Performance Bounds for Queueing Networks and Scheduling Policies[*][†]

Sunil Kumar and P. R. Kumar[‡]

Abstract

Except for the class of queueing networks and scheduling policies admitting a product form solution for the steady–state distribution, little is known about the performance of such systems. For example, if the priority of a part depends on its class (e.g., the buffer that the part is located in), then there are no existing results on performance, or even stability. However, in most applications such as manufacturing systems, one has to choose a control or scheduling policy, i.e., a priority discipline, that optimizes a performance objective.

In this paper we introduce a new technique for obtaining upper and lower bounds on the performance of Markovian queueing networks and scheduling policies. Assuming stability, and examining the consequence of a steady–state for general quadratic forms, we obtain a set of linear equality constraints on the mean values of certain random variables that determine the performance of the system. Further, the conservation of time and material gives an augmenting set of linear equality and inequality constraints. Together, these allow us to bound the performance, either above or below, by solving a linear program.

We illustrate this technique on several typical problems of interest in manufacturing systems. For an open re–entrant line modeling a semiconductor plant, we plot a bound on the mean delay (called cycle–time) as a function of line loading. We show that the Last Buffer First Serve policy is almost optimal in light traffic. For another such line, we show that it dominates the First Buffer First Serve policy. For a set of open queueing networks, we compare our lower bounds with those obtained by another method of Ou and Wein. For a closed queueing network, we bracket the performance of all buffer priority policies, including the suggested priority policy of Harrison and Wein. We also study the asymptotic heavy traffic limits of the lower and upper bounds. For a manufacturing system with machine failures, we show how the performance changes with failure and repair rates. For systems with finite buffers, we show how to bound the

[*] Please address all correspondence to the second author.

[†] The research reported here has been supported in part by the National Science Foundation under Grants Nos. ECS-90-25007 and ECS-92-16487, and in part by the Joint Services Electronics Program under Contract No. N00014-90-J1270.

[‡] Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory, University of Illinois, 1308 W. Main Street, Urbana, IL 61801.

throughput. Finally, we illustrate the application of our method to $GI/GI/1$ queues. We obtain analytic bounds which improve upon Kingman's bound for $E_2/M/1$ queues.

1 Introduction

Except for a narrow class of systems admitting a product form solution, see Kelly [1], little is known concerning the performance of queueing networks and scheduling policies. Yet, it is important in many applications, such as manufacturing systems, to choose a control or scheduling policy that optimizes a performance measure, such as the mean number in the system. However, if the priority of a part under a scheduling policy depends on its buffer location (i.e., its class), as it most certainly should, then the system is not of product form type and little is known concerning performance, see [2]. Though a theory of stability of bursty deterministic models has recently been developed for the class of all Least Slack policies and some buffer priority policies, see [3], little is known concerning the performance of such policies.

In most such situations, it is doubtful that closed form solutions for the steady–state distribution will ever be determined. As a second best solution, it would be desirable to at least have upper and lower bounds on the performance measures of interest. We refer the reader to Ou and Wein [4] and Baccelli and Makowski [5] for some other approaches to obtaining bounds.

In this paper we propose a technique for obtaining such upper and lower bounds on performance. It is applicable to a broad class of Markovian queueing networks and scheduling policies. Briefly, assuming stability, we study the consequences of a steady–state for general quadratic forms. This yields a set of linear equalities satisfied by the means of the pairwise products of certain random variables. Additionally, from the conservation of time and material, we obtain supplementing sets of linear equalities and inequalities. Together, these constraints allows one to bound performance measures, either above or below, by solving a linear program.

This technique can be regarded as an extension of the idea of Meyn and Down [6] and

Meyn, described in [2], where the square of the workload is studied. In prior work, Bertsimas, Paschalidis and Tstisiklis [7] have extended this by considering general "potential" functions i.e., several Lyapunov functions. The method of Section 3 is the same as the "non–parametric method" of Section 4.2 of Bertsimas, Paschalidis and Tstisiklis [8]; both were obtained simultaneously and independently.

We illustrate our approach on several typical situations of interest in manufacturing systems in general, and semiconductor manufacturing in particular. For an open re–entrant line modeling a semiconductor manufacturing plant, we plot a lower bound on the so called "actual–to–theoretical" ratio, Mean delay/Mean total processing time, as a function of line loading. We also compare our bounds with the bounds obtained by Ou and Wein [4], on three of their systems. For an open re-entrant line, we show that the Last Buffer First Serve (LBFS) policy of [3] is almost optimal in light traffic. In the first example modeling a semiconductor plant, we show that the upper bound on delay under the LBFS policy is less than the lower bound on delay under the First Buffer First Serve (FBFS) policy. For a closed re–entrant line modeling the so called "closed loop" release policy of Wein [9] (see also Glassey and Resende [10]), we bracket the performance of all the buffer priority scheduling policies, including the one conjectured to be optimal in heavy traffic by Harrison and Wein [11]. For another closed queueing network, we compare our bounds on the throughput with the simulation results in [11]. We also study the behavior of the bounds in the heavy traffic limit. For a manufacturing system with machine failures, we show how the performance changes with the failure and repair rates. For a finite buffer system, we outline how one may bound the throughput. Finally, we illustrate the application of our method to $GI/GI/1$ queues. We obtain an analytic expression for a bound that is better than that in Kingman [12] for a generalization of $E_2/M/1$ queues.

By combining the various modules above, e.g., GI arrivals or services, machine failures, closed routes, etc., one can thus obtain bounds for general Markovian queueing networks.

2 The Basic Open Re–Entrant Line

Let us begin by considering an open re-entrant line, i.e., a Kelly network with one fixed route. In later sections, we show how this can be generalized.

The system consists of S machines $\{1, 2, \ldots, S\}$, see Figure 1. Parts enter the system according to a Poisson process of rate λ. They first visit machine $\sigma(1) \in \{1, 2, \ldots, S\}$, where they are stored in a buffer labeled b_1. Then they visit machine $\sigma(2)$, where they are stored in buffer b_2, etc. Let buffer b_L at machine $\sigma(L)$ be the last buffer visited. The sequence $\{\sigma(1), \ldots, \sigma(L)\}$ is the *route* of the part, and the corresponding buffers are $\{b_1, \ldots, b_L\}$. Suppose that the service times for parts in buffer b_i are exponentially distributed with mean $\frac{1}{\mu_i}$. We assume that a machine can work on only one part at a time, but that service can be preempted. We also assume that the service and inter–arrival times are independent. Let $B(k) := \{b_i \mid \sigma(i) = \sigma(k)\}$ be the set of all buffers competing for the same machine as b_k. Also, let

$$I(k) := \{i \mid \sigma(i) = \sigma(k)\} \tag{1}$$

be the corresponding set of indices.

Suppose the performance measure of interest is the mean number of parts in the system, or equivalently their mean delay through the system. Its minimization can be formulated as a dynamic programming problem, see [13]. However, the non–finiteness of the state–space renders it impossible to obtain the explicit solution from the dynamic programming equations, except in special cases (see [14] for the solution of a routing problem).

3 Linear Equality Constraints Implied by a Steady–State with Finite Second Moment

Consider any control or scheduling policy which is:

(i) Non–idling. That is, if any buffer at a machine is nonempty, then the machine cannot stay idle.

(ii) Stationary. That is, the control decisions are purely a function of the system state, which for us is the vector of all buffer queue lengths.

Let us rescale time so that $\lambda + \sum_{i=1}^{L} \mu_i = 1$. Throughout this paper we shall resort to *uniformization*, see Lippman [15]. This is a method of sampling a continuous time system to obtain a discrete time system with the same steady state behavior. Let us pretend that every buffer is always being worked on, and sample the system at all service completion times, as well as at the arrival times of new parts to the system. If the buffer is not really being worked on, we shall call the service completion a "virtual service completion"; otherwise it is a real service completion. The incorporation of such virtual service completion times ensures that we do not undersample the buffers not being worked on. Let $\{\tau_n\}$ be the sequence of such random sampling times, and let $\mathcal{F}_{\tau_n}$ denote the $\sigma-$field generated by the events up to time τ_n.

Let $X_i(t)$ denote the number of parts in buffer b_i at time t (including any in service). Also, let $W_i(t) = 1$ if the machine at machine $\sigma(i)$ is working on buffer b_i at time t, and 0 otherwise. We take all processes to be right continuous, and thus $X_i(\tau_n)$ is the state after the n-th event, while due to the stationarity of the scheduling policy, $W_i(\tau_n) = 1$ implies that machine $\sigma(i)$ is busy working on buffer b_i in the time interval $[\tau_n, \tau_{n+1})$.

Denote $X^T(\tau_n) := (X_1(\tau_n), \ldots, X_L(\tau_n))$. Consider any stationary control policy under which the system is in steady–state. In particular, suppose that

$$E[X^T(\tau_{n+1})QX(\tau_{n+1})] = E[X^T(\tau_n)QX(\tau_n)] \quad \text{for every symmetric matrix } Q. \tag{2}$$

Alternately, we could suppose that the system is not started in steady–state, but that, $E[X^T(\tau_{n+1})QX(\tau_{n+1})] - E[X^T(\tau_n)QX(\tau_n)] \to 0$ as $n \to \infty$, and then all the results in this paper apply to the steady state distribution. Either way, one presumes that the steady–state distribution has a finite second moment on the total number of parts.

We will now show that (2) yields a set of $L(L+1)/2$ linear equality constraints satisfied by the L^2 variables $E[X_i(\tau_n)W_j(\tau_n)]$. Clearly, for (2) to hold, we need,

$$E[X_i(\tau_{n+1})X_j(\tau_{n+1})] = E[X_i(\tau_n)X_j(\tau_n)] \quad \text{for } 1 \le i, j \le L. \tag{3}$$

In fact, (3) is equivalent to (2).

Let us begin by considering the implication of the equality,

$$E[X_1^2(\tau_{n+1})] = E[X_1^2(\tau_n)]. \tag{4}$$

Note that (see Figure 2),

$$\begin{aligned} X_1(\tau_{n+1}) &= X_1(\tau_n)+1 && \text{if the event at } \tau_{n+1} \text{ is an arrival,} \\ &= X_1(\tau_n)-1 && \text{if the event at } \tau_{n+1} \text{ is a real service completion} \\ & && \text{from } b_1, \text{ i.e., } \sigma(1) \text{ is working on } b_1 \text{ in } [\tau_n,\tau_{n+1}), \\ &= X_1(\tau_n) && \text{if the event at } \tau_{n+1} \text{ is a virtual service completion} \\ & && \text{from } b_1, \text{ i.e., } \sigma(1) \text{ is not working on } b_1 \text{ in } [\tau_n,\tau_{n+1}), \\ &= X_1(\tau_n) && \text{otherwise.} \end{aligned}$$

Since the conditional probability, given $\mathcal{F}_{\tau_n}$, of the first possibility is λ, the second $\mu_1 W_1(\tau_n)$, the third $\mu_1(1-W_1(\tau_n))$, and that of the fourth is $\mu_2+\ldots+\mu_L$, we obtain

$$\begin{aligned} E[X_1^2(\tau_{n+1}) \mid \mathcal{F}_{\tau_n}] &= \lambda(X_1(\tau_n)+1)^2 + \mu_1 W_1(\tau_n)(X_1(\tau_n)-1)^2 + \mu_1(1-W_1(\tau_n))X_1^2(\tau_n) \\ &+ \sum_{i=2}^{L} \mu_i X_1^2(\tau_n). \end{aligned}$$

Hence

$$\begin{aligned} E[X_1^2(\tau_{n+1}) \mid \mathcal{F}_{\tau_n}] &= \left(\lambda + \sum_{i=1}^{L} \mu_i\right) X_1^2(\tau_n) + 2\lambda X_1(\tau_n) + \lambda - 2\mu_1 W_1(\tau_n) X_1(\tau_n) \\ & \quad + \mu_1 W_1(\tau_n). \end{aligned}$$

Noting $\lambda + \sum_{i=1}^{L} \mu_i = 1$, and taking the unconditional expectation, we obtain

$$E[X_1^2(\tau_{n+1})] = E[X_1^2(\tau_n)] + 2\lambda E[X_1(\tau_n)] + \lambda - 2\mu_1 E[W_1(\tau_n)X_1(\tau_n)] + \mu_1 E[W_1(\tau_n)].$$

By the steady–state assumption (4), we obtain

$$2\lambda E[X_1(\tau_n)] + \lambda - 2\mu_1 E[W_1(\tau_n)X_1(\tau_n)] + \mu_1 E[W_1(\tau_n)] \quad = \quad 0.$$

However, in order for a steady–state to exist, one must have

$$\mu_1 E[W_1(\tau_n)] \quad = \quad \lambda,$$

since the average work brought per unit time to b_1 is completed by the machine. Thus,

$$2\lambda E[X_1(\tau_n)] + 2\lambda - 2\mu_1 E[W_1(\tau_n)X_1(\tau_n)] = 0. \tag{5}$$

Note that, by the non–idling nature of the policy, recalling the notation (1),

$$X_1(\tau_n) > 0 \;\;\Rightarrow\;\; \sum_{i \in I(1)} W_i(\tau_n) = 1.$$

As a consequence,

$$X_1(\tau_n) = \sum_{i \in I(1)} W_i(\tau_n) X_1(\tau_n). \tag{6}$$

Let us now define the variables,

$$z_{ij} \;\; := \;\; E[W_i(\tau_n)X_j(\tau_n)].$$

Then (5) yields,

$$2\lambda \sum_{i \in I(1)} z_{i1} + 2\lambda - 2\mu_1 z_{11} \;\; = \;\; 0. \tag{7}$$

Thus, we have obtained a linear equality constraint on the z_{ij}'s from the steady–state relation (4).

We can similarly obtain more linear equality constraints from the equalities,

$$E[X_j^2(\tau_{n+1})] = E[X_j^2(\tau_n)] \quad \text{for } j = 2, \ldots, L. \tag{8}$$

As in Figure 2,

$$\begin{aligned} X_j(\tau_{n+1}) \;&=\; X_j(\tau_n) + 1 \quad \text{if the event at } \tau_{n+1} \text{ is a real service completion at } b_{j-1}, \\ &\qquad\qquad\qquad\;\; \text{i.e., } W_{j-1}(\tau_n) = 1, \\ &=\; X_j(\tau_n) - 1 \quad \text{if the event at } \tau_{n+1} \text{ is a real service completion} \\ &\qquad\qquad\qquad\;\; \text{from } b_j, \text{ i.e., } W_j(\tau_n) = 1, \\ &=\; X_j(\tau_n) \qquad\;\;\, \text{otherwise.} \end{aligned}$$

As above, we obtain

$$2\mu_{j-1} z_{j-1,j} + 2\lambda - 2\mu_j z_{j,j} = 0 \qquad \text{for } j = 2, \ldots, L. \tag{9}$$

Similarly, from the steady–state equalities,

$$\begin{aligned}
E[X_1(\tau_{n+1})X_2(\tau_{n+1})] &= E[X_1(\tau_n)X_2(\tau_n)], \\
E[X_1(\tau_{n+1})X_j(\tau_{n+1})] &= E[X_1(\tau_n)X_j(\tau_n)] \quad \text{for } j = 3,\ldots,L, \\
E[X_i(\tau_{n+1})X_{i+1}(\tau_{n+1})] &= E[X_i(\tau_n)X_{i+1}(\tau_n)] \quad \text{for } i = 2,\ldots,L-1, \\
E[X_i(\tau_{n+1})X_j(\tau_{n+1})] &= E[X_i(\tau_n)X_j(\tau_n)] \quad \text{for } i = 2,\ldots,L-2 \text{ and } j = i+2,\ldots,L,
\end{aligned}$$

we obtain, respectively, the equality constraints,

$$\lambda \sum_{j\in I(2)} z_{j2} - \lambda - \mu_1(z_{12} - z_{11}) - \mu_2 z_{21} = 0, \tag{10}$$

$$\lambda \sum_{i\in I(j)} z_{ij} - \mu_1 z_{1j} - \mu_j z_{j1} + \mu_{j-1} z_{j-1,1} = 0 \quad \text{for } j = 3,\ldots,L, \tag{11}$$

$$\mu_{i-1} z_{i-1,i+1} - \mu_i z_{i,i+1} - \lambda + \mu_i z_{ii} - \mu_{i+1} z_{i+1,i} = 0 \quad \text{for } i = 2,\ldots,L-1, \text{ and} \tag{12}$$

$$\mu_{i-1} z_{i-1,j} - \mu_i z_{i,j} + \mu_{j-1} z_{j-1,i} - \mu_j z_{ji} = 0 \quad \text{for } i = 2,\ldots,L-2 \text{ and } j = i+2,\ldots,L. \tag{13}$$

Thus, from (7,9,10,11,12,13), we have $L(L+1)/2$ linear equality constraints on the L^2 variables $\{z_{ij} : 1 \le i,j \le L\}$.

4 The Non–Idling Inequality Constraints

By the non–idling nature of the policy, we know that just as in (6),

$$X_i(\tau_n) = \sum_{j\in I(i)} W_j(\tau_n) X_i(\tau_n) \quad \text{for } i = 1,\ldots,L. \tag{14}$$

Also, for any σ, $\sum_{\{j|\sigma(j)=\sigma\}} W_j(\tau_n) \le 1$, and so,

$$X_i(\tau_n) \;\ge\; \sum_{\{j|\sigma(j)=\sigma\}} W_j(\tau_n) X_i(\tau_n) \quad \text{for } i = 1,\ldots,L \text{ and } \sigma = 1,\ldots,S.$$

Thus, we obtain the inequality constraints,

$$\sum_{\{j|\sigma(j)=\sigma\}} z_{ji} \le \sum_{j\in I(i)} z_{ji} \text{ for } i = 1,\ldots,L \text{ and } \sigma = 1,\ldots,S \text{ with } \sigma \ne \sigma(i); \tag{15}$$

a total of $L(S-1)$ inequality constraints.

5 The Linear Program Bounds

From (14), the mean number of parts in the system is,

$$E[\sum_{i=1}^{L} X_i(\tau_n)] = \sum_{i=1}^{L} \sum_{j \in I(i)} z_{ji}. \tag{16}$$

In many systems, this is the primary performance objective of interest.

We have obtained the following theorem.

Theorem 1 *Consider any non–idling, stationary policy with a steady–state distribution that has a bounded second moment. Then the mean total number of parts in the system is bounded below by the value of the following linear program:*

$$Min \sum_{i=1}^{L} \sum_{j \in I(i)} z_{ji}$$

subject to the equality constraints (7,9,10,11,12,13), the inequality constraints (15), and the nonnegativity constraints,

$$z_{ij} \geq 0 \quad \textit{for all } i, j = 1, \ldots, L. \tag{17}$$

We note that the existence of an optimal stationary policy can be proved as in Theorem 3.1 of Borkar [16]. For the specific cost criterion $\sum_{i=1}^{L} X_i$ considered here, it can be chosen to be non-idling, which can be proved by dynamic programming induction arguments as in [14], or by sample path arguments as in [17]. However, for a cost criterion $\sum_{i=1}^{L} c_i X_i$, where $c_{i+1} > c_i$ for some i, every optimal optimal policy may require some idling.

Instead of a lower bound on the mean number of parts under a non–idling stationary policy, one can also obtain an upper bound (i.e., on worst case behavior) by replacing "Min" by "Max" in Theorem 1.

Theorem 2 *Consider any non–idling, stationary policy with a steady–state distribution that has a bounded second moment. An upper bound on the mean total number of parts in the system is given by the value of the linear program:*

$$Max \sum_{i=1}^{L} \sum_{j \in I(i)} z_{ji}$$

subject to the same constraints as in Theorem 1.

We note that there are additional constraints $z_{jj} \geq \frac{\lambda}{\mu_j}$ for every j, that could be appended to the linear programs. However such constraints are deducible from (7,9) here (and by similar equations for the other systems in this paper), and are hence omitted.

6 Minimum Possible Delay in an Open Re–entrant Line

We illustrate the method on an example.

Example 1 *An Open Re–entrant Line.*

Consider the open re–entrant line of Figure 3. The service rate for buffer b_i is μ_i. The arrival rate is λ. We scale time so that $\lambda + \mu_1 + \mu_2 + \mu_3 + \mu_4 = 1$. The inter–arrival and service times are all exponentially distributed, and independent.

The linear program to lower bound the mean number of parts in the system is:

$$\text{Min } [z_{11} + z_{31} + z_{22} + z_{42} + z_{13} + z_{33} + z_{24} + z_{44}]$$

subject to the equality constraints,

$$\begin{aligned}
\rho_1 + \rho_1(z_{11} + z_{31}) - z_{11} &= 0 \\
\rho_1(z_{22} + z_{42} - 1) - z_{12} + z_{11} - \frac{\mu_2}{\mu_1} z_{21} &= 0 \\
\rho_1(z_{13} + z_{33}) - z_{13} + \frac{\mu_2}{\mu_1} z_{21} - \frac{\mu_3}{\mu_1} z_{31} &= 0 \\
\rho_1(z_{24} + z_{44}) - z_{14} + \frac{\mu_3}{\mu_1} z_{31} - \frac{\mu_4}{\mu_1} z_{41} &= 0 \\
\rho_2 + \frac{\mu_1}{\mu_2} z_{12} - z_{22} &= 0 \\
\frac{\mu_1}{\mu_2} z_{13} - z_{23} - \rho_2 + z_{22} - \frac{\mu_3}{\mu_2} z_{32} &= 0 \\
\frac{\mu_1}{\mu_2} z_{14} - z_{24} + \frac{\mu_3}{\mu_2} z_{32} - \frac{\mu_4}{\mu_2} z_{42} &= 0 \\
\rho_3 + \frac{\mu_2}{\mu_3} z_{23} - z_{33} &= 0
\end{aligned}$$

$$\begin{aligned}\frac{\mu_2}{\mu_3}z_{24} - z_{34} + z_{33} - \rho_3 - \frac{\mu_4}{\mu_3}z_{43} &= 0\\ \rho_4 + \frac{\mu_3}{\mu_4}z_{34} - z_{44} &= 0,\end{aligned}$$

the inequality constraints,

$$\begin{aligned} z_{21} + z_{41} &\leq z_{11} + z_{31}\\ z_{12} + z_{32} &\leq z_{22} + z_{42}\\ z_{23} + z_{43} &\leq z_{13} + z_{33}\\ z_{14} + z_{34} &\leq z_{24} + z_{44},\end{aligned}$$

and the nonnegativity constraints,

$$z_{ij} \geq 0 \quad \text{for} \quad i, j = 1, 2, 3, 4.$$

Above, $\rho_i := \lambda/\mu_i$. From the lower bound on the mean number of parts in the system, one can use Little's law to obtain the lower bound on mean delay. Figure 4 plots the lower bound on the ratio Mean Delay/Mean Total Processing Time, as a function of the line loading $\rho := \rho_1 + \rho_3 = \rho_2 + \rho_4$, for the special case $\mu_1 = \mu_2 = \mu_3 = \mu_4$. This ratio is called the "actual–to–theoretical" ratio, and is an important measure of the performance of a semiconductor manufacturing plant. □

Example 2 *Comparison of Lower Bounds With Those of Ou and Wein.*

Ou and Wein [4] have proposed a method for obtaining a confidence interval on a lower bound on achievable cost. Their approach first constructs a modified system whose cost lower bounds the cost of all scheduling policies in the original system. Then they simulate the modified system, obtaining a confidence interval on the lower bound.

For the three open re–entrant lines shown in Figures 5 (a),(b) and (c), the 95% confidence intervals of their bounds are compared with our lower bounds in Table 1 (obtained using a slight extension of Theorem 1 to account for two arrival processes). For all 12 imbalanced systems, our bounds are better than the lower limits of their confidence intervals. In fact, under very heavy loading, our bounds are far superior. For the balanced systems, their bounds are better at lighter loadings, while ours are better at very heavy loadings. In [4], the authors also propose scheduling policies, obtaining confidence intervals on their cost. For all three imbalanced systems under very heavy loading, their confidence intervals are below our lower bound on achievable cost, showing that the confidence intervals are overly optimistic. □

7 Bounds for Buffer Priority Policies: Buffer Priority Constraints and Tandem Constraints

Let us now consider how to bound the performance of *buffer priority policies*, defined as in Lu and Kumar [3]. For each service station σ, one orders the buffers $B_\sigma := \{b_i \mid \sigma(i) = \sigma\}$ served at the machine. Let $r : B_\sigma \to Z$ be an ordering of the buffers (r is one–to–one), and suppose that preemptive priority service is provided to buffers lower in the ordering. The choice $r(b_i) = i$ gives the First Buffer First Serve Policy (FBFS) of [3], while the choice $r(b_i) = L + 1 - i$ gives the Last Buffer First Serve Policy (LBFS).

Thus,

$$W_i(\tau_n) = 1 \quad \Leftrightarrow \quad X_i(\tau_n) \geq 1 \text{ and } X_j(\tau_n) = 0 \text{ for all } b_j \in B_{\sigma(i)} \text{ with } r(b_j) < r(b_i).$$

The consequence is that $E[W_i(\tau_n)X_j(\tau_n)] = 0$ for all $b_j \in B_{\sigma(i)}$ with $r(b_j) < r(b_i)$. Hence, we simply obtain more equality constraints to be appended to the linear program of Theorem 1:

$$z_{ij} = 0 \quad \text{for all} \quad i, j \text{ with } b_j \in B_{\sigma(i)} \text{ and } r(b_j) < r(b_i).$$

Note that for each machine σ, we obtain $|B(\sigma)|(|B(\sigma)| - 1)$ additional equality constraints.

Example 3 *LBFS is Nearly Optimal in Light Traffic.*

Consider the system of Figure 3, with $\mu_1 = \mu_2 = \mu_3 = \mu_4 =: \mu$. Figure 6 plots the ratio Upper Bound on the Mean Delay Under LBFS/Lower Bound on Optimal Mean Delay, as a function of system loading. As seen, LBFS is almost optimal at light loading. For the specific value of $\lambda = \mu/20$, i.e., $\rho = 0.1$, the lower bound on the minimum mean number in the system, and the lower and upper bounds for LBFS, are shown in Table 2. LBFS is no more than 8.335% away from optimal.

It may be noted that currently we do not know of any such analytic result concerning dominance, even for a specific example. □

For the FBFS policy, we can obtain additional equality constraints from the tandem behavior of the first few buffers, as shown in the following example.

Example 4 *LBFS is Better Than FBFS, and the Tandem Constraints for FBFS.*

Consider the system of Figure 3. Let $\mu_1 = \mu_2 = \mu_3/7 = \mu_4/7$, and let λ be such that the load factors are $\frac{\lambda}{\mu_1} + \frac{\lambda}{\mu_3} = \frac{\lambda}{\mu_2} + \frac{\lambda}{\mu_4} = 0.4$. We append the constraints $z_{13} = 0$ and $z_{24} = 0$, to describe the LBFS policy. By maximizing the objective function, i.e., $\max[z_{11} + z_{31} + z_{22} + z_{42} + z_{13} + z_{33} + z_{24} + z_{44}]$, we obtain an upper bound on the number in the system under LBFS, while from $\min[z_{11} + z_{31} + z_{22} + z_{42} + z_{13} + z_{33} + z_{24} + z_{44}]$ we obtain a lower bound on the number in the system under LBFS. Similarly, using the constraints $z_{31} = 0$, $z_{42} = 0$, we describe the FBFS policy. In addition, under the FBFS policy, the first two buffers b_1 and b_2 act like two $M/M/1$ queues in tandem. From the known mean queue lengths, and their independence (Burke's Theorem), we have the following constraints:

$$\begin{aligned}
z_{11} &= \frac{\lambda}{\mu_1 - \lambda}, \quad z_{22} = \frac{\lambda}{\mu_2 - \lambda}, \\
z_{12} &= E[W_1(\tau_n) X_2(\tau_n)] = E[W_1(\tau_n)] E[X_2(\tau_n)] = \frac{\lambda}{\mu_1} z_{22}, \text{ and similarly,} \\
z_{21} &= \frac{\lambda}{\mu_2} z_{11}.
\end{aligned}$$

Table 3 summarizes the results obtained with as well as without these "tandem" constraints. The upper bound on the mean number in the system under LBFS is less than the lower bound on the mean number under the FBFS policy. This conclusively proves that LBFS is the better policy for this example.

8 Closed Re–Entrant Lines: Conservation of Material Constraints

In semiconductor manufacturing, a "closed-loop release policy" is one in which the number of parts in the system is maintained constant by releasing a new part into the system whenever a finished part exits from it. The reader is referred to Wein [9] for a thorough simulation study of this policy, and to Miller [18] for implementation results. In more classical terminology, it simply gives rise to a closed queueing network, with a fixed population of size N.

Typically, the goal is to maximize the throughput of the system by choosing a scheduling policy. Let us denote the throughput of the system by λ, noting that $\mu_i E[W_i(\tau_n)] = \lambda$ for $i = 1, 2, \ldots, L$. We rescale time so that $\sum_{i=1}^{L} \mu_i = 1$. The dynamics of the system is described by,

$$\begin{array}{rcll} X_i(\tau_{n+1}) & = & X_i(\tau_n) + 1 & \text{if the event at } \tau_{n+1} \text{ is a real service} \\ & & & \text{completion from } b_{i-1}, \text{ i.e., } \sigma(i-1) \text{ is working on } b_{i-1}, \\ & = & X_i(\tau_n) - 1 & \text{if the event at } \tau_{n+1} \text{ is a real service} \\ & & & \text{completion from } b_i, \text{ i.e., } \sigma(i) \text{ is working on } b_i, \\ & = & X_1(\tau_n) & \text{otherwise.} \end{array}$$

Above, we interpret the indices modulo L, i.e., by b_0 we mean b_L, and by b_{L+1} we mean b_1. By considering the equalities (3) we obtain the equality constraints (20),(21) and (22) shown below. Additionally, we can obtain $L+1$ equality constraints due to the conservation of material. Since the number of parts in the system is fixed at N, we obtain the constraint (18). Finally, since $E[W_i(X_1 + X_2 + \cdots + X_L)] = N\frac{\lambda}{\mu_i}$, for every buffer b_i, we obtain the equality constraints (19).

Theorem 3 *The maximal throughput*[1] *in the closed re–entrant line is bounded above by the value of the following linear program:*

$$Max \quad \lambda$$

[1] Note that due to the finiteness of the state–space, there exists an optimal stationary policy, which can also be shown to be non–idling for this cost criterion.

subject to the equality constraints,

$$\begin{aligned}
\sum_{i=1}^{L} \sum_{j \in I(i)} z_{ji} &= N && && (18)\\
\sum_{j=1}^{L} z_{ij} &= N \frac{\lambda}{\mu_i}, && \textit{for all } i && (19)\\
2\mu_{j-1} z_{j-1,j} + 2\lambda - 2\mu_j z_{j,j} &= 0 && \textit{for all } j && (20)\\
\mu_{i-1} z_{i-1,i+1} - \mu_i z_{i,i+1} - \lambda + \mu_i z_{ii} - \mu_{i+1} z_{i+1,i} &= 0 && \textit{for all } i && (21)\\
\mu_{i-1} z_{i-1,j} - \mu_i z_{i,j} + \mu_{j-1} z_{j-1,i} - \mu_j z_{ji} &= 0 && \textit{for all } i, j \neq i+1, && (22)
\end{aligned}$$

the inequality constraints (15), and the nonnegativity constraints (17).

Note that an additional constraint on λ is that it is limited by the bottleneck throughput, i.e. $\sum_{\{j|\sigma(j)=\sigma\}} \frac{\lambda}{\mu_j} \leq 1, \quad \sigma = 1, \ldots, S$. However, these inequalities are contained in the constraints (18,19,15) of Theorem 3, and have therefore been omitted.

For closed queueing networks with two machines, as in Figure 7, Harrison and Wein [11] have analyzed the Brownian network associated with a heavy traffic scenario (N large), and conjecture that a particular buffer priority policy provides maximal throughput as $N \to \infty$. To describe their policy, it is convenient to label an arbitrary buffer, say b_L, as the "exit" buffer, even though it is not semantically correct since parts never exit a closed system. Consider a part in buffer b_i. It may return several times to the same machine before "exiting" from b_L. Then, $M_{\sigma,i} := \sum_{\{j|i \leq j \leq L: \text{ and } \sigma(j)=\sigma\}} \frac{1}{\mu_j}$ is the mean amount of work on a part in buffer b_i, still remaining to be done by the machine σ prior to the part's "exit" from b_L. Also, let $\eta_\sigma := M_{\sigma,1}$ be the "relative utilization" of machine σ. At the first machine, $\sigma = 1$, rank the buffers to give priority to a buffer b_j with the smallest value of $\eta_2 M_{1j} - \eta_1 M_{2j}$. At the second machine, $\sigma = 2$, rank the buffers to give priority to a buffer b_j with the smallest value of $\eta_1 M_{2j} - \eta_2 M_{1j}$. Let us call this the "Balanced" buffer priority policy, in line with the arguments of [11].

Example 5 *A Closed Re-Entrant Line.*

Consider the system of Figure 7. For $\mu_1 = \frac{1}{3}$, $\mu_2 = \frac{1}{3.5}$, $\mu_3 = 1$ and $\mu_4 = 2$, the four buffer priority policies are the following:

- LBFS (b_3 has higher priority over b_1, and b_4 over b_2): $z_{13} = 0$ and $z_{24} = 0$.
- FBFS (b_1 has higher priority over b_3, and b_2 over b_4): $z_{31} = 0$ and $z_{42} = 0$.
- Harrison and Wein's Balanced Policy (b_1 has higher priority over b_3, and b_4 over b_2): $z_{31} = 0$ and $z_{24} = 0$.
- The "Unbalanced" Policy (b_3 has higher priority over b_1, and b_2 over b_4): $z_{13} = 0$ and $z_{42} = 0$.

For population sizes of $N = 20$ and $N = 100$, the bounds on the throughput in Table 4 are obtained. We see that in both cases, the Balanced policy outperforms the Unbalanced policy, since the lower bound on its throughput exceeds the other's upper bound. □

The following Theorem studies the heavy traffic behavior of the bounds as $N \to \infty$.

Theorem 4 *Let r be a fixed buffer priority policy, and let λ_N be the throughput of the closed network with a population of size N. Let λ_∞ be any limit point of λ_N as $N \to \infty$. Then*

$$\underline{\lambda} \leq \lambda_\infty \leq \bar{\lambda}.$$

Above, $\bar{\lambda}$ is the bottleneck throughput, $\bar{\lambda} := \left[\max_{1\leq\sigma\leq S} \sum_{\{j|\sigma(j)=\sigma\}} \frac{1}{\mu_j}\right]^{-1}$, while $\underline{\lambda}$ is the value of the linear program:

$$Min\ \lambda \tag{23}$$

subject to

$$\begin{aligned}
\sum_{i=1}^{L} \sum_{j\in I(i)} \bar{z}_{ji} &= 1 && && (24)\\
\sum_{j=1}^{L} \bar{z}_{ij} &= \frac{\lambda}{\mu_i}, && \textit{for } i = 1, \ldots, L && (25)\\
2\mu_{j-1}\bar{z}_{j-1,j} - 2\mu_j \bar{z}_{j,j} &= 0 && \textit{for } j = 1, \ldots, L && (26)\\
\mu_{i-1}\bar{z}_{i-1,i+1} - \mu_i \bar{z}_{i,i+1} + \mu_i \bar{z}_{ii} - \mu_{i+1}\bar{z}_{i+1,i} &= 0 && \textit{for } i = 1, \ldots, L && (27)\\
\mu_{i-1}\bar{z}_{i-1,j} - \mu_i \bar{z}_{i,j} + \mu_{j-1}\bar{z}_{j-1,i} - \mu_j \bar{z}_{ji} &= 0 && \textit{for } i = 1, \ldots, L-2 \textit{ and } j = i+2, \ldots, L, && \\
& && \textit{such that } j \not\equiv i \pm 1 \pmod{L}, && (28)
\end{aligned}$$

$$\bar{z}_{ij} = 0 \textit{ for all } i,j \textit{ with } j \in I(i) \textit{ and } r(b_j) < r(b_i), \tag{29}$$

$$\sum_{\{j|\sigma(j)=\sigma\}} \bar{z}_{ji} \leq \sum_{j \in I(i)} \bar{z}_{ji} \textit{ for } i = 1, \ldots, L \textit{ and } \sigma = 1, \ldots, S \textit{ with } \sigma \neq \sigma(i), \tag{30}$$

$$\textit{and } \bar{z}_{ij} \geq 0 \quad \textit{for all } i,j. \tag{31}$$

Proof. The bound $\lambda_\infty \leq \bar{\lambda}$ is immediate, and so we show $\lambda_\infty \geq \underline{\lambda}$. Define $\bar{z}_{ij} := \frac{z_{ij}}{N}$, and then λ_N is bounded below by the linear program (23,24,25,28,29,30,31) above, with (26,27) replaced by

$$\begin{aligned} 2\mu_{j-1}\bar{z}_{j-1,j} + 2\frac{\lambda}{N} - 2\mu_j\bar{z}_{j,j} &= 0 \quad \text{for all } j, \\ \mu_{i-1}\bar{z}_{i-1,i+1} - \mu_i\bar{z}_{i,i+1} - \frac{\lambda}{N} + \mu_i\bar{z}_{ii} - \mu_{i+1}\bar{z}_{i+1,i} &= 0 \quad \text{for all } i. \end{aligned}$$

Note that all variables $\bar{z}_{ji}$ and λ are bounded. Let $\lambda_N \to \lambda_\infty$ along some subsequence $\{N_k\}$. Then along some further subsequence of $\{N_k\}$, all the $\bar{z}_{ji}$'s will have limits satisfying (24-31). By the continuity of the constraint equations, this shows that λ_∞ is a feasible value for (23-31), and thus is no smaller than $\underline{\lambda}$. □

Note that if $\underline{\lambda} = \bar{\lambda}$, then this would serve both as a proof that the asymptotic throughput in the heavy traffic limit is the bottleneck throughput, and also that in heavy traffic our 100% confidence intervals have vanishingly small lengths. Indeed this is the case for every non–idling policy in the system of Example 5. However it is not the case for the following "imbalanced" system. (It is also not the case for one of the buffer priority policies in Example 7, which is a balanced system).

Example 6

Consider the closed re–entrant line of Figure 7. Let $\mu_1 = 2$, $\mu_2 = 1$, $\mu_3 = 4$ and $\mu_4 = 3$. Let the buffer priority policy be FBFS. Then $\underline{\lambda} = \frac{9}{13} < \frac{3}{4} = \bar{\lambda}$. In fact, numerical results indicate that the lower bounds converge to $\underline{\lambda}$. □

We note that to actually show that the limit of the values of these linear programs, for either the upper or the lower bounds, is the value of the corresponding limiting program, we need only verify the regularity condition that the set of all optimal solution vectors of the dual of the limiting program is compact, see Theorem 4.23 of Murty [19].

9 Bernoulli Splitting

In some applications, the part routes appear "random." One models this as though the route of a part is chosen randomly on–line. Note that by incorporating probabilistic routes, we allow the general routing features of Kelly networks; see Section 3.1 of [1].

Suppose every buffer b_i has an exogenous arrival process, which is Poisson with rate λ_i. Also suppose that with probability q_{ij}, a part leaving b_i goes next to b_j, and leaves the system with probability $1 - \sum_{j=1}^{L} q_{ij}$. For simplicity of presentation, suppose that $q_{ii} = 0$. As earlier, μ_i is the service rate at b_i. The routing choices are independent of the service times and arrival processes. Let $\{\rho_i\}$ be the solution to the set of traffic equations,

$$\lambda_i + \sum_{k=1}^{L} q_{ki}\mu_k\rho_k = \mu_i\rho_i \text{ for } i = 1, \ldots, L. \tag{32}$$

Let us suppose that the network is open, and that there is an unique solution to (32).

We suppose that $\sum_{i=1}^{L} \lambda_i + \mu_i = 1$, and sample at all arrival or departure epochs, real or virtual. Then, from the state transition diagram in Figure 8, we have

$$\begin{aligned} X_i(\tau_{n+1}) &= X_i(\tau_n) + 1 && \text{if there is an exogenous arrival to } b_i \text{ at } \tau_{n+1}, \\ &= X_i(\tau_n) + 1 && \text{if there is a real service completion from } b_j \text{ at } \tau_{n+1}, \\ & && \text{and if the part moves from } b_j \text{ to } b_i \\ &= X_i(\tau_n) - 1 && \text{if the event at } \tau_{n+1} \text{ is a real service completion} \\ & && \text{from } b_i, \\ &= X_i(\tau_n) && \text{otherwise.} \end{aligned}$$

Proceeding as in Section 3, we obtain the equality constraints

$$2\lambda_i \left(\sum_{j \in I(i)} z_{ji} \right) + 2\sum_{j=1}^{L} \mu_j q_{ji} z_{ji} - 2\mu_i z_{ii} + 2\mu_i\rho_i = 0, \tag{33}$$

$$\begin{aligned} \lambda_i \left(\sum_{k \in I(j)} z_{kj} \right) + \lambda_j \left(\sum_{k \in I(i)} z_{ki} \right) + \sum_{k \neq j} \mu_k q_{ki} z_{kj} + \sum_{k \neq i} \mu_k q_{kj} z_{ki} + \mu_j q_{ji}(z_{jj} - z_{ji} - \rho_j) \\ + \mu_i q_{ij}(z_{ii} - z_{ij} - \rho_i) - \mu_i(1 - q_{ij}) z_{ij} - \mu_j(1 - q_{ji}) z_{ji} = 0. \end{aligned} \tag{34}$$

Theorem 5 *Consider any stationary, non–idling scheduling policy with a steady–state distribution possessing a finite second moment, for the system above. Then the mean number* *in the system is bounded above by* $\max \sum_i \sum_{j\in\sigma(i)} z_{ji}$*, and below by* $\min \sum_i \sum_{j\in\sigma(i)} z_{ji}$ *subject to the equality constraints (33,34), and the same inequality and nonnegativity constraints as in Theorem 1.*

We can adapt the theorem above to closed networks with Bernoulli splitting. Then $\lambda_i \equiv 0$ and $\sum_j q_{ij} = 1$ for $i = 1, \ldots, L$, and so even under irreducibility, (32) possesses a solution unique only up to scaling. Any one of the utilization factors ρ_i can be maximized to maximize the throughput. We also have the additional constraints (18,19).

Example 7 *The Harrison–Wein Network.*

Consider the closed queueing network shown in Figure 9. The nonzero routing probabilities are $q_{12} = q_{34} = q_{45} = q_{56} = 1$ and $q_{21} = q_{23} = q_{61} = q_{63} = 0.5$. We take $\mu_1 = 0.25$, $\mu_2 = 1$, $\mu_3 = 0.125$, $\mu_4 = 1/6$, $\mu_5 = 0.5$ and $\mu_6 = 1/7$. Harrison and Wein [11] conjecture that their Balanced buffer priority policy ($z_{13} = 0$, $z_{15} = 0$, $z_{35} = 0$, $z_{42} = 0$, $z_{46} = 0$, $z_{62} = 0$) yields optimal throughput as $N \to \infty$. The upper and lower bounds for the throughput obtained with the Balanced policy, and the simulation result from [11] (obtained there only for $N = 7$), are shown in Table 5 for population sizes of 7 and 30.

Also, the solution of the limiting linear program of Theorem 4 (for this policy) is $\underline{\lambda} = \bar{\lambda}$. Hence, as $N \to \infty$, our upper and lower bounds converge to the bottleneck throughput, thus yielding vanishingly small lengths for the 100% confidence intervals in heavy traffic. (However, for the "Imbalanced" policy, $\underline{\lambda} < \bar{\lambda}$). □

10 Networks with Machine Failures: Constraints Due to One of a Set of Buffers Always Being Busy

In many applications such as manufacturing systems, machine failures are a significant contributor to delay, see Wein [9]. This has motivated much work in scheduling under failures, see Kimemia and Gershwin [20], Akella and Kumar [21] and Bielecki and Kumar [22]. If the failures are preemptive, then one can just model them as "fictitious" parts whose priority preempts the "real" parts already in service. (In the notation of [23], we are considering "time dependent" rather than "operation dependent" failures). Such systems are thus a special case of Section 4, where multiple routes were allowed[2]. The preemption constraints due to the machine failures give rise to a buffer priority policy, and thus give additional equality constraints. Note that there is exactly one customer in the sub–network modeling the failure, which is always being worked on by one of the machines in the sub–network. Hence we obtain an additional set of equality constraints due to one of the two buffers always being busy. We present an illustrative example.

Example 8 *The Effect of Failure and Repair Rates.*

Consider a re-entrant line with two machines. Machine 1 serves buffers b_1 and b_3, while machine 2 serves b_2. Suppose that machine 1 is subject to random failures. The times to failure are exponentially distributed with mean $1/\mu_0$, and repair times are exponentially distributed with mean $1/\mu_4$. This is modeled by adding two buffers b_0 and b_4 as shown in Figure 10. Buffer b_4 is served by the machine, with the highest priority, at a rate μ_4. The additional buffer b_0 is exclusively served by a server of rate μ_0. Buffers b_0 and b_4 form a closed network with exactly one part.

The equality constraints corresponding to (3) are, $2\lambda(z_{11} + z_{31} + z_{41}) + 2\lambda - 2\mu_1 z_{11} = 0$, $2\mu_1 z_{12} - 2\mu_2 z_{22} + 2\lambda = 0$, $2\mu_2 z_{23} - 2\mu_3 z_{33} + 2\lambda = 0$, $2\mu_0 z_{04} + 2\mu_0\mu_4/(\mu_0 + \mu_4) - 2\mu_4 z_{44} = 0$, $2\mu_4 z_{40} + 2\mu_0\mu_4/(\mu_0 + \mu_4) - 2\mu_0 z_{00} = 0$, $\lambda z_{22} + \mu_1 z_{11} - \mu_1 z_{12} - \lambda - \mu_2 z_{21} = 0$, $\lambda(z_{13} + z_{33} + z_{43}) - \mu_1 z_{13} - \mu_3 z_{31} + \mu_2 z_{21} = 0$, $\lambda(z_{14} + z_{34} + z_{44}) - \mu_1 z_{14} - \mu_4 z_{41} + \mu_0 z_{01} = 0$, $\lambda z_{00} - \mu_1 z_{10} - \mu_0 z_{01} + \mu_4 z_{41} = 0$, $\mu_1 z_{13} + \mu_2 z_{22} - \mu_2 z_{23} - \lambda - \mu_3 z_{32} = 0$, $\mu_1 z_{14} - \mu_2 z_{24} + \mu_0 z_{02} - \mu_4 z_{42} = 0$, $\mu_1 z_{10} - \mu_2 z_{20} + \mu_4 z_{42} - \mu_0 z_{02} = 0$, $\mu_2 z_{24} - \mu_3 z_{34} + \mu_0 z_{03} - \mu_4 z_{43} = 0$, $\mu_2 z_{20} - \mu_3 z_{30} + \mu_4 z_{43} - \mu_0 z_{03} = 0$ and $\mu_0 z_{00} - \mu_0 z_{04} - 2\mu_0\mu_4/(\mu_0 + \mu_4) + \mu_4 z_{44} - \mu_4 z_{40} = 0$. The additional constraints due to the priority of the fictitious parts are

$$z_{14} = 0, \text{ and } z_{34} = 0.$$

Since there is only one part in the sub–network modeling the failure, we have

$$z_{40} = 0, z_{04} = 0, \text{ and } z_{00} + z_{44} = 1.$$

[2] This is an example of a *mixed* network in which one route is open while another is closed.

Since either b_0 or b_4 must be worked on at all times (which corresponds to the fact that the real machine is either up or down at all times),[3] we obtain $W_0 + W_4 \equiv 1$. From this we obtain the additional equality constraints,

$$z_{02} + z_{42} = z_{22},\ z_{01} + z_{41} = z_{11} + z_{31} + z_{41},\ \text{and } z_{03} + z_{43} = z_{13} + z_{33} + z_{43}.$$

Let us consider the LBFS policy, $z_{13} = 0$. The non–idling inequality constraints are: $z_{11} + z_{31} + z_{41} \geq z_{21}$, $z_{11} + z_{31} + z_{41} \geq z_{01}$, $z_{22} \geq z_{12} + z_{32} + z_{42}$, $z_{22} \geq z_{02}$, $z_{44} \geq z_{24}$, $z_{44} \geq z_{04}$, $z_{33} + z_{43} \geq z_{23}$, $z_{33} + z_{43} \geq z_{03}$, $z_{00} \geq z_{20}$ and $z_{00} \geq z_{10} + z_{30} + z_{40}$. We note that three of the above inequalities are rendered redundant by the equality constraints resulting from either b_0 or b_4 always being busy. The objective function of interest is the mean number in the system excluding the fictitious part modeling failure, $z_{11} + z_{31} + z_{41} + z_{22} + z_{13} + z_{33} + z_{43}$.

For $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$ and $\lambda = \mu/4$, with the normalization $\lambda + \sum_{i=0}^{4} \mu_i = 1$, the bounds on the throughput, as μ_0/μ_4, the ratio Mean Time To Repair/Mean Time To Failure is varied, are shown in Table 6. We note that the upper bounds are finite until the machine capacity is reached, as one would desire. □

11 Networks with Finite Buffers

There is much interest in networks with finite buffers, see Sevastyanov [24] and Gershwin [25]. Such finiteness constraints are an important feature of most flow shops and assembly lines.

Suppose that in the basic re–entrant line of Section 2, buffer b_i has a capacity of C_i. When it is full, no part can leave b_{i-1}. To simplify presentation, suppose that a blocked part simply undergoes another round of service.

Let us introduce the variables $z_{ijkl} := E[W_i(\tau_n)X_j(\tau_n)1_{X_k(\tau_n)=l}]$, and $y_{ijk} := E[W_i(\tau_n)1_{X_j(\tau_n)=k}]$. A typical equality constraint resulting from (8) is,

$$\mu_{i-1}[2\sum_{k=0}^{C_i-1} z_{i-1,i,i,k} + \sum_{k=0}^{C_i-1} y_{i-1,i,k}] + \mu_i[-2\sum_{k=1}^{C_{i+1}-1} z_{i,i,i+1,k} + \sum_{k=1}^{C_{i+1}-1} y_{i,i+1,k}] = 0.$$

[3] More generally, if there is a set of buffers one of which is always being worked on, then the sum of the W_i's over the set is no less than one. This can be used to obtain additional inequality constraints by multiplying the sum by every X_j.

Also, there are equality constraints relating the z_{ijkl}'s and the y_{ikl}'s. This approach requires more examination to see if it is competitive with existing methods.

12 $GI/GI/1$ Queues: Constraints from Method of Stages

The method of stages, see Kleinrock [26], can be used to represent by a queueing network any $GI/GI/1$ queue whose inter–arrival and service distributions have rational Laplace transforms, say $A^*(s) = \sum_{i=1}^{M} \alpha_i \left(\frac{r_i\lambda_i}{s+r_i\lambda_i}\right)^{r_i}$ and $B^*(s) = \sum_{i=1}^{N} \beta_i \left(\frac{t_i\mu_i}{s+t_i\mu_i}\right)^{t_i}$, respectively. The sub–network for the service distribution consists of N parallel routes, each of which is chosen by a part with probability $\frac{\beta_i}{\sum_{i=1}^{N}\beta_i}$, the ith route having t_i exponential servers, each of rate $t_i\mu_i$, in series. It has the special feature that at most one part can be in the system at any time. To model the arrival process also, a similar sub–network with unit population is used.

Example 9 *The $E_2/M/1$ Queue.*

Consider the closed sub–network of machine M_0 with buffers b_1 and b_2, which is a part of Figure 11. It models the arrivals to machine M_1. If the total population of this sub–network is set to 1, then the time between consecutive arrivals to b_3 is the sum of two exponential random variables with means $\frac{1}{\lambda_1}$ and $\frac{1}{\lambda_2}$. Hence in the special case $\lambda_1 = \lambda_2$, Figure 10 models an $E_2/M/1$ queue. However we consider the slight generalization where λ_1 need not be equal to λ_2, while still referring to it as an $E_2/M/1$ queue. As before, we could rescale time so that $\lambda_1 + \lambda_2 + \mu = 1$, but that is unnecessary as the equations below show.

The constraints for the resulting linear program are given below[4]. The equality constraints are: $\lambda_2 z_{21} - \lambda_1 z_{11} + \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} = 0$, $-\lambda_1 z_{12} + \lambda_1 z_{11} - \lambda_2 z_{21} + \lambda_2 z_{22} - 2\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} = 0$, $\lambda_1 z_{12} - \lambda_2 z_{22} + \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} = 0$, $\lambda_2 z_{23} - \mu z_{33} + \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} = 0$, $\lambda_2 z_{22} - \lambda_2 z_{23} + \lambda_1 z_{13} - \mu z_{32} - \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} = 0$, and $-\lambda_1 z_{13} + \lambda_2 z_{23} + \lambda_2 z_{21} - \mu z_{31} + \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} = 0$. The constraint that there is exactly one part in buffers b_1 and b_2 together, gives,

$$z_{12} = 0, \text{ and } z_{21} = 0.$$

[4] Note that the equations are not linearly independent.

Since one of the buffers b_1 or b_2 is always being worked on by machine M_0, we obtain

$$z_{13} + z_{23} \quad = \quad z_{33}.$$

The non–idling constraints are, $z_{13} + z_{23} \leq z_{33}$ (this being subsumed by the above equality constraint), $z_{31} \leq z_{11}$ and $z_{32} \leq z_{22}$. The non-negativity constraints are $z_{ij} \geq 0$ for all i, j.

The linear programs for the upper and lower bounds can be analytically solved to obtain explicit expressions for the bounds. Note that the mean delay for this $GI/M/1$ queue can be determined exactly by finding the roots of a certain polynomial. We compare our upper bound with Kingman's upper bound (see [12]) for the mean sojourn time in a $GI/GI/1$ queue, $\frac{\sigma_a^2+\sigma_b^2}{2\bar{t}(1-\rho)} + \bar{x}$, where σ_a^2, σ_b^2, $\bar{t}$, $\bar{x}$ and ρ are the variance of the inter–arrival time, the variance of the service time, the mean inter–arrival time, the mean service time, and the load factor, respectively.

Theorem 6 *(i) The mean number in an $E_2/M/1$ is bounded below by $z_{33,min}$ and above by $z_{33,max}$ where*

$$\begin{aligned}
z_{33,min} \quad &:= \quad \text{Max } \{\rho, \frac{\mu\rho^2}{\lambda_2(1-\rho)}, \frac{\rho}{1-\rho} - \frac{\mu^2\rho^2}{\lambda_1\lambda_2(1-\rho)}\} \\
&= \quad \rho \quad \textit{if } \lambda_1 \leq \mu, \\
&= \quad \frac{\mu\rho^2}{\lambda_2(1-\rho)} \quad \textit{if } \lambda_2 \leq \mu \leq \lambda_1, \\
&= \quad \frac{\rho}{1-\rho} - \frac{\mu^2\rho^2}{\lambda_1\lambda_2(1-\rho)} \quad \textit{if } \mu \leq \lambda_2,
\end{aligned}$$

$$\begin{aligned}
z_{33,max} \quad &:= \quad \text{Min } \{\frac{\rho}{1-\rho}, \frac{\rho}{1-\rho} - \frac{\mu^2\rho^2}{\lambda_1\lambda_2(1-\rho)} + \frac{\lambda_2}{\lambda_1+\lambda_2}\} \\
&= \quad \frac{\rho}{1-\rho} \quad \textit{if } \lambda_1 \leq \mu, \\
&= \quad \frac{\rho}{1-\rho} - \frac{\mu^2\rho^2}{\lambda_1\lambda_2(1-\rho)} + \frac{\lambda_2}{\lambda_1+\lambda_2} \quad \textit{if } \mu \leq \lambda_1.
\end{aligned}$$

Above, $\rho := \frac{\lambda_1\lambda_2}{\mu(\lambda_1+\lambda_2)}$ is the load factor. (Both bounds above are sharper if we take $\lambda_2 \leq \lambda_1$, which is without loss of generality).

(ii) The upper and lower bounds differ by no more than $\frac{\lambda_2}{\lambda_1+\lambda_2}$, which will be no more than $1/2$, since we can take $\lambda_2 \leq \lambda_1$, without loss of generality.

(iii) The upper bound is always tighter than Kingman's upper bound by at least $\frac{\rho}{2}$.

Proof. From the equality constraints, we can solve for all the z_{ij}'s either explicitly or in terms of z_{33}. This provides simple inequality constraints on z_{33}, from which the minimum and maximum of the linear program are obtained by inspection. Kingman's upper bound for the mean number of parts is $z_{33,king} = \frac{\left(\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\right)^2\left(\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\mu^2}\right)}{2\left(1-\frac{\lambda_1\lambda_2}{\mu(\lambda_1+\lambda_2)}\right)} + \frac{\lambda_1\lambda_2}{\mu(\lambda_1+\lambda_2)}$. The rest follows by calculation. □

Table 7 shows the results for an $E_2/M/1$ queue with $\lambda_1 = \lambda_2$, for several values of the load factor ρ.

13 Conclusions

We have presented a methodology for obtaining bounds on the performance of a variety of queueing networks and scheduling policies. The technique allows a fair degree of generality in the type of system that can be considered. We have illustrated it's application to several interesting queueing networks and scheduling policies, of particular relevance in semiconductor manufacturing.

Several interesting questions remain. First, it is of considerable interest to add more equality and inequality constraints and thus improve the bounds. This appears to be a non–trivial task. Second, it would be useful to study the equality and inequality constraints to obtain qualitative or analytical results. Third, we have presumed that the underlying system is stable and has a finite second moment. To resolve the problem of stability is non–trivial. It is addressed in [27]. Finally, it would be useful to develop a software package that, given just a network description, derives the corresponding linear programs, and then solves them to provide performance bounds, thus automating all the tedious and error prone chores.

References

[1] F. P. Kelly, *Reversibility and Stochastic Networks.* New York, NY: John Wiley and Sons, 1979.

[2] P. R. Kumar, "Re–entrant lines," *Queueing Systems: Theory and Applications: Special Issue on Queueing Networks*, vol. 13, pp. 87–110, May 1993.

[3] S. H. Lu and P. R. Kumar, "Distributed scheduling based on due dates and buffer priorities," *IEEE Transactions on Automatic Control*, vol. 36, pp. 1406–1416, December 1991.

[4] J. Ou and L. M. Wein, "Performance bounds for scheduling queueing networks," *Annals of Applied Probability*, vol. 2, pp. 460–480, 1992.

[5] F. Baccelli and A. M. Makowski, "Queueing models for systems with synchronization constraints," *IEEE Proceedings*, vol. 77, pp. 138–161, 1989.

[6] S. P. Meyn and D. Down, "Stability of generalized Jackson networks." To appear in *Annals Applied Prob.*, 1993.

[7] D. Bertsimas, I. Ch. Paschalidis and J. N. Tsitsiklis, "Scheduling of multiclass queueing networks: Bounds on achievable performance," in *Workshop on Hierarchical Control for Real–Time Scheduling of Manufacturing Systems*, (Lincoln, New Hampshire), October 16–18, 1992.

[8] D. Bertsimas, I. Ch. Paschalidis and J. N. Tsitsiklis, "Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance." Laboratory for Information and Decision Systems and Operations Research Center, M. I. T., December 1992.

[9] L. M. Wein, "Scheduling semiconductor wafer fabrication," *IEEE Transactions on Semiconductor Manufacturing*, vol. 1, pp. 115–130, August 1988.

[10] C. R. Glassey and M. Resende, "Closed-loop job release control for VLSI circuit manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 1, pp. 36–46, February 1988.

[11] J. M. Harrison and L. M. Wein, "Scheduling networks of queues: Heavy traffic analysis of a two-station closed network," *Operations Research*, vol. 38, no. 6, pp. 1052–1064, 1990.

[12] J. F. C. Kingman, "Inequalities in the theory of queues," *Journal of the Royal Statistical Society, Series B*, vol. 32, pp. 102–110, 1970.

[13] P. R. Kumar and P. P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control.* Englewood Cliffs, NJ: Prentice-Hall, 1986.

[14] W. Lin and P. R. Kumar, "Optimal control of a queueing system with two heterogeneous servers," *IEEE Trans. Automat. Control*, vol. AC-29, pp. 696–703, August 1984.

[15] S. Lippman, "Applying a new device in the optimization of exponential queueing systems," *Operations Research*, vol. 23, pp. 687–710, 1975.

[16] V. S. Borkar, "Controlled Markov chains and stochastic networks," *SIAM Journal on Control and Optimization*, vol. 21, no. 4, pp. 652–666, 1983.

[17] J. Walrand, "A note on "Optimal control of a queueing system with two heterogeneous servers"," *Systems and Control Letters*, vol. 4, pp. 131–134, May 1984.

[18] D. J. Miller, "Simulation of a semiconductor manufacturing line," *Communications of the ACM*, vol. 33, no. 10, pp. 98–108, 1990.

[19] K. G. Murty, *Linear Programming.* New York, NY: John Wiley & Sons, 1983.

[20] J. Kimemia and S. B. Gershwin, "An algorithm for the computer control of a flexible manufacturing system," *IIE Transactions*, vol. 15, pp. 353–362, December 1983.

[21] R. Akella and P. R. Kumar, "Optimal control of production rate in a failure prone manufacturing system," *IEEE Trans. Automat. Control*, vol. AC-31, pp. 116–126, February 1986.

[22] T. Bielecki and P. R. Kumar, "Optimality of zero-inventory policies for unreliable manufacturing systems," *Operations Research*, vol. 36, pp. 532–541, July-August 1988.

[23] J. A. Buzacott and L. E. Hanifin, "Models of automatic transfer lines with inventory banks – a review and comparison," *AIEE Transactions*, vol. 10, pp. 197–207, June 1978.

[24] B. A. Sevastyanov, "Influence of storage bin capacity on the standstill time of a production line," *Theory of Probability and Its Applications*, vol. 7, pp. 429–438, 1962.

[25] S. B. Gershwin, "An efficient decomposition algorithm for unreliable tandem queueing systems with finite buffers," in *Queueing Networks with Blocking* (H. G. Perros and T. Altiok, eds.), pp. 127–146, Amsterdam: North Holland, 1989.

[26] L. Kleinrock, *Queueing Systems, Volume 1: Theory.* New York, NY: Wiley–Interscience, 1975.

[27] P. R. Kumar and S. P. Meyn, "Stability of queueing networks and scheduling policies," tech. rep., C. S. L., University of Illinois, 1993.