Modelling and analysis of Markovian continuous flow systems with a finite buffer

Citation: Tan, Barıș, and Stanley B. Gershwin. "Modelling and Analysis of Markovian Continuous Flow Systems with a Finite Buffer." Annals of Operations Research 182.1 (2009) : 5-30. Copyright © 2009, Springer Science+Business Media, LLC

As Published: http://dx.doi.org/10.1007/s10479-009-0612-6

Publisher: Springer

Persistent URL: http://hdl.handle.net/1721.1/65420

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Modelling and Analysis of Markovian Continuous Flow Systems with a Finite Buffer

Barış Tan
Graduate School of Business
Koç University
Rumeli Feneri Yolu, Sarıyer
Istanbul, Turkey
btan@ku.edu.tr

Stanley B. Gershwin
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307
USA
gershwin@mit.edu

July 21, 2009

Abstract

In this study, a Markovian fluid flow system with two stages separated by a finite buffer is considered. Fluid flow models have been analyzed extensively to evaluate the performance of production, computer, and telecommunication systems. Recently, we developed a methodology to analyze general Markovian continuous flow systems with a finite buffer. The flexibility of this methodology allows us to analyze a wide range of systems by specifying the transition rates and the flow rates associated with each state of each stage. In this study, in order to demonstrate the applicability of our methodology, we model and analyze a range of models studied in the literature. The examples we analyze as special cases of our general model include systems with phase-type failure and repair-time distributions, systems with machines that have multiple up and down states, and systems with multiple unreliable machines in series or parallel in each stage. For each case, the Markovian model is developed, the transition and flow rates are determined, and representative numerical results are obtained by using our methodology.

1 Introduction

In this study, we consider the modelling and analysis of various two-stage continuous flow systems with a finite capacity buffer. The dynamics of each stage are described by a continuous-time, discrete-state Markov chain where a different flow rate is associated with each state (Figure 1).

By determining the state transition rates for each stage and the flow rates associated with each state, this model can represent a wide range of systems. For example, it may represent a portion of a factory in which a stage represents an unreliable machine that may have phase-type up- and down-time distributions; or a machine with variable quality; or multiple stations in series or in parallel

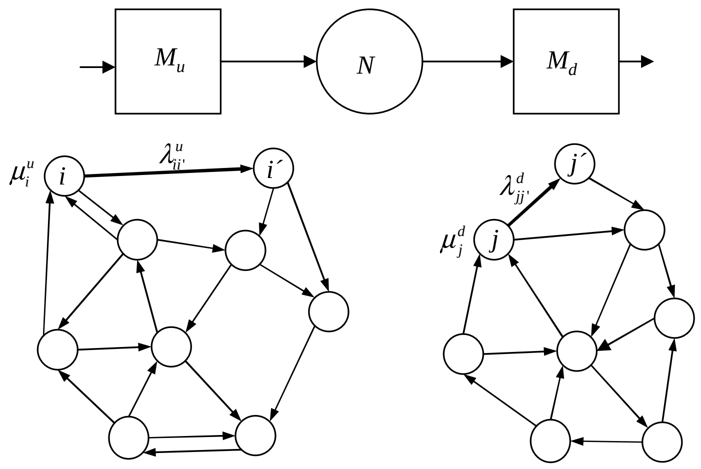

Figure 1: A Single Buffer Fluid Flow System with Two Stages

without intermediate buffers. For another example, it can represent a communications network in which message flow rates change according to Markov processes (e.g. Anick, Mitra, and Sondhi 1982, Elwalid and Mitra 1991, and Mandjes, Mitra, and Scheinhardt 2003). In the following, we use the terms *stage* and *machine* interchangeably.

Recently, Tan and Gershwin (2007, 2009) presented a methodology to analyze a general two-stage Markovian continuous flow system with a finite buffer. This methodology allows the analysis of wide variety of systems by only identifying their transition rates and flow rates associated with each state of each stage. In this study, we model and analyze various systems studied in the literature to demonstrate the applicability of this methodology as a general tool to analyze the performance of continuous flow systems.

In the last four decades, a vast number of papers that analyze single-buffer two-stage Markovian continuous flow systems appeared in the literature starting with Sevast'Yanov (1962). From the modelling perspective, the main difference in these studies is the way the transition and flow rates are identified. For a given system, these studies use an analytical method that is based on the special structure of the system. By using our methodology, all of these models can be analyzed directly as special cases.

The majority of the papers that focus on analysis of continuous flow production systems consider models with two unreliable stations and a finite buffer. In the simplest case, each unreliable machine has only two states: a single up state that represents the condition of a fully productive machine and a single down state that represent the condition where the machine is not productive due to a failure and the failure and repair times are exponential random variables. There is a very large literature on the analysis of this special case (e.g. Wijngaard 1979, Gershwin and Schick 1980, Dubois and Forestier 1982, Yeralan, Franck, and Quasem 1986, Yeralan and Tan 1997, among others).

In the performance evaluation of computer and telecommunication systems, there exist different methodologies to analyze general fluid flow models of computer and telecommunication systems with a finite buffer, (e.g. Serucola 2001, Ahn and Ramaswami 2003, Mandjes, Mitra, and Scheinhardt 2003, Ahn, Jeon, and Ramaswami 2005, Soares and Latouche 2006). Due to the operation-dependent failure mechanism observed in a production setting, the methodologies developed for telecommunication and computer systems cannot be used directly. Similarly, in the production literature, Koster (1989) also presented a framework to analyze general two-stage production systems with time-dependent failures.

In the operation-dependent failure case, an idle machine that is blocked or starved cannot fail. If a machine is partially blocked or partially starved and operating at a reduced rate, its failure rate will be lower than its rate when the buffer is partially full. As a result, the boundary processes when the buffer is empty or full are not the same as the interior process. In order to analyze the operation-dependent failure mechanism, the methodology presented in this paper analyzes the interior process and the boundary processes that are governed by different rate matrices jointly.

The two-stage single buffer system is often used as a building block in the decomposition methods that are used to evaluate the performance of multi-station production systems. In order to improve the accuracy of the decomposition method, the basic two-stage model has been extended to approximate non-exponential repair time distributions. For example, Dallery and Bihan (1999) use a model with exponential failure time and generalized exponential repair time. Bihan and Dallery (2000) use a model with exponential failure time and two-stage hyper-exponential repair time. Levantesi, Matta, and Tolio (2003) considers a model with multiple failure modes that is equivalent to a system with exponential failure time and hyper-exponential repair time. Özdoğru and Altıok (2003) analyze a system with exponential failure time and two-stage Coxian repair time distribution.

All the studies discussed above evaluate the performance of a production system in terms of *quantity* of output produced. Recently, new models have been developed to investigate not only quantity but also *quality* of output produced. These models allow examining quality and quantity issues jointly in the design and operation of production systems (Kim and Gershwin 2005, Poffe and Gershwin 2005). In the quality-quantity models, there are multiple up states associated with different quality production. Accordingly, there are different down states associated with each up state as well as common down states associated with system failures and maintenance. For example, Poffe and Gershwin (2005) consider a continuous flow production system where the first stage has two up and three down states and the second stage has one up and one down state. The flexibility of our model allows analysis of more elaborate quality-quantity models with more number of states describing the behavior of each stage.

Models with multiple unreliable machines in series or parallel in each stage and separated by a finite buffer also received some attention in the literature. Forestier (1980) described the model and Mitra (1988) provided a general approach to analyze two-stage continuous flow systems with identical parallel stations in each stage. Tan (2001), Helber and Jusic (2004) and Diamantidis, Papadopoulos, and Vidalis (2004) consider merge structures where the upstream stage has two unreliable machines in parallel and the downstream station has one unreliable machine.

In this study, we demonstrate how our methodology can be used to analyze the models summarized above as special cases by specifying the transition rates for each stage and the flow rates associated with each state. The flexibility of our model shows that it can be used as a general tool to analyze Markovian fluid flow systems with a finite buffer. Therefore, the main contribution of this method is to allow researchers to focus on developing models that describe the behavior of production systems by removing the burden of analysis.

The organization of the remainder part of the paper is as follows: In Section 2, we summarize the methodology to analyze general Markovian two-stage continuous flow systems with a finite buffer. We give an example in Section 3 to show how this methodology can be used to analyze a given system. In Section 4, we discuss the analysis of different classes of models by using the methodology, including: models with unreliable stations and phase-type failure and repair time distributions; models to analyze quality-quantity interactions; and models with series, parallel, and merge structures. Finally, conclusions are given in Section 5.

2 Analysis of the General Model

In this section, we summarize the methodology we developed to analyze general Markovian continuous flow systems with a finite buffer and state the equations that yield the steady-state probabilities and the desired performance measures. The complete presentation of the methodology with the derivation of all the results are given in Tan and Gershwin (2007) and Tan and Gershwin (2009).

2.1 Model Description

We consider a continuous flow system with two stages separated by a buffer with capacity N (Figure 1). The state of the system at time t is $s(t) = (X, \alpha_u, \alpha_d)$ where $0 \leq X \leq N$ is the buffer level, $\alpha_u \in \{1, ..., I_u\}$ is the state of the upstream stage M_u and $\alpha_d \in \{1, ..., I_d\}$ is the state of the downstream stage M_d. There are $I_u I_d$ discrete states in the machine state space S_M, $(\alpha_u, \alpha_d) \in S_M$.

The maximum processing rate of M_u in state i is $\mu_i^u \geq 0$ and the maximum processing rate of M_d in state j is $\mu_j^d \geq 0$. The vectors $\mathbf{m}^u = \{\mu_i^u \mid 1 \leq i \leq I_u\}$ and $\mathbf{m}^d = \{\mu_j^d \mid 1 \leq j \leq I_d\}$ contain these processing rates for M_u and M_d respectively.

The machines operate at their maximum rates unless they are starved or blocked. When the buffer is empty in the machine state $(\alpha_u, \alpha_d) = (i, j)$ with $\mu_i^u = 0$ and $\mu_j^d > 0$ then M_d is said to be completely starved and it is forced to stop. However, when the buffer is empty and $\mu_j^d > \mu_i^u > 0$, M_d is said to be partially starved and it can continue its production at a reduced rate of μ_i^u. When the buffer is full in machine state $(\alpha_u, \alpha_d) = (i, j)$ with $\mu_i^u > 0$ and $\mu_j^d = 0$ then M_u is said to be completely blocked and the flow into the buffer is stopped. However, in the same state if $\mu_i^u > \mu_j^d > 0$, M_u is said to be partially blocked and it can continue its production at a reduced rate of μ_j^d. We assume that M_u is never starved and M_d is never blocked.

We partition the discrete states of the system into three sets depending on how the buffer level changes when $0 < X < N$: Υ in which the buffer level goes up ($(i, j) \in \Upsilon$ if $\mu_i^u > \mu_j^d$); Δ in which it goes down ($(i, j) \in \Delta$ if $\mu_i^u < \mu_j^d$); or Z in which it stays the same ($(i, j) \in Z$ if $\mu_i^u = \mu_j^d$) in that state. The number of states in each of these sets are $I_\Upsilon = |\Upsilon|$, $I_\Delta = |\Delta|$, and $I_Z = |Z|$ respectively and $I_\Upsilon + I_\Delta + I_Z = I_u I_d$.

For M_u, when $0 < X < N$, the transition time from state i to state i' is an exponential random variable with rate $\lambda_{ii'}^u$. Similarly for M_d, the transition time from state j to state j' is an exponential random variable with rate $\lambda^d_{jj'}$. When M_u is partially blocked, the transition time from state i to state i' is also an exponential random variable with rate $\psi^u_{ii'}$. Similarly, when M_d is partially starved, the transition rate from state j to state j' is $\psi^d_{jj'}$. We do not make any assumptions regarding the transition rates $\psi^u_{ii'}$ and $\psi^d_{jj'}$. With this general setting, operation dependent failure mechanisms can be modelled easily as shown in Section 4.

The time-dependent probability density while the buffer is neither full nor empty is

$$f(x,i,j,t) = \frac{\partial}{\partial x}\text{prob}[X(t) \le x, \alpha_u(t) = i, \alpha_d(t) = j] \text{ for } 0 < x < N.$$

We assume that the process is ergodic and the steady-state probabilities exist. The steady-state density functions are defined as $f(x,i,j) = \lim_{t\to\infty} f(x,i,j,t)$ for $0 < x < N$. The probability of state $(0,i,j)$ at time t when the buffer is empty is denoted by $p(0,i,j,t)$ and the probability of state (N,i,j) at time t when the buffer is full is denoted by $p(N,i,j,t)$. The steady-state probabilities at the empty and full buffer states when $(\alpha_u,\alpha_d) = (i,j)$ are $p(0,i,j) = \lim_{t\to\infty} p(0,i,j,t)$ and $p(N,i,j) = \lim_{t\to\infty} p(N,i,j,t)$ respectively.

2.2 Analysis of Interior and Boundary Processes

In the rest of this paper, we only consider steady-state behavior, and we suppress the t argument. Our solution methodology requires only matrices $\lambda^u = \{\lambda^u_{ii'}\}$, $\lambda^d = \{\lambda^d_{jj'}\}$, $\psi^u = \{\psi^u_{ii'}\}$, $\psi^d = \{\psi^d_{jj'}\}$, vectors $\mu^u = \{\mu^u_i\}$, $\mu^d = \{\mu^d_j\}$, and the buffer size N as its inputs. In the operation dependent failure case, ψ^u and ψ^d are functions of λ^u and λ^d, and the flow rate vectors $\mathbf{m}^u$ and $\mathbf{m}^d$. As a result, once the transition rate matrices λ^u and λ^d, and the flow rate vectors $\mathbf{m}^u$ and $\mathbf{m}^d$ are given, the methodology summarized below yields the desired performance measures.

We first determine the differential equations that describe the dynamics of the system when the buffer is in the interior $(0 < X < N)$ and when the buffer is at the boundary, i.e. when the buffer is empty $(X = 0)$ or full $(X = N)$.

2.2.1 Interior Process

The probability density functions satisfy the following set of equations in steady state:

$$\begin{aligned}(\mu^u_i - \mu^d_j)\frac{\partial f(x,i,j)}{\partial x} &= -f(x,i,j)\left(\sum_{\substack{i'=1\\ i'\neq i}}^{I_u} \lambda^u_{ii'} + \sum_{\substack{j'=1\\ j'\neq j}}^{I_d} \lambda^d_{jj'}\right) \\ &+ \sum_{\substack{i'=1\\ i'\neq i}}^{I_u} f(x,i',j)\lambda^u_{i'i} + \sum_{\substack{j'=1\\ j'\neq j}}^{I_d} f(x,i,j')\lambda^d_{j'j},\ (i,j)\in S_M. \qquad (1)\end{aligned}$$

The above equation is written in the matrix form as

$$\left[\begin{array}{c} \left[\begin{array}{c} \frac{\partial \mathbf{f}_{\Upsilon}(x)}{\partial x} \\ \frac{\partial \mathbf{f}_{\Delta}(x)}{\partial x} \end{array}\right] \\ \mathbf{0} \end{array}\right] = \left[\begin{array}{cc} A_1 & A_2 \\ A_3 & A_4 \end{array}\right] \left[\begin{array}{c} \left[\begin{array}{c} \mathbf{f}_{\Upsilon}(x) \\ \mathbf{f}_{\Delta}(x) \end{array}\right] \\ \mathbf{f}_{Z}(x) \end{array}\right] \tag{2}$$

where $\mathbf{f}_S(x) = \{f(x,i,j)\}$ for $(i,j) \in S$, $S = \Upsilon, \Delta, Z$, A_1 is a square matrix of size $(I_\Upsilon + I_\Delta) \times (I_\Upsilon + I_\Delta)$, A_4 is a square matrix of size $I_Z \times I_Z$, A_2 is a matrix of size $(I_\Upsilon + I_\Delta) \times I_Z$, A_3 is a matrix of size $I_Z \times (I_\Upsilon + I_\Delta)$, and $\mathbf{0}$ is a column vector of zeroes of length I_Z. These matrices are determined by λ^u, λ^d, $\mathbf{m}^u$, and $\mathbf{m}^d$.

The solution of the set of differential and algebraic equations given in Equation 2 is

$$\left[\begin{array}{c} \mathbf{f}_{\Upsilon}(x) \\ \mathbf{f}_{\Delta}(x) \end{array}\right] = e^{\Lambda x}\mathbf{w} \tag{3}$$

and

$$\mathbf{f}_Z(x) = \Omega e^{\Lambda x}\mathbf{w}. \tag{4}$$

where $\Lambda = A_1 - A_2 A_4^{-1} A_3$, $e^{\Lambda x}$ is a matrix exponential determined by matrix Λ, $\Omega = -A_4^{-1} A_3$ and $\mathbf{w}$ is a column vector of length $I_\Upsilon + I_\Delta$.

When vector $\mathbf{w}$ is determined, all the density functions are determined by Equations (3) and (4). Since the length of $\mathbf{w}$ is $I_\Upsilon + I_\Delta$, $I_\Upsilon + I_\Delta$ equations are needed to determine the weights uniquely. We determine these equations by analyzing the boundary processes in the following.

2.2.2 Empty Buffer Process

As the buffer level decreases in states $(i,j) \in \Delta$, the buffer eventually becomes empty if no other transition occurs first. Once the buffer becomes empty, it stays empty until the system makes a transition to a state $(i,j) \in \Upsilon$. When the buffer is empty, the set of states where the buffer stays empty is $S_0 = \Delta \cup Z$ and $I_{S_0} = |S_0| = I_\Delta + I_Z$.

Let t_k^0 be the kth time the buffer becomes empty and $\pi(0,i,j,t_k^0 + \tau)$ be the probability that $X = 0$ and $(\alpha_u, \alpha_d) = (i,j)$ at time $t_k^0 + \tau$ given that the buffer became empty at time t_k^0 and has been empty during $[t_k^0, t_k^0 + \tau]$. The dynamics of the system during an interval when the buffer stays empty are given by the following equations:

$$\begin{aligned} \frac{d\pi(0,i,j,\tau)}{d\tau} &= -\pi(0,i,j,\tau)\left(\sum_{\substack{i'=1 \\ i' \neq i}}^{I_u} \lambda^u_{ii'} + \sum_{\substack{j'=1 \\ j' \neq j}}^{I_d} \psi^d_{jj'}\right) \\ &+ \sum_{\substack{i'=1 \\ i' \neq i \\ (i',j) \in S_0}}^{I_u} \pi(0,i',j,\tau)\lambda^u_{i'i} + \sum_{\substack{j'=1 \\ j' \neq j \\ (i,j') \in S_0}}^{I_d} \pi(0,i,j',\tau)\psi^d_{j'j}, \quad (i,j) \in S_0. \end{aligned} \tag{5}$$

Equation (5) can be written in matrix form as

$$\frac{d\pi^0_{S_0}(\tau)}{d\tau} = A_0 \pi^0_{S_0}(\tau) \tag{6}$$

where $\pi^0_{S_0}(\tau) = \{\pi(0,i,j,\tau)\}$ for $(i,j) \in S_0$ and A_0 is a $I_{S_0} \times I_{S_0}$ square matrix.

The empty buffer process ends with a transition into a state where the buffer level starts increasing. The rate at which the process enters into the state $(i,j) \in \Upsilon$ after a time period of length τ following the buffer becoming empty is

$$q(0,i,j,\tau) = \sum_{\substack{i'=1 \\ i' \neq i \\ (i',j) \in S_0}}^{I_u} \pi(0,i',j,\tau)\lambda^u_{i'i} + \sum_{\substack{j'=1 \\ j' \neq j \\ (i,j') \in S_0}}^{I_d} \pi(0,i,j',\tau)\psi^d_{j'j}, \quad (i,j) \in \Upsilon. \tag{7}$$

Equation (7) can be written in matrix form as

$$\mathbf{q}^0_{\Upsilon}(\tau) = B_0 \pi^0_{S_0}(\tau) \tag{8}$$

where $\mathbf{q}^0_{\Upsilon}(\tau) = \{q(0,i,j,\tau)\}$ for $(i,j) \in \Upsilon$ and B_0 is a $I_{\Upsilon} \times I_{S_0}$ matrix.

We use the relationship between the probability that the buffer becomes empty while the machines are in a particular state $(i,j) \in \Delta$ and the probability that the process exits the empty buffer state with a transition into state $(i,j) \in \Upsilon$ to derive a set of boundary equations. We express the probability that the buffer becomes empty and the process exits the empty buffer state as ratios of the number of level crossings in the corresponding states to the total number of level crossings. Only the matrices A_0 and B_0 are required to express the resulting boundary equation:

$$\begin{bmatrix} \text{diag}(\mathbf{m}_{\Upsilon}) & 0_{I_{\Upsilon} \times I_{\Delta}} \end{bmatrix} \mathbf{w} = G_0 \begin{bmatrix} 0_{I_{\Delta} \times I_{\Upsilon}} & \text{diag}(\mathbf{m}_{\Delta}) \end{bmatrix} \mathbf{w} \tag{9}$$

where G_0 is a $I_{\Upsilon} \times I_{\Delta}$ matrix that is obtained by eliminating the columns of $-B_0 A_0^{-1}$ corresponding to states in Z, $\mathbf{m}_{\Upsilon} = \{(\mu^u_i - \mu^d_j)| \ (i,j) \in \Upsilon\}$, $\mathbf{m}_{\Delta} = \{(\mu^d_j - \mu^u_i)| \ (i,j) \in \Delta\}$, and the notation diag($\mathbf{a}$) represents a diagonal matrix formed with the elements of vector $\mathbf{a}$ and $0_{k \times l}$ is a $k \times l$ matrix of zeros.

In Equation (9), the left-hand side is the probability vector that the process exits the empty-buffer process in a state $(i,j) \in \Upsilon$. The matrix G_0 is the matrix of conditional probabilities that the empty buffer process exits in a particular state $(i,j) \in \Upsilon$ given that it starts in one of the states $(i',j') \in S_0$ where the buffer stays empty. G_0 is obtained by eliminating the columns of $-B_0 A_0^{-1}$ corresponding to states in Z. The term that is multiplied by G_0 is the probability vector that the buffer becomes empty while the machine is in a state $(i,j) \in \Delta$. Since the probability that the process exits the empty buffer process in one of the states in Υ is one, Equation (9) gives $I_{\Upsilon} - 1$ linearly independent equations that will be used to determine $\mathbf{w}$. Once $\mathbf{w}$ is determined, the steady state distribution of the states when the buffer is empty can be determined directly.

Due to the ergodicity of the process, the probability that $X = 0$ and $(\alpha_u, \alpha_d) = (i,j)$ is also the fraction of the total time the process stays in this state in a given time period the long run.

We can determine the total time the process stays in state $(i,j) \in S_0$ while $X = 0$ in a given time period by determining the number of times the buffer becomes empty and the time the process stays in this state for each time the buffer becomes empty in the same time period. Given that the machine states $(\alpha_u, \alpha_d) = (i', j') \in \Delta$ at the time the buffer becomes empty, the expected time that the machine states (α_u, α_d) stay in $(i,j) \in S_0$ before exiting to a state $(\alpha_u, \alpha_d) \in \Upsilon$ is denoted by $E[T^0_{(i,j),(i',j')}]$.

Using the solution of the density functions given in Equation (3), the steady-state empty buffer probability distribution can be written as

$$\mathbf{p}_0 = E[T^0] \left[\begin{array}{cc} 0_{I_\Delta \times I_\Upsilon} & \text{diag}(\mathbf{m}_\Delta) \end{array} \right] \mathbf{w} \tag{10}$$

where $\mathbf{p}_0 = \{p(0,i,j)\}$ and $E[T^0]$ is an $I_{S_0} \times I_\Delta$ matrix that is obtained by eliminating the columns of $-A_0^{-1}$ corresponding to states in Z. In Equation (10), the term multiplied with $E[T^0]$ is the number of times the buffer becomes empty per unit time which is equal to the number of upward level crossings at $x = 0^-$ per unit time expressed in terms of the interior process densities and flow rates.

2.2.3 Full Buffer Process

The dynamics of the system when the buffer stays full in state $(i,j) \in S_N = \Upsilon \cup Z$,the set of states where the buffer stays full, are given below:

$$\begin{aligned} \frac{d\pi(N,i,j,\tau)}{d\tau} &= -\pi(N,i,j,\tau) \left(\sum_{\substack{j'=1 \\ j' \neq j}}^{I_d} \lambda^d_{jj'} + \sum_{\substack{i'=1 \\ i' \neq i}}^{I_u} \psi^u_{ii'} \right) \\ &+ \sum_{\substack{j'=1 \\ j' \neq j \\ (i,j') \in S_N}}^{I_d} \pi(N,i,j',\tau)\lambda^d_{j'j} + \sum_{\substack{i'=1 \\ i' \neq i \\ (i',j) \in S_N}}^{I_u} \pi(N,i',j,\tau)\psi^u_{i'i}, \quad (i,j) \in S_N. \end{aligned} \tag{11}$$

The above equation can be written in matrix form as

$$\frac{\pi^N_{S_N}(\tau)}{d\tau} = A_N \pi^N_{S_N}(\tau) \tag{12}$$

where $\pi(N,i,j,t^N_k + \tau)$ is the probability that $X = N$ and $(\alpha_u, \alpha_d) = (i,j)$ at time $t^N_k + \tau$ given that the buffer became full at time t^N_k and has been full during $[t^N_k, t^N_k + \tau]$ and t^N_k is the kth time the buffer becomes full, $\pi^N_{S_N}(\tau) = \{\pi(N,i,j,\tau)\}$ for $(i,j) \in S_N$, $I_{S_N} = |S_N|$, and A_N is a $I_{S_N} \times I_{S_N}$, square matrix.

The full buffer process ends with a transition into a state where the buffer level starts decreasing. The rate vector at which the process enters into one of the states $(i,j) \in \Delta$ is

$$q(N,i,j,\tau) = \sum_{\substack{j'=1 \\ j' \neq j \\ (i,j') \in S_N}}^{I_d} \pi(N,i,j',\tau)\lambda^d_{j'j} + \sum_{\substack{i'=1 \\ i' \neq i \\ (i',j) \in S_N}}^{I_u} \pi(N,i',j,\tau)\psi^u_{i'i}, \quad (i,j) \in \Delta \tag{13}$$

or in matrix form

$$\mathbf{q}^N_\Delta(\tau) = B_N \pi^N_{S_N}(\tau) \tag{14}$$

where $\mathbf{q}^N_\Delta(\tau) = \{q(N,i,j,\tau)\}$ for $(i,j) \in \Delta$ and B_N is a $I_\Delta \times I_{S_N}$ matrix.

Similar to the empty-buffer case, only the matrices A_N and B_N are required to write the second boundary equation that relates the entry and exit probabilities when the buffer is full:

$$\left[\begin{array}{cc} 0_{I_\Delta \times I_\Upsilon} & \text{diag}(\mathbf{m}_\Delta) \end{array}\right] e^{\Lambda N}\mathbf{w} = G_N \left[\begin{array}{cc} \text{diag}(\mathbf{m}_\Upsilon) & 0_{I_\Upsilon \times I_\Delta} \end{array}\right] e^{\Lambda N}\mathbf{w} \tag{15}$$

where G_N is a $I_\Delta \times I_\Upsilon$ matrix that is obtained by eliminating the columns of $-B_N A_N^{-1}$ corresponding to states $S_N \setminus \Upsilon$.

Since the probability that the process exits the full buffer state with a transition into one of the states in Δ is one, Equation (15) gives $I_\Delta - 1$ linearly independent equations that will be used to determine $\mathbf{w}$.

The full-buffer steady-state distribution is expressed in terms of the solution of the interior process, the flow rate, and the full-buffer process dynamics:

$$\mathbf{p}_N = E[T^N] \left[\begin{array}{cc} \text{diag}(\mathbf{m}_\Upsilon) & 0_{I_\Upsilon \times I_\Delta} \end{array}\right] e^{\Lambda N}\mathbf{w}. \tag{16}$$

where $\mathbf{p}_N = \{p(N,i,j)\}$ and $E[T^N]$ is an $I_{S_N} \times I_\Upsilon$ matrix that is obtained by eliminating the columns of $-A_N^{-1}$ corresponding to states in $S_N \setminus \Upsilon$. In Equation (16), the term multiplied with $E[T^N]$ is the expected number of times the buffer becomes full while the machine state is in a state in Υ per unit time in the long run.

2.2.4 Determination of the Probability Densities

Once the weight vector $\mathbf{w}$ is determined, all the steady-state probabilities are also determined. Since there are $I_\Upsilon + I_\Delta$ weights and Equations (9) and (15) give a total of $I_\Upsilon + I_\Delta - 2$ equations, two additional equations are required to uniquely determine $\mathbf{w}$.

The first equation is the equivalence of the total upward and downward crossings in the interior region:

$$\left[\begin{array}{cc} \mathbf{m}_\Upsilon & -\mathbf{m}_\Delta \end{array}\right] \left(\int_0^N e^{\Lambda x} dx\right) \mathbf{w} = 0. \tag{17}$$

The second equation is the normalization equation:

$$\sum_{i=1}^{I_u}\sum_{j=1}^{I_d} \left(p(0,i,j) + p(N,i,j)\right) + \int_0^N \sum_{i=1}^{I_u}\sum_{j=1}^{I_d} f(x,i,j)dx = 1 \tag{18}$$

or in matrix form

$$\left(u_{I_{S_0}} E[T^0] \left[\begin{array}{cc} 0_{I_\Delta \times I_\Upsilon} & \mathrm{diag}(\mathbf{m}_\Delta) \end{array} \right] + \nu \left(\int_0^N e^{\Lambda x} dx \right) + u_{I_{S_N}} E[T^N] \left[\begin{array}{cc} \mathrm{diag}(\mathbf{m}_\Upsilon) & 0_{I_\Upsilon \times I_\Delta} \end{array} \right] e^{\Lambda N} \right) \mathbf{w} = 1 \tag{19}$$

where $\nu = (u_{I_\Upsilon + I_\Delta} + u_{I_Z}\Omega)$.

Now Equations (9) and (15) with Equations (17) and (19) give $I_\Upsilon + I_\Delta$ linearly independent equations that uniquely determine $\mathbf{w}$. Therefore all the steady-state probability distributions that describe the dynamics of the system are determined by these equations.

2.2.5 Performance Measures

When the probability densities are determined, all performance measures of interest can be calculated. In a production setting, the main performance measures of interest are the production rate and the expected buffer level.

The production rate is the amount of material processed per unit time in the long run. The production rate of the first stage can be written as

$$\Pi = \sum_{(i,j)\in S_0} \mu_i^u p(0,i,j) + \sum_{(i,j)\in S_M} \int_0^N \mu_i^u f(x,i,j)dx + \sum_{(i,j)\in S_N} \mu_j^d p(N,i,j). \tag{20}$$

The last term in the above equation reflects the reduced processing rate of the first stage due to blocking. Note that the amount of material processed by both stages is the same in the long run.

The expected buffer level is determined as

$$E[X] = \sum_{i=1}^{I_u} \sum_{j=1}^{I_d} \left(\int_0^N x f(x,i,j)dx + Np(N,i,j) \right). \tag{21}$$

Once the steady-state distribution is determined, other performance measures of interest such as the blocking and starvation probabilities can also be evaluated directly.

3 Analysis of an Example

In this section, we present the detailed analysis of a two-station continuous flow system with exponential failure and repair times to illustrate the methodology. All the variables defined in Section 2 are given explicitly for this model. Tan and Gershwin (2007) give explicit analysis of another example where each stage has multiple up and down states.

The upstream machine is unreliable and has one up (State 1) and one down state (State 0). The processing rate of the upstream machine is μ^u and the failure and repair times are exponential random variables with rates p^u and r^u respectively. Similarly, the downstream machine is also unreliable and has one up (State $1'$) and one down state (State $0'$).The processing rate of the downstream machine is μ^d and the failure and repair rates are also exponential random variables with rates p^d and r^d respectively.

We assume that we have operation-dependent failure in this model. Operation-dependent failure can be modelled as a case where there is a reduction in the transition rates at the boundaries which is proportional to the reduction in the processing rate. This is the assumption that is made in many papers in the literature (e.g. Gershwin and Schick 1980). That is, when the buffer is empty and M_d is producing at a reduced rate of μ_i^u, we have $\psi_{jj'}^d = \frac{\mu_i^u}{\mu_j^d}\lambda_{jj'}^d$ for failure transitions. This implies that when $\mu_i^u = 0$ and $x = 0$, we have $\psi_{jj'}^d = 0$ and therefore it is not possible for M_d to make a failure transition when it is completely starved.

Similarly, when the buffer is full and M_u is producing at a reduced rate of μ_j^d, we have $\psi_{ii'}^u = \frac{\mu_j^d}{\mu_i^u}\lambda_{ii'}^d$ for failure transitions. When $\mu_j^d = 0$ and $x = N$, we have $\psi_{ii'}^u = 0$ and therefore M_u cannot have a failure transition when it is completely blocked.

3.1 Model Inputs

Our solution methodology requires only matrices $\lambda^u = \{\lambda_{ii'}^u\}$, $\lambda^d = \{\lambda_{jj'}^d\}$, $\psi^u = \{\psi_{ii'}^u\}$, $\psi^d = \{\psi_{jj'}^d\}$, vectors $\mu^u = \{\mu_i^u\}$, $\mu^d = \{\mu_j^d\}$, and the buffer size N as its inputs. In this specific example, since $\psi_{ii'}^u = \frac{\mu_j^d}{\mu_i^u}\lambda_{ii'}^d$ and $\psi_{jj'}^d = \frac{\mu_i^u}{\mu_j^d}\lambda_{jj'}^d$, ψ^u and ψ^d are defined by the other inputs.

We first order the states of M_u as $\{1, 0\}$. The transition rate matrix of M_u is given as

$$\lambda^u = \begin{bmatrix} -p_u & p_u \\ r_u & -r_u \end{bmatrix}. \tag{22}$$

The processing rates in states $\{1, 0\}$ are

$$\mu^u = \begin{bmatrix} \mu_u & 0 \end{bmatrix}.$$

Similarly, the states of M_d are ordered as $\{1', 0'\}$. The transition rate matrix of M_d is given as

$$\lambda^d = \begin{bmatrix} -p_d & p_d \\ r_d & -r_d \end{bmatrix}. \tag{23}$$

In states $\{1', 0'\}$ the processing rates of M_d are given as

$$\mu^d = \begin{bmatrix} \mu_d & 0 \end{bmatrix}.$$

3.2 Analysis of the Model

Once these inputs are given, we can specify matrices A_1, A_2, A_3, A_4, A_0, B_0, A_N, B_N and vectors $\mathbf{m}_\Upsilon$, $\mathbf{m}_\Delta$, and $\mathbf{m}_Z$ directly. Once these matrices and vectors are specified, the methodology outlined in the preceding sections yields the desired performance measures directly.

The table given in (24) lists the states, the corresponding processing rates, and the classification of each state in sets Υ, Δ, and Z depending on μ_u and μ_d. In this section only the case $\mu_u > \mu_d$ is discussed in detail.

State	State					S	
M_u	M_d	α_u	α_d	$\mathbf{m}_S$	$\mu_u > \mu_d$	$\mu_u = \mu_d$	$\mu_u < \mu_d$
1	$1'$	1	1	$\mu_u - \mu_d$	Υ	Z	Δ
1	0	2	1	μ_u	Υ	Υ	Υ
0	$1'$	1	2	$-\mu_d$	Δ	Δ	Δ
0	$0'$	2	2	0	Z	Z	Z

(24)

There are 4 discrete states in the state space. When $\mu_u > \mu_d$, $I_\Upsilon = 2$, $I_\Delta = 1$, and $I_Z = 1$. In this case,

$$\mathbf{m}_\Upsilon = \left[\begin{array}{cc} \mu_u - \mu_d & \mu_u \end{array} \right],$$

$$\mathbf{m}_\Delta = \left[\begin{array}{c} \mu_d \end{array} \right],$$

$$\mathbf{m}_Z = \left[\begin{array}{c} 0 \end{array} \right].$$

For this specific case, the submatrices A_1, A_2, A_3, and A_4 are

$$A_1 = \left[\begin{array}{ccc} \frac{-p_u-p_d}{\mu_u-\mu_d} & \frac{r_d}{\mu_u-\mu_d} & \frac{r_u}{\mu_u-\mu_d} \\ \frac{p_d}{\mu_u} & \frac{-p_u-r_d}{\mu_u} & 0 \\ -\frac{p_u}{\mu_d} & 0 & \frac{r_u+p_d}{\mu_d} \end{array} \right], \tag{25}$$

$$A_2 = \left[\begin{array}{ccc} 0 & \frac{r_u}{\mu_u} & -\frac{r_d}{\mu_d} \end{array} \right]^{\mathrm{T}}, \tag{26}$$

$$A_3 = \left[\begin{array}{ccc} 0 & p_u & p_d \end{array} \right], \tag{27}$$

$$A_4 = \left[\begin{array}{c} -r_u - r_d \end{array} \right]. \tag{28}$$

The submatrices for the cases $\mu_u = \mu_d$ and $\mu_u < \mu_d$ can be written similarly.

When $\mu_u \neq \mu_d$, the buffer level does not change when both machines are down. Since these states cannot be reached when the buffer is empty or full, $S_0 = \Delta$ and $S_N = \Upsilon$. Therefore $I_{S_0} = I_\Delta = 1$ and $I_{S_N} = I_\Upsilon = 2$.

For the empty buffer process, since M_d is completely starved in all transient states, the matrices A_0 and B_0 for the empty buffer process are

$$A_0 = \left[\begin{array}{c} -r_u \end{array} \right] \tag{29}$$

and

$$B_0 = \left[\begin{array}{c} r_u \\ 0 \end{array} \right]. \tag{30}$$

Since $S_0 = \Delta$, $E[T_0] = -A_0^{-1} = \frac{1}{r_u}$ and $G_0 = -B_0 A_0^{-1} = [-1, 0]^{\mathrm{T}}$.

For the full buffer process, M_u is partially blocked in state $(1,1')$ and completely blocked in state $(1,0')$. Then the matrices A_N and B_N are

$$A_N = \begin{bmatrix} -p_u\frac{\mu_d}{\mu_u} - p_d & p_d \\ r_d & -p_d \end{bmatrix} \tag{31}$$

$$B_N = \begin{bmatrix} p_u\frac{\mu_d}{\mu_u} & 0 \end{bmatrix}. \tag{32}$$

Since $S_N = \Upsilon$, $E[T_N] = -A_N^{-1}$, $G_N = -B_N A_N^{-1}$.

4 Modelling of Various Systems

In this section, we will model different systems to illustrate the application of our methodology in the analysis of different production systems. We first discuss models with series or parallel stations at each stage. This case also includes merge structures analyzed in the literature. We then discuss unreliable stations with phase-type failure and repair time distributions. Next, we model systems with exponential up time and hyper-exponential down time, with Erlang-type up and down times, and exponential up and phase-type down time distributions. Finally, we discuss models to analyze quality-quantity interactions. In all the examples, we assume operation dependent failures.

In each example, we show how these systems are modelled and we specify the inputs of the methodology. Our solution methodology requires only matrices $\lambda^u = \{\lambda^u_{ii'}\}$, $\lambda^d = \{\lambda^d_{jj'}\}$, $\psi^u = \{\psi^u_{ii'}\}$, $\psi^d = \{\psi^d_{jj'}\}$, vectors $\mu^u = \{\mu^u_i\}$, $\mu^d = \{\mu^d_j\}$, and the buffer size N as its inputs. Consequently, in the operation dependent failure case, ψ^u and ψ^d are defined by λ^u and λ^d, and the flow rate vectors $\mathbf{m}^u$ and $\mathbf{m}^d$. As a result, in the examples, we specify the transition rate matrices λ^u and λ^d, and the flow rate vectors $\mathbf{m}^u$ and $\mathbf{m}^d$ for each case and use the methodology to analyze the performance of each model.

4.1 Multiple Stations in Each Stage

4.1.1 A Model with Parallel Stations

We now model a system where M_u has m_u and M_d has m_d identical stations in parallel similar to the one described in Forestier (1980) and analyzed for the time-dependent failure case in Mitra (1988). Each station is unreliable and has one up and one down state. In the upstream stage, the processing rate of each station is μ^u and the failure and repair times are exponential random variables with rates p^u and r^u respectively. In the downstream stage, the processing rate of each station is μ^d and the failure and repair rates are also exponential random variables with rates p^d and r^d respectively.

In this model M_u has $m_u + 1$ and M_d has $m_d + 1$ states. In state i of M_u, i stations are operational and the effective processing rate is $i\mu^u$, $0 \le i \le m_u$. Similarly, in state j of M_d, j stations are operational and the effective processing rate is $j\mu^d$, $0 \le j \le m_d$.

Accordingly, the possible transitions for M_u are

- from state i to state $i-1$ with rate ip^u for $i = 1, ..., m_u$, and
- from state i to state $i+1$ with rate $(m_u - i)r^u$ for $i = 0, ..., m_u - 1$.

Similarly, possible transitions for M_d are

- from state j to state $j-1$ with rate jp^d for $j = 1, ..., m_d$ and
- from state j to state $j+1$ with rate $(m_d - j)r^d$ for $j = 0, ..., m_d - 1$.

Figure 2 depicts the state transitions for M_u and M_d for a specific case where M_u has $m_u = 3$ stations and M_d has $m_d = 2$ stations in parallel.

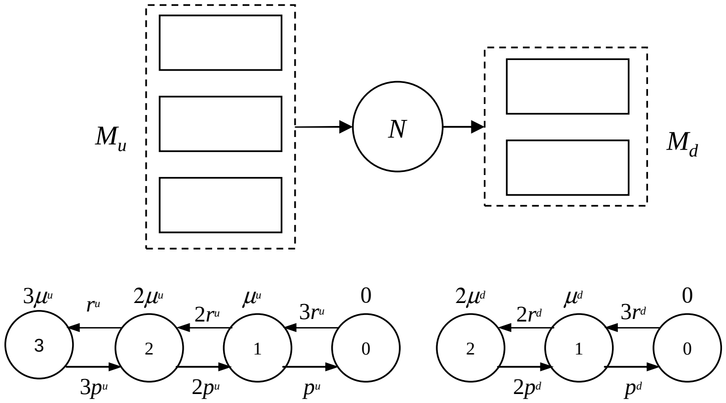

Figure 2: A system with parallel stations

The matrices λ^u and λ^d and the vectors $\mathbf{m}^u$ and $\mathbf{m}^d$ for this specific case are given below:

$$\lambda^u = \begin{bmatrix} -3r^u & 3r^u & 0 & 0 \\ p^u & -p^u - 2r^u & 2r^u & 0 \\ 0 & 2p^u & -2p^u - r^u & r^u \\ 0 & 0 & 3p^u & -3p^u \end{bmatrix} \tag{33}$$

where the states are ordered as $\{0, 1, 2, 3\}$. The processing rates in these states are

$$\mathbf{m}^u = \begin{bmatrix} 0 & \mu^u & 2\mu^u & 3\mu^u \end{bmatrix}.$$

Similarly,

$$\lambda^d = \begin{bmatrix} -2r^d & 2r^d & 0 \\ p^d & -p^d - r^d & r^d \\ 0 & 2p^d & -2p^d \end{bmatrix} \tag{34}$$

where the states are ordered as $\{0, 1, 2\}$. In these states the processing rates of M_d are given as

5
10
0
5
10
0.5
0.6
0.7
0.8
0.9
0.82
0.822
0.824
0.826
0.828

$$\mathbf{m}^d = \left[\begin{array}{ccc} 0 & \mu^d & 2\mu^d \end{array} \right].$$

There are a total of twelve states in the state space. Once these inputs are given, the methodology described above yields the desired performance measures directly. Figure 3 shows the effect of the number of parallel stations on the production rate and the expected buffer level. In this specific case, the production rate of the second stage in isolation is kept equal to the production rate of the first stage in isolation as the number of parallel stations in the second stage increases. The figures shows that as the number of parallel stations increase both the production rate and the expected buffer level increases.

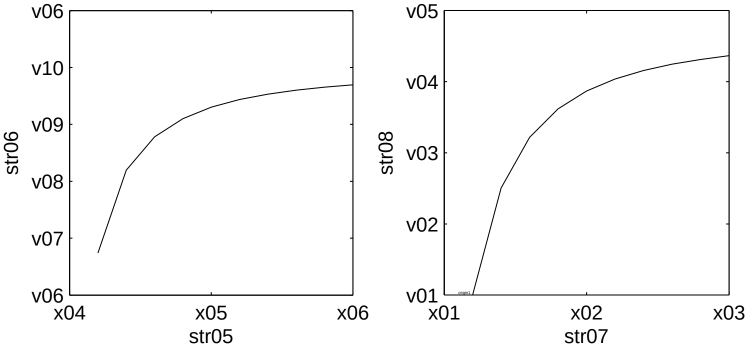

Figure 3: Effect of the number of parallel stations ($\mu^u = 1$, $p^u = 0.01$, $r^u = 0.09$, $m_u = 1$, $\mu^d = \mu^u \frac{m_u}{m_d}$, $p^d = 0.01$, $r^d = 0.09$, $N = 1$)

4.1.2 A Model with a Merge Structure

We now consider a three station merge system with a shared buffer. This system was analyzed in detail in (Tan 2001). Helber and Jusic (2004) also analyzes a similar system. In the upstream stage, there are two unreliable stations with processing rates μ_1 and μ_2. In the downstream stage, there is only one station with processing rate μ_3. The failure and repair rates for each station are p_i and r_i for $i = 1, 2, 3$. Figure 4 depicts the state transitions for M_u and M_d for this specific case.

Similar to the first example, we will specify the matrices λ^u and λ^d and the vectors μ^u and μ^d as the inputs of the solution methodology. The transition rates for M_u are given as

$$\lambda^u = \left[\begin{array}{cccc} -p_1 - p_2 & p_2 & p_1 & 0 \\ r_2 & -p_1 - r_2 & 0 & p_1 \\ r_1 & 0 & -p_2 - r_1 & p_2 \\ 0 & r_1 & r_2 & -r_1 - r_2 \end{array} \right] \tag{35}$$

where the states are ordered as $\{11, 10, 01, 00\}$. The processing rates in these states are

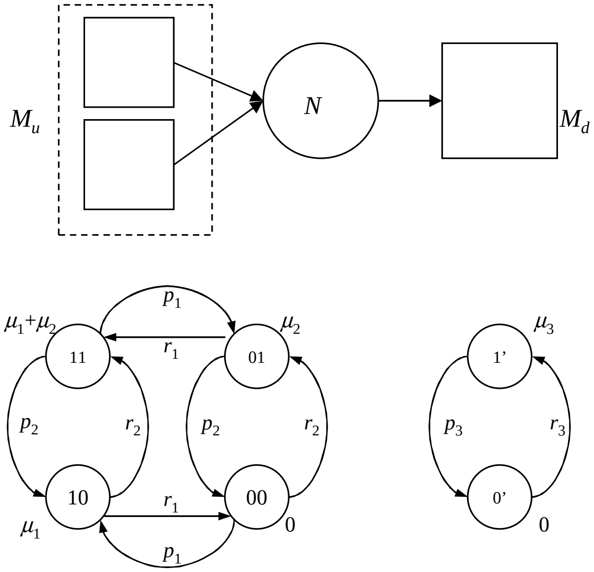

Figure 4: A system with a shared buffer

$$\mathbf{m}^u = \left[\begin{array}{cccc} \mu_1 + \mu_2 & \mu_1 & \mu_2 & 0 \end{array} \right].$$

Similarly,

$$\lambda^d = \left[\begin{array}{cc} -p_3 & p_3 \\ r_3 & -r_3 \end{array} \right] \tag{36}$$

where the states are ordered as $\{1, 0\}$. In these states the processing rates of M_d are given as

$$\mathbf{m}^d = \left[\begin{array}{cc} \mu_3 & 0 \end{array} \right].$$

There are eight discrete states in the state space. Once these inputs are given, the methodology described above yields the desired performance measures directly. We compare this case with the results given in (Tan 2001). Since a specific case with hot standby is analyzed in (Tan 2001), the method described above is modified accordingly. Figure 5 shows the effect of μ_3 on the production rate and the expected buffer level obtained by using the methodology given here and the results in (Tan 2001) that are equal to each other.

4.1.3 A Model with Series Stations

We now consider a production line where M_u has m_u and M_d has m_d stations in series. The stations are indexed from 1 to $m_u + m_d$. Each station is unreliable and has one up and one down state. The

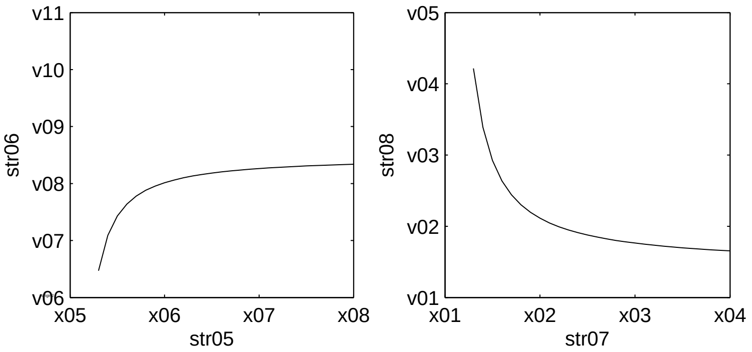

Figure 5: Effect of the processing rate ($\mu_1 = 1.2$, $\mu_2 = 1$, $p_1 = 0.1$, $p_2 = 0.1$, $p_3 = 0.2$, $r_1 = 0.9$, $r_2 = 0.9$, $r_3 = 0.9$, $N = 1$)

processing rate of station k is μ_k. The failure and repair times of station k are exponential random variables with rates p_k and r_k, $k = 1, ..., m_u + m_d$.

The state of the upstream stage is a vector of length m_u with its ith element is 1 if station i is operational and 0 otherwise, $1 \leq i \leq m_u$. Similarly, the state of the downstream stage is a vector of length m_d with its jth element is 1 if station $m_u + j$ is operational and 0 otherwise, $1 \leq j \leq m_d$. Accordingly, M_u has 2^{m_u} states and M_d has 2^{m_d} states.

Since each stage is operational only when all the stations are up, M_u produces at the maximum rate of $\mu^u = \min\{\mu_1, ..., \mu_{m_u}\}$ when all the stations are up and it can not produce if one of the machines is down. Similarly, M_d produces at the maximum rate of $\mu^d = \min\{\mu_{m_u+1}, ..., \mu_{m_u+m_d}\}$ when all the stations are up and it cannot produce when one of the stations is down.

When all the stations of M_u are up, each station can fail with rate $p_i \frac{\mu^u}{\mu_i}$, $i = 1, ..., m_u$ due to operational failures. Similarly, when one station is down, none of the other up stations can fail since they will be forced to stop due to the down station. As a result, the only possible transition when station k is down is the repair of station k with rate r_k. Therefore although there are 2^{m_u} states for M_u, only $m_u + 1$ of them will be non-transient. The case for M_d is similar.

Figure 6 depicts the state transitions for M_u and M_d for a specific case where M_u has 3 stations and M_d has 2 stations in series.

The matrices λ^u and λ^d and the vectors $\mathbf{m}^u$ and $\mathbf{m}^d$ for this specific case are given below:

$$\lambda^u = \begin{bmatrix} -\mu^u(\frac{p_1}{\mu_1} + \frac{p_2}{\mu_2} + \frac{p_3}{\mu_3}) & p_1\frac{\mu^u}{\mu_1} & p_2\frac{\mu^u}{\mu_2} & p_3\frac{\mu^u}{\mu_3} \\ r_1 & -r_1 & 0 & 0 \\ r_2 & 0 & -r_2 & 0 \\ r_3 & 0 & 0 & -r_3 \end{bmatrix} \tag{37}$$

where the states are ordered as $\{(1,1,1), (0,1,1), (1,0,1), (1,1,0)\}$. The processing rates in these states are

$$\mathbf{m}^u = \begin{bmatrix} \mu^u & 0 & 0 & 0 \end{bmatrix}$$

where $\mu^u = \min\{\mu_1, \mu_2, \mu_3\}$. Similarly

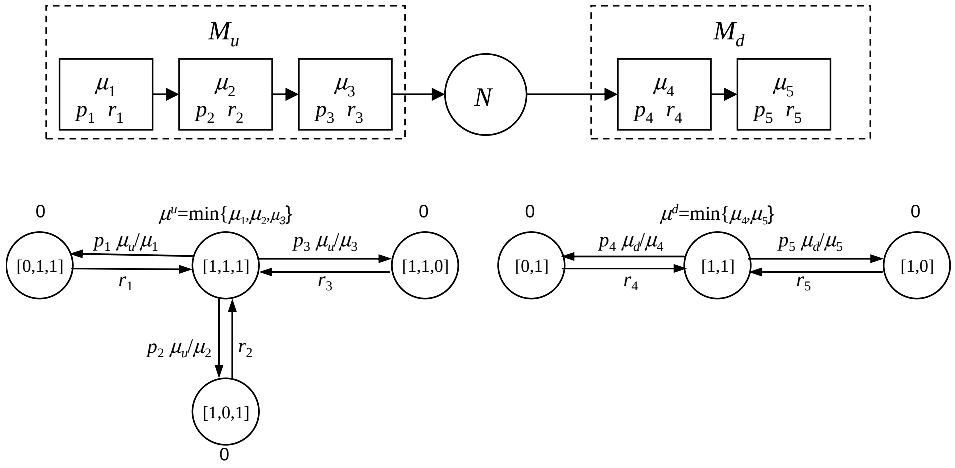

Figure 6: A system with series stations in each stage

$$\lambda^d = \begin{bmatrix} -\mu^d(\frac{p_4}{\mu_4} + \frac{p_5}{\mu_5}) & p_4\frac{\mu^d}{\mu_4} & p_5\frac{\mu^d}{\mu_5} \\ r_4 & -r_4 & 0 \\ r_5 & 0 & -r_5 \end{bmatrix} \tag{38}$$

where the states are ordered as $\{(1,1),(1,0),(0,1)\}$. In these states the processing rates of M_d are given as

$$\mathbf{m}^d = \begin{bmatrix} \mu^d & 0 & 0 \end{bmatrix}$$

where $\mu^d = \min\{\mu_4, \mu_5\}$. There are a total of twelve states in the state space. Once these inputs are given, the methodology described above yields the desired performance measures directly.

Consider the problem of locating a finite buffer in a continuous material flow production line with no interstation buffers. This problem has not been addressed in the literature before. Once the buffer is located between machine k and $k+1$, the line is divided into two stages. The resulting two-stage system can be analyzed by using the methodology outlined above. Figure 7 shows the effect of the buffer placement on the production rate for a production line with ten identical stations. As expected, for this homogeneous system placing the buffer in the middle, between Machine 5 and 6 maximizes the production rate.

However, when the stations are not identical, the buffer location that maximizes the production rate can be different. Figure 8 shows the effect of the buffer placement on the production rate for a production line with ten non-identical stations. In this case, placing the buffer between station 5 and 6 maximizes the production rate.

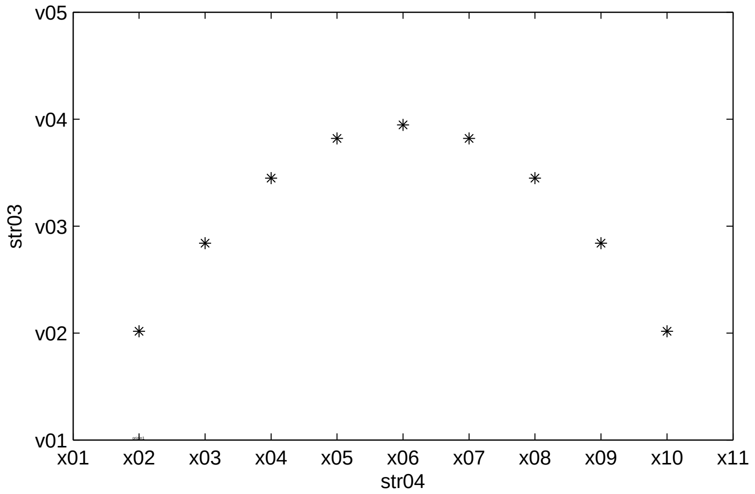

Figure 7: Effect of the buffer placement on the production rate ($\mu_i = 1$, $p_i = 0.01$, $r_i = 0.9$, $i = 1, ..., 10$, $N = 1$)

4.2 Phase-Type Failure and Repair Time Distributions

Models of unreliable production systems include states that describe when the machine is up and operating and down. Our methodology does not classify states as up and down states. States with zero flow rates can be labelled as down states in this context. Similarly a transition that reduces the processing rate can be labelled as a failure and a transition that increases the processing rate can be labelled as a repair. With this view the general model depicted in Figure 1 can include systems with phase-type failure and repair time distributions.

Most of the studies that focus on unreliable production systems assume exponential failure and exponential repair times. Although the exponential failure time assumption can be justified, observed repair time distributions are not generally exponential. Therefore analyzing production systems with general failure and repair time distributions is of interest. Using phase-type distributions allows us to handle a wide range of probability distributions within the framework of continuous time Markov chains.

Another motivation for studying two stage production systems with phase-type distributions is to develop a building block that can be used to evaluate the performance of a multistation production system approximately. One effective approximation method to evaluate the performance of production systems is decomposition. The decomposition approach basically considers each buffer of a given system and approximates the upstream and downstream flow dynamics of this buffer by building a two-stage single buffer system. The parameters of the upstream and the downstream station are then determined by relating the solution of one subsystem to another in a consecutive way until a convergence criterion is satisfied.

In this section, we first present the general model for a system with exponential failure and phase-type repair time distribution. Then we present the model discussed above: a system with

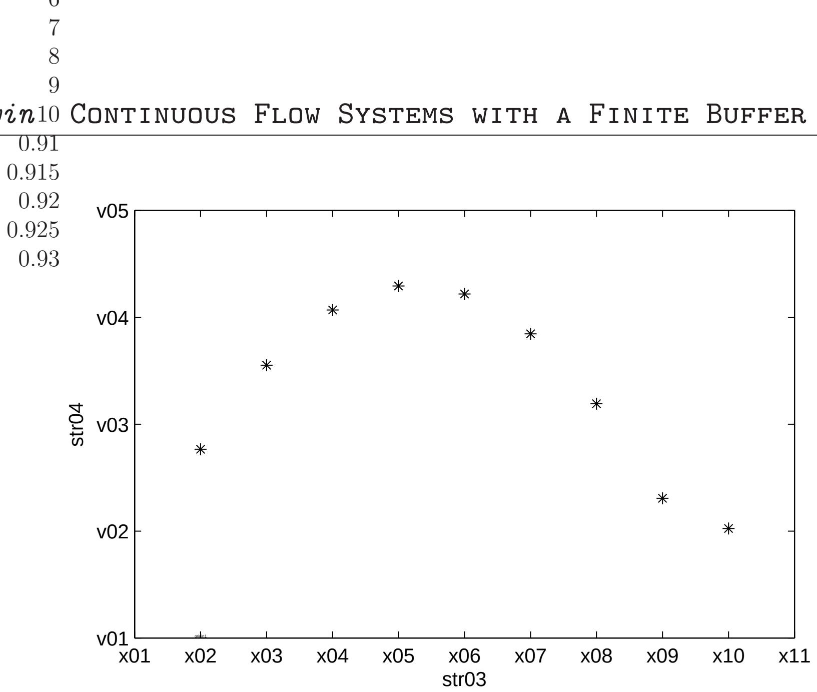

Figure 8: Effect of the buffer placement on the production rate ($p_i = 0.01$, $r_i = 0.9$, $i = 1 : 10$, $\mu_j = 1$, $j = 1 : 8$, $\mu_k = 4$, $k = 9, 10$, $N = 1$)

exponential failure and hyper-exponential repair time distribution with different number of stages. Finally, a model with Erlang-type failure and repair time distributions is discussed.

4.2.1 Exponential Up and Phase-type Down Time

Let us first consider a model with exponential up and phase-type down times. In this model, there is a single up state and a number of down states. The number of down states in M_u is κ^u and the number of down states in M_d is κ^d. When M_u is in its up state, it produces with rate μ^u and when M_d is in its up state, it produces with rate μ^d. The up times of M_u and M_d are exponential random variables with rates p^u and p^d respectively. The probability that M_u fails with a transition to down state i is q^u_i and the probability that M_d fails with a transition to down state j is q^d_j. The time spent at each down state is an exponential random variable. The transition rate from down state i to the up state of M_u is r^u_i and the transition rate from down state j to the up state of M_d is r^d_j. The transition rate from down state i to down state i' is $\lambda^u_r = \{\lambda^u_{ii'}\}$ for M_u. Similarly the transition rate from down state j to down state j' is $\lambda^d_r = \{\lambda^d_{jj'}\}$ for M_d. Figure 9 depicts the system.

The general structure of the matrices λ^u, λ^d and the vectors $\mathbf{m}^u$ and $\mathbf{m}^d$ are given below:

$$\lambda^u = \left[\begin{array}{c|cccc} -p^u & q^u_1 p^u & q^u_2 p^u & \cdots & q_{\kappa^u} p^u \\ \hline r^u_1 & & & & \\ r^u_2 & & \lambda^u_r & & \\ \vdots & & & & \\ r^u_{\kappa^u} & & & & \end{array} \right], \tag{39}$$

$$\mathbf{m}^u = \left[\begin{array}{cccc} \mu^u & 0 & \cdots & 0 \end{array} \right]$$

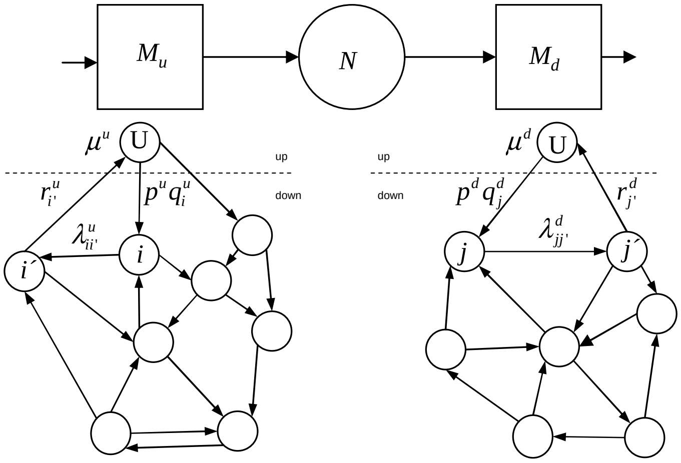

Figure 9: A system with Exponential Up and Phase-type Down times

where the states are ordered with the up states first and then the down states. Similarly,

$$\lambda^d = \left[\begin{array}{c|cccc} -p^d & q^d_1 p^d & q^d_2 p^d & \cdots & q_{\kappa^d} p^d \\ \hline r^d_1 & & & & \\ r^d_2 & & & & \\ \vdots & & \lambda^d_r & & \\ r^d_{\kappa^d} & & & & \end{array} \right], \tag{40}$$

$$\mathbf{m}^d = \left[\begin{array}{cccc} \mu^d & 0 & \cdots & 0 \end{array} \right].$$

4.2.2 Exponential Up and Hyper-Exponential Down Time

A special case of the model with exponential up and phase-type down time distribution is a model where the up times are exponentially distributed random variables and the repair times are hyper-exponential random variables with different number of stages and rates. For example, a machine with multiple down states associated with different type failures and a single up state is an example for this case.

Bihan and Dallery (2000) present a decomposition method for a continuous material flow production line with exponential failure and repair times. In order to capture the behavior of downstream flow of a given buffer, they analyze a two-machine building block where each machine has exponential up and two-stage hyper-exponential down time distribution. They set the parameters of the two-stage hyper-exponential distribution to fit the first three moments of a given repair time distribution. Levantesi, Matta, and Tolio (2003) also present a similar model with exponential up and hyper-exponential down time with an arbitrary number of stages and analyze this system exactly. Their model exploits the special structure of the rate matrix and yields an exact solution for system with up to 1000 failure modes per station.

Let the upstream machine M_u have κ^u different failure modes. The failure rate to mode i is $p_i^u = p^u q_i^u$ and the repair rate in mode i is r_i^u. When the machine is operational, the processing rate when it is not blocked or starved is μ^u. Similarly, the downstream machine M_d has κ^d different failure modes. The failure rate to mode j is $p_j^d = p^d q_j^d$ and the repair rate in mode j is r_j^d. When the machine is operational, its processing rate when it is not blocked or starved is μ^d. Figure 10 depicts an example of this system.

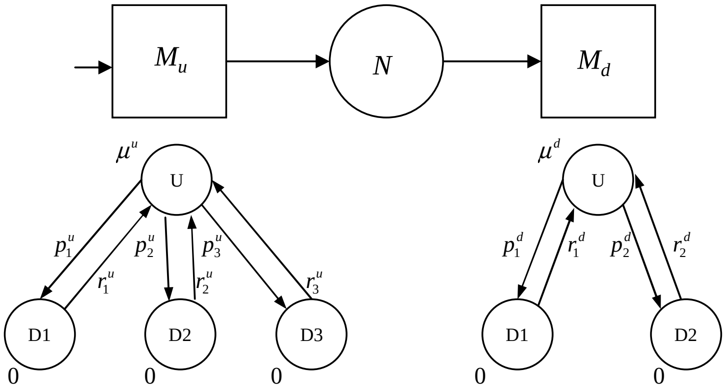

Figure 10: A system with Exponential Up and Hyper-exponential Down times

The possible transitions for M_u are

- from state U to state D_i with rate p_i^u, $i = 1, ..., \kappa^u$,
- from state D_i to state U with rate r_i^u, $i = 1, ..., \kappa^u$.

Similarly, the possible transitions for M_d are

- from state U to state D_j with rate p_j^d, $j = 1, ..., \kappa^d$,
- from state D_j to state U with rate r_j^d, $j = 1, ..., \kappa^d$.

For example, in the specific case depicted in Figure 10 with $\kappa^u = 3$, $\kappa^d = 2$, the matrices λ^u and λ^d and the vectors μ^u and μ^d are given below:

$$\lambda^u = \begin{bmatrix} -p_1^u - p_2^u - p_3^u & p_1^u & p_2^u & p_3^u \\ r_1^u & -r_1^u & 0 & 0 \\ r_2^u & 0 & -r_2^u & 0 \\ r_3^u & 0 & 0 & -r_3^u \end{bmatrix}, \tag{41}$$

$$\mathbf{m}^u = \begin{bmatrix} \mu^u & 0 & 0 & 0 \end{bmatrix},$$

$\kappa^u \backslash \kappa^d$	5	10	15	20	25	30	35	40	45	50
5	0.4	1.1	0.4	0.5	0.6	0.6	0.4	0.5	0.4	0.3
10	0.8	1.0	1.3	0.8	1.2	0.9	1.5	1.2	1.1	1.1
15	0.6	1.4	0.9	1.5	1.6	1.1	1.4	1.3	1.6	1.4
20	1.5	1.1	1.0	2.0	1.7	1.4	1.7	1.2	1.7	2.2
25	0.6	0.9	1.6	0.9	2.0	1.6	1.3	1.8	3.5	2.4
30	1.3	1.3	2.6	1.6	1.8	1.6	3.7	3.1	2.1	3.1
35	0.5	1.6	1.1	1.5	3.0	1.6	2.1	3.1	2.6	3.1
40	0.9	1.1	1.6	2.2	2.0	2.1	1.7	2.2	3.2	2.9
45	0.7	0.9	0.8	1.7	1.4	2.4	3.5	2.5	4.0	2.1
50	0.9	0.7	1.3	1.4	2.6	2.0	2.0	2.2	2.4	2.7

Figure 11: Accuracy of the Two-stage Hyper-Exponential Approximation

$$\lambda^d = \begin{bmatrix} -p_1^d - p_2^d & p_1^d & p_2^d \\ r_1^d & -r_1^d & 0 \\ r_2^d & 0 & -r_2^d \end{bmatrix}, \tag{42}$$

$$\mathbf{m}^d = \begin{bmatrix} \mu^d & 0 & 0 \end{bmatrix}.$$

By using the exact solution of the two-stage continuous flow production system with an exponential up and a hyper-exponential repair time distribution obtained by the proposed methodology, we can assess the accuracy of the two-stage approximation used by Bihan and Dallery (2000).

In order to evaluate the accuracy of the two-stage approximation, we generated ten random two-stage systems for each case with different number of stages for the upstream and the downstream stations with $2 \leq \kappa^u \leq 50$, $2 \leq \kappa^d \leq 50$. For each case we normalized $\mu^u = \mu^d = 1$, and $p^u = 1$ and generated other parameters randomly with $N \sim$ Uniform$[1, 10]$, $p^d \sim$ Uniform$[0, 1]$, $x_i^u \sim$ Uniform$[0, 1]$ and $q_i^u = \frac{x_i^u}{\sum_{i=1}^{\kappa^u x_i^u}}$ $x_i^u \sim$ Uniform$[0, 1]$ and $q_i^u = \frac{x_i^u}{\sum_{i=1}^{\kappa^u} x_i^u}$ for $i = 1, ..., \kappa^u$, $x_j^d \sim$ Uniform$[0, 1]$ and $q_j^d = \frac{x_j^d}{\sum_{j=1}^{\kappa^d} x_j^d}$ for $j = 1, ..., \kappa^d$.

The table shown in Figure 11 gives the absolute percentage error of the two-stage approximation for the production rate $\epsilon_\Pi = 100\frac{|\Pi - \Pi_a|}{\Pi}$ (Π_a is the production rate with two-stage approximation and Π is the exact production rate) for systems with different number of stages. As the table shows, the two-stage approximation may yield errors up to 4% for systems with high number of stages.

Note also that in the model with $\kappa^u = 50$ and $\kappa^d = 50$ stages, there are 2601 machine states and the analysis requires the solution of 100 differential and 2501 algebraic equations subject to boundary conditions. Consequently, the exact solution described here takes longer than Bihan and Dallery's approximation when the number of stages is large.

4.2.3 A Model with Erlang Up and Down Times

We now model a production system where the failure and repair times are Erlang-type random variables. We assume that the failure time of M_u is an Erlang random variable with κ_f^u stages. The expected failure time is $\frac{1}{p^u}$ and the squared coefficient of variation of the failure time is $scv_f^u = \frac{1}{\kappa_f^u}$. The repair time of M_u is also an Erlang random variable with κ_r^u stages. The expected failure time is $\frac{1}{r^u}$ and the squared coefficient of variation of the failure time is $scv_r^u = \frac{1}{\kappa_r^u}$.

Similarly, failure time of M_d is an Erlang random variable with κ_f^d stages. The expected failure time is $\frac{1}{p^d}$ and the squared coefficient of variation of the failure time is $scv_f^d = \frac{1}{\kappa_f^d}$. The repair time of M_d is also an Erlang random variable with κ_r^d stages. The expected repair time is $\frac{1}{r_d}$ and the squared coefficient of variation of the repair time is $scv_r^d = \frac{1}{\kappa_r^d}$.

The processing rates of M_u and M_d are μ^u and μ^d respectively. In this model M_u has $\kappa_f^u + \kappa_r^u$ states and M_d has $\kappa_f^d + \kappa_r^d$ states. The states of M_u are indexed from 1 to $\kappa_f^u + \kappa_r^u$ and ordered such that states $i = 1, ..., \kappa_f^u$ are for the up states and states $i = \kappa_f^u + 1, ..., \kappa_f^u + \kappa_r^u$ are for the down states of M_u. Similarly the states of M_d are indexed from 1 to $\kappa_f^d + \kappa_r^d$ and ordered such that states $i = 1, ..., \kappa_f^d$ are for the up states and states $i = \kappa_f^d + 1, ..., \kappa_f^d + \kappa_r^d$ are for the down states of M_u.

The possible transitions for M_u are

- from state i to state $i+1$ with rate $\kappa_f^u p^u$, $i = 1, ..., \kappa_f^k$,
- from state i to state $i+1$ with rate $\kappa_r^u r^u$, $i = \kappa_f^u + 1, ..., \kappa_f^u + \kappa_r^u - 1$,
- from state $\kappa_f^u + \kappa_r^u$ to state 1 with rate $\kappa_r^u r^u$.

Similarly, the possible transitions for M_d are

- from state j to state $j+1$ with rate $\kappa_f^d p^d$, $j = 1, ..., \kappa_f^d$,
- from state j to state $i+1$ with rate $\kappa_r^d r_d$, $j = \kappa_f^d + 1, ..., \kappa_f^d + \kappa_r^d - 1$,
- from state $\kappa_f^d + \kappa_r^d$ to state 1 with rate $\kappa_r^d r_d$.

For example, let us consider a specific case with $\kappa_f^u = 2$, $\kappa_r^u = 2$, $\kappa_f^d = 1$, and $\kappa_r^u = 3$. For this specific system, Figure 12 depicts the state transition diagram.

The matrices λ^u and λ^d and the vectors $\mathbf{m}^u$ and $\mathbf{m}^d$ for this specific case are given below:

$$\lambda^u = \begin{bmatrix} -\kappa_f^u p^u & \kappa_f^u p^u & 0 & 0 \\ 0 & -\kappa_f^u p^u & \kappa_f^u p^u & 0 \\ 0 & 0 & -\kappa_r^u r^u & \kappa_r^u r^u \\ \kappa_r^u r^u & 0 & 0 & -\kappa_r^u r^u \end{bmatrix}, \tag{43}$$

$$\mathbf{m}^u = \begin{bmatrix} \mu^u & \mu^u & 0 & 0 \end{bmatrix},$$

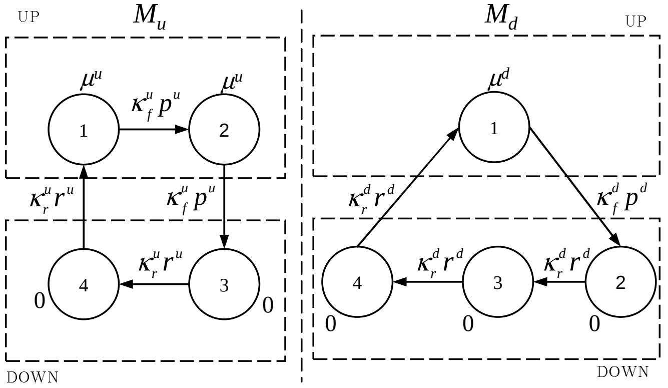

Figure 12: A system with Erlang Up and Down times

$$\lambda^d = \begin{bmatrix} -\kappa_f^d p^d & \kappa_f^d p^d & 0 & 0 \\ 0 & -\kappa_r^d r_d & \kappa_r^d r_d & 0 \\ 0 & 0 & -\kappa_r^d r_d & \kappa_r^d r_d \\ \kappa_r^d r_d & 0 & 0 & -\kappa_r^d r_d \end{bmatrix}, \tag{44}$$

$$\mathbf{m}^d = \begin{bmatrix} \mu^d & 0 & 0 & 0 \end{bmatrix}.$$

Figures 13 and 14 show the effects of the failure and repair time variabilities of each stage on the production rate and the expected buffer level. Figure 13 shows that as the coefficient of variation of the failure times of first and the second stages increase, the production rate decreases. On the other hand, a decrease in the variability of the failure time of the upstream machine results in an increase in the expected buffer level. Similarly, Figure 14 shows the effect of the repair time variability of the firs and the second stage on the production rate and the expected buffer level. A decrease in repair time variability of either stage increases the production rate. On the other hand, a decrease of the repair time variability of only the first stage increases the expected buffer level.

4.3 Quality-Quantity Models

In this section, we consider a production system with two unreliable machines with multiple up and down states and a finite buffer that is an extension of the one studied by Poffe and Gershwin (2005).

In the system we consider, the both stages has two up (State 1 and State -1 for M_u and State 1' and State -1' for M_d) and three down states (State D_1, D_{-1}, and D_Q for M_u and $D_{1'}$, $D_{-1'}$, and $D_{Q'}$ for M_d). In States 1 and 1', both machines produce products with no quality problems but when M_u is in State -1, the quality of the products produced by M_u is not perfect. Similarly, when M_d is in State -1', the quality of the products produced by M_d is not perfect. Furthermore, the

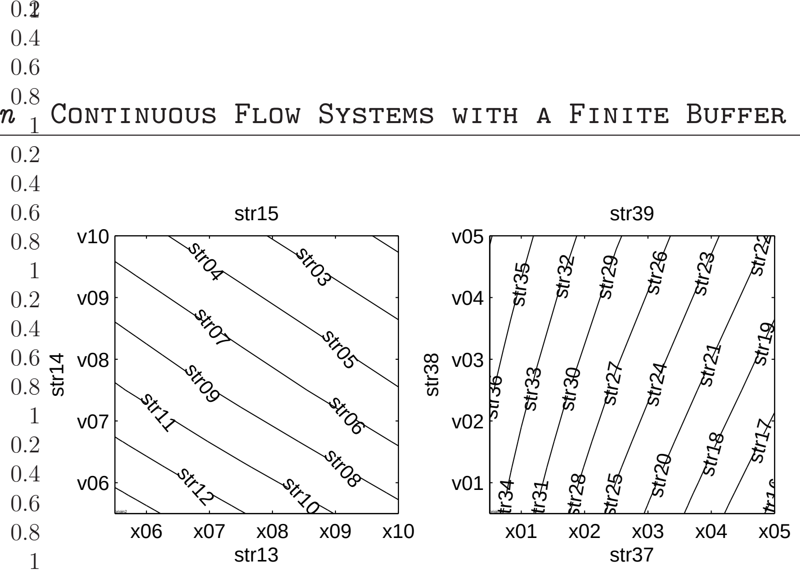

Figure 13: Effect of the failure time variability ($\mu^u = 1$, $\mu^d = 1$, $p^u = 0.005$, $p^d = 0.01$, $r^u = 0.15$, $r^d = 0.1$, $N = 10$)

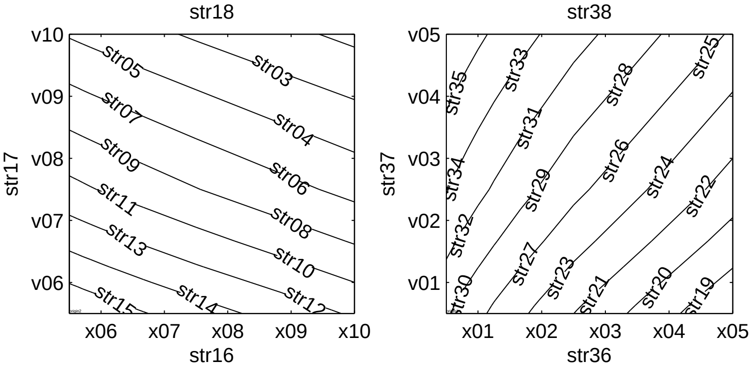

Figure 14: Effect of the repair time variability ($\mu^u = 1$, $\mu^d = 1$, $p^u = 0.005$, $p^d = 0.01$, $r^u = 0.15$, $r^d = 0.1$, $N = 10$)

machines are subject to two different failures: operational failures (States D_1, D_{-1}, $D_{1'}$, and $D_{-1'}$) and quality failures (States D_Q and $D_{Q'}$) and they have different mean times to repair. Since these failures are different in nature, they cannot be modelled with a single down state.

The processing rates of the upstream stage in both of the up states are equal to μ^u; the processing rate of the downstream stage in both of its up states are μ^d; and the processing rates of all the down states for both stages are equal to 0. Figure 4.3 depicts the state transitions for M_u and M_d for this model.

The states of M_u are ordered as $\{1, -1, D_1, D_{-1}, D_Q\}$ and numbered from 1 to $I_u = 5$. Similarly, the states of M_d are ordered as $\{1', -1', D_{1'}, D_{-1'}, D_{Q'}\}$ and numbered from 1 to $I_d = 5$.

The matrices λ^u and λ^d and the vectors $\mathbf{m}^u$ and $\mathbf{m}^d$ for this model are given below:

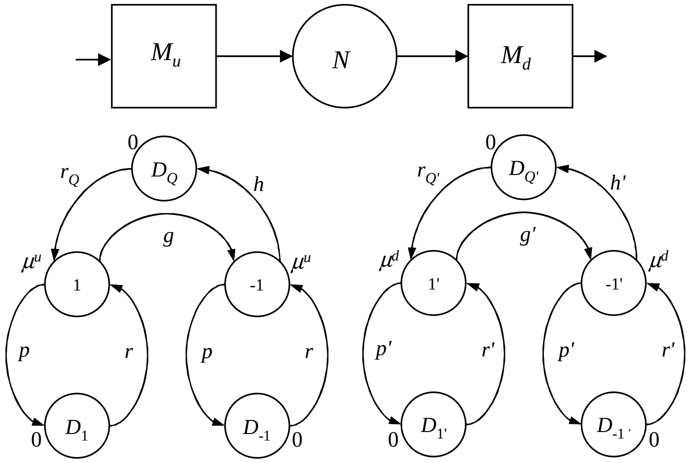

Figure 15: A System with multiple up and down states

$$\lambda^u = \begin{bmatrix} -g-p & g & p & 0 & 0 \\ 0 & -p-h & 0 & p & h \\ r & 0 & -r & 0 & 0 \\ 0 & r & 0 & -r & 0 \\ r_Q & 0 & 0 & 0 & -r_Q \end{bmatrix}. \tag{45}$$

$$\lambda^d = \begin{bmatrix} -g'-p' & g' & p' & 0 & 0 \\ 0 & -p'-h' & 0 & p' & h' \\ r' & 0 & -r' & 0 & 0 \\ 0 & r' & 0 & -r' & 0 \\ r_{Q'} & 0 & 0 & 0 & -r_{Q'} \end{bmatrix}. \tag{46}$$

$$\mathbf{m}^u = \begin{bmatrix} \mu^u & \mu^u & 0 & 0 \end{bmatrix}.$$

$$\mathbf{m}^d = \begin{bmatrix} \mu^d & \mu^d & 0 & 0 \end{bmatrix}.$$

5 Conclusion

The methodology we developed allows us to analyze general Markovian continuous-flow material flow two stage-single buffer production systems. A wide range of models can be analyzed by our methodology directly by determining the transition rates of each stage and the flow rates associated with the discrete states of each stage.

We illustrated the generality of our method by showing how a number of different models analyzed in the literature can be handled as special cased of our general model. We validated all the results with simulation.

The run time of the methodology is fast and affected by the number of discrete states of the system and not by the buffer size. In general, it is known that spectral methods have some accuracy and stability issues and this is not a limitation to this paper. In that case the size matters. In our case this can be an issue only when we use this methodology in a decomposition where it is needed to model the downstream and the upstream processes of a buffer by using a two-stage building block. In this case, if all the stations in the production system are modelled by using a given set of assumptions that correspond to a particular structure, this structure can be exploited to improve the solution efficiency. In other words, a more efficient computational method can be devised to implement the general methodology to analyze a given system.In all the other two-stage models previously published in the literature and analyzed as special cases of our methodology, the method works without any stability issues.

The ability to analyze two-stage systems with general structures yields devising new decomposition methods for multistation production systems with different machines and finite buffers. More accurate two-stage building blocks can be built to characterize the dynamics of the flows in and out of a given buffer by using mixtures of phase-type distributions corresponding to repairs of upstream and downstream stations. Using the general two-station building block to devise decomposition methods for general production systems is left for future research.

The main contribution of this method is allowing researchers to focus on developing models that describe the behavior of production systems and analyzing these models easily by using our methodology. Therefore we present our methodology as a general tool to analyze Markovian fluid flow systems with a finite buffer.

Acknowledgement

The first author acknowledges the support from TÜBA and TÜBİTAK. The second author acknowledges the support from the Singapore-MIT Alliance and the General Motors Research and Development Center.

References

Ahn, S., J. Jeon, and V. Ramaswami (2005). Steady state analysis for finite fluid flow models using finite QBDs. *Queueing Systems 49*, 223259.

Ahn, S. and V. Ramaswami (2003). Fluid flow models and queues: a connection by stochastic coupling. *Stochastic Models 19*(3), 325–348.

Anick, D., D. Mitra, and M. Sondhi (1982). Stochastic theory of a data handling system with multiple sources. *Bell System Technology Journal 61*, 1871–1894.

Bihan, H. L. and Y. Dallery (2000). A robust decomposition method for the analysis of production lines with unreliable machines and finite buffers. *Annals of Operations Research 93*, 265–297.

Dallery, Y. and H. L. Bihan (1999). An improved decomposition method for the analysis of production lines with unreliable machines and finite buffers. *International Journal of Production Research 37*(5), 1093–1117.

Diamantidis, A., C. Papadopoulos, and M. Vidalis (2004). Exact analysis of a discrete material three-station one-buffer merge system with unreliable machines. *International Journal of Production Research 42*(4), 651–675.

Dubois, D. and J. P. Forestier (1982). Productivité et en-cours moyens d'un ensemble de deux machines séparées par une zone de stockage. *RAIRO Automatique 16*, 105–132.

Elwalid, A. and D. Mitra (1991). Analysis and design of rate-based congestion control of high speed networks, i: stochastic fluid models, access regulation. *Queueing Systems Theory Applications 9*, 19–64.

Forestier, J. (1980). Modelisation stochastique et comportement asymptotique d'un systeme automatise de production. *RAIRO Automatique 14*(2), 127–143.

Gershwin, S. B. and I. C. Schick (1980). Continuous model of an unreliable two-machine material flow system with a finite interstage buffer. Report LIDS-R-1039, MIT Laboratory for Information and Decision Systems.

Helber, S. and H. Jusic (2004). A new decomposition approach for non-cyclic continuous material flow lines with a merging flow of material. *Annals of Operations Research 125*, 117139.

Kim, J. and S. B. Gershwin (2005). Integrated quality and quantity modeling of a production line. *OR Spectrum 2*, 287–314.

Koster, M. B. M. D. (1989). *Capacity Oriented Analysis and Design of Production Systems*. Berlin: Springer Verlag.

Levantesi, R., A. Matta, and T. Tolio (2003). Performance evaluation of continuous production lines with machines having different processing times and multiple failure modes. *Performance Evaluation 51*, 247–268.

Mandjes, M., D. Mitra, and W. Scheinhardt (2003). Models of network access using feedback fluid queues. *Queueing Systems 44*, 365–398.

Mitra, D. (1988). Stochastic theory of a fluid model of multiple failure-susceptible producers and consumers coupled by a buffer. *Advances in Applied Probability 20*, 646–676.

Özdoğru, U. and T. Altıok (2003). Analysis of two-valve fluid flow systems with general repair times. In *Analysis and Modeling of Manufacturing Systems*, pp. 255–288. Kluwer Publications.

Poffe, A. and S. Gershwin (2005). Integrating quality and quantity modelling in a production line. Technical Report ORC-377-05, Massachusetts Institute of Technology Operations Research Center Working Paper Series.

Serucola, B. (2001). A finite buffer fluid queue driven by a markovian queue. *Queueing Systems 38*, 213–220.

Sevast'Yanov, B. (1962). Infuence of stage bin capacity on the average standstill time of a production line. *Theory of Probability and Its Applications 7*(4), 429–438.

Soares, A. D. S. and G. Latouche (2006). Matrix-analytic methods for fluid queues with finite buffers. *Performance Evaluation 63*, 295314.

Tan, B. (2001). A three station continuous materials flow merge system with unreliable stations and a shared buffer. *Mathematical and Computer Modelling 33*(8-9), 1011–1026.

Tan, B. and S. B. Gershwin (2007). Modelling and analysis of markovian continuous flow production systems with a finite buffer: A general methodology and applications. Technical Report ORC-381-07, Massachusetts Institute of Technology Operations Research Center Working Paper Series.

Tan, B. and S. B. Gershwin (2009). Analysis of a general markovian two-stage continuous flow production system with a finite buffer. *International Journal of Production Research, to appear*.

Wijngaard, J. (1979). The effect of interstage buffer storage on the output of two unreliable production units in series with different production rates. *AIIE Transactions 11*(1), 42–47.

Yeralan, S., W. Franck, and M. A. Quasem (1986). A continuous materials flow production line model with station breakdown. *European Journal of Operational Research 27*, 289–300.

Yeralan, S. and B. Tan (1997). A station model for continuous materials flow production. *International Journal of Production Research 35*(9), 2525–2541.