Queueing Models and
Service Management
Vol. 8, No. 3, page 1–32, 2025

QMSM
©AU 2023

Characterizing the Output Process of Two-Stage Flow Lines with Unreliable Parallel Machines and Finite Intermediate Buffer

B. Madhu Rao[1,*] and Shaukat Brah[2]

[1]Department of Business Systems and Analytics
School of Business Administration
Stetson University, DeLand, FL 32723, USA

[2]Management Consultant
787 C, Phase VI, DHA, Lahore, Pakistan

(*Received April 2024; accepted April 2025*)

Abstract: This paper analyzes two-stage flow lines where raw material is processed sequentially in two stages to produce finished units. There are multiple machines at each stage that randomly break down and require repair. There is a limited amount of storage space (*buffer*) between the two stages. When the buffer is full, some or all machines at the first stage may be *blocked* (i.e., forced to idle due to the inability to unload a finished unit), and when the buffer is empty, some or all machines at the second stage may *starve* (i.e., forced to idle due to a lack of jobs for processing). The state changes in the system can be described by a continuous-time Markov chain when processing times, times between machine failures, and repair times are exponentially distributed. The study focuses on the variability and autocorrelation structure of the output stream of finished products from stage two. Efficient algorithms are developed to compute steady-state system characteristics using matrix analytical methods. The paper presents detailed numerical results highlighting the qualitative features of system behavior for a wide range of parameter values. Our key finding is that the output process of the system approximates a Poisson process for buffer size as small as one, and the numbers of machines at the two stages as small as two.

Keywords: Autocorrelation, blocking, matrix analytical methods, output process, production flow lines.

1. Introduction

A *flow line* (also referred to as *automatic transfer* or *production line*) is a manufacturing system consisting of workstations separated by intermediate storage areas or *buffers*, where material flows in a fixed sequence, visiting each work center exactly once. Multiple parallel machines may be used at one or more workstations to balance production capacity across different stages of production. Intermediate buffer space decouples successive stages of

* Corresponding author
Email: bmrao@stetson.edu

production and reduces the effect of machine failures. Flow lines are employed in high-volume, multi-stage production or service facilities. Manufacturing examples of flow lines include the production of automobiles and electronics, food packaging, furniture production, and flexible manufacturing systems. Service examples of flow lines include restaurants, hotels, banks, customer service centers, and telecommunication services.

The design of such systems consists of selecting numbers of parallel machines at each stage, size of buffer spaces, and the capacity to repair failed machines to achieve the desired average output rate at minimal overall cost. Capacity planning for flow lines often involves trade-offs between capital cost and the desired service or output level. Understanding flow line behavior is crucial in evaluating the economic implications at the design and operational stages. The two most important measures of performance of a flow line are the rate and variability of the output. Variability in output is important because even with a sufficient mean rate of output, high variability can lead to times of excess inventories with the associated inventory carrying costs, and times of shortages resulting in lost sales and the resulting damage to customer goodwill.

Literature on modeling and analyzing flow lines is vast and covers systems with and without buffers, with and without workstation failures, with single or multiple machines at each workstation, with possible scrapping of units, with random or deterministic processing times, using exact or approximate analysis, and using discrete or continuous time analysis. In this paper, we limit our literature review to studies focusing exclusively on variability and correlation structure of the output process from flow lines. Studies dealing solely with throughput rate analysis are included only when relevant to the current study. For a comprehensive review of flow line models, readers can refer to Dallery and Gershwin [6], Li et al. [17], and Tan [29].

Several studies highlight the importance of understanding the variability of output from a flow line and its autocorrelation structure for effective management of manufacturing and service systems. Tan [29] reports that data from a consumer durable manufacturer shows a standard deviation of daily production as high as 10.8% of the mean. Assaf [2] reports that in an engine-block production line the standard deviation of daily production over 10 days was observed to be 11.5% of the mean. Assaf [2] demonstrated that reducing the output variability can significantly improve service level for a wide range of demand patterns. Betterton and Silver [3] show that knowledge of the variance of interdeparture times can help identify bottlenecks in serial production lines. Inman [14] observed autocorrelation in the interarrival times to workstations in automotive body welding lines. Tan [29] reported significant autocorrelation in the interdeparture times of cars leaving an assembly line of a Japanese automotive manufacturer. Altiok and Melamed [1] found that autocorrelations at various manufacturing stages negatively affect overall performance. Dizbin and Tan [8] show that ignoring these correlations can result in severe overestimation or underestimation of key performance measures.

Study of the variability and correlation structure of the output from flow lines has received a great deal of attention. Miltenburg [19], Gershwin [9], and Dincer and Deler [7], studied output variability in terms of the variance of $N(t)$, the number of units produced in an arbitrary time interval $(0, t)$. More recent studies by He et al. [10], and Shin and Moon [23], used a Markovian arrival process (MAP) (Latouche and Ramaswami [15]) to study the variance of $N(t)$. Tan [27, 28] used a Markov reward model to develop a recursive method to determine the mean and the variance of $N(t)$ for a two-stage production line with a finite intermediate buffer. These studies do not address the correlation structure of the output process.

Hendrix [11] and Hendricks and McClain [12] studied interdeparture times of the output from flow lines with several stages, single reliable machine at each stage, general processing times, and finite intermediate buffers with possible blocking. They adopted simulation to study the variance and correlation structure of the output process. For similar systems with phase type service times, Tan and Lagershausen [30] obtain analytical results for mean and variance of $N(t)$ for two-stage systems and propose algorithms to compute autocorrelations of inter-departure times. Shin and Moon [25] studied the transient behavior of $N(t)$ of a two-stage flow line with a single unreliable machine at each stage and obtained expressions for the variance and distribution of the time to n^{th} departure.

Output variability is the result of factors that are internal or external to the organization. Internal factors such as randomness in production, machine failure and repair are under the control of the management and may be considered in the design stage. External factors such as demand and supply variability are typically not under the control of management. Understanding the effect of both internal and external sources of variability is crucial for the success of a manufacturing operation. This paper contributes to the creation of tools to describe the output process, including variability and autocorrelation structure, to manage a two-stage flow line effectively.

In this paper, we consider a two-stage flow line with multiple machines at each stage and a finite buffer for units processed in the first stage and waiting to be processed in the second stage. When the finite buffer is full, one or more machines in the first stage may be *blocked* (i.e., do not have space in the buffer to unload completed jobs). Similarly, when the buffer is empty, one or more machines in the second stage may *starve* (i.e., do not have jobs to process). Machines are prone to random failure and require regular repair. Ample repair capacity is assumed, and machines do not wait for repair. We assume that machines do not fail when blocked or starved. Abundant supply of jobs at the first stage and sufficient demand (or storage space) for the output of the second stage are assumed. The throughput rate of similar systems is studied by Liu et al. [18] and Shin and Moon [24]. We use matrix analytic methods to develop algorithms to compute the moments, density function and autocorrelations of the interdeparture times for completed jobs. Extensive numerical results are presented and interpreted to provide insight into the impact of the number of machines at each stage and the size of the buffer on the system behavior.

The main contribution of this paper is the complete characterization of the interdeparture times. The algorithmic methodology proposed in this paper is easy to implement and flexible enough to accommodate many variations to the basic model. Our analysis of the numerical results also yielded valuable insights into the effects of buffer size, and numbers of machines at the two stages, on the throughput rate and output variability. Numerical results indicate that the output stream of finished products approximates a Poisson process for buffer sizes as small as one and the number of machines at each stage as small as two.

The remainder of the paper is organized as follows: A formal mathematical model for the system is presented in Section 2, using which the output process is characterized in Section 3. Details of the algorithmic implementation are presented in Section 4. Numerical results with a detailed discussion of the qualitative characteristics of the system are presented in Section 5 followed by concluding remarks in Section 6.

In the following discussion, we assume that states of a Markov chain are arranged in lexicographic order. Unless stated otherwise, subscripts 1 and 2 refer to the two stages of production. The dimensions of vectors and matrices are generally clear from a given mathematical expression and are provided only when necessary for clarity. A summary of definitions for all relevant terms/symbols is provided in Table 1 for ready reference.

Table 1. Summary of Notation

Symbol	**Description**
	All vectors are denoted by bold letters (e.g., $\mathbf{x}$*,* $\boldsymbol{\alpha}$*,* $\boldsymbol{\beta}$*).* *Unless stated otherwise, are defined as row vectors.*
$\mathbf{e}$	A vector of 1's of appropriate dimension
$\mathbf{0}$	A vector of 0's of appropriate dimension
I	An identity matrix
Q	Infinitesimal generator describing system dynamics
$\mathbf{x}$	Row vector of steady state probabilities at an arbitrary time
$\mathbf{y}$	Row vector of steady state probabilities at departure instants
M	Buffer size
N_1, N_2	Numbers of machines at stages 1 and 2
μ_1, μ_2	Production rates for *each* machine at stages 1 and 2
λ_1, λ_2	Failure rates for *each* machine at stages 1 and 2
θ_1, θ_2	Repair rates for *each* machine at stages 1 and 2
ϕ_1, ϕ_2	Design production rates at stages 1 and 2
ϕ	Design throughput rate
ϕ_a	Actual throughput rate
Z_{b1}	Random variable describing the number of blocked machines at stage 1
Z_{s2}	Random variable describing the number of starving machines at stage 2
Z_{bb}	Random variable describing the number of units in the buffer
U	Random variable describing an interdeparture time
$f_U(u)$	Probability density function of interdeparture time
$\mu_u = E(U)$	Mean interdeparture time
$\sigma_u^2 = Var(U)$	Variance of interdeparture time
CV_u	Coefficient of variation of interdeparture time
r_1	Lag 1 autocorrelation of the output stream of finished jobs

2. Mathematical Model

The system consists of N_1 $[N_2]$ identical machines at the first [second] stage of production. There is storage space for M items processed at the first stage and waiting to be processed at the second stage. Each machine at the first [second] stage can process jobs at the rate of μ_1 $[\mu_2]$ items per unit time with exponentially distributed processing times. Each machine at the first [second] stage of production alternates between exponentially distributed *operating* time with parameter λ_1 $[\lambda_2]$ and exponentially distributed *down* time (under repair) with parameter θ_1 $[\theta_2]$. Ample (infinite) repair capacity is assumed at both stages of production so that there is no waiting for repair. This assumption is quite natural for manufacturing systems where the operator of a machine is typically responsible for its routine maintenance. Model modifications to eliminate this assumption are discussed in Section 6.

A machine is assumed to fail only when processing an item and the unit in-process remains on the machine (or stored in a location different from the buffer). When a failed machine is restored to operating condition, it continues processing the item that was in process at the time of failure. Thus, when a machine fails or is restored to operating condition, there will be no change in the number of units in the buffer. Modification to the model to incorporate possible scrapping of units in process at the time of a machine failure is discussed in Section 6.

Let ϕ_1 $[\phi_2]$ denote the steady state production rate at stage 1 [stage 2], *if it were operating independently*, and will be referred to as the *design production rate* at stage 1 [stage 2]. ϕ_1 $[\phi_2]$ can be obtained by recognizing that each machine in stage 1 [stage 2] alternates between an exponential operating time with parameter λ_1 $[\lambda_2]$, and an exponential repair time with parameter θ_1 $[\theta_2]$. From the properties of alternating renewal processes, the proportion of time a machine at stage 1 [stage 2] will be operational is given by $\frac{\theta_1}{\lambda_1+\theta_1}$ $\left[\frac{\theta_2}{\lambda_2+\theta_2}\right]$. Hence, $\phi_1 = N_1\mu_1\left(\frac{\theta_1}{\lambda_1+\theta_1}\right)$, and $\phi_2 = N_2\mu_2\left(\frac{\theta_2}{\lambda_2+\theta_2}\right)$.

A two-stage flow line is defined as *balanced* if $\phi_1 = \phi_2$ and *unbalanced* otherwise. For unbalanced systems, the stage with the lower production rate will be referred to as the *bottleneck* stage. $\phi = min\{\phi_1, \phi_2\}$ yields the upper bound on the throughput rate from the flow shop and will be referred to as *design throughput rate* of the flow line. ϕ_a, the *actual steady state throughput rate* of the flow line, will be less than ϕ due to blocking and starving. Understanding the effects of various system parameters on the difference between ϕ and ϕ_a is a key objective of the paper.

A Markovian description of the system state is given by $\{k, \mathcal{C}\}$, where k represents the number of units in the buffer (referred to as *level*) and $\mathcal{C}$ describes the configuration of the machines in the two stages. There will be $M + 3$ levels as follows:

- Levels 0 to M consisting of states with 0 to M units in the buffer and no blocking or starving.
- Level 0_s, consisting of states when the buffer is empty ($k = 0$) *and* one or more machines at stage 2 are starving.
- Level M_b, consisting of states when the buffer is full ($k = M$) *and* one or more

machines at stage 1 are blocked.

Possible configurations ($\mathcal{C}$) at each level are as follows:

- At levels 0 to M, configuration $\mathcal{C}$ can be described by the two tuple $\{i, j\}$, i and j denoting the number of machines in operating condition and processing jobs (with the rest of the machines under repair) at stages 1 and 2, respectively. n, the number of states at each level k is given by all possible combinations of i $(= 0, 1, 2, \cdots N_1)$, and j $(= 0, 1, 2, \cdots N_2)$, and is equal to $(N_1 + 1) * (N_2 + 1)$;
- At level 0_s, configuration $\mathcal{C}$ is defined by the three tuple $\{i, j, j_s\}$, where i and j are as defined above, and j_s is the number of starving machines in stage 2, meaning that $(j - j_s)$ machines are processing jobs. At level 0_s,
  - i can take values $\{0, 1, 2, \cdots N_1\}$,
  - j can only take values $\{1, 2, \cdots N_2\}$, because states with $j = 0$ will have no starving machines and will be part of level 0, and
  - j_s can take values $\{1, 2, \cdots j\}$.

  For each value of i, total number possible combinations of $\{j, j_s\}$ is given by $\{N_2 + (N_2 - 1) + \cdots + 2 + 1\}$, which sums to $\frac{N_2 * (N_2 + 1)}{2}$. n_s, the total number of states at level 0_s is equal to $\frac{N_2 * (N_2 + 1) * (N_1 + 1)}{2}$;
- At level M_b, configuration $\mathcal{C}$ is defined by the three tuple $\{i, i_b, j\}$, where i and j are defined as above, and i_b is the number of machines blocked in stage 1, meaning that $(i - i_b)$ machines are processing jobs. At level M_b,
  - j can take values $\{0, 1, 2, \cdots N_2\}$,
  - i can only take values $\{1, 2, \cdots N_1\}$, because states with $i = 0$ will have no blocked machines, and will be part of level M, and
  - i_b can take values $\{1, 2, \cdots i\}$,

  For each value of j, total number possible configurations of $\{i, i_b\}$ is given by $\{N_1 + (N_1 - 1) + \cdots + 2 + 1\}$, which sums to $\frac{N_1 * (N_1 + 1)}{2}$. n_b, the total number of states at level M_b is equal to $\frac{N_1 * (N_1 + 1) * (N_2 + 1)}{2}$.

Arranging the states in lexicographic order, dynamics of the system under study can be described by a continuous time Markov chain (CTMC) with infinitesimal generator Q is shown in Figure 1 with states partitioned into $M + 3$ levels. Matrices A_0, A_1 and A_2 are square matrices of order n. Matrices B_0, B_1 and B_2 are of order $(n_s \times n)$, $(n_s \times n_s)$ and $(n \times n_s)$ respectively. Similarly, matrices C_0, C_1, and C_2 are of order $(n \times n_b)$, $(n_b \times n_b)$ and $(n_b \times n)$ respectively. L, the dimension of matrix Q, is given by $n_s + M * n + n_b$. Complete description of the structure of the matrices A_0, A_1, A_2, B_0, B_1, B_2, C_0, C_1, and C_2 for N_1= 2 and N_2= 2 is provided in Appendix A (Figures A1-A9). All elements of Q are positive except the diagonal elements, which are negative and are equal in magnitude to the sum of all other elements in that row. We use the notation of displaying the diagonal elements by an ”*,” as in matrices A_1, B_1, and C_1 displayed in the Appendix.

$$
Q = \begin{array}{c|ccccccccc|}
 & 0_s & 0 & 1 & 2 & \cdot & \cdot & M-1 & M & M_b \\
\hline
0_s & B_1 & B_0 & & & & & & & \\
0 & B_2 & A_1 & A_0 & & & & & & \\
1 & & A_2 & A_1 & A_0 & & & & & \\
2 & & & A_2 & A_1 & A_0 & & & & \\
\cdot & & & & \cdot & \cdot & \cdot & & & \\
\cdot & & & & & \cdot & \cdot & \cdot & & \\
M-1 & & & & & & A_2 & A_1 & A_0 & \\
M & & & & & & & A_2 & A_1 & C_0 \\
M_b & & & & & & & & C_2 & C_1
\end{array}
$$

Figure 1. Infinitesimal Generator.

$$
Q_1 = \begin{array}{c|ccccccc|}
 & 0 & 1 & 2 & 3 & \cdot & N_1-1 & N_1 \\
\hline
0 & * & N_1\theta_1 & & & & & \\
1 & \lambda_1 & * & (N_1-1)\theta_1 & & & & \\
2 & & 2\lambda_1 & * & \cdot & & & \\
\cdot & & & \cdot & * & \cdot & & \\
N_1-2 & & & & \cdot & * & 2\theta_1 & \\
N_1-1 & & & & & (N_1-1)\lambda_1 & * & \theta_1 \\
N_1 & & & & & & N_1\lambda_1 & *
\end{array}
$$

Figure 2. Infinitesimal Generator.

When there is no blocking or starving, the two stages function as independent subsystems. Hence, A_0, A_1, and A_2 can be expressed in terms of two independently operating Markov chains describing the changes in values of i and j. Let Q_1 and Q_2 denote the infinitesimal generators for the two Markov chains. Q_1 has $N_1 + 1$ states and has the structure shown in Figure 2. Let $\boldsymbol{\mu}_1$ denote the column vector of production rates corresponding to the $(N + 1)$ states of the generator Q_1. $\boldsymbol{\mu}_1 = [0\ \mu_1\ 2\mu_1\ \cdots (N_1 - 1)\mu_1\ N_1\mu_1]'$, where the superscript $'$ indicates a transpose. Q_2 and $\boldsymbol{\mu}_2$ (not shown here) are of order $N_2 + 1$ and have a similar structure with subscript 1 replaced by 2. Matrices A_0, A_1 and A_2 can be expressed in terms Q_1, Q_2, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ as follows.

$$
\begin{aligned}
A_1 &= I_1 \otimes (Q_2 - \Delta(\boldsymbol{\mu}_2)) + (Q_1 - \Delta(\boldsymbol{\mu}_1)) \otimes I_2, \\
A_0 &= \Delta(\boldsymbol{\mu}_1) \otimes I_2, \text{and} \\
A_2 &= I_1 \otimes \Delta(\boldsymbol{\mu}_2),
\end{aligned}
$$

where I_1 and I_2 are identity matrices of size $(N_1 + 1)$ and $(N_2 + 1)$ respectively. For any vector **a**, $\Delta(\mathbf{a})$ represents a diagonal matrix with elements given by the elements of **a**. $\otimes$ is the Kronecker product operator. It is, in general, not possible to express the boundary matrices B_0, B_1, B_2, C_0, C_1 and C_2 in such a compact form.

Let $x_{k,\mathcal{C}}$ denote the unique equilibrium probability of the system being in state $(k,\mathcal{C})$, and let **x** denote a row vector with $x_{k,\mathcal{C}}$ as elements arranged in the same order as the states of the Markov chain. **x** is the unique vector satisfying the equations $\mathbf{x}Q = \mathbf{0}$, and $\mathbf{xe} = 1$, where **0** and **e** respectively are column vectors of *zeros* and *ones* of size L.

To take advantage of the sparsity of Q, Gauss-Seidel iterative method is adopted in the evaluation of **x**. Convergence of the iterative process can be accelerated by using the aggregation/disaggregation approach discussed in Heyman and Goldsmith [13] and others. Implementation details are presented in Section 4.

2.1. Measures of system behavior

Average, variability, and autocorrelation structure of the output stream of finished products from stage two are the focus of this paper. Section 3 details the algorithms to evaluate these characteristics. Expressions for other useful measures of system behavior are presented below in an easily implementable form, in terms of sub-vectors $\mathbf{x}_k, k = 0_s, 0, 1, \cdots M - 1, M$, and M_b. These subvectors are portions of the steady state probability vector **x** corresponding to levels $0_s, 0, 1, \cdots M-1, M, M_b$. $\mathbf{x}_k, k = 0, 1, 2, \cdot M$ are of order n, and $\mathbf{x}_{0_s}$ and $\mathbf{x}_{M_b}$ are of order n_s and n_b respectively.

- Probability distribution of Z_{b1}, the number of blocked machines at stage 1, and $E(Z_{b1})$, the expected number of blocked machines at stage 1, are given by,

$$\begin{aligned}
P(Z_{b1} = l) &= \sum_{i=l}^{N_1}\sum_{j=0}^{N_2} \mathbf{x}_{M_b}(i,l,j), \quad l = 1, 2, \cdots N_1, \\
P(Z_{b1} = 0) &= 1 - \sum_{l=1}^{N_1} P(Z_{b1} = l), \\
E(Z_{b1}) &= \sum_{l=1}^{N_1} l \; P(Z_{b1} = l).
\end{aligned}$$

- Probability distribution of Z_{s2}, the number of starving machines at stage 2, and $E(Z_{s2})$, the expected number of starving machines at stage 2 are given by,

$$\begin{aligned}
P(Z_{s2} = l) &= \sum_{i=0}^{N_1}\sum_{j=l}^{N_2} \mathbf{x}_{0_s}(i,j,l), \quad l = 1, 2, \cdots N_2, \\
P(Z_{s2} = 0) &= 1 - \sum_{l=1}^{N_2} P(Z_{b_2} = l), \\
E(Z_s) &= \sum_{l=1}^{N_2} l \; P(Z_{b_2} = l).
\end{aligned}$$

- Probability distribution of Z_{bb}, the number of units in the buffer, and $E(Z_{bb})$ the expected number of occupied buffer spaces are given by,

$$\begin{aligned}
P(Z_{bb}=0) &= \mathbf{x}_{0_s}\mathbf{e}+\mathbf{x}_0\mathbf{e},\\
P(Z_{bb}=l) &= \mathbf{x}_l\mathbf{e},\ l=1,2,\cdots M-1,\\
P(Z_{bb}=M) &= \mathbf{x}_M\mathbf{e}+\mathbf{x}_{M_b}\mathbf{e},\\
E(Z_{bb}) &= \sum_{l=1}^{M} l\ P(Z_{bb}=l).
\end{aligned}$$

3. Output Process from the System

The first step in characterizing the stream of finished parts leaving the system is to determine steady state departure instant probabilities (i.e., immediately after the instants of job completions at stage 2). These probabilities, together with the infinitesimal generator described in Figure 1 can be used to characterize the distribution of time between two successive departures from the system, or the *interdeparture time*. This is further extended to study the autocorrelation structure of interdeparture times.

3.1. Departure instant probabilities

Let t denote an arbitrary point in time when the system is operating under steady state conditions. $y_{k,\mathcal{C}}$, the steady state probability that a departure leaves the system in state $(k,\mathcal{C})$ can be expressed as,

$$y_{k,\mathcal{C}}=\frac{q_{k,\mathcal{C}}\ dt}{\gamma\ dt}\qquad k=0_s,0,1,\cdots M-1,M,M_b,\ \ \vee\mathcal{C},$$

where, $(q_{k,\mathcal{C}}\ dt)$ is the probability of a departure in the interval $(t,t+dt)$ leaving the system in state $(k,\mathcal{C})$ at $t+dt$, and $(\gamma\ dt)$ is the normalizing probability of a departure in the arbitrary infinitesimal interval $(t,t+dt)$.

Under steady state conditions, possible departures from the system during the infinitesimal interval $(t,t+dt)$, and the corresponding changes in the system state are summarized below.

- The system is in level 0_s, (i.e., at least one of the machines at stage 2 is starving) at time t, and a service completion occurs in stage 2 during the interval $(t,t+dt)$. After departure, the system remains at level 0_s and the number of starving machines is increased by 1.
- The system is in level 0 at time t, and a service completion occurs in stage 2 during the interval $(t,t+dt)$. After departure, the system moves from level 0 to level 0_s, with exactly one starving machine at stage 2.
- The system is in level k $(1\le k\le M)$ at time t, and a service completion occurs at stage 2 during the interval $(t,t+dt)$. After departure, the system moves to level $k-1$ because one of the units from the buffer is loaded into the machine with service completion.

- The system is in level M_b with exactly one blocked machine at stage 1 at time t, and a service completion takes place in stage 2 during the interval $(t, t+dt)$. After departure, the system moves to level M.
- The system is in level M_b with more than one blocked machines at stage 1 at time t, and a service completion occurs in stage 2 during the interval $(t, t + dt)$. After departure, the system remains in level M_b and the number of blocked machines is reduced by 1.

To simplify the evaluation of **y**, we split matrix B_1 $[C_1]$ into B_{11} and B_{12} $[C_{11}$ and $C_{12}]$ where B_{12} $[C_{12}]$ contains all the elements from B_1 $[C_1]$ with μ_2 (thus representing the rate at which a service completions take place from the respective states) and B_{11} $[C_{11}]$ contains all other elements. $\gamma\ dt$ can be expressed as follows.

$$\gamma\ dt = \mathbf{x}_{0_s} B_{12}\mathbf{e}\ dt + \mathbf{x}_0 B_2\mathbf{e}\ dt + (\sum_{i=1}^{M} \mathbf{x}_i A_2\mathbf{e})\ dt + \mathbf{x}_{M_b} C_2\mathbf{e}\ dt + \mathbf{x}_{M_b} C_{12}\mathbf{e}\ dt.$$

Let **y** denote the vector of departure instant probabilities arranged in the same manner as **x**. The subvectors of **y** can be expressed as follows:

$$\begin{aligned} \mathbf{y}_{0_s} &= (\mathbf{x}_{0_s} B_{12} dt + \mathbf{x}_0 B_2 dt)/(\gamma dt) = (\mathbf{x}_{0_s} B_{12} + \mathbf{x}_0 B_2)/\gamma, \\ \mathbf{y}_i &= (\mathbf{x}_{i+1} A_2)/\gamma \ \text{ for } i = 0, 1, 2, \cdots M-1, \\ \mathbf{y}_M &= (\mathbf{x}_{M_b} C_2)/\gamma, \\ \mathbf{y}_{M_b} &= (\mathbf{x}_{M_b} C_{12})/\gamma. \end{aligned}$$

γ describes the equilibrium departure/throughput rate from the system and is equivalent to ϕ_a, the *actual throughput rate* of the flow line, defined in Section 2.

3.2. Characteristics of the interdeparture times

When the system is operating in steady state, an interval between two successive departures can be described as a phase type random variable by structuring the interval as the time till absorption in a CTMC. This can be accomplished by modifying the infinitesimal generator Q (Figure 1) by adding an absorbing state and diverting all transitions that lead to departures from the system (i.e., transitions due to service completions at stage 2) to the absorbing state. This idea is based on the original work by Neuts [20] and adopted by Rao and Posner [21] to develop algorithms for the moments, density function, and correlation structure of interdeparture intervals under steady state conditions.

Based on the detailed description of departure resulting transitions in Section 3.1, infinitesimal generator Q can be modified by adding an absorbing state D_1 and described by S_1^* (Figure 3).

Let U denote the random variable describing an arbitrary inter-departure interval. Under steady state conditions, U starts in one of the states in S_1^* (excluding state D_1) with probabilities given by the vector **y**, and ends with absorption in state D_1. U has a phase type probability distribution with L $(= n_s + M * n + n_b)$ phases and representation $(\boldsymbol{\beta}_1, T_1)$

$$S_1^* = \begin{array}{c||ccccccccc|c||}
 & 0_s & 0 & 1 & 2 & \cdot & \cdot\ \cdot & M-1 & M & M_b & D_1 \\
0_s & B_{11} & B_0 & & & & & & & & B_{12}\mathbf{e} \\
0 & & A_1 & A_0 & & & & & & & B_2\mathbf{e} \\
1 & & & A_1 & A_0 & & & & & & A_2\mathbf{e} \\
2 & & & & A_1 & A_0 & & & & & A_2\mathbf{e} \\
\cdot & & & & & \cdot\ \cdot & & & & & \cdot \\
\cdot & & & & & & \cdot\ \cdot & & & & \cdot \\
M-1 & & & & & & & A_1 & A_0 & & A_2\mathbf{e} \\
M & & & & & & & & A_1 & C_0 & A_2\mathbf{e} \\
M_b & & & & & & & & & C_{11} & C_2\mathbf{e} + C_{12}\mathbf{e} \\
\hline
D_1 & & & & & \mathbf{0} & & & & & 0
\end{array} = \begin{bmatrix} T_1 & \mathbf{T}_1^0 \\ \mathbf{0} & 0 \end{bmatrix}$$

Figure 3. Time between two departures ($M > 1$).

(Neuts [20]), where $\boldsymbol{\beta}_1 = \mathbf{y}$ and T_1 is as defined in Figure 3. The density function and the moments of U can be written down as below.

$$\begin{aligned}
f_U(u) &= \boldsymbol{\beta}_1\, e^{T_1 u}\, \mathbf{T}_1^0, \quad u > 0 && (1)\\
E(U) &= \mu_u = -\boldsymbol{\beta}_1 T_1^{-1}\mathbf{e}, && (2)\\
E(U^2) &= 2\boldsymbol{\beta}_1 T_1^{-2}\mathbf{e}, && (3)\\
Var(U) &= \sigma_u^2 = E(U^2) - (E(U))^2, && (4)
\end{aligned}$$

where $\mathbf{T}_1^0 = -T_1\mathbf{e}$. The departure rate (i.e., the throughput rate) γ is given by $1/\mu_u = \phi_a$.

Computing moments and density function of U respectively require the evaluation of the inverse and exponential of T_1. Unless N_1, N_2 and M are small, T_1 will be large, making computing inverse or exponential of T_1, inefficient and numerically hazardous. As with the evaluation of **x**, iterative methods present a practical alternative. Implementation details are presented in Section 4.

3.3. Autocorrelation of interdeparture times

The lag one autocorrelation of the output process can be studied by considering the time to two successive departures, starting at a departure instant. Let U and V denote two successive inter-departure intervals and let $\psi_2 = U + V$. ψ_2 can be described as the time to absorption in a CTMC with infinitesimal generator S_2^* with $2L + 1$ states, partitioned into two sets of L states and an absorbing state D_2, as displayed in Figure 4. Transition from the first set of L states (U-states) to the second set of L states (V-states) indicates the end of U and the start of V, and absorption into D_2 indicates the end of V.

S_2^* is displayed in partitioned form in Figure 5, where T_1 describes the transitions during the interdeparture interval and T_2 describes the transition from U to V. T_1 and T_2 are square matrices of dimension L.

ψ_2 has a phase type probability distribution with $2L$ phases and representation $(\boldsymbol{\alpha}_2, S_2)$. $\boldsymbol{\alpha}_2 = [\boldsymbol{\beta}_1\ \mathbf{0}]$, indicating that ψ_2 always starts with U. The joint density function of U and V

$$S_2^* = \begin{array}{c||ccccccc|ccccccc|c||}
 & 0_s & 0 & 1 & \cdot & M-1 & M & M_b & 0_s & 0 & 1 & \cdot & M-1 & M & M_b & D_2 \\
\hline
0_s & B_{11} & B_0 & & & & & & B_{12} & & & & & & & \\
0 & & A_1 & A_0 & & & & & B_2 & & & & & & & \\
1 & & & A_1 & A_0 & & & & & A_2 & & & & & & \\
\cdot & & & & \cdot & \cdot & & & & & \cdot & & & & & \mathbf{0} \\
M-1 & & & & & A_1 & A_0 & & & & & A_2 & & & & \\
M & & & & & & A_1 & C_0 & & & & & A_2 & & & \\
M_b & & & & & & & C_{11} & & & & & & C_2 & C_{12} & \\
\hline
0_s & & & & & & & & B_{11} & B_0 & & & & & & B_{12}\mathbf{e} \\
0 & & & & & & & & & A_1 & A_0 & & & & & B_2\mathbf{e} \\
1 & & & & & & & & & & A_1 & A_0 & & & & A_2\mathbf{e} \\
\cdot & & & & & & & & & & & \cdot & \cdot & & & \cdot \\
M-1 & & & & & & & & & & & & A_1 & A_0 & & A_2\mathbf{e} \\
M & & & & & & & & & & & & & A_1 & C_0 & A_2\mathbf{e} \\
M_b & & & & & & & & & & & & & & C_{11} & C_2\mathbf{e} + C_{12}\mathbf{e} \\
\hline
D_2 & & & & \mathbf{0} & & & & & & & \mathbf{0} & & & & 0
\end{array}$$

Figure 4. Times between three successive departures.

$$S_2^* = \left[\begin{array}{cc} S_2 & \mathbf{S}_2^0 \\ \mathbf{0} & 0 \end{array}\right] = \left[\begin{array}{cc|c} T_1 & T_2 & \mathbf{0} \\ 0 & T_1 & \mathbf{T}_1^0 \\ \hline \mathbf{0} & \mathbf{0} & 0 \end{array}\right].$$

Figure 5. Partitioned Matrix S_2^*.

can be expressed as

$$\begin{aligned} f_{U,V}(u,v) &= [f_U(u)] . [f_{V|U}(v|u)] \\ &= \left[-\boldsymbol{\beta}_1\, e^{T_1 u}\, T_1\, \mathbf{e}\right] . \left[-\boldsymbol{\beta}_2^*(u)\, e^{T_1 v}\, T_1\, \mathbf{e}\right]. \end{aligned} \tag{5}$$

$\boldsymbol{\beta}_2(u) = \boldsymbol{\beta}_1 e^{T_1 u}\, T_2$, is the vector of probabilities of transitions from U-states to V-states, given that $U = u$ [20, 21]. $\boldsymbol{\beta}_2(u)$, normalized by the sum of its elements (displayed as $\boldsymbol{\beta}_2^*(u)$ in equation 5) represents the vector of initial probabilities for V, given $U = u$. By recognizing that $T_2\mathbf{e} = -T_1\mathbf{e}$, we see that the sum of the elements of $\boldsymbol{\beta}_2(u)$ is equal to $\boldsymbol{\beta}_2(u)\mathbf{e} = \boldsymbol{\beta}_1 e^{T_1 u}\, T_2\mathbf{e} = -\boldsymbol{\beta}_1\, e^{T_1 u}\, T_1\, \mathbf{e} = f_U(u)$. $f_{U,V}(u,v)$ can now be written as,

$$f_{U,V}(u,v) = -\boldsymbol{\beta}_2(u)\, e^{T_1 v}\, T_1\, \mathbf{e} = -\boldsymbol{\beta}_1\, e^{T_1 u}\, T_2\, e^{T_1 v}\, T_1\, \mathbf{e}. \tag{6}$$

It is easy to verify that $f_{U,V}(u,v)$ given by equation (6) leads to identical marginal distributions for U and V, consistent with equation (1). Using equation (6), $E(UV)$ can be obtained as,

$$E(UV) \;=\; -\int_{u=0}^{\infty}\int_{v=0}^{\infty} u\, v\, \boldsymbol{\beta}_1\, e^{T_1 u}\, T_2\, e^{T_1 v}\, T_1\, \mathbf{e}\, du\, dv$$

$$= \; -\boldsymbol{\beta}_1\, T_1^{-2}\, T_2\, T_1^{-1}\, \mathbf{e},$$

from which r_1,the autocorrelation at lag 1, can be obtained as follows.

$$r_1 = \frac{Cov(UV)}{\sqrt{Var(U)}\sqrt{Var(V)}} \;=\; \frac{E(UV)-E(U)E(V)}{\sqrt{Var(U)}\sqrt{Var(V)}} = \frac{E(UV)-\mu_u^2}{\sigma_u^2}. \qquad (7)$$

Details of the numerical implementation of equations (2), (3), (4), and (7) to compute $E(U)$, $Var(U)$, and r_1 are discussed in Section 4.

Autocorrelations at higher lags can be obtained by considering the total time for correspondingly increased number of departures. For example, r_2, the lag 2 autocorrelation, can be obtained by considering the time to 3 successive departures, starting at a departure instant. Let U, V and W denote three successive inter-departure intervals. $\psi_3 = U+V+W$, can be described as the time to absorption in a CTMC with $3L+1$ states and infinitesimal generator S_3^* displayed in partitioned form in Figure 6.

$$S_3^* = \begin{bmatrix} S_3 & \mathbf{S}_3^0 \\ \mathbf{0} & 0 \end{bmatrix} = \left[\begin{array}{ccc|c} T_1 & T_2 & 0 & \mathbf{0} \\ 0 & T_1 & T_2 & \mathbf{0} \\ 0 & 0 & T_1 & \mathbf{T}_1^0 \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 \end{array}\right].$$

Figure 6. Partitioned Matrix S_3^*.

ψ_3 has a phase type probability distribution with $3L$ phases and representation $(\boldsymbol{\alpha}_3, S_3)$, with $\boldsymbol{\alpha}_3 = [\boldsymbol{\beta}_1\ \mathbf{0}\ \mathbf{0}] = [\mathbf{y}\ \mathbf{0}\ \mathbf{0}]$. The joint density function of U, V and W can be expressed as

$$\begin{aligned} f_{U,V,W}(u,v,w) &= [f_U(u)] \,.\, \left[f_{V|U}(v|u)\right] \,.\, \left[f_{W|U,V}(w|u,v)\right] \\ &= \left[-\boldsymbol{\beta}_1\, e^{T_1 u}\, T_1\, \mathbf{e}\right] . \left[-\boldsymbol{\beta}_2^*(u)\, e^{T_1 v}\, T_1\, \mathbf{e}\right] . \left[-\boldsymbol{\beta}_3^*(u,v)\, e^{T_1 w}\, T_1\, \mathbf{e}\right] \\ &= \left[-\boldsymbol{\beta}_1\, e^{T_1 u}\, T_2\, e^{T_1 v}\, T_1\, \mathbf{e}\right] . \left[-\boldsymbol{\beta}_3^*(u,v)\, e^{T_1 w}\, T_1\, \mathbf{e}\right] . \end{aligned}$$

$\boldsymbol{\beta}_3^*(u,v)$ is the normalized version of $\boldsymbol{\beta}_3(u,v) = \boldsymbol{\beta}_1\, e^{T_1 u}\, T_2\, e^{T_1 v}\, T_2$, the vector of probabilities of transitions from U-states to W-states (through the V-states), given that $U = u$ and $V = v$. Using $T_2\mathbf{e} = -T_1\mathbf{e}$, the sum of elements of $\boldsymbol{\beta}_3(u)$ can be expressed as $\boldsymbol{\beta}_3(u,v)\ \mathbf{e} = \boldsymbol{\beta}_1\, e^{T_1 u}\, T_2\, e^{T_1 v}\, T_2\ \mathbf{e} = -\boldsymbol{\beta}_1\, e^{T_1 u}\, T_2\, e^{T_1 v}\, T_1\ \mathbf{e}$. $f_{U,V,W}(u,v,w)$ can now be written as,

$$f_{U,V,W}(u,v,w) = -\boldsymbol{\beta}_3(u,v)\, e^{T_1 w}\, T_1\, \mathbf{e} = -\boldsymbol{\beta}_1\, e^{T_1 u}\, T_2\, e^{T_1 v}\, T_2\, e^{T_1 w}\, T_1\, \mathbf{e}. \qquad (8)$$

Using equation (8), $E(UW)$ is obtained as,

$$\begin{aligned} E(UW) &= -\int_{u=0}^{\infty}\int_{v=0}^{\infty}\int_{w=0}^{\infty} u\, w\, \boldsymbol{\beta}_1\, e^{T_1 u}\, T_2\, e^{T_1 v}\ \ T_2\, e^{T_1 w}\, T_1\, \mathbf{e}\, du\, dv\, dw \\ &= -\boldsymbol{\beta}_1\, T_1^{-2}\, T_2\, T_1^{-1}\, T_2\, T_1^{-1}\mathbf{e}. \end{aligned}$$

r_2, the autocorrelation at lag 2 can be obtained as $r_2 = [E(UW) - \mu_u^2]/\sigma_u^2$. Extension to higher order autocorrelations is tedious but routine. In the interests of brevity, we limit the computations to autocorrelation at lag 1 in this paper.

4. Algorithmic Considerations

In this Section, we provide detailed descriptions of the numerical procedures used in computing the system characteristics in the sequence in which the computations need to be carried out.

4.1. Computation of $\mathbf{x}$

$\mathbf{x}$ is the unique solution of the equation $\mathbf{x}Q = \mathbf{0}$ and $\mathbf{xe} = 1$ and can be evaluated using the iterative scheme described above. After each iteration, the vector $\mathbf{x}$ needs to be scaled so that the elements of the vector sum to 1 to satisfy equation $\mathbf{xe} = 1$.

The number of iterations required for convergence can be greatly reduced by adopting the aggregation/disaggregation approach discussed in Heyman and Goldsmith [13] and others. $\mathbf{x}Q = \mathbf{0}$ can be expressed as follows.

$$\begin{aligned}
\mathbf{x}_{0_s}B_1 + \mathbf{x}_0B_2 &= \mathbf{0}, && (9)\\
\mathbf{x}_{0_s}B_0 + \mathbf{x}_0A_1 + \mathbf{x}_1A_2 &= \mathbf{0}, && (10)\\
\mathbf{x}_{k-1}A_0 + \mathbf{x}_kA_1 + \mathbf{x}_{k+1}A_2 &= \mathbf{0}, \quad k = 1, 2, \cdots M-1, && (11)\\
\mathbf{x}_{M-1}A_0 + \mathbf{x}_MA_1 + \mathbf{x}_{M_b}C_2 &= \mathbf{0}, && (12)\\
\mathbf{x}_MC_0 + \mathbf{x}_{M_b}C_1 &= \mathbf{0}. && (13)
\end{aligned}$$

Post-multiplying equations (10-14) by $\mathbf{e}$ and simplifying, we obtain

$$\begin{aligned}
\mathbf{x}_{0_s}\mathbf{b}_0 &= \mathbf{x}_0\left(\mathbf{e}_1 \otimes \boldsymbol{\mu}_2\right), && (14)\\
\mathbf{x}_{i-1}\left(\boldsymbol{\mu}_1 \otimes \mathbf{e}_2\right) &= \mathbf{x}_i\left(\mathbf{e}_1 \otimes \boldsymbol{\mu}_2\right), \quad i = 1, 2, 3, \cdots M, && (15)\\
\mathbf{x}_M\left(\boldsymbol{\mu}_1 \otimes \mathbf{e}_2\right) &= \mathbf{x}_{M_b}\mathbf{c}_2, && (16)
\end{aligned}$$

where, $\mathbf{b}_0 = B_0\mathbf{e}$ and $\mathbf{c_2} = C_2\mathbf{e}$ and $\mathbf{e}_1$ and $\mathbf{e}_2$ are column vectors of 1s of dimension N_1 are N_1 respectively. In the implementation of the Gauss-Seidel method, at the end of each iteration, subvectors of $\mathbf{x}$ can be individually scaled to satisfy the macro balance equations (14-16) prior scaling them to satisfy $\mathbf{xe} = 1$. Specific details of computational efficiency are presented in Section 5.2. Measure of system behavior (Section 2.1) and $\mathbf{y}$ (Section 3.1) can be computed using vector $\mathbf{x}$.

4.2. Computation of $E(U)$*,* $Var(U)$*, and* r_1

Inversion of matrix T_1 required by the direct implementation of equations (2), (3), (4), and (7) in computing $E(U)$, $Var(U)$, and r_1 can be avoided by structuring the computations as solutions of sparse systems of equations as follows.

1. $E(U)$ in equation (2) can be expressed as $\boldsymbol{\omega}_1\mathbf{e}$ and $\boldsymbol{\omega}_1(= -\mathbf{y}T_1^{-1})$ can be computed by solving the system of equations $\boldsymbol{\omega}_1T_1 = -\mathbf{y}$.
2. $E(U^2)$ in equation (3) can be expressed as $2\boldsymbol{\omega}_2\mathbf{e}$ and $\boldsymbol{\omega}_2(= -\boldsymbol{\omega}_1T_1^{-1})$ can be computed by solving the system of equations $\boldsymbol{\omega}_2T_1 = -\boldsymbol{\omega}_1$. $Var(U)$ can be computed using equation (4).

3. r_1 can be computed as follows:
   - Define $\boldsymbol{\omega}_4 = \boldsymbol{\omega}_3 T_1^{-1}$, where $\boldsymbol{\omega}_3 = \boldsymbol{\omega}_2 T_2$.
   - $\boldsymbol{\omega}_4$ can be computed by solving the system of equations $\boldsymbol{\omega}_4 T_1 = \boldsymbol{\omega}_3$.
   - Using equation (4), $E(UV)$ can be computed as $-\boldsymbol{\omega}_4 \mathbf{e}$. r_1 can then be computed using equation (7).

Vectors $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, $\boldsymbol{\omega}_3$, and $\boldsymbol{\omega}_4$ are typically very large, but they can be computed efficiently using iterative methods and exploiting the sparsity of matrices T_1 and T_2. High-precision computation is recommended to avoid loss of significance due to the successive evaluation of $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, $\boldsymbol{\omega}_3$ and $\boldsymbol{\omega}_4$.

4.3. Computation of the density function $f_U(u)$

Efficient and numerically stable algorithms for the computation of the density function $f_U(u)$ are discussed in Section 4.3.

Computing the probability density function of U using equation (1) requires the evaluation of $e^{T_1 u}$. The presence of negative diagonal elements in T_1 makes the direct computation of $e^{T_1 u}$ numerically hazardous and is not recommended. For computational stability and error control, the recommended method is *uniformization* (Latouche and Ramaswami [15]), where computations are performed in terms of a corresponding discrete time Markov chain, embedded in a Poisson process with rate τ equal to the absolute value of the most negative diagonal element in T_1. Let the matrix K be defined as,

$$K = \frac{1}{\tau} S_1^* + I = \begin{bmatrix} \frac{1}{\tau} T_1 + I & \frac{1}{\tau}\mathbf{T}_1^0 \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} P & \mathbf{P}^0 \\ \mathbf{0} & 1 \end{bmatrix},$$

where S_1^* is as defined in Figure 3. We then have [15],

$$e^{T_1 u} \;=\; \sum_{k=0}^{\infty} e^{-\tau u} \frac{(\tau u)^k}{k!} P^k.$$

The density functions can be expressed as,

$$f_U(u) \;= \mathbf{y}\, e^{T_1 u}\, \mathbf{T}_1^0 = \sum_{k=0}^{\infty} e^{-\tau u} \frac{(\tau u)^k}{k!}\, \mathbf{y}\, P^k\, \mathbf{T}_1^0.$$

This equation can be expressed as follows for efficient algorithmic implementation.

$$\begin{aligned}
f_U(0) &= \mathbf{y}\mathbf{T}_1^0, \\
f_U(u) &= \sum_{k=0}^{\infty} a_k \boldsymbol{\psi}_k \mathbf{T}_1^0 \ \text{ for } \ u > 0, \\
\boldsymbol{\psi}_0 &= \mathbf{y}, \ \text{ and } \ a_0 = e^{-\tau u} \\
\boldsymbol{\psi}_k &= \boldsymbol{\psi}_{k-1} P, \ \text{ and } \ a_k = a_{k-1}(\frac{\tau u}{k}), \ \text{ for } k = 1, 2, \cdots.
\end{aligned}$$

Only two vectors $\boldsymbol{\psi}_k$ need to be stored as they are calculated recursively and the scalar values of $f_U(u)$ are accumulated. In the present case, the matrix P and vector $\mathbf{P}^0$ are very sparse so that the computations can be organized efficiently without actually generating and storing P and $\mathbf{P}^0$.

Evaluation of $f_U(u)$ would require the truncation of an infinite series. The truncation point $k = k^*$, can be determined such that $\sum_{k=0}^{k^*} \frac{e^{-\tau u}(\tau u)^k}{k!} \leq 1 - \epsilon$. This will ensure that the overall error in the computation of $f_U(u)$ will be bounded by ϵ. Since K is a probability matrix, the elements of P, and its powers, are positive and are uniformly bounded by 1.

5. Summary of Numerical Results

This Section presents observations on system behavior based on extensive numerical results from the implementation of the algorithms developed in this paper. Sections 5.1 and 5.2 present details of the parameter values used in implementing the algorithms, and some notes on computations. Section 5.3 presents observations on the system behavior based on the numerical results.

5.1. Values of parameters used

In order to limit a potentially large parameter space, we decided to keep the ratios $\frac{\lambda_1}{\theta_1}$ and $\frac{\lambda_2}{\theta_2}$ fixed at 10. Values of λ_1 and λ_2 are chosen from 1, 5, and 10, and the values of θ_1 and θ_2 are computed using the fixed ratio. This implies that, on average, each machine is operational 90.91% ($= \frac{10}{10+1}$) and does not represent a significant loss of generality because decreasing [increasing] the failure rate and increasing [decreasing] the repair rate have the same effect on the *design production rates* ϕ_1 and ϕ_2. With a fixed ratio of $\frac{\lambda_1}{\theta_1}$ [$\frac{\lambda_2}{\theta_2}$], larger values of λ_1 [λ_2] imply that stage 1 [stage 2] machines fail more often and are repaired more quickly, relative to smaller values of λ_1 [λ_2] with *design production rates* ϕ_1 and ϕ_2 remaining unaffected.

To enable easy distinction and comparison of balanced and unbalanced flow lines, we specify ϕ_1 and ϕ_2 (each chosen from the five values of 80, 90, 100, 110, and 120) instead of specifying μ_1 and μ_2. $\phi_1 = \phi_2$ yields a *balanced* flow line, and $\phi_1 > \phi_2$ [$\phi_1 < \phi_2$] yields an *unbalanced* flow line with higher [lower] production rate at stage 1. Values of N_1 and N_2 are chosen from the four values of 1, 2, 5, and 10, and μ_1 and μ_2 are computed using the relationships $\phi_1 = N_1\mu_1\left(\frac{\theta_1}{\lambda_1+\theta_1}\right)$, and $\phi_2 = N_2\mu_2\left(\frac{\theta_2}{\lambda_2+\theta_2}\right)$. Values of M are chosen from the eight values of 0, 1, 2, 3, 5, 10, 20, and 50.

A representative subset of results are summarized in Tables 2-5 and Figures 7-16. For simplicity in presentation, Tables 2 and 3 present results for systems with $N_1 = N_2$ and $\lambda_1 = \lambda_2$. Additional results for systems with $N_1 \neq N_2$ and $\lambda_1 \neq \lambda_2$ are presented in Tables 4 and 5. Summary statements describing the effects of various parameters on the system behavior are based on the understanding gained from the complete set of runs.

Table 2. Departure Rate and Average Buffer contents.
$(N_1 = N_2 = N, M = 5, \lambda_1 = \lambda_2 = 1, \theta_1 = \theta_2 = 10)$

	Departure Rate (ϕ_a)				E(Z_{bb} = Number in buffer)			
M ↓	$N = 1$	$N = 2$	$N = 5$	$N = 10$	$N = 1$	$N = 2$	$N = 5$	$N = 10$
$\phi_1 = 100$ and $\phi_2 = 100$								
0	64.879	73.983	82.882	87.707	0.000	0.000	0.000	0.000
1	72.114	78.422	85.076	88.936	0.500	0.500	0.500	0.500
2	76.443	81.341	86.717	89.927	1.000	1.000	1.000	1.000
3	79.355	83.434	87.998	90.745	1.500	1.500	1.500	1.500
5	83.088	86.292	89.890	92.025	2.500	2.500	2.500	2.500
10	87.749	90.158	92.674	94.050	5.000	5.000	5.000	5.000
20	91.661	93.556	95.242	96.044	10.000	10.000	10.000	10.000
50	95.596	96.800	97.675	98.026	25.000	25.000	25.000	25.000
$\phi_1 = 100$ and $\phi_2 = 110$								
0	67.916	77.392	86.575	91.467	0.000	0.000	0.000	0.000
1	75.433	81.979	88.807	92.684	0.456	0.442	0.413	0.380
2	79.904	84.976	90.460	93.653	0.895	0.869	0.810	0.743
3	82.892	87.108	91.739	94.443	1.322	1.283	1.193	1.093
5	86.682	89.986	93.598	95.652	2.138	2.077	1.921	1.750
10	91.306	93.784	96.229	97.465	4.011	3.890	3.540	3.178
20	95.044	96.936	98.410	99.002	7.285	6.926	6.042	5.262
50	98.503	99.406	99.822	99.917	14.737	12.577	9.591	7.767
$\phi_1 = 110$ and $\phi_2 = 100$								
0	67.916	77.392	86.575	91.467	0.000	0.000	0.000	0.000
1	75.433	81.979	88.807	92.684	0.545	0.558	0.587	0.621
2	79.904	84.976	90.460	93.653	1.105	1.131	1.190	1.257
3	82.892	87.108	91.739	94.443	1.678	1.717	1.807	1.907
5	86.682	89.986	93.598	95.652	2.862	2.923	3.079	3.250
10	91.306	93.784	96.229	97.465	5.989	6.110	6.460	6.822
20	95.044	96.936	98.410	99.002	12.715	13.074	13.958	14.738
50	98.503	99.406	99.822	99.917	35.263	37.423	40.409	42.233

Table 3. Interdeparture time properties.
$(N_1 = N_2 = N, M = 5, \lambda_1 = \lambda_2 = 1, \theta_1 = \theta_2 = 10)$

	Coefficient of variation (CV_u)				Autocorrelation at lag 1 (r_1)			
M↓	$N = 1$	$N = 2$	$N = 5$	$N = 10$	$N = 1$	$N = 2$	$N = 5$	$N = 10$
$\phi_1 = 100$ and $\phi_2 = 100$								
0	***1.500***	1.022	0.988	0.992	-0.048	-0.018	-0.015	-0.009
1	***1.629***	1.055	0.996	0.995	-0.023	0.008	-0.006	-0.006
2	***1.696***	1.074	1.002	0.998	-0.012	0.022	0.000	-0.003
3	***1.736***	1.085	1.006	0.999	-0.007	0.030	0.004	-0.001
5	***1.778***	1.099	1.011	1.002	-0.003	0.039	0.008	0.001
10	***1.803***	1.110	1.017	1.005	0.001	0.046	0.014	0.004
20	***1.789***	1.113	1.020	1.008	0.002	0.049	0.017	0.007
50	***1.745***	1.110	1.022	1.009	0.002	0.049	0.018	0.008
$\phi_1 = 100$ and $\phi_2 = 110$								
0	***1.524***	1.025	0.988	0.992	-0.047	-0.017	-0.014	-0.009
1	***1.656***	1.058	0.997	0.995	-0.022	0.009	-0.005	-0.005
2	***1.726***	1.077	1.002	0.998	-0.012	0.023	0.000	-0.003
3	***1.768***	1.089	1.006	1.000	-0.007	0.031	0.004	-0.001
5	***1.813***	1.102	1.011	1.002	-0.002	0.040	0.009	0.001
10	***1.848***	1.115	1.017	1.005	0.001	0.048	0.014	0.005
20	***1.851***	1.120	1.020	1.007	0.003	0.052	0.017	0.007
50	***1.835***	1.121	1.022	1.008	0.004	0.053	0.019	0.008
$\phi_1 = 110$ and $\phi_2 = 100$								
0	***1.523***	1.026	0.989	0.993	-0.046	-0.017	-0.013	-0.008
1	***1.653***	1.058	0.998	0.997	-0.021	0.009	-0.005	-0.004
2	***1.721***	1.077	1.003	0.999	-0.012	0.023	0.001	-0.002
3	***1.760***	1.088	1.007	1.001	-0.007	0.031	0.005	0.000
5	***1.798***	1.101	1.012	1.003	-0.002	0.039	0.009	0.002
10	***1.811***	1.110	1.018	1.007	0.001	0.046	0.015	0.006
20	***1.778***	1.111	1.021	1.009	0.002	0.048	0.017	0.008
50	***1.712***	1.106	1.023	1.010	0.001	0.047	0.019	0.009

Table 4. Effect of N_1 and N_2.
$M = 5$, $\lambda_1 = \lambda_2 = 1$, $\phi_1 + \phi_2 = 200$

(N_1, N_2)	$\phi_1 = 90, \phi_2 = 110$				$\phi_1 = 100, \phi_2 = 100$				$\phi_1 = 110, \phi_2 = 90$			
$\downarrow$	ϕ_a	$E(Z_{bb})$	CV_u	r_1	ϕ_a	$E(Z_{bb})$	CV_u	r_1	ϕ_a	$E(Z_{bb})$	CV_u	r_1
(1, 1)	80.989	1.761	***1.769***	-0.002	83.088	2.500	***1.778***	-0.003	80.989	3.240	***1.739***	-0.002
(1, 2)	82.911	1.638	***1.448***	0.032	84.697	2.444	***1.402***	0.036	82.062	3.232	***1.333***	0.040
(2, 1)	82.062	1.768	***1.484***	0.006	84.697	2.556	***1.537***	0.001	82.911	3.362	***1.555***	-0.001
(2, 2)	83.943	1.639	1.099	0.040	86.292	2.500	1.099	0.039	83.943	3.361	1.096	0.039
(2, 5)	85.763	1.308	1.047	0.032	88.172	2.298	1.034	0.023	85.235	3.289	1.026	0.018
(5, 2)	85.235	1.711	1.058	0.024	88.172	2.702	1.070	0.029	85.763	3.692	1.082	0.035
(5, 5)	86.830	1.349	1.012	0.010	89.890	2.500	1.011	0.008	86.830	3.651	1.015	0.011
(2, 10)	86.889	1.002	1.030	0.025	89.497	2.085	1.018	0.014	86.138	3.203	1.011	0.009
(10, 2)	86.138	1.798	1.055	0.021	89.497	2.915	1.072	0.029	86.889	3.998	1.086	0.037
(10, 10)	88.261	1.061	1.003	0.003	92.025	2.500	1.002	0.001	88.261	3.940	1.006	0.004

Table 5. Effect of λ_1 and λ_2.
$(N_1 = 2, N_2 = 2, M = 5, \lambda_1/\theta_1 = \lambda_2/\theta_2 = 10)$

	ϕ_1=90, ϕ_2=110			ϕ_1=100, ϕ_2=100			ϕ_1=100, ϕ_2=90		
	$\lambda_1 = 1$	$\lambda_1 = 5$	$\lambda_1 = 10$	$\lambda_1 = 1$	$\lambda_1 = 5$	$\lambda_1 = 10$	$\lambda_1 = 1$	$\lambda_1 = 5$	$\lambda_1 = 10$
ϕ_a									
λ_2= 1	83.943	84.386	84.519	86.292	87.011	87.213	83.943	84.733	84.933
λ_2= 5	84.733	85.237	85.390	87.011	87.829	88.054	84.386	85.237	85.444
λ_2= 10	84.933	85.444	85.600	87.213	88.054	88.284	84.519	85.390	85.600
CV_u									
λ_2= 1	1.099	1.058	1.050	1.099	1.068	1.063	1.096	1.078	1.076
λ_2= 5	1.075	1.031	1.024	1.063	1.030	1.026	1.053	1.033	1.031
λ_2= 10	1.067	1.023	1.015	1.052	1.018	1.014	1.039	1.019	1.017
r_1									
λ_2= 1	0.040	0.020	0.016	0.039	0.025	0.022	0.039	0.031	0.031
λ_2= 5	0.031	0.009	0.004	0.022	0.006	0.003	0.016	0.007	0.006
λ_2= 10	0.027	0.004	-0.001	0.016	-0.001	-0.004	0.008	-0.001	-0.003

5.2. Notes on computations

The aggregation/disaggregation approach used in the iterative solutions dramatically reduced the number of iterations relative to direct iterations in most cases. The ratio of the number of iterations required to achieve the desired degree of convergence with direct iterations to the corresponding number of iterations using the aggregation/disaggregation approach, ranged from approximately 1 (indicating almost no improvement) to 151.43 (reduction of number of iterations from 4543 to 30). The average ratio for all the runs was 43.39, indicating the significant computational efficiency due to the use of the aggregation/disaggregation method. The ratio was higher for smaller values of M and decreased with increasing value of M. Convergence of the iterative methods for computing ω_1, ω_2, ω_3, and ω_4 was faster than the convergence for vector $\mathbf{x}$. This was not surprising because in these iterations, probability mass flows only in one direction.

5.3. Description of system behavior

Since the focus of this paper is on the departure process, the steady state system behavior is described in terms of the mean (μ_u), the coefficient of variation (CV_u), the lag 1 autocorrelation (r_1), and the density function of interdeparture times within the chosen ranges of parameter values. Sections 5.3.1 and 5.3.2 summarize the effects of various system parameters on the departure rate, the average buffer contents, the coefficient of variation, and lag 1 autocorrelation of the interdeparture time. Section 5.3.3 discusses the interdeparture time density function.

5.3.1. Throughput rate and average buffer contents

Table 2 provides a comprehensive summary of the effects of M, N_1 and N_2 on ϕ_a and $E(Z_{bb})$, for balanced ($\phi_1 = \phi_2 = 100$) and unbalanced systems ($\phi_1 = 100$ and $\phi_2 = 110$; and $\phi_1 = 110$ and $\phi_2 = 100$). As M increases, throughput rate (ϕ_a) increases, gradually approaching ϕ. This is expected because the buffer decouples the two stages of manufacture and increases the throughput rate by reducing the likelihood of *blocking* in stage 1 and *starving* in stage 2. The marginal improvement in ϕ_a decreased with increasing M, with most of the benefit obtained for values of M under 5. Increasing N_1 and/or N_2 also resulted in similar improvement in ϕ_a, because splitting the capacity among several smaller capacity machines moderates the inherent randomness of the production process, resulting in an increased throughput rate. As with M, the marginal benefit of increasing N_1 and N_2 decreases with increasing values of N_1 and N_2, with most of the benefit obtained when N_1 and N_2 are increased from 1 to 2. Improvements in ϕ_a due to increasing M, N_1, and N_2 are additive.

Additional details on the influence of N_1 and N_2 individually are provided in Table 4 and Figures 7-10. Increasing either N_1 or N_2 resulted in increases in throughput rate, with progressively decreasing marginal improvement. For unbalanced systems, increasing the number of machines at the non-bottleneck stage yielded marginally better improvement in throughput rate compared to corresponding increase at the bottleneck stage. For a given total production capacity between the two stages of production, balanced flow lines have a slightly higher throughput rate (Table 2), because the upper bound on the throughput rate is

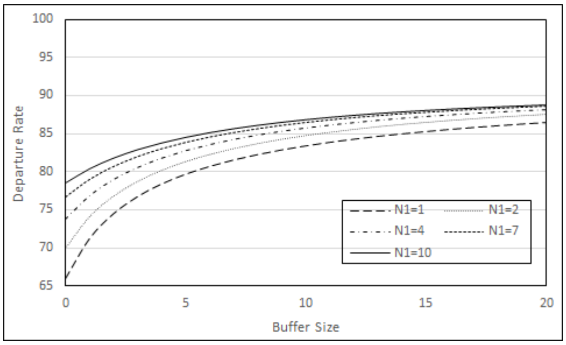

Figure 7. Effect of M and N_1 on ϕ_a, $\phi_1 = 100, \phi_2 = 90, N_2 = 2$.

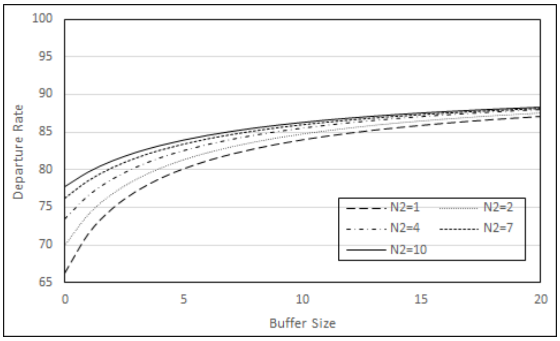

Figure 8. Effect of M and N_2 on ϕ_a, $\phi_1 = 90, \phi_2 = 100, N_2 = 2$

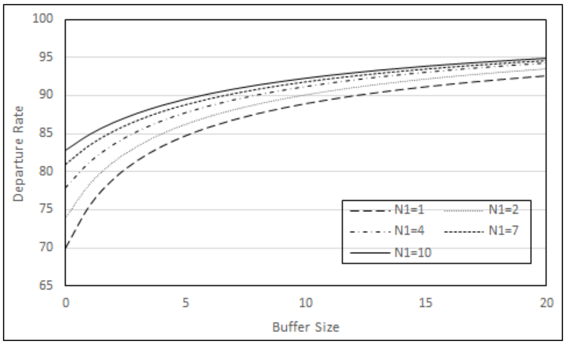

Figure 9. Effect of M and N_1 on ϕ_a, $\lambda_1 = \lambda_2 = 1, \theta_1 = \theta_2 = 10$, $\phi_1 = 100, \phi_2 = 100, N_2 = 2$.

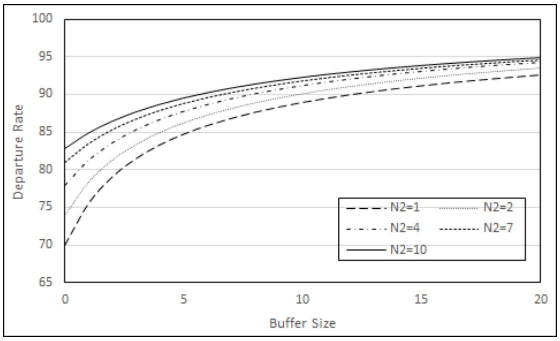

Figure 10. Effect of M and N_2 on ϕ_a, $\lambda_1 = \lambda_2 = 1, \theta_1 = \theta_2 = 10$, $\phi_1 = 100, \phi_2 = 100, N_1 = 2$.

determined by the bottleneck stage in the flow line.

For the two unbalanced flow lines displayed in Table 2 (with values of ϕ_1 and ϕ_2 reversed), throughput rates were identical, even though the values of $E(Z_{bb})$ (and CV_u and r_1 in Table 3) are different. This outcome is the result of the choice of $\lambda_1 = \lambda_2$. When $\lambda_1 \neq \lambda_2$ (and ϕ_1 and ϕ_2 values are reversed), throughput rates were different (very slightly) for the two unbalanced systems (Table 5). Larger λ for the non-bottleneck stage had a slightly larger impact than a corresponding value for λ for the bottleneck stage.

The probability distribution of number in buffer (Z_{bb}) varied predictably. When $\phi_1 > \phi_2$ probability mass is concentrated at the higher buffer values, resulting in a smaller probability of starving, larger probability of blocking, and $E(Z_{bb}) > 0.5M$. When $\phi_1 < \phi_2$, more probability mass is concentrated at the lower values of M resulting in smaller probability of blocking, larger probability of starving, and $E(Z_{bb}) < 0.5M$. For balanced systems, the probability mass is evenly distributed and $E(Z_{bb}) \approx 0.5M$. In the interests of brevity, probability distributions of Z_{bb} are not presented in this paper.

5.3.2. Coefficient of variation and lag 1 autocorrelation of interdeparture time

In all cases, except when $N_1 = 1$ and/or $N_2 = 1$, the coefficient of variation of the interdeparture time (CV_u) is very close to 1. When $N_1 = N_2 = 1$ (Table 3) or when $N_1 = 1$

or $N_2 = 1$ (Table 4), (displayed in ***bold italics***), CV_u ranged between 1.5 to 1.8. CV_u decreased dramatically to a value close to 1 when N_1 and N_2 increased to 2. Further increases in N_1 or N_2 resulted only in small reductions in CV_u. Increasing M did not have a significant effect on CV_u.

When $\lambda_1 \neq \lambda_2$, the coefficient of variation (CV_u) and the lag 1 autocorrelation (r_1) did not differ significantly relative to systems with $\lambda_1 = \lambda_2$ (Table 5). Larger λ for the bottleneck stage had a slightly larger impact than a corresponding increase in λ for the non-bottleneck stage.

For large values of M, a small increase in CV_u is observed (Table 3). Detailed analysis (not presented here for brevity) indicated that while both μ_u and Var_u decreased with increasing M, the rate of decrease in μ_u at large values of M was smaller than corresponding decrease in Var_u, resulting in the small increase in CV_u.

For most of the parameter space considered, the lag 1 autocorrelation of the inter-departure time are positive and very close to zero. For smaller values of M, some negative autocorrelations are observed. Given the extremely small magnitude, this observation is not of much practical significance.

5.3.3. Interdeparture time density function

When N_1 and N_2 are greater than 1, lag 1 autocorrelation of the interdeparture times are very close to zero, and the values of CV_u are very close to one, suggesting an approximation of the output process by a Poisson process.

Interdeparture time density functions were graphed for a wide range of parameter values, and visually compared with the exponential density function with a rate equal to ϕ_a. Except when N_1 or N_2 are equal to 1, there is a close agreement between the interdeparture time density function and the corresponding exponential density function. This approximation gets better as the values of N_1 and N_2 increase. A small subset of the graphs analyzed are presented in Figures 11-16, where the horizontal scale describing the interdeparture time is displayed only up to 0.05 to highlight the contract between the two density functions. This means that for systems with the numbers of machines at each stage as low as 2, the output process can be effectively approximated by a Poisson process. This observation provides a strong justification for several earlier successful approximations of flow line models based on decomposition methods.

6. Concluding Remarks

This paper considers a two-stage flow line with multiple unreliable machines at each stage and finite intermediate buffer and develop algorithms to fully characterize the departure process in terms of mean, variance, density function, and lag 1 autocorrelation of the interdeparture times. The following are some key observations in the paper which offer managerial insight in evaluating economic implications at the design and operational stages.

- The output process is a very close approximation to a Poisson process for buffer sizes as small as 1, and the number of machines at each stage as small as 2. This is a very meaningful and valuable observation in the study of flow lines with several stages

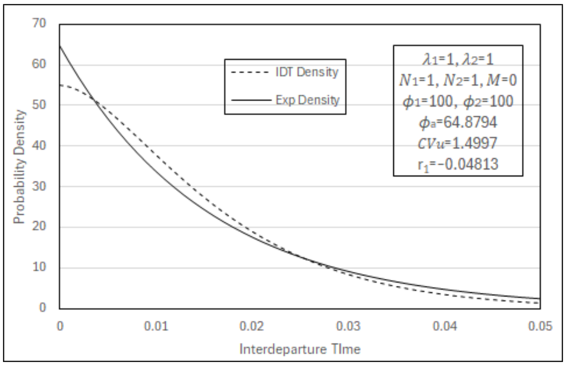

Figure 11. IDT Density Function 1.

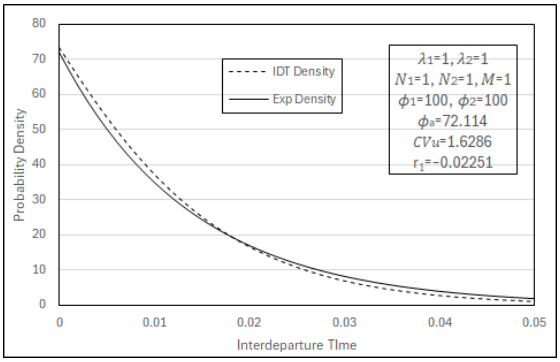

Figure 12. IDT Density Function 2.

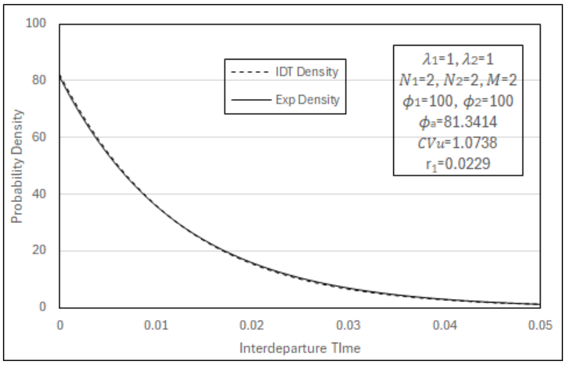

Figure 13. IDT Density Function 3.

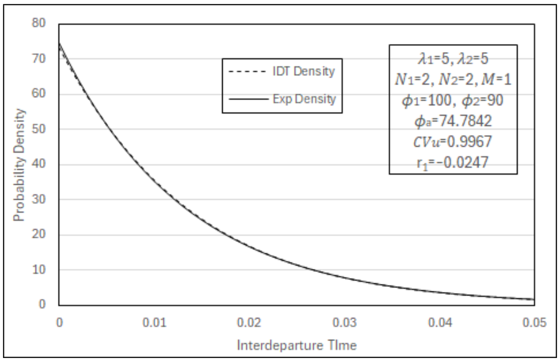

Figure 14. IDT Density Function 4.

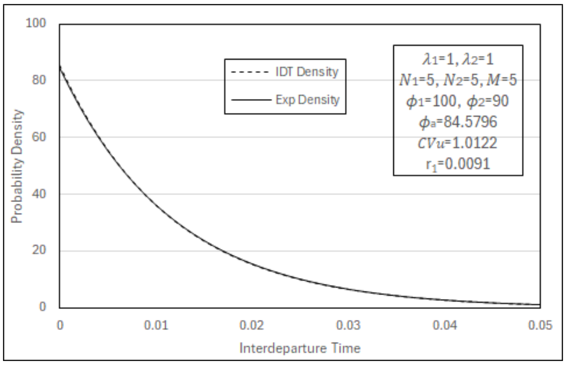

Figure 15. IDT Density Function 5.

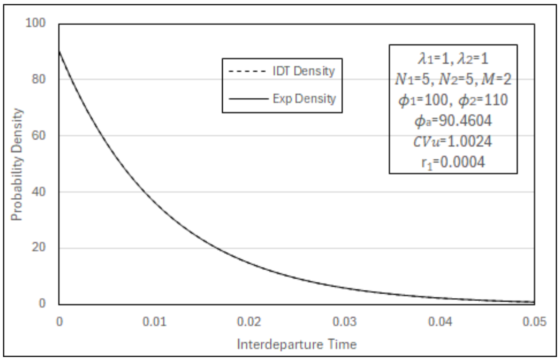

Figure 16. IDT Density Function 6.

where output from one stage forms the input to the next stage.

- The benefit due to additional buffer capacity between two stages drops very quickly after the initial few spaces. This observation is valuable in allocating limited buffer capacity among several stages of a multistage manufacturing process.
- For a given capacity at any stage, having several machines rather than a single machine with the same total capacity at either stage increases throughput rate and reduces process variability. Splitting the capacity among several machines at stage 2 yields marginally better improvement in system performance than a similar change in stage 1. This insight is particularly useful in designing flow lines.

The algorithmic methodology developed in this paper is easy to implement and fully exploits the special structure of the model. It also permits easy incorporation of minor variations to the basic model. The following subsections present two examples.

6.1. Limited repair resources

The model presented in this paper assumes ample repair capacity at each stage of production so that repair of failed machines starts immediately after failure and with no waiting. The methodology proposed in this paper can be adapted to systems with limited repair resources, as long as each stage of production has dedicated repair facilities. Modification to the model when only one machine at each stage can be repaired at a time is presented below. Extension to systems where more than one machine can be repaired at a time (at each stage) will be obvious from the following. Considering the submatrices for the example with $N_1 = 2$ and $N_2 = 2$, shown in the Appendix, Matrix A_1 needs to be modified, as shown in Figure 17. Matrices A_0 and A_2 are unaffected. Similar adjustments need to be made for matrices at boundary level 0_s and M_b.

$$A_1 = \begin{array}{r||ccc|ccc|ccc||}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline
(0,0)\ 1 & * & \theta_2 & & \theta_1 & & & & & \\
(0,1)\ 2 & \lambda_2 & * & \theta_2 & & \theta_1 & & & & \\
(0,2)\ 3 & & 2\lambda_2 & * & & & \theta_1 & & & \\
\hline
(1,0)\ 4 & \lambda_1 & & & * & \theta_2 & & \theta_1 & & \\
(1,1)\ 5 & & \lambda_1 & & \lambda_2 & * & \theta_2 & & \theta_1 & \\
(1,2)\ 6 & & & \lambda_1 & & 2\lambda_2 & * & & & \theta_1 \\
\hline
(2,0)\ 7 & & & & 2\lambda_1 & & & * & \theta_2 & \\
(2,1)\ 8 & & & & & 2\lambda_1 & & \lambda_2 & * & \theta_2 \\
(2,2)\ 9 & & & & & & 2\lambda_1 & & 2\lambda_2 & *
\end{array}$$

Figure 17. Matrix A_1 when repair resources are limited.

6.2. Scrapping of units

The model presented in this paper assumes that when a machine fails, the unit in-process remains on the machine, and when the machine is restored to operating condition, the machine will continue processing the item that was in process at the time of failure. Shin and Moon [26] developed approximations for the throughput rate with phase-type processing times and three possible service-failure interactions, namely, resume interrupted service; restart service as a new unit; or scrap the unit. With exponential service times, the first two interactions will be identical. A more general version of the third type of interaction was considered by Shanthikumar and Tien [22], where the unit in-process at the time of failure is scrapped with a certain probability. This variation can easily be incorporated into the model. Let a be the probability that a unit in service at the time of failure needs be scrapped and let $b = 1 - a$ be the probability with which the unit need not be scrapped. Because service times are exponential, this change will have no effect on the state of the system for failures at stage 1. In stage 2, when a failed machine is restored to operating condition, the buffer contents will decrease by 1 with probability a and remain the same with probability b. For the example matrices shown in the Appendix, this change can be easily incorporated by modifying matrices A_1 and A_2 as shown in Figures 18 and 19. Matrix A_0 remains the same. Similar adjustments need to be made for matrices at boundary level 0_s and M_b.

$$
A_1 = \begin{array}{r||ccc|ccc|ccc||}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\text{(0,0) 1} & * & 2b\theta_2 & & 2\theta_1 & & & & & \\
\text{(0,1) 2} & \lambda_2 & * & b\theta_2 & & 2\theta_1 & & & & \\
\text{(0,2) 3} & & 2\lambda_2 & * & & & 2\theta_1 & & & \\
\hline
\text{(1,0) 4} & \lambda_1 & & & * & 2b\theta_2 & & \theta_1 & & \\
\text{(1,1) 5} & & \lambda_1 & & \lambda_2 & * & b\theta_2 & & \theta_1 & \\
\text{(1,2) 6} & & & \lambda_1 & & 2\lambda_2 & * & & & \theta_1 \\
\hline
\text{(2,0) 7} & & & & 2\lambda_1 & & & * & 2b\theta_2 & \\
\text{(2,1) 8} & & & & & 2\lambda_1 & & \lambda_2 & * & b\theta_2 \\
\text{(2,2) 9} & & & & & & 2\lambda_1 & & 2\lambda_2 & *
\end{array}
$$

Figure 18. Matrix A_1 with scrapping of units at stage 2.

6.3. Implications of model assumptions to real-world applications

One of the key assumptions in this and many other papers on the topic, is the exponential distribution of the processing times of jobs and operating and repair times of machines. This assumption is motivated by the Markovian property and the resulting analytical tractability offered by exponential distribution.

Case studies from the automotive industry (Inman [14], Colledani, Ekvall, Lundholm, Moriggi, Polato, and Tolio [5]) and water bottling production lines (Assaf [2]) suggest that exponential assumption is reasonable for times between failures and times to repair. The

$$A_2 = \begin{array}{r||ccc|ccc|ccc||}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline
(0,0)\ 1 & 0 & 2a\theta_2 & & & & & & & \\
(0,1)\ 2 & & \mu_2 & a\theta_2 & & & & & & \\
(0,2)\ 3 & & & 2\mu_2 & & & & & & \\
\hline
(1,0)\ 4 & & & & 0 & 2a\theta_2 & & & & \\
(1,1)\ 5 & & & & & \mu_2 & a\theta_2 & & & \\
(1,2)\ 6 & & & & & & 2\mu_2 & & & \\
\hline
(2,0)\ 7 & & & & & & & 0 & 2a\theta_2 & \\
(2,1)\ 8 & & & & & & & & \mu_2 & a\theta_2 \\
(2,2)\ 9 & & & & & & & & & 2\mu_2
\end{array}$$

Figure 19. Matrix A_2 with scrapping of units at stage 2.

evidence, however, is not so conclusive for processing times. In many fully automatic flow lines, processing times have a finite minimum with very small allowance for variation and are therefore largely deterministic. There are, however, situations where flow line processing times are more appropriately modeled as random variables. One example of potential variability in processing times is a flow line employed in producing customized jobs (e.g., automotive flow line where two and four door models are manufactured on the same line). Another example is the *process drift* discussed by Chincholkar and Herrmann [4], where process parameters degrade over time, requiring additional inspections and rework, leading to variability in processing times. In some manufacturing processes, products make several (random number of) passes through work centers or *rework loops* (e.g., painting shop in an automotive facility) adding to the variability of processing time (Li [16]). Empirical studies (Inman [14], Colledani, Ekvall, Lundholm, Moriggi, Polato, and Tolio [5]) show that exponential distribution is often a poor fit for processing times. Despite this observation, Inman [14] states that models using this assumption do not necessarily lead to inaccurate results. The user must exercise caution and use the results after evaluating the sensitivity of assumptions. Erlang (sum of exponential variables yielding coefficients of variation < 1) and hyperexponential (linear combination of exponential variables yielding coefficients of variation > 1) distributions, which retain some analytical tractability, offer reasonable alternatives. This is supported by [14] and [5], where Erlang distribution appears to fit some processing times well.

Acknowledgment

We are very grateful to the referees and the associate editor for their painstaking review of the manuscript and for their meaningful comments. The manuscript has been greatly improved as a result of their suggestions.

References

[1] Altiok, T., & Melamed, B. (2001). The case for modeling correlation in manufacturing systems. *IIE Transactions*, 33(9), 779–791.

[2] Assaf, R. (2012). Analysis of the output variability in multi-stage manufacturing systems. Ph. D. Thesis, Politecnico di Milano, Department of Mechanical Engineering, Milano, Italy.

[3] Betterton, C. E., & Silver, E. J. (2012). Detecting bottlenecks in serial production lines – a focus on interdeparture time variance. *International Journal of Production Research*, 50, 4158-–4174.

[4] Chincholkar, M., & Herrmann, J. W. (2008). Estimating manufacturing cycle time and throughput in flow shops with process drift and inspection. *International Journal of Production Research*, 46, 7057–7072.

[5] Colledani, M., Ekvall, M., Lundholm, T., Moriggi, P., Polato, A., & Tolio, T. (2010). Analytical methods to support continuous improvements at Scania. *International Journal of Production Research*, 48, 1913–1945.

[6] Dallery, Y., & Gershwin, S. B. (1992). Manufacturing flow line systems: A review of models and analytical results. *Queueing Systems*, 12, 3–94.

[7] Dincer, C., & Deler, B. (2000). On the distribution of throughput of transfer lines. *Journal of the Operational Research Society*, 51, 1170-1178.

[8] Dizbin, N. M., & Tan, B. (2019). Modelling and analysis of the impact of correlated inter-event data on production control using Markovian arrival processes. *Flexible Services and Manufacturing Journal*, 31, 1042–1076.

[9] Gershwin, S. B. (1993). Variance of Output of a tandem production system. *Queueing Networks with Finite Capacity*, 291–304.

[10] He, X. F., Wu, S., & Li, Q. L. (2007). Production Variability of Production Lines. *International Journal of Production Economics*, 107, 78–87.

[11] Hendricks, K. B. (1992). The output processes of serial production lines of exponential machines with finite buffers. *Operations Research*, 40, 1139–1147.

[12] Hendricks, K. B., & McClain, J. O. (1993). The output processes of serial production lines of general machines with finite buffers. *Management Science*, 39, 1194-1201.

[13] Heyman, D. P., & Goldsmith, M. J. (1995). Comparisons between aggregation/disaggregation and a direct algorithm for computing the stationary probabilities of a Markov chain. *ORSA Journal on Computing*, 7, 101–108.

[14] Inman, R. R. (1999). Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. *Production and Operations Management*, 8, 409 –- 432.

[15] Latouche, G., & Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Statistics and Applied Mathematics.

[16] Li, J. (2004). Throughput analysis in automotive paint shops: a case study. *IEEE Transactions on Automation Science and Engineering*, 1, 90-98.

[17] Li, J., Blumenfeld, D. E., Huang, N., & Alden J. M. (2009). Throughput analysis of production systems: recent advances and future topics. *International Journal of Production Research*, 47, 3823–3851.

[18] Liu J., Yang S., Wu A., S., & Hu J. (2012). Multi-state throughput analysis of a two-stage manufacturing system with parallel unreliable machines and a finite buffer. *European Journal of Operational Research*, 219, 296–304.

[19] Miltenburg, G. J. (1987). Variance of the number of units produced on a transfer line with buffer inventories during a period of length T. *Naval Research Logistics Quarterly*, 34, 811–822.

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

[21] Rao, B. M., & Posner, M. J. M. (1984). On the output process of an $M/M/1$ queue with randomly varying system parameters. *Operations Research Letters*, 3, 191–197.

[22] Shanthikumar, J. G., & Tien, C. C. (1983). An algorithmic solution to two-stage transfer lines with possible scrapping of units. *Management Science*, 29, 1069–1086.

[23] Shin, Y W., & Moon, D. H. (2016). Variability of output in two-node tandem production line. *Proceedings of 11th International Conference on Queueing Theory and Network Applications*, 1–4.

[24] Shin, Y W., & Moon, D. H. (2017a). Throughput of Flow Lines with Unreliable Parallel-Machine Workstations and Blocking. *Journal of Industrial and Management Optimization*, 13, 901–916.

[25] Shin, Y. W., & Moon, D. H. (2017b). Variance of Departure Process in Two-Node Tandem Queue with Unreliable Servers and Blocking. *Proceedings of 6th International Conference on Operations Research and Enterprise Systems*, 258–264.

[26] Shin, Y. W., & Moon, D. H. (2021). A unified approach for an approximation of tandem queues with failures and blocking under several types of service-failure interactions. *Computers and Operations Research*, 127, 105161.

[27] Tan, B. (1999). Variance of the output as a function of time: Production line dynamics. *European Journal of Operational Research*, 117, 470–484.

[28] Tan, B. (2000). Asymptotic variance rate of the output in production lines with finite buffers. *Annals of Operations Research*, 93, 385–403.

[29] Tan, B. (2013). Modeling and analysis of output variability in discrete material flow production systems. In *Handbook of Stochastic Models and Analysis of Manufacturing System Operations*, 287–311, New York, NY: Springer New York.

[30] Tan, B., & Lagershausen, S. (2017). On the output dynamics of production systems subject to blocking. *IISE Transactions*, 49, 268–284.

Appendix

A. Detailed Description of Submatrices of Q (Figure 1) ($N_1 = 2, N_2 = 2$)

$$A_0 = \begin{array}{r||ccc|ccc|ccc||}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
(0,0)\ 1 & 0 & & & & & & & & \\
(0,1)\ 2 & & 0 & & & & & & & \\
(0,2)\ 3 & & & 0 & & & & & & \\
\hline
(1,0)\ 4 & & & & \mu_1 & & & & & \\
(1,1)\ 5 & & & & & \mu_1 & & & & \\
(1,2)\ 6 & & & & & & \mu_1 & & & \\
\hline
(2,0)\ 7 & & & & & & & 2\mu_1 & & \\
(2,1)\ 8 & & & & & & & & 2\mu_1 & \\
(2,2)\ 9 & & & & & & & & & 2\mu_1
\end{array}$$

Figure A1. Matrix A_0.

$$A_1 = \begin{array}{r||ccc|ccc|ccc||}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
(0,0)\ 1 & * & 2\theta_2 & & 2\theta_1 & & & & & \\
(0,1)\ 2 & \lambda_2 & * & \theta_2 & & 2\theta_1 & & & & \\
(0,2)\ 3 & & 2\lambda_2 & * & & & 2\theta_1 & & & \\
\hline
(1,0)\ 4 & \lambda_1 & & & * & 2\theta_2 & & \theta_1 & & \\
(1,1)\ 5 & & \lambda_1 & & \lambda_2 & * & \theta_2 & & \theta_1 & \\
(1,2)\ 6 & & & \lambda_1 & & 2\lambda_2 & * & & & \theta_1 \\
\hline
(2,0)\ 7 & & & & 2\lambda_1 & & & * & 2\theta_2 & \\
(2,1)\ 8 & & & & & 2\lambda_1 & & \lambda_2 & * & \theta_2 \\
(2,2)\ 9 & & & & & & 2\lambda_1 & & 2\lambda_2 & *
\end{array}$$

Figure A2. Matrix A_1.

$$A_2 = \begin{array}{r||ccc|ccc|ccc||}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\text{(0,0) 1} & 0 & & & & & & & & \\
\text{(0,1) 2} & & \mu_2 & & & & & & & \\
\text{(0,2) 3} & & & 2\mu_2 & & & & & & \\
\hline
\text{(1,0) 4} & & & & 0 & & & & & \\
\text{(1,1) 5} & & & & & \mu_2 & & & & \\
\text{(1,2) 6} & & & & & & 2\mu_2 & & & \\
\hline
\text{(2,0) 7} & & & & & & & 0 & & \\
\text{(2,1) 8} & & & & & & & & \mu_2 & \\
\text{(2,2) 9} & & & & & & & & & 2\mu_2
\end{array}$$

Figure A3. Matrix A_2.

$$B_0 = \begin{array}{r||ccc|ccc|ccc||}
 & \text{(0,0)} & \text{(0,1)} & \text{(0,2)} & \text{(1,0)} & \text{(1,1)} & \text{(1,2)} & \text{(2,0)} & \text{(2,1)} & \text{(2,2)} \\
\text{(0,11):1} & & & & & & & & & \\
\text{(0,21):2} & & & & & & & & & \\
\text{(0,22):3} & & & & & & & & & \\
\hline
\text{(1,11):4} & & & & & \mu_1 & & & & \\
\text{(1,21):5} & & & & & & \mu_1 & & & \\
\text{(1,22):6} & & & & & & & & & \\
\hline
\text{(2,11):7} & & & & & & & & 2\mu_1 & \\
\text{(2,21):8} & & & & & & & & & 2\mu_1 \\
\text{(2,22):9} & & & & & & & & &
\end{array}$$

Figure A4. Matrix B_0.

$$B_1 = \begin{array}{r||ccc|ccc|ccc||}
 & \text{(0,11)} & \text{(0,21)} & \text{(0,22)} & \text{(1,11)} & \text{(1,21)} & \text{(1,22)} & \text{(2,11)} & \text{(2,21)} & \text{(2,22)} \\
\text{(0,11):1} & * & & \theta_2 & 2\theta_1 & & & & & \\
\text{(0,21):2} & \lambda_2 & * & \mu_2 & & 2\theta_1 & & & & \\
\text{(0,22):3} & & & * & & & 2\theta_1 & & & \\
\hline
\text{(1,11):4} & \lambda_1 & & & * & & \theta_2 & \theta_1 & & \\
\text{(1,21):5} & & \lambda_1 & & \lambda_2 & * & \mu_2 & & \theta_1 & \\
\text{(1,22):6} & & & \lambda_1 & & \mu_1 & * & & & \theta_1 \\
\hline
\text{(2,11):7} & & & & 2\lambda_1 & & & * & & \theta_2 \\
\text{(2,21):8} & & & & & 2\lambda_1 & & \lambda_2 & * & \mu_2 \\
\text{(2,22):9} & & & & & & 2\lambda_1 & & 2\mu_1 & *
\end{array}$$

Figure A5. Matrix B_1.

$$B_2 = \begin{array}{c||ccc|ccc|ccc||}
 & (0,11) & (0,21) & (0,22) & (1,11) & (1,21) & (1,22) & (2,11) & (2,21) & (2,22) \\
\hline
(0,0){:}1 & & & & & & & & & \\
(0,1){:}2 & \mu_2 & & & & & & & & \\
(0,2){:}3 & & 2\mu_2 & & & & & & & \\
\hline
(1,0){:}4 & & & & & & & & & \\
(1,1){:}5 & & & & \mu_2 & & & & & \\
(1,2){:}6 & & & & & 2\mu_2 & & & & \\
\hline
(2,0){:}7 & & & & & & & & & \\
(2,1){:}8 & & & & & & & \mu_2 & & \\
(2,2){:}9 & & & & & & & & 2\mu_2 & \\
\end{array}$$

Figure A6. Matrix B_2.

$$C_0 = \begin{array}{c||ccc|ccc|ccc||}
 & (11,0) & (11,1) & (11,2) & (21,0) & (21,1) & (21,2) & (22,0) & (22,1) & (22,2) \\
\hline
(0,0){:}1 & & & & & & & & & \\
(0,1){:}2 & & & & & & & & & \\
(0,2){:}3 & & & & & & & & & \\
\hline
(1,0){:}4 & \mu_1 & & & & & & & & \\
(1,1){:}5 & & \mu_1 & & & & & & & \\
(1,2){:}6 & & & \mu_1 & & & & & & \\
\hline
(2,0){:}7 & & & & 2\mu_1 & & & & & \\
(2,1){:}8 & & & & & 2\mu_1 & & & & \\
(2,2){:}9 & & & & & & 2\mu_1 & & & \\
\end{array}$$

Figure A7. Matrix C_0.

$$C_1 = \begin{array}{c||ccc|ccc|ccc||}
 & (11,0) & (11,1) & (11,2) & (21,0) & (21,1) & (21,2) & (22,0) & (22,1) & (22,2) \\
\hline
(11,0){:}1 & * & 2\theta_2 & & \theta_1 & & & & & \\
(11,1){:}2 & \lambda_2 & * & \theta_2 & & \theta_1 & & & & \\
(11,2){:}3 & & 2\lambda_2 & * & & & \theta_1 & & & \\
\hline
(21,0){:}4 & \lambda_1 & & & * & 2\theta_2 & & \mu_1 & & \\
(21,1){:}5 & & \lambda_1 & & \lambda_2 & * & \theta_2 & & \mu_1 & \\
(21,2){:}6 & & & \lambda_1 & & 2\lambda_2 & * & & & \mu_1 \\
\hline
(22,0){:}7 & & & & & & & * & 2\theta_2 & \\
(22,1){:}8 & & & & & \mu_2 & & \lambda_2 & * & \theta_2 \\
(22,2){:}9 & & & & & & 2\mu_2 & & 2\lambda_2 & * \\
\end{array}$$

Figure A8. Matrix C_1.

$$C_2 = \begin{array}{r||ccc|ccc|ccc||}
 & (0,0) & (0,1) & (0,2) & (1,0) & (1,1) & (1,2) & (2,0) & (2,1) & (2,2) \\
(11,0){:}\ 1 & & & & & & & & & \\
(11,1){:}\ 2 & & & & & \mu_2 & & & & \\
(11,2){:}\ 3 & & & & & & 2\mu_2 & & & \\
\hline
(21,0){:}\ 4 & & & & & & & & & \\
(21,1){:}\ 5 & & & & & & & & \mu_2 & \\
(21,2){:}\ 6 & & & & & & & & & 2\mu_2 \\
\hline
(22,0){:}\ 7 & & & & & & & & & \\
(22,1){:}\ 8 & & & & & & & & & \\
(22,2){:}\ 9 & & & & & & & & &
\end{array}$$

Figure A9. Matrix C_2.