A & B & C & & & & \\
& A & B & C & & & \\
& & & \cdot & \cdot & \cdot & \\
& & & & \cdot & \cdot & \cdot \\
& & & & & \cdot & \cdot & \cdot \\
& & & & & A & B & C \\
& & & & & & A & B_M
\end{bmatrix}
\tag{2}
$$

where $B_0 \underline{1}^T + C \underline{1}^T = A \underline{1}^T + B \underline{1}^T + C \underline{1}^T = A \underline{1}^T + B_M \underline{1}^T = \underline{0}$.

This will be the definition of QBD in this paper. An interesting observation is that the structure of equation (2) has been already shown to be of great use in the modeling of production lines in the seminal work by Yeralan and Muth [4]. Three works following this lead, Tan [5], Yeralan and Tan [6], Fadiloglu and Yeralan [7] exploit this structure in finding the steady-state probabilities by applying a matrix-polynomial solution procedure. The computational effort in the procedure is independent of the cardinality of the counting set of the QBD. Thus, the QBD models have the additional advantage of being amenable to an efficient solution procedure. Furthermore, the spectral theory based on the procedure carries the prospect of furthering the understanding of the behavior modeled systems manifest.

2 QBD Models for Production Lines

Although the QBD platform is quite general and can be used for many stochastic models that pertain to different areas, in this paper, examples of QBD models pertaining to production lines are examined. A similar effort can be conducted for other areas, which give rise to QBDs, but the scope of the present research is limited to the models of production lines, which have been studied extensively by many scholars due to the importance of the subject [8].

In this section production line models are presented in a natural progression. First, basic models, and then, more complex models that mimic further behavior of real production systems are introduced. The submatrices that define the QBD for each model will be specified. Once these are produced, the models are ready for application of the matrix polynomial procedure.

Model 1: Exponential Server with Limited Buffer

A graphical representation for the model is given in Figure 3. This is most probably the first model to be encountered in any queuing theory book, the quintessential M/M/1 model. Here, the only deviation from this basic model is that the buffer–or the queue–has a limited size, M. Usually this is named M/M/1/M, with the last M representing the size of the queue. In this model, parts arrive to the server as a Poisson process with a mean rate of λ, and they are serviced with exponential service times with mean μ.
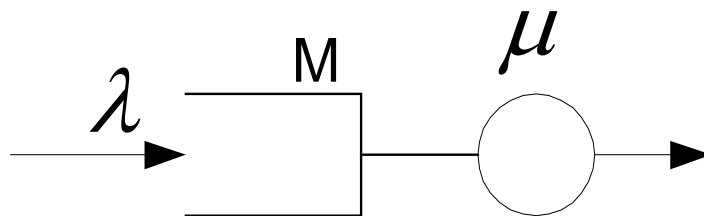
Figure 3 Exponential server with limited buffer and its relevant parameters

One can obtain the Markov chain diagram shown in Figure 4 using this description for the model. As one can observe from the diagram, this model is actually a simple birth-death process. But since QBD is a generalization of birth-death process this model still qualifies for the name QBD. Actually it forms one of the simplest QBDs possible. One can easily write the steady-state equations for this model by looking at the diagram:

$$-\lambda P(0) + \mu P(1) = 0$$

$$-(\lambda + \mu) P(i) + \mu P(i+1) + \lambda P(i-1) = 0 \quad for\ 1 \le i \le M - 1$$

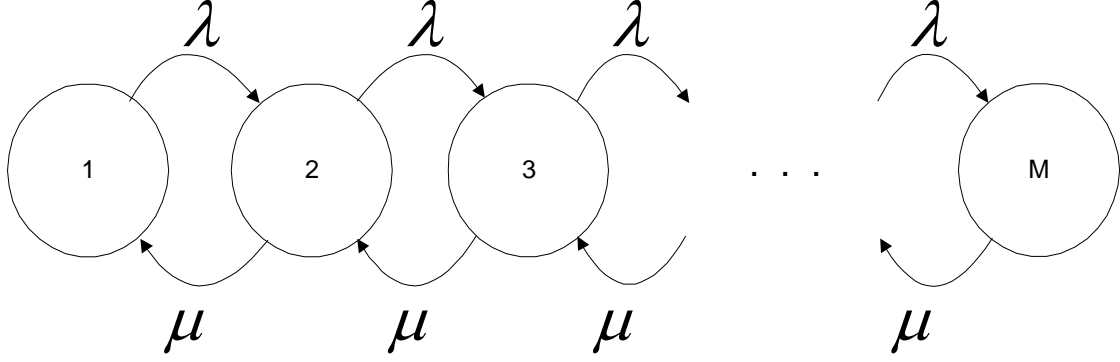$$-\mu P(M) + \lambda P(M-1) = 0$$

Figure 4 Markov transition diagram of Model 1

One can obtain the following transition rate matrix from the preceding equations:

$$R = \begin{bmatrix} -\lambda & \lambda & & & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & & & \\ & \mu & -(\lambda+\mu) & . & & & & \\ & & \mu & . & . & & & \\ & & & . & . & . & & \\ & & & & . & . & \lambda & \\ & & & & & . & -(\lambda+\mu) & \lambda \\ & & & & & & \mu & -\mu \end{bmatrix}$$

Clearly, transition rate matrix fits the QBD platform as defined. By comparing this transition rate matrix with the generic transition rate matrix defined by equation (2), one can give the specific values of the submatrices used in the formal definition:

$$A = \mu \, , \; B = -(\lambda + \mu) \, , \; C = \lambda$$

$$B_0 = -\lambda \, , \; B_M = -\mu$$

It is easy to observe that for this model the submatrices in the definition of the QBD manifest themselves as simple scalars. Although the application of the matrix polynomial method for this case is overkill, it is helpful to understand basic mechanisms of the procedure.

Model 2: Exponential Server with Breakdown and Repair, Limited Buffer

A graphical representation for the model is given in Figure 5. This model builds on Model 1 by incorporating breakdown and repair capability. If the server is

operating, it breaks down with a time to breakdown distribution that is exponential with parameter α. Once the breakdown occurs, the repair starts. The time to the end of repair is also exponentially distributed with parameter β. During the breakdown, the part that was being processed at the time breakdown occurred is not processed. Processing starts on the same part after the repair.
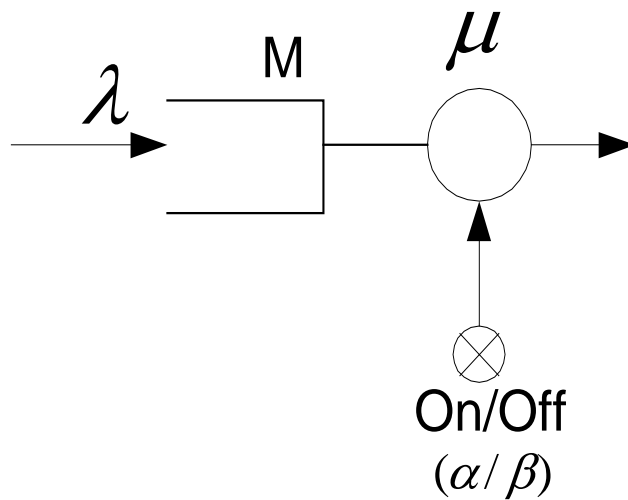

Figure 5 Exponential server with breakdown and repair and its relevant parameters.

For the described process the Markov transition diagram is given in Figure 6. It is clear that the state space is the Kroenicker product of the counting set with a set of two elements {W,R}. W corresponds to the states for which the server is working and R corresponds to the states for which the station is broken and the repair is taking place.
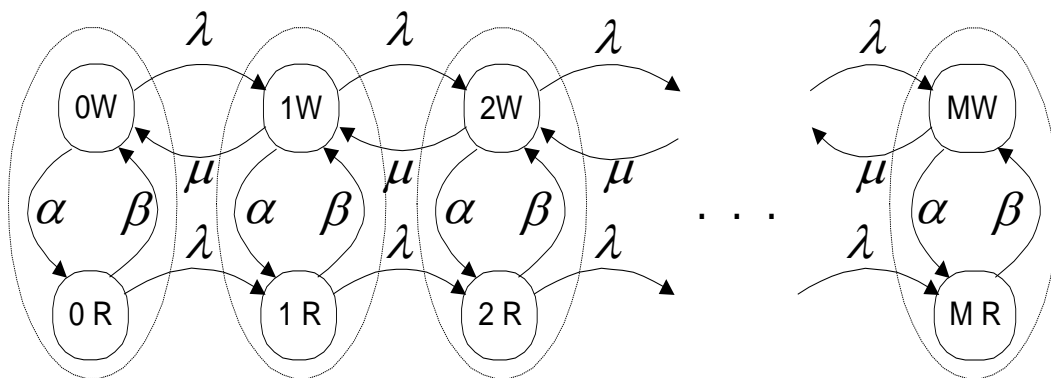

Figure 6 Markov transition diagram of Model 2.

The steady-state equations for the process are

$$-(\alpha + \lambda)P(0,W) + \beta P(0,R) + \mu P(1,W) = 0$$

$$-(\beta + \lambda)P(0,R) + \alpha P(0,W) = 0$$

$$-(\alpha + \lambda + \mu)P(i,W) + \beta P(i,R) + \lambda P(i-1,W) + \mu P(i+1,W) = 0 \qquad for\ 1 \le i \le M-1$$

$$-(\lambda + \mu)P(i,R) + \alpha P(i,W) + \lambda P(i-1,R) = 0 \qquad for\ 1 \le i \le M-1$$

$$-(\alpha + \mu)P(M,W) + \beta P(M,R) + \lambda P(M-1,W) = 0$$

$$-\beta P(M,R) + \alpha P(M,W) + \lambda P(M-1,R) = 0$$

One can obtain the following transition rate matrix from the preceding equations:

$$
R =
\begin{bmatrix}
-(\alpha+\lambda) & \alpha & \lambda & 0 & & & & & & \\
\beta & -(\beta+\lambda) & 0 & \lambda & & & & & & \\
\mu & 0 & -(\alpha+\lambda+\mu) & \alpha & . & . & & & & \\
0 & 0 & \beta & -(\beta+\lambda) & . & . & & & & \\
& & & & . & . & . & . & & \\
& & & & . & . & . & . & & \\
& & & & . & . & . & . & & \\
& & & & . & . & . & . & & \\
& & & & . & . & -(\alpha+\lambda+\mu) & \alpha & \lambda & 0 \\
& & & & . & . & \beta & -(\beta+\lambda) & 0 & \lambda \\
& & & & & & \mu & 0 & -(\alpha+\mu) & \alpha \\
& & & & & & 0 & 0 & \beta & -\beta
\end{bmatrix}
$$

Clearly, this transition rate matrix conforms to the definition of a QBD as stated. The dimension of the submatrices is 2x2 and the submatrices can be easily extracted:

$$
A = \begin{bmatrix} \mu & 0 \\ 0 & 0 \end{bmatrix}, \quad
B = \begin{bmatrix} -(\alpha+\lambda+\mu) & \alpha \\ \beta & -(\beta+\lambda) \end{bmatrix}, \quad
C = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}
$$

$$
B_0 = \begin{bmatrix} -(\alpha+\lambda) & \alpha \\ \beta & -(\beta+\lambda) \end{bmatrix}, \quad
B_M = \begin{bmatrix} -(\alpha+\mu) & \alpha \\ \beta & -\beta \end{bmatrix}
$$

Model 3: Server With Erlang-2 Service Times, Limited Buffer

The only difference between this model, depicted in Figure 7, and Model 1 is the statistical distribution of the service times. In the M/M/1 model the distribution is exponential, whereas here it is Erlang which is the sum of two exponential distributions with the same rate. In order to accommodate for this non-exponential

distribution, one needs to incorporate an additional state for each counting level. When this idea is applied, one gets the Markov transition diagram shown in Figure 8.
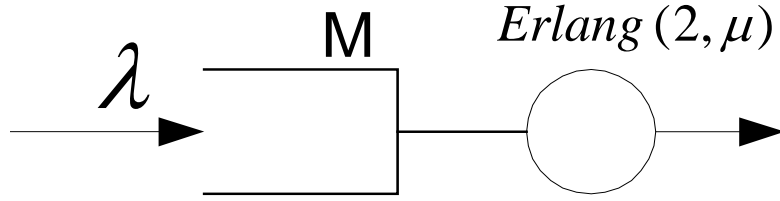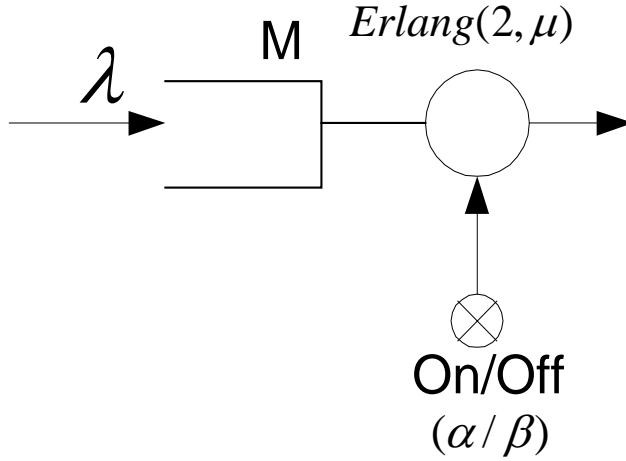

Figure 7 Server with Erlang-2 service times, limited buffer and its relevant parameters.

The state space is the Kroenicker product of the counting set and the set {1,2}, which keeps track of the Erlang stage in the processing of parts. Here, only one state corresponds to the element 0 of the counting set. Consequently, this element cannot be incorporated in the QBD and has to be handled separately. The QBD starts from element 1. This has to be taken into account while applying the matrix polynomial solution procedure.
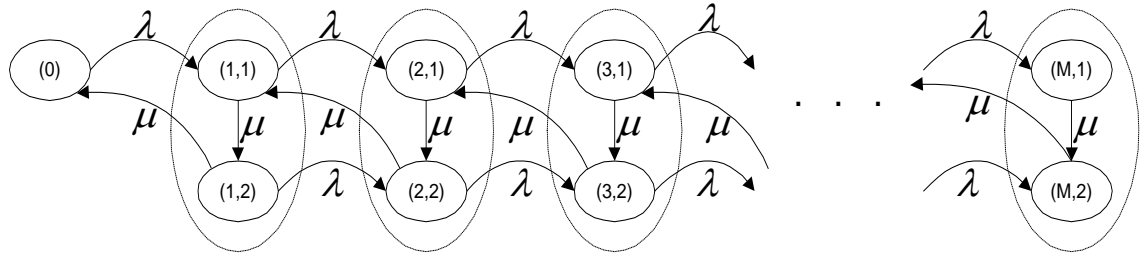
Figure 8 Markov transition diagram of Model 3

The steady-state equations for the process are

$$-\lambda P(0) + \mu P(1,2) = 0$$

$$-(\lambda + \mu) P(1,1) + \lambda P(0) + \mu P(2,2) = 0$$

$$-(\lambda + \mu) P(1,1) + \mu P(1,2) = 0$$

$$-(\lambda + \mu) P(i,1) + \lambda P(i-1,1) + \mu P(i+1,2) = 0 \qquad for \ 2 \leq i \leq M-1$$

$$-(\lambda + \mu) P(i,2) + \lambda P(i-1,2) + \mu P(i,1) = 0 \qquad for \ 2 \leq i \leq M-1$$

$$-\mu P(M,1) + \lambda P(M-1,1) = 0$$

$$-\mu P(M,2) + \lambda P(M-1,2) + \mu P(M,1) = 0$$

After substituting the value of $P(0)$ obtained from the first equation into the second equation, one can obtain the following transition rate matrix from the preceding equations:

$$
R = \begin{bmatrix}
-(\lambda+\mu) & \mu & \lambda & 0 & & & & & & \\
\mu & -(\beta+\lambda) & 0 & \lambda & & & & & & \\
0 & 0 & -(\lambda+\mu) & \mu & . & . & & & & \\
\mu & 0 & 0 & -(\lambda+\mu) & . & . & & & & \\
& & & & . & . & . & . & & \\
& & & & . & . & . & . & & \\
& & & & . & . & . & . & & \\
& & & & . & . & . & . & & \\
& & & & . & . & -(\lambda+\mu) & \mu & \lambda & 0 \\
& & & & . & . & 0 & -(\lambda+\mu) & 0 & \lambda \\
& & & & & & 0 & 0 & -\mu & \mu \\
& & & & & & \mu & 0 & 0 & -\mu
\end{bmatrix}
$$

Clearly, this transition rate matrix conforms to the definition of a QBD as stated. The dimension of the submatrices is 2x2 and the submatrices can be easily extracted:

$$
A = \begin{bmatrix} 0 & 0 \\ \mu & 0 \end{bmatrix}, \quad
B = \begin{bmatrix} -(\lambda+\mu) & \mu \\ 0 & -(\lambda+\mu) \end{bmatrix}, \quad
C = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}
$$

$$
B_1 = \begin{bmatrix} -(\lambda+\mu) & \mu \\ \mu & -(\lambda+\mu) \end{bmatrix}, \quad
B_M = \begin{bmatrix} -\mu & \mu \\ 0 & -\mu \end{bmatrix}
$$

Model 4: Server With Erlang-2 Service Times, Breakdown and Repair, Limited Buffer

This model, depicted in Figure 9, combines the characteristics of Model 2 and Model 3. There is breakdown and repair mechanism just as in Model 2. However, the production times are distributed with Erlang-2 distribution as the case is in Model 3. To accommodate for these, one needs three states corresponding to each counting set level. The Markov transition diagram for the process is given in Figure 10.

Figure 9 Server with Erlang-2 service times, breakdown and repair, limited buffer and its relevant parameters.

This time the state space is the Kroenicker product of the counting set and the set {1,2,B}. This second set is to specify the Erlang stage in the processing while the machine is up or to specify that the machine is broken. But only two states correspond to the element zero of the counting set. This element has to be handled separately and the QBD starts from element 1.

The steady-state equations for the process are

$$-(\alpha + \lambda)P(0,W) + \beta P(0,B) + \mu P(1,2) = 0$$

$$-(\beta + \lambda)P(0,B) + \alpha P(0,W) = 0$$

$$-(\alpha + \lambda + \mu)P(1,1) + \lambda P(0,W) + \beta P(1,B) + \mu P(2,2) = 0$$

$$-(\alpha + \lambda + \mu)P(1,2) + \mu P(1,1) = 0$$

$$-(\beta + \lambda)P(1,B) + \lambda P(0,B) + \alpha P(1,1) + \alpha P(1,2) = 0$$

$$-(\alpha + \lambda + \mu)P(i,1) + \lambda P(i-1,1) + \beta P(i,B) + \mu P(i+1,2) = 0 \ \ for \ 2 \le i \le M-1$$

$$-(\alpha + \lambda + \mu)P(i,2) + \lambda P(i-1,2) + \mu P(i,1) = 0 \qquad\qquad for \ 2 \le i \le M-1$$

$$-(\beta + \lambda)P(i,B) + \lambda P(i-1,B) + \alpha P(i,1) + \alpha P(i,2) = 0 \qquad for \ 2 \le i \le M-1$$

$$-(\alpha + \mu)P(M,1) + \lambda P(M-1,1) + \beta P(M,B) = 0$$

$$-(\alpha + \mu)P(M,2) + \lambda P(M-1,2) + \mu P(M,1) = 0$$

$$-\beta P(M,B) + \lambda P(M-1,B) + \alpha P(M,1) + \alpha P(M,2) = 0$$
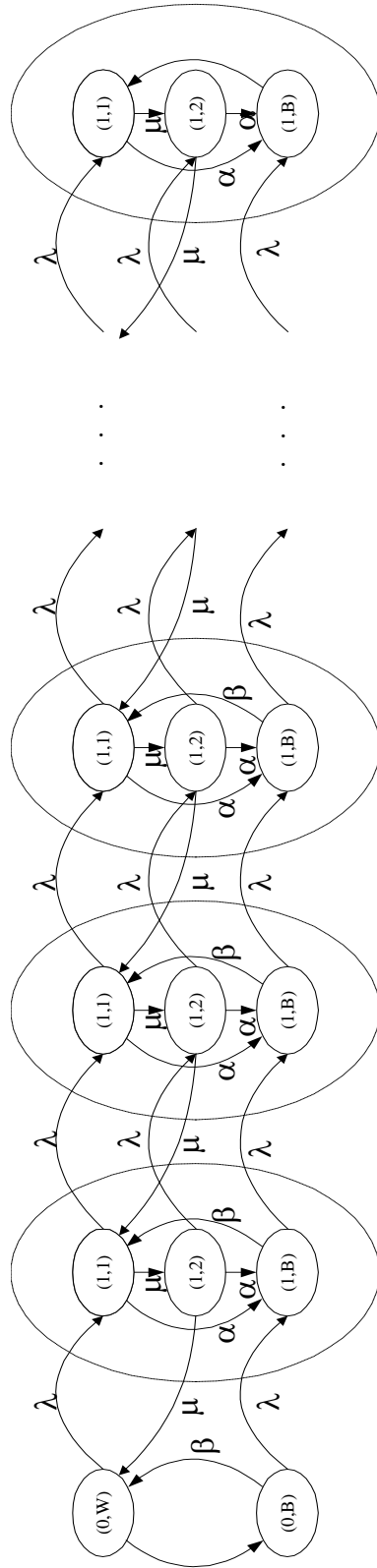
Figure 10 Markov transition diagram of Model 4

After substituting the values of $P(0,W)$ and $P(0,B)$ obtained from the first two equations into the third and fifth equations, one can obtain the transition rate matrix which is of the form given in Equation (2). The submatrices should be defined as

$$A = \begin{bmatrix} 0 & 0 & 0 \\ \mu & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -(\alpha+\lambda+\mu) & \mu & \alpha \\ 0 & -(\alpha+\lambda+\mu) & \alpha \\ \beta & 0 & -(\beta+\lambda) \end{bmatrix}$$

$$C = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}, \quad B_1 = \begin{bmatrix} -(\alpha+\lambda+\mu) & \mu & \alpha \\ \dfrac{\mu(\beta+\lambda)}{(\alpha+\lambda+\mu)} & -(\alpha+\lambda+\mu) & \dfrac{\alpha\mu}{(\alpha+\lambda+\mu)}+\alpha \\ \beta & 0 & -(\beta+\lambda) \end{bmatrix}$$

$$B_M = \begin{bmatrix} -(\alpha+\mu) & \mu & \alpha \\ 0 & -(\alpha+\mu) & \alpha \\ \beta & 0 & -\beta \end{bmatrix} .$$

Model 5: Two Exponential Servers with One Intermediate Limited Buffer

In this model shown in Figure 11, two exponential servers are present and the buffer is placed between the servers. The typical two-server model in the literature assumes an infinite source for the server, which is the stochastic equivalent of Model 1. The input is a Poisson process. One can assume that there is a second buffer at the input of the system. But the size of this buffer would be limited to one. In following models, the difficulties faced when a generic sized second buffer is incorporated in the models are illustrated. The Markov transition diagram for the process is given in Figure 12.
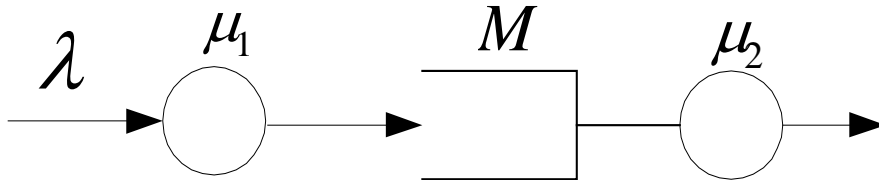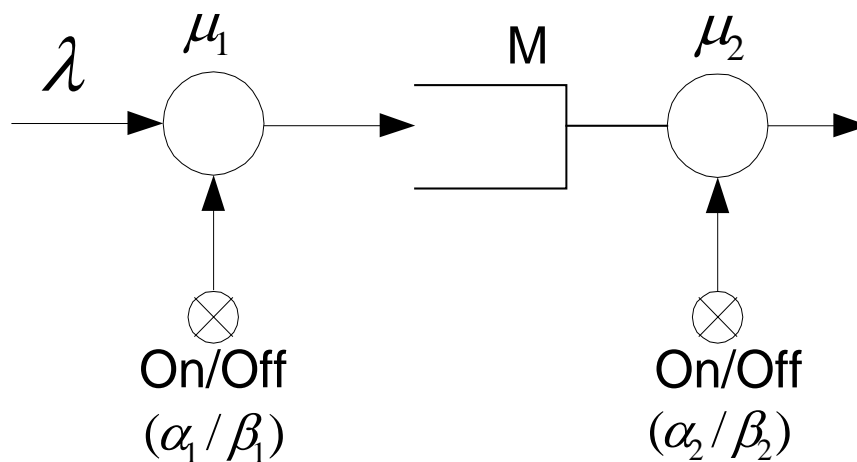

Figure 11 Two exponential servers with one intermediate limited buffer and its relevant parameters.

The state space is the Kroenicker product of the counting set and the set {0,1}. The counting set relates to the buffer level as usual. The set {0,1} can be considered as the level of an imaginary first buffer whose size is one; in other words it tells whether there is a part being processed in server one.
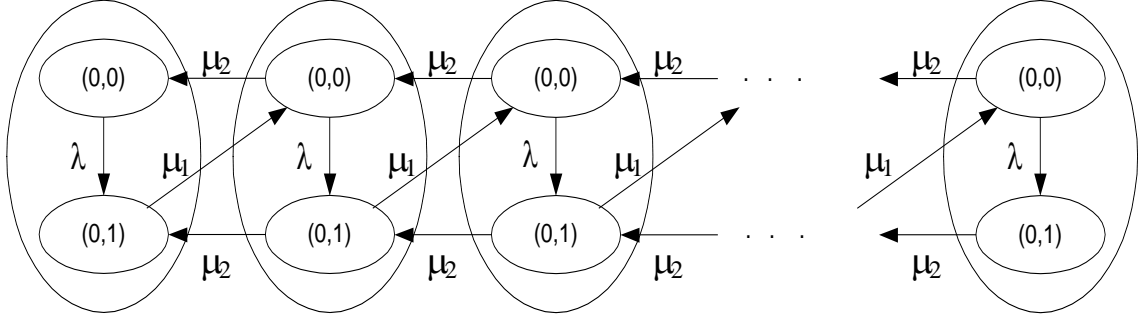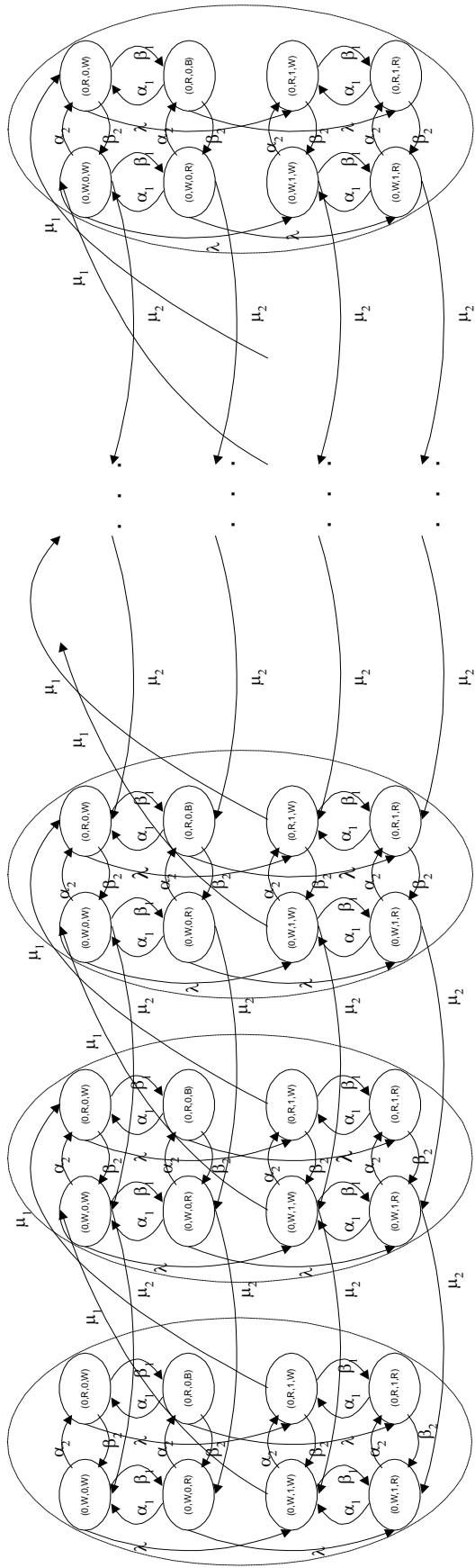

Figure 12 Markov transition diagram of Model 5

The steady-state equations for the process are

$$-\lambda P(0,0) + \mu_2 P(1,0) = 0$$

$$-\mu_1 P(0,1) + \lambda P(0,0) + \mu_2 P(1,1) = 0$$

$$-(\lambda + \mu_2)P(i,0) + \mu_1 P(i-1,1) + \mu_2 P(i+1,0) = 0 \qquad \text{for } 1 \le i \le M-1$$

$$-(\mu_1 + \mu_2)P(i,1) + \lambda P(i,0) + \mu_2 P(i+1,1) = 0 \qquad \text{for } 1 \le i \le M-1$$

$$-(\lambda + \mu_2)P(M,0) + \mu_1 P(M-1,1) = 0$$

$$-\mu_2 P(M,1) + \lambda P(M,0) = 0$$

This process conforms to the definition of a QBD as stated. Thus all one has to do for formally defining the QBD is to state the relevant submatrices whose sizes are 2x2:

$$A = \begin{bmatrix} \mu_2 & 0 \\ 0 & \mu_2 \end{bmatrix}, \quad B = \begin{bmatrix} -(\lambda + \mu_2) & \lambda \\ 0 & -(\mu_1 + \mu_2) \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ \mu_1 & 0 \end{bmatrix}$$

$$B_0 = \begin{bmatrix} -\lambda & \lambda \\ 0 & -\mu_1 \end{bmatrix}, \quad B_M = \begin{bmatrix} -(\lambda + \mu_2) & \lambda \\ 0 & -\mu_2 \end{bmatrix}$$

Model 6: Two Exponential Servers With One Intermediate Limited Buffer, Breakdown and Repair

This model depicted in Figure 13 is an extension of Model 4. The breakdown and repair mechanisms are added to two basic servers of Model 4. Both servers become non-operational when a breakdown occurs at the server, up to the time that the repair ends. Thus, the production process is hindered by these occurrences. These breakdowns and repairs occur independently in both servers. The modeling of this system requires a large number of states for each intermediate buffer level as illustrated in the Markov transition diagram of the process in Figure 14.

Figure 13 Two exponential servers with one intermediate limited buffer, breakdown and repair and its relevant parameters

The state space for this model is the Kroenicker product of four sets: the counting set C–which appears in all the QBDs–modeling the intermediate buffer level, the set $S_1 = \{W,R\}$ modeling the repair status of the first server, the set $L_1 = \{0,1\}$ modeling number of parts in server one, and the set $S_2 = \{W,R\}$ modeling the repair status of the second server.

Figure 14 Markov transition diagram of Model 6

The steady-state equations for the process are

$$-(\lambda+\alpha_1+\alpha_2)P(0,W,0,W)+\beta_1 P(0,W,0,R)+\beta_2 P(0,R,0,W)+\mu_2 P(1,W,0,W)=0$$

$$-(\lambda+\alpha_2+\beta_2)P(0,W,0,R)+\alpha_1 P(0,W,0,W)+\beta_2 P(0,R,0,R)+\mu_2 P(1,W,0,R)=0$$

$$-(\alpha_1+\alpha_2+\mu_1)P(0,W,1,W)+\lambda P(0,W,0,W)+\beta_1 P(0,W,1,R)$$
$$+\beta_2 P(0,R,1,W)+\mu_2 P(1,W,1,W)=0$$

$$-(\alpha_2+\beta_1)P(0,W,1,R)+\lambda P(0,W,0,R)+\alpha_1 P(0,W,1,W)$$
$$+\beta_2 P(0,R,1,R)+\mu_2 P(1,W,1,R)=0$$

$$-(\lambda+\alpha_1+\beta_2)P(0,R,0,W)+\alpha_2 P(0,W,0,W)+\beta_1 P(0,R,0,R)=0$$

$$-(\lambda+\beta_1+\beta_2)P(0,R,0,R)+\alpha_1 P(0,R,0,W)+\alpha_2 P(0,W,0,R)=0$$

$$-(\alpha_1+\beta_2+\mu_1)P(0,R,1,W)+\lambda P(0,R,0,W)+\alpha_2 P(0,W,1,W)+\beta_1 P(0,R,1,R)=0$$

$$-(\beta_1+\beta_2)P(0,R,1,R)+\lambda P(0,R,0,R)+\alpha_1 P(0,R,1,W)+\alpha_2 P(0,W,1,R)=0$$

$$-(\lambda+\alpha_1+\alpha_2+\mu_2)P(i,W,0,W)+\beta_1 P(i,W,0,R)+\beta_2 P(i,R,0,W)$$
$$+\mu_2 P(i+1,W,0,W)+\mu_1 P(i-1,W,1,W)=0 \qquad \textit{for } 1\le i \le M-1$$

$$-(\lambda+\alpha_2+\beta_2+\mu_2)P(i,W,0,R)+\alpha_1 P(i,W,0,W)+\beta_2 P(i,R,0,R)$$
$$+\mu_2 P(i+1,W,0,R)=0 \qquad \textit{for } 1\le i \le M-1$$

$$-(\alpha_1+\alpha_2+\mu_1+\mu_2)P(i,W,1,W)+\lambda P(i,W,0,W)+\beta_1 P(i,W,1,R)$$
$$+\beta_2 P(i,R,1,W)+\mu_2 P(i+1,W,1,W)=0 \qquad \textit{for } 1\le i \le M-1$$

$$-(\alpha_2+\beta_1+\mu_2)P(i,W,1,R)+\lambda P(i,W,0,R)+\alpha_1 P(i,W,1,W)$$
$$+\beta_2 P(i,R,1,R)+\mu_2 P(i+1,W,1,R)=0 \qquad \textit{for } 1\le i \le M-1$$

$$-(\lambda+\alpha_1+\beta_2)P(i,R,0,W)+\alpha_2 P(i,W,0,W)+\beta_1 P(i,R,0,R)$$
$$+\mu_1 P(i-1,R,1,W)=0 \qquad \textit{for } 1\le i \le M-1$$

$$-(\lambda+\beta_1+\beta_2)P(i,R,0,R)+\alpha_1 P(i,R,0,W)+\alpha_2 P(i,W,0,R)=0 \quad \textit{for } 1\le i \le M-1$$

$$-(\alpha_1+\beta_2+\mu_1)P(i,R,1,W)+\lambda P(i,R,0,W)+\alpha_2 P(i,W,1,W)$$
$$+\beta_1 P(i,R,1,R)=0 \qquad \textit{for } 1\le i \le M-1$$

$$-(\beta_1+\beta_2)P(i,R,1,R)+\lambda P(i,R,0,R)+\alpha_1 P(i,R,1,W)$$
$$+\alpha_2 P(i,W,1,R)=0 \qquad \textit{for } 1\le i \le M-1$$

$$-(\lambda+\alpha_1+\alpha_2+\mu_2)P(M,W,0,W)+\beta_1 P(M,W,0,R)+\beta_2 P(M,R,0,W)$$
$$+\mu_1 P(M-1,W,1,W)=0$$

$$-(\lambda+\alpha_2+\beta_2+\mu_2)P(M,W,0,R)+\alpha_1 P(M,W,0,W)+\beta_2 P(M,R,0,R)=0$$

$$-(\alpha_1+\alpha_2+\mu_2)P(M,W,1,W)+\lambda P(M,W,0,W)+\beta_1 P(M,W,1,R)$$
$$+\beta_2 P(M,R,1,W)=0$$

$$-(\alpha_2 + \beta_1 + \mu_2)P(M,W,1,R) + \lambda P(M,W,0,R) + \alpha_1 P(M,W,1,W)$$
$$+\beta_2 P(M,R,1,R) = 0$$

$$-(\lambda + \alpha_1 + \beta_2)P(M,R,0,W) + \alpha_2 P(M,W,0,W) + \beta_1 P(M,R,0,R)$$
$$+\mu_1 P(M-1,R,1,W) = 0$$

$$-(\lambda + \beta_1 + \beta_2)P(M,R,0,R) + \alpha_1 P(M,R,0,W) + \alpha_2 P(M,W,0,R) = 0$$

$$-(\alpha_1 + \beta_2)P(M,R,1,W) + \lambda P(M,R,0,W) + \alpha_2 P(M,W,1,W) + \beta_1 P(M,R,1,R) = 0$$

$$-(\beta_1 + \beta_2)P(M,R,1,R) + \lambda P(M,R,0,R) + \alpha_1 P(M,R,1,W) + \alpha_2 P(M,W,1,R) = 0$$

This process conforms to the definition of a QBD as defined. Thus, all one has to do for formally defining the QBD is to state the relevant submatrices whose sizes are 8x8:

$$A = \begin{bmatrix}
\mu_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \mu_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \mu_2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mu_2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}, \quad
C = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mu_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

$$B = \begin{bmatrix}
-(\lambda+\alpha_1+\alpha_2+\mu_2) & \alpha_1 & \lambda & 0 & \alpha_2 & 0 & 0 & 0 \\
\beta_1 & -(\lambda+\alpha_2+\beta_1+\mu_2) & 0 & \lambda & 0 & \alpha_2 & 0 & 0 \\
0 & 0 & -(\alpha_1+\alpha_2+\mu_1+\mu_2) & \alpha_1 & 0 & 0 & \alpha_2 & 0 \\
0 & 0 & \beta_1 & -(\alpha_2+\beta_1+\mu_2) & 0 & 0 & 0 & \alpha_2 \\
\beta_2 & 0 & 0 & 0 & -(\lambda+\alpha_1+\beta_2) & \alpha_1 & \lambda & 0 \\
0 & \beta_2 & 0 & 0 & \beta_1 & -(\lambda+\beta_1+\beta_2) & 0 & \lambda \\
0 & 0 & \beta_2 & 0 & 0 & 0 & -(\alpha_1+\beta_2+\mu_1) & \alpha_1 \\
0 & 0 & 0 & \beta_2 & 0 & 0 & \beta_1 & -(\beta_1+\beta_2)
\end{bmatrix}$$

$$B_0 = \begin{bmatrix}
-(\lambda+\alpha_1+\alpha_2) & \alpha_1 & \lambda & 0 & \alpha_2 & 0 & 0 & 0 \\
\beta_1 & -(\lambda+\alpha_2+\beta_1) & 0 & \lambda & 0 & \alpha_2 & 0 & 0 \\
0 & 0 & -(\alpha_1+\alpha_2+\mu_1) & \alpha_1 & 0 & 0 & \alpha_2 & 0 \\
0 & 0 & \beta_1 & -(\alpha_2+\beta_1) & 0 & 0 & 0 & \alpha_2 \\
\beta_2 & 0 & 0 & 0 & -(\lambda+\alpha_1+\beta_2) & \alpha_1 & \lambda & 0 \\
0 & \beta_2 & 0 & 0 & \beta_1 & -(\lambda+\beta_1+\beta_2) & 0 & \lambda \\
0 & 0 & \beta_2 & 0 & 0 & 0 & -(\alpha_{11}+\beta_2+\mu_1) & \alpha_1 \\
0 & 0 & 0 & \beta_2 & 0 & 0 & \beta_1 & -(\beta_1+\beta_2)
\end{bmatrix}$$

$$
B_M =
\begin{bmatrix}
-(\lambda+\alpha_1+\alpha_2+\mu_1) & \alpha_1 & \lambda & 0 & \alpha_2 & 0 & 0 & 0 \\
\beta_1 & -(\lambda+\alpha_2+\beta_1+\mu_1) & 0 & \lambda & 0 & \alpha_2 & 0 & 0 \\
0 & 0 & -(\alpha_1+\alpha_2+\mu_1) & \alpha_1 & 0 & 0 & \alpha_2 & 0 \\
0 & 0 & \beta_1 & -(\alpha_2+\beta_1+\mu_1) & 0 & 0 & 0 & \alpha_2 \\
\beta_2 & 0 & 0 & 0 & -(\lambda+\alpha_1+\beta_2) & \alpha_1 & \lambda & 0 \\
0 & \beta_2 & 0 & 0 & \beta_1 & -(\lambda+\beta_1+\beta_2) & 0 & \lambda \\
0 & 0 & \beta_2 & 0 & 0 & 0 & -(\alpha_1+\beta_2) & \alpha_1 \\
0 & 0 & 0 & \beta_2 & 0 & 0 & \beta_1 & -(\beta_1+\beta_2)
\end{bmatrix}
$$

Model 7: Server With Erlang-N Service Times, Limited Buffer

In Model 3, the case when the server has Erlang-2 service times is studied. In this model depicted in Figure 15 this model is extended to Erlang-N service times. The number of states in this model increases both with M and N. Yet the matrix polynomial solution procedure does not depend on the size of the counting space which is the buffer size here. But the larger the Erlang parameter is, the larger the submatrices are. Thereby, the computational effort increases with the parameter N. The Markov transition diagram of the process is given below:
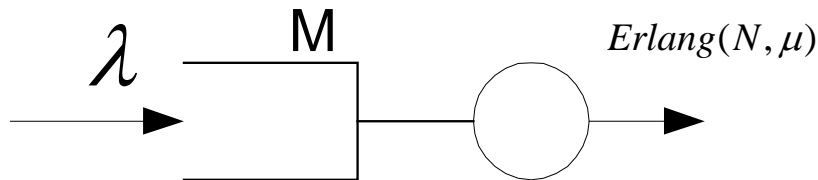

Figure 15 Server with Erlang-N service times, limited buffer and its relevant parameters.

The state space for this model is the Kroenicker product of two sets: the counting set C that corresponds to the buffer level and the Erlang stage set S. The Erlang stage set itself, is also a counting set. Thus one can have two different matrix representations for the QBD: one using the buffer level as the counting set of the QBD and one using the Erlang stage set. But since there is a transition from the Erlang stage N to Erlang stage 1, the choice of Erlang stage set would not yield a QBD as defined.

The steady-state equations for the process are

$$-\lambda P(0) + \mu P(1,N) = 0$$

$$-(\lambda + \mu)P(1,1) + \lambda P(0) + \mu P(2,N) = 0$$

$$-(\lambda + \mu)P(1,j) + \mu P(1,j-1) = 0 \qquad\qquad for \ \ 1 \le j \le N$$

$$-(\lambda + \mu)P(i,1) + \lambda P(i-1,1) + \mu P(i,N) = 0 \qquad\qquad for \ \ 2 \le i \le M-1$$

$$-(\lambda + \mu)P(i,j) + \lambda P(i-1,j) + \mu P(i,j-1) = 0$$

$$for \ \ 2 \le i \le M-1 \ \ and \ \ 2 \le j \le N$$

$$-\mu P(M,1) + \lambda P(M-1,1) = 0$$

$$-\mu P(M,j) + \lambda P(M-1,j) + \mu P(M,j-1) = 0 \qquad\qquad 2 \le j \le N$$
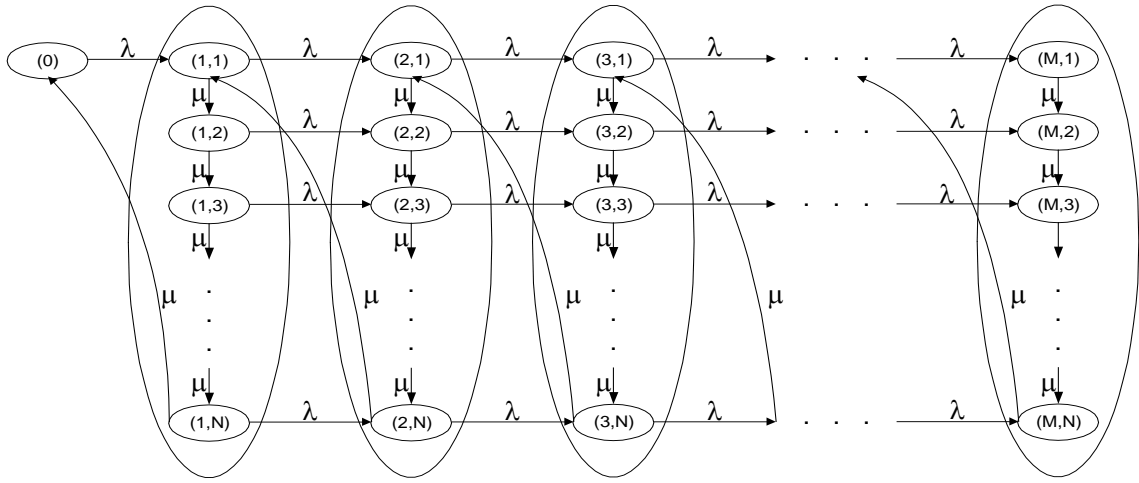
Figure 16 Markov transition diagram of Model 7

After substituting the value of $P(0)$ obtained from the equations, one can obtain the transition rate matrix for this process. The transition rate matrix conforms to the definition of the QBDs as one could expect. This time the submatrices employed are of size NxN:

$$A = \begin{bmatrix} 0 & 0 & . & . & 0 \\ . & . & & & . \\ . & . & & & . \\ 0 & 0 & . & . & 0 \\ \mu & 0 & . & . & 0 \end{bmatrix} , \quad B = \begin{bmatrix} -(\lambda+\mu) & \mu & & & \\ & -(\lambda+\mu) & . & & \\ & & . & . & \\ & & & . & \mu \\ & & & & -(\lambda+\mu) \end{bmatrix}$$

$$C = \begin{bmatrix} \lambda & & & & \\ & \lambda & & & \\ & & . & & \\ & & & . & \\ & & & & \lambda \end{bmatrix} , \quad B_1 = \begin{bmatrix} -(\lambda+\mu) & \mu & & & \\ & -(\lambda+\mu) & . & & \\ & & . & . & \\ & & & . & \mu \\ \mu & & & & -(\lambda+\mu) \end{bmatrix}$$

$$B_M = \begin{bmatrix} -\mu & \mu & & & \\ & -\mu & . & & \\ & & . & . & \\ & & & . & \mu \\ & & & & -\mu \end{bmatrix}$$

If one wanted to use Erlang stage set as the counting set, he would not get the a QDP in the strict sense of the word, but something quite close to it. And the matrix polynomial solution procedure for obtaining the steady-state probabilities would still be applicable since the deviation from the QDP definition would only occur in the boundary equations. This is due to the fact that the matrix polynomial solution procedure is based on the structure of the inner equations. It does not preclude a deviation from the definition in boundary equations.

The transition rate matrix obtained with such a choice would be of the form

$$R = \begin{bmatrix} B_0 & C & & & & & & & X_M \\ A & B & C & & & & & & \\ & A & B & . & & & & & \\ & & A & . & . & & & & \\ & & & . & . & & & & \\ & & & & . & B & C & & \\ & & & & & A & B & C & \\ X_0 & & & & & & A & B_M \end{bmatrix} \qquad (3)$$

This one is a generalization of the transition rate matrix given in equation (2). One can observe that two impurities are allowed at the corners of the matrix.

Once this transition rate matrix is defined, one can write the transition rate matrix of the process with the Erlang stage used as the counting set within this platform. Indeed to define the submatrices in the (3) would suffice:

$$B = \begin{bmatrix} -(\lambda+\mu) & \lambda & & & \\ & -(\lambda+\mu) & . & . & \\ & & . & . & \\ & & & \lambda & \\ & & -(\lambda+\mu) & \lambda \\ & & & -\mu \end{bmatrix} , \quad B_1 = B$$

$$
B_M = \begin{bmatrix} -\mu & \mu & & & & & \\ & -\mu & \cdot & \cdot & & & \\ & & \cdot & \cdot & & & \\ & & & & \mu & & \\ & & & & -\mu & \mu \\ & & & & & -\mu \end{bmatrix} , \quad C = \begin{bmatrix} \mu & & & & \\ & \mu & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \mu \end{bmatrix} , \quad A = \underline{\underline{0}}
$$

$$
X_1 = \begin{bmatrix} \mu & & & & & \\ 0 & 0 & & & & \\ & \mu & 0 & & & \\ & & \mu & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & \cdot & 0 \\ & & & & \mu & 0 \end{bmatrix} , \quad X_M = \underline{\underline{0}}
$$

Model 8: Two Exponential Servers with Two Limited Buffers

This model depicted in Figure 17 is an extension on the Model 5. In Model 5, the first server does not have a buffer for the arrivals. Thereby, those arrivals that occur while the server is occupied are lost. In this model, those arrivals are stored in a buffer whose size is M_1. This model is quite important because it is a tandem queue, which happens to form the basic building block of many models for modern production systems. Yet, a theoretically sound methodology has not developed to tackle the stochastic analysis of this queue class up to this point.

Figure 17 Two exponential servers with two limited buffers and their relevant parameters

To perform stochastic analysis of this model, one can generate a Markov chain using a large state-space to incorporate the second queue. Though this is a valid

option, it is not viable for many real models since the computational effort increases with the size of the state space. Moreover it is not conducive to obtain parametric results from which one can obtain an idea about the behavior of such systems.

Within the platform of matrix polynomial solution procedure, one can solve QBDs with an effort independent of the size of the counting set. Yet, for the tandem queues one does not have only one counting set but many of them. For this model with two tandem queues, there are two counting sets, each corresponding to the buffer levels of the two servers.
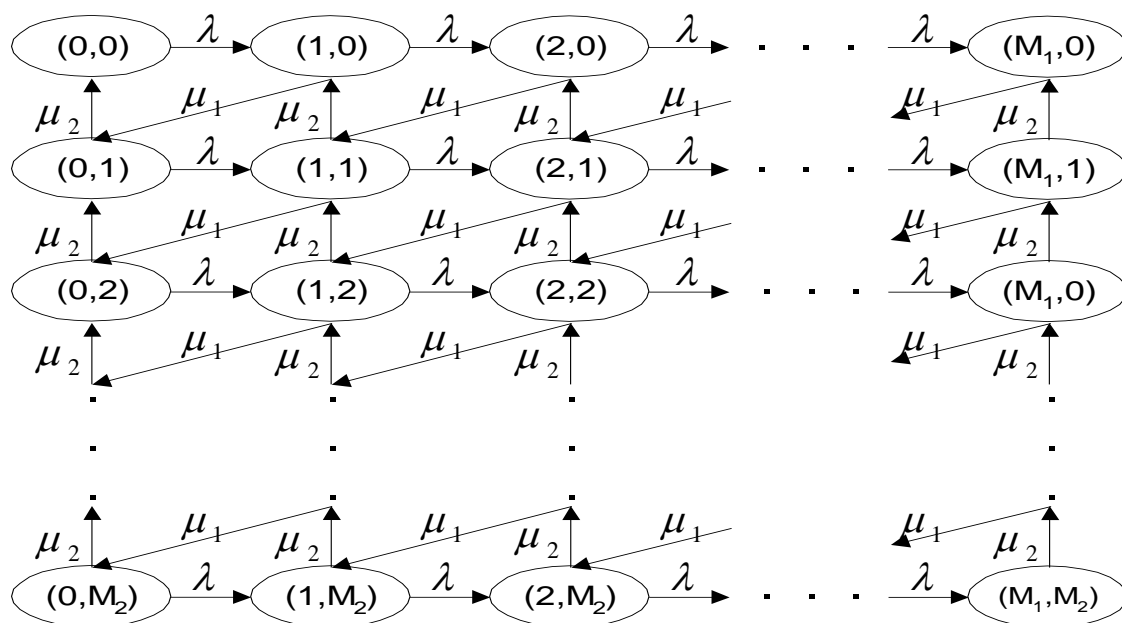
Figure 18 Markov transition diagram of Model 8

Since the matrix polynomial methodology only allows the exploitation of the structure for a single counting set, one cannot make use of the second counting set. One point to consider is the choice of the principal counting set of the QBD. The logical strategy for this is to choose the one with higher cardinality. Then, the computational effort that relates to the solution decreases.

The Markov transition diagram pertaining to the model is given in Figure 18. One can clearly observe by examining the diagram that the state space is the Kroenicker product of two counting sets: the first buffer level set $C_1 = \{1,2,3, \ldots, M_1\}$ and the second buffer level set $C_2 = \{1,2,3, \ldots, M_2\}$. The first index of the states is an

element of C_1 and the second index of C_2. One also observes that there are $M_1.M_2$ states in the model.

The steady-state equations for the process are

$$-\lambda P(0,0) + \mu_2 P(0,1) = 0$$

$$-(\lambda + \mu_1)P(i,0) + \lambda P(i-1,0) + \mu_2 P(i,1) = 0 \qquad \text{for } 1 \le i \le M_1 - 1$$

$$-\mu_1 P(M_1,0) + \lambda P(M_1 - 1,0) + \mu_2 P(M_1,1) = 0$$

$$-(\lambda + \mu_2)P(0,j) + \mu_1 P(1, j-1) + \mu_2 P(0, j+1) = 0 \qquad \text{for } 1 \le j \le M_2 - 1$$

$$-(\lambda + \mu_2)P(0, M_2) + \mu_1 P(1, M_2 - 1) = 0$$

$$-(\lambda + \mu_1 + \mu_2)P(i, j) + \lambda P(i-1, j) + \mu_1 P(i+1, j-1) + \mu_2 P(i, j+1) = 0$$

$$\text{for } 1 \le i \le M_1 - 1 \ \ 1 \le j \le M_2 - 1$$

$$-(\lambda + \mu_2)P(i, M_2) + \lambda P(i-1, M_2) + \mu_1 P(i+1, M_2 - 1) = 0 \ \text{ for } 1 \le j \le M_2 - 1$$

$$-(\mu_1 + \mu_2)P(M_1, j) + \lambda P(M_1 - 1, j) + \mu_2 P(M_1, j+1) = 0 \ \text{ for } 1 \le j \le M_2 - 1$$

$$-\mu_2 P(M_1, M_2) + \lambda P(M_1 - 1, M_2) = 0$$

Lucidly, this process conforms to the definition of the QBD. Moreover, one has two choices for the counting set. If the set C_1 is elected as the counting set, the submatrices that define the QBD are of the dimension $M_2 x M_2$. These relevant submatrices are

$$A = \begin{bmatrix} 0 & \mu_1 & & & & \\ & 0 & \mu_1 & & & \\ & & 0 & . & & \\ & & & . & . & \\ & & & & . & \mu_1 \\ & & & & & 0 \end{bmatrix} \quad , \quad C = \begin{bmatrix} \lambda & & & & \\ & \lambda & & & \\ & & . & & \\ & & & . & \\ & & & & \lambda \end{bmatrix}$$

$$B = \begin{bmatrix} -(\lambda + \mu_1) & & & & & \\ \mu_2 & -(\lambda + \mu_1 + \mu_2) & & & & \\ & \mu_2 & -(\lambda + \mu_1 + \mu_2) & & & \\ & & \mu_2 & . & & \\ & & & . & . & \\ & & & & . & -(\lambda + \mu_1 + \mu_2) \\ & & & & \mu_2 & -(\lambda + \mu_2) \end{bmatrix}$$

$$B_0 = \begin{bmatrix} -\lambda & & & & & & \\ \mu_2 & -(\lambda+\mu_2) & & & & & \\ & \mu_2 & -(\lambda+\mu_2) & & & & \\ & & \mu_2 & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & \cdot & -(\lambda+\mu_2) & \\ & & & & & \mu_2 & -(\lambda+\mu_2) \end{bmatrix}$$

$$B_{M_1} = \begin{bmatrix} -\mu_1 & & & & & & \\ \mu_2 & -(\mu_1+\mu_2) & & & & & \\ & \mu_2 & -(\mu_1+\mu_2) & & & & \\ & & \mu_2 & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & \cdot & -(\mu_1+\mu_2) & \\ & & & & & \mu_2 & -(\mu_2) \end{bmatrix}$$

If the set C_2 is elected as the counting set, the submatrices that define the QBD are of the dimension $M_1 \times M_1$. These relevant submatrices are

$$A = \begin{bmatrix} \mu_2 & & & & \\ & \mu_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \mu_2 \end{bmatrix} , \quad C = \begin{bmatrix} 0 & & & & \\ \mu_2 & 0 & & & \\ & \mu_2 & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 0 \\ & & & & \mu_2 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} -(\lambda+\mu_2) & \lambda & & & & \\ & -(\lambda+\mu_1+\mu_2) & \lambda & & & \\ & & -(\lambda+\mu_1+\mu_2) & \cdot & & \\ & & & \cdot & \cdot & \\ & & & \cdot & \lambda & \\ & & & & -(\lambda+\mu_1+\mu_2) & \lambda \\ & & & & & -(\mu_1+\mu_2) \end{bmatrix}$$

$$B_0 = \begin{bmatrix} -\lambda & \lambda & & & & \\ & -(\lambda+\mu_1) & \lambda & & & \\ & & -(\lambda+\mu_1) & \cdot & & \\ & & & \cdot & \cdot & \\ & & & \cdot & \lambda & \\ & & & & -(\lambda+\mu_1) & \lambda \\ & & & & & -\mu_1 \end{bmatrix}$$

$$
B_{M_2} = \begin{bmatrix}
-(\lambda + \mu_2) & \lambda & & & & & \\
& -(\lambda + \mu_2) & \lambda & & & & \\
& & -(\lambda + \mu_2) & . & & & \\
& & & . & . & & \\
& & & & . & \lambda & \\
& & & & & -(\lambda + \mu_2) & \lambda \\
& & & & & & -\mu_2
\end{bmatrix}
$$

The interesting phenomena that one should observe is that the block tridiagonal structure that is the insignia of the QBD is also exhibited in each of the submatrices that has been presented. Indeed, it is possible to show that this phenomenon generalizes for tandem queues having more than two buffers. For each buffer, there would be another level of trigonal structure embedded in the blocks belonging to the previous level.

One can relate this phenomenon to fractals. The same kind of macro-structure appears in the elements of the whole. As one gets to examine these submatrices, the same structure exhibits itself. And the amount that this self-similarness occurs is the number of stations–and thereby the number of buffers–in the tandem queue.

Although no methodology to fully exploit this marvelous structure -- in order to obtain the nullspace of the transition rate matrix—exists at this point, it can still be quite useful in generating the transition rate matrix for a given process in an efficient fashion.

4 Conclusion

In this paper, the suitability of QBD platform in the stochastic modeling of the production lines is illustrated. First, an introduction to QBDs is made. Then, 8 examples of how production lines with different characteristics can be modeled as QBDs, are presented. It is obvious that it would not be difficult to augment the number of these examples by using the same principles. The more complicated the nature of the modeled system is, more states corresponding to each element of the counting space would be needed.

By using the methodology described in the papers by Fadiloglu and Yeralan [7], and Yeralan and Tan [6] one can find the steady-state probabilities for these models irrespective of the counting set's cardinality. In the models, the counting set

always corresponds to a buffer of choice. Yet when more than one buffer exists, one comes up with a structure we can call *Multi-Dimensional Quasi-Birth-Death Process*. This process is still a QBD, thus amenable to the same solution methodologies. Yet, there is some additional structure in these processes as has been illustrated in Model 8. The exploitation of this structure would be an important addition to this line of research.

REFERENCES

[1] Evans, R.V. (1967), "Geometric Distribution in Some Two-dimensional Queuing Systems," *Operations Research*, Vol. 15, pp. 830-846.

[2] Wallace, V. (1969), "The Solution of Quasi Birth and Death Processes Arising from Multiple Access Computer Systems," Ph.D. Thesis, Systems Engineering Laboratory, University of Michigan, Ann Arbor.

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

[4] Yeralan, S. and Muth, E.J. (1987), "A General Model of a Production Line with Intermediate Buffer and Station Breakdown," *IIE Transactions*, Vol. 19, No. 2, pp. 130-139.

[5] Tan, B. (1994), "A Decomposition Method for Multistation Production Systems", Ph.D. Dissertation, Industrial and Systems Engineering, University of Florida, Gainesville.

[6] Yeralan, S. and Tan, B. (1997), "Analysis of Multistation Production Systems with Limited Buffer Capacity Part 1: The Subsystem Model," *Mathematical and Computer Modelling*, Vol. 25, No. 7, pp.109-122.

[7] Fadiloglu, M. M., and Yeralan, S. (1998), "A General Theory on Spectral Properties of State-Homogeneous Quasi-Birth-Death Processes", to appear in *European Journal of Operational Research*

[8] Dallery, Y., and Gershwin, S.B. (1992), "Manufacturing Flow Line Systems: A Review of Models and Analytical Results," *Queuing Systems*, Vol. 12, No. 3, pp. 3-94.