Linear programming performance bounds for Markov chains with polyhedrally translation invariant transition probabilities and applications to unreliable manufacturing systems and enhanced wafer fab models *†

James R. Morrison‡
IBM Microelectronics, Manufacturing Methods
and

P. R. Kumar§
University of Illinois at Urbana–Champaign

October 12, 2001

Abstract

Our focus is on a class of Markov chains which have a polyhedral translation invariance property for the transition probabilities. This class can be used to model several applications of interest which feature complexities not found in usual models of queueing networks, for example failure prone manufacturing systems which are operating under hedging point policies, or enhanced wafer fab models featuring batch tools and setups or affine index policies. We present a new family of performance bounds which is more powerful both in expressive capability as well as the quality of the bounds than some earlier approaches.

Keywords: Performance evaluation, wafer fabs, manufacturing systems, semiconductor manufacturing plants, queueing networks, scheduling, hedging point policies, batch tools, set-up times.

*The research reported here has been supported in part by NSF Grant No. DMI-9743165, SRC Contract No. 97-FJ-489, EPRI and the USARP under subcontract to Cornell University under Contract Nos. WO8333-04 and 35352-6086. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the above agencies.

†Please address all correspondence to the second author.

‡IBM Microelectronics, Manufacturing Methods, Burlington, VT. Email: jrmorr@us.ibm.com

§University of Illinois, Coordinated Science Laboratory, 1308 West Main Street, Urbana, IL 61801. Email: prkumar@uiuc.edu. Tel: (217) 333-7476. Web: http://black.csl.uiuc.edu/∼prkumar.

1 Introduction

Markov chains are used to model a variety of communication networks, manufacturing systems, and transportation systems. In almost all cases these models feature complexities not explicitly analyzable in closed form solution. In this paper we consider a special class of Markov chains which have translation invariance of the transition probabilities within each polyhedron in a class covering the integer lattice state space. Many problems of interest, such as unreliable manufacturing systems with failure prone machines operating under hedging point policies, or wafer fabs with batch tools and set up times, fit within this framework. For this class of Markov chains we develop a method for obtaining performance bounds. These bounds allow a greater expressive power for system modeling. In addition, the bounds also improve on some earlier methods, as we illustrate in the sequel.

As an example of the type of situation that our approach can be applied to, consider a manufacturing system with failure prone machines. In a pioneering work, [1] modeled the system by aggregating it as a single entity in which the capacity region is subject to random changes, and elucidated very coherently the role played by hedging point policies. There are few examples of such systems which can be explicitly solved. Reference [2] showed how to solve a simple example of one unreliable machine making one product under a discounted absolute value type cost criterion, while [3] showed how to solve the average cost case by the computing steady state distribution and also how to verify the optimality of the result. Over the years these approaches have been exploited to solve some additional scenarios, see [4, 5, 6, 7, 8, 9, 10]. The methods developed here allow us to obtain bounds on system performance for hedging point policies as well as linear switch curve policies (see for example [6]), for systems with random production times as well as random inter-demand times. The methods allow us to model the internal network details of the production system (as for example in [11]), without requiring us to view the system as an aggregate. Thus, details of scheduling within the factory can also be modeled. These methods thus allow us to more routinely model the types of complexities often present in manufacturing systems, and one therefore hopes

that methods such as these will be incorporated into the arsenal of methods for studying manufacturing systems, alongside the more established simulation modeling.

As noted above, the methods developed here allow us to obtain bounds for queueing networks. In this field too, the class of systems for which such explicit solutions are known is very limited [12], being essentially limited to the class of "product form networks" where certain buffer levels are independent in steady state. Motivated by this, [13] and [14] developed methods to computationally obtain upper and lower bounds on the performance. These methods allowed the treatment of certain complexities such as buffer priorities and closed networks. These methods have also been extended over the years; see [15, 16, 17, 18, 19, 20, 21, 22]. The methodology we develop here allows us to analyze a richer class of systems modeling several complexities which arise in applications. It allows us, for example, to model the presence of batch tools which are common in wafer fabs. It also allows us to analyze manufacturing systems where set-up times are incurred in changing between part-types. The types of scheduling policies that can be modeled is also rich, allowing for example affine index policies, which dovetail naturally with our methodology, as shown below.

All this is made possible via a recognition of the underlying structure in the inequality relaxation of the average cost equation. We show that common to all these formulations is a certain polyhedral translation invariance of the transition probabilities. Since polyhedra can be described by the conjunction of several linear inequalities, a common bound for a set of polyhedra covering the state space is obtainable by employing a performance-to-performance duality which guarantees the satisfaction of the desired average cost inequality. The power of the bounds is also enhanced by the consideration of piecewise quadratic functions, one for each polyhedron, which are naturally accommodated by the duality mentioned above, and elucidated in the sequel.

The rest of this paper is organized as follows. In Section 2, we recall the average cost inequality which is the starting point for the development of the LP performance bounds. For discrete–time Markov chains whose transition probabilities are translation invariant on polyhedra we propose a certain form for the surrogate of the differential cost function. The

unifying LP performance bound theorem is obtained by appealing to a performance–to–performance duality to ensure the average cost inequality on the polyhedra. The application of the general theorem to queueing networks operating under affine index policies is presented in Section 3. Here we define the affine index policies and demonstrate how they subsume many stationary policies of interest, including those addressed in the previous LP performance bound works mentioned above. We also mention some techniques which may be applied to improve the bounds thereby obtained and reduce the computational complexity of the LPs. In Section 4, we analyze the quality of the bounds for an unreliable manufacturing system with random production times and random demands. In Section 5 we show how to analyze batch tools which are important for the modeling of wafer fabs, and in Section 6 we show how to obtain bounds for systems with set-up times. The treatment of potential applications is far from exhaustive as the unified LP performance bound theorem applies to any discrete–time Markov chain exhibiting polyhedral translational invariance of its transition probabilities. Finally, concluding remarks are presented in Section 7.

2 Performance bounds for discrete–time Markov chains with translation invariant transition probabilities on covering polyhedra

Consider a discrete–time, time–homogeneous Markov chain with state space $\mathcal{S} \subseteq \mathcal{Z}_+^K$, transition probabilities $[p_{x,y}]$, a one-step cost function $c(x)$, with $x(k)$ denoting the state at time k. Throughout we suppose that $x(0)$ is deterministic, and that $E|c(x(k))| < +\infty$ for every k. By relaxing the average cost equation of dynamic programming to an inequality, one can obtain a bound on the performance.

Lemma 2.1. Average cost inequality performance bounds.

(i) *If W bounded above and a real J are such that*

$$J + W(x) \;\leq\; c(x) + \sum_{y \in S} p_{x,y} W(y), \text{ for all } x \in \mathcal{S}, \tag{1}$$

then

$$J \;\leq\; \liminf_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} E[c(x(k))]. \tag{2}$$

(ii) *If W is instead bounded below, and the inequality in (1) is reversed, then*

$$J \;\geq\; \limsup_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} E[c(x(k))]. \tag{3}$$

Proof: The inequality of (1) can be rewritten as

$$J + W(x(k)) \;\leq\; c(x(k)) + E[W(x(k+1)) \mid x(k)].$$

Taking the unconditional expectation,

$$J + E[W(x(k))] \leq E[c(x(k))] + E[W(x(k+1))].$$

If $E[W(x(k+1))] = -\infty$, then one will have $E[W(x(k))] = -\infty$ too, and by recursion $E[W(x(0))] = W(x(0)) = -\infty$, which is a contradiction. Hence, $E[W(x(k))]$ is finite for all k. Now, summing the above inequalities, and telescoping from $k = 0$ to $k = T-1$ we have

$$J \;\leq\; \frac{1}{T} \sum_{k=0}^{T-1} c(x(k)) + \frac{EW(x(T)) - W(x(0))}{T}.$$

Since $W(x(T))$ is assumed bounded above in (i), we may take the limit to yield the result of (2). The proof of (ii) is similar. □

The subclass of the above Markov chains which we address is detailed below. Let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ denote the K dimensional unit vector with a 1 in the i^{th} position for $i = 1, \ldots, K$, while e_{K+1} will be taken as the 0 vector. Let $H^m := \{x \in S : A^m x \geq b^m\}$ for $1 \leq m \leq M$ be polyhedra which cover the state space, i.e., $\cup_{m=1}^{M} H^m = \mathcal{S}$; see Figure 1, and such that the "relative" transition probabilities are the same for all states in each polyhedron; that is, for some $f^m(\cdot)$, $p_{x,y} = f^m(y - x)$ for all $x \in H^m$. We say that the Markov chain exhibits *translation invariance of its transition probabilities.* We will suppose throughout that $c(x) = c^T x$.

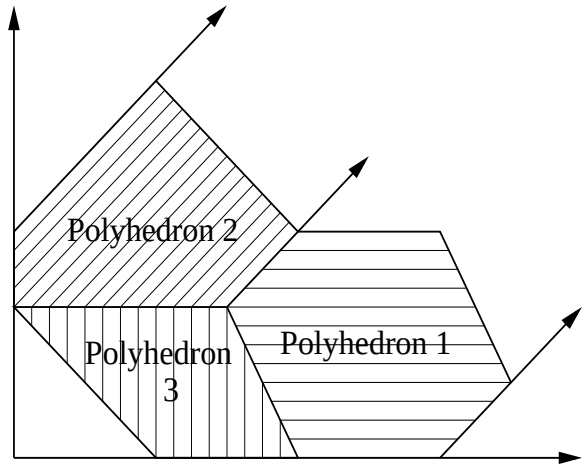

Figure 1: The relative transition probabilities are constant on polyhedron m.

Suppose for simplicity that we seek a quadratic surrogate $W(x) = p^T x + (1/2)x^T Q x$, where p and $Q = Q^T$ are to be determined. As a consequence of the translation invariance of the transition probabilities on polyhedra, by substituting in (1), we obtain

$$J \leq c^T x + \sum_{\{z \in \mathcal{Z}^K : f^m(z) \neq 0\}} f^m(z)[p^T z + \frac{1}{2} z^T Q z + z^T Q x], \text{ for } x \in H^m,$$

Thus,

$$J \leq d^m + r^{m^T} x, \text{ for all } x \in H^m, \text{ for } m = 1, 2, \ldots, M. \tag{4}$$

where

$$d^m := \sum_{\{z \in \mathcal{Z}^K : f^m(z) \neq 0\}} f^m(z)[p^T z + \frac{1}{2} z^T Q z], \tag{5}$$

$$r^m := c + \sum_{\{z \in \mathcal{Z}^K : f^m(z) \neq 0\}} f^m(z) Q z. \tag{6}$$

To proceed further, one should note the following features. The constant d^m and the vector r^m depend *linearly* upon the variables p and Q to be determined. Additionally, the right hand side in (4) depends linearly on x. The inequalities in (4) thus have an overall form which is *bilinear* in x for $x \in H^m$, and in (p, Q), for each fixed $m \in \{1, 2, \ldots, M\}$. Now our problem is to choose J and (p, Q) such that (4) is satisfied. To find (p, Q), we exploit the structural feature that the average cost inequality (4) has a single *fixed* linear form in

x, for all states $x \in H^m$. It is this invariance of the form of the average cost inequality on polyhedra which is the key to obtaining the bounds, and this fact will be further exploited later when we allow piecewise quadratic surrogates for the differential cost function. To solve the above problem and to render it into a purely linear problem which then allows us to use linear programming to obtain bounds, we invoke duality.

Theorem 2.2: Polyhedral translational invariance LP performance bounds.

(i) *Let $(J, p, Q,$ and vectors y^m,for $m \in \{1, \ldots, M\})$ be the decision variables in the following linear program $\underline{T}$:*

$$Max\ J \tag{7}$$

subject to

$$\begin{aligned} A^{m^T} y^m &\leq r^m, \\ b^{m^T} y^m &\geq J - d^m, \\ y^m &\geq 0,\ \textit{for all } m \in \{1, \ldots, M\}. \end{aligned} \tag{8}$$

If $W(x)$ corresponding to a feasible solution is bounded above for all $x \geq 0$, then J is a lower bound for the average cost as in (2).

(ii) *Let $(J, p, Q,$ and vectors y^m,for $m \in \{1, \ldots, M\})$ be the decision variables in the following linear program $\overline{T}$:*

$$Min\ J \tag{9}$$

subject to

$$\begin{aligned} A^{m^T} y^m &\leq -r^m, \\ b^{m^T} y^m &\geq d^m - J, \\ y^m &\geq 0,\ \textit{for all } m \in \{1, \ldots, M\}. \end{aligned} \tag{10}$$

for all $m \in \{1, \ldots, M\}$. If $W(x)$ corresponding to a feasible solution is bounded below for all $x \geq 0$, then J is an upper bound on the average cost as in (3).

Proof: For the lower bound (i), consider the following inequality in the region H^m:

$$J \leq d^m + r^{m^T} x, \qquad \forall x \in \{x : A^m x \geq b^m, x \geq 0\}. \tag{11}$$

Note that we relax x to allow it to lie in the *real* nonnegative orthant $\mathcal{R}_+^K$, rather than just the nonnegative integer lattice. The inequalities of (11) hold if and only if J, d^m, and r^m are such that the linear program $\{\underset{x}{\text{Min}}\ (r^{m^T} x + d^m - J) : A^m x \geq b^m, x \geq 0\}$, has value greater than or equal to 0. Note that this latter LP is feasible since H^m is nonempty. Its value is therefore nonnegative if and only if the following feasible linear program, denoted by L, has value $VL \geq J - d^m$:

$$\text{Min } r^{m^T} x, \text{ subject to } A^m x \geq b^m \text{ and } x \geq 0.$$

This in turn will hold if and only if the following dual of the linear program L, which we shall denote by D, has value $VD \geq J - d^m$:

$$\text{Max } b^{m^T} y^m, \text{ subject to } A^{m^T} y^m \leq r^m \text{ and } y^m \geq 0.$$

We have thus demonstrated that the inequalities of (11) hold if and only if there exists a $y^m \geq 0$ such that $A^{m^T} y^m \leq r^m$ and $b^{m^T} y^m \geq J - d^m$. Similarly, repeating the argument for each H^m, for all $m \in \{1, \ldots, M\}$, we deduce that the average cost inequalities (4) hold for all $x \in S$ if and only if there exists, for each $m \in \{1, \ldots, M\}$, a $y^m \geq 0$ such that $A^{m^T} y^m \leq r^m$, and $b^{m^T} y^m \geq J - d^m$. Thus, the average cost inequality will hold if there exist J, p, Q (symmetric), and vectors $y^m \geq 0, \forall m \in \{1, \ldots, M\}$, such that the constraints of (8) are satisfied, i.e., for any feasible solution (J, p, Q (symmetric), and y^m for $m \in \{1, \ldots, M\}$) of (8). If the resulting $W(x)$ is bounded above, then the result follows from Lemma 2.1. A similar proof shows (ii). □

3 Queueing networks operating under affine index policies

To illustrate the applicability of the above approach we first turn to open re–entrant line models. We consider a class of stationary scheduling policies, which we term affine index

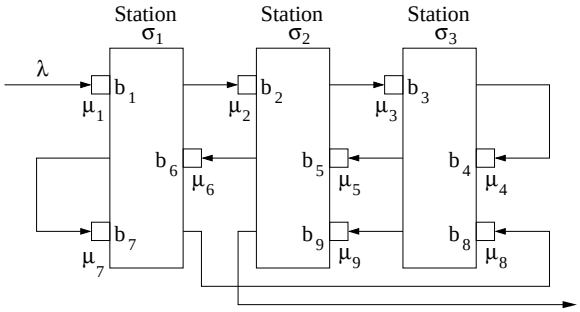

Figure 2: An open re–entrant line.

policies. Such policies subsume many policies of interest including buffer priority policies [23], the Fluctuation Smoothing Policy for the Mean Cycle Time (FSMCT) for scheduling wafer fabs [24], some policies based on Brownian motion models [25, 26], some policies based on fluid models [27, 28, 29, 30, 31], policies based on hedging point policies [1], and those based on linear switching curves [6]. We show how one can identify polyhedral regions with translation invariant transition probabilities. We also show how to improve the bounds via consideration of piecewise quadratic functions.

3.1 Open re–entrant lines

Consider an open re–entrant line as in [14], shown in Figure 2. It consists of E stations $\{\sigma_1, \ldots, \sigma_E\}$, and L buffers $\{b_1, \ldots, b_L\}$. Customers in b_i are serviced at a station $\sigma(i) \in \{\sigma_1, \ldots, \sigma_E\}$ (we use $i \in \sigma$ to indicate that $\sigma(i) = \sigma$), with an exponentially distributed service time of mean $1/\mu_i$. Customers arrive to the network at buffer b_1 as a Poisson process of rate λ. After receiving service in buffer b_i, it next moves to buffer b_{i+1}, unless $i = L$, in which case the customer exits the network. We assume that all service times and interarrival times are independent.

Let $x(t) = (x_1(t), \ldots, x_L(t))^T$ denote the state, where $x_i(t)$ is the number of customers in buffer b_i at time t including any in service. Let $u_i(t) = 1$ if a customer from buffer b_i is receiving service, and $u_i(t) = 0$, otherwise. A station can serve at most one customer at a time. We assume that the scheduling policy is stationary, i.e., $u(t)$ is of the form $u(x(t))$,

nonidling, i.e., a station must work on some customer if any are present in its buffers, and priorities are implemented in a preempt–resume manner. We denote the class of such policies by $\mathcal{U}$. One can convert this into a discrete time controlled Markov chain by uniformization [32]. Rescale time so that $\lambda+\sum_{i=1}^{L}\mu_i = 1$, and suppose that every buffer not receiving service has a fictitious virtual customer in service. Sample the system at the times $\{\tau_k\}$ with $\tau_0 = 0$, corresponding to arrivals, real service completions, or virtual service completions, and define $x_i(k) := x_i(\tau_k)$, $u_i(k) := u_i(\tau_k)(= u_i(x(k))$. Letting e_i be the unit vector with a 1 in the i^{th} coordinate, and e_{L+1} be the zero vector, the transition probabilities are $p_{x,x+e_1}(u(x)) = \lambda$, $p_{x,x-e_i+e_{i+1}}(u(x)) = \mu_i u_i(x)$ for $i = 1, 2, \ldots, L$, and $p_{x,x}(u(x)) = 1 - \lambda - \sum_{i=1}^{L} u_i(x)\mu_i$.

3.2 Affine index policies

Many well known policies are special cases of the following class of affine index policies.

Definition 3.1: Affine index policies. *Assign to buffer b_i an affine* index η_i*:*

$$\eta_i(x) := k_i + \sum_{j=1}^{L} m_j^i x_j,$$

where k_i and m_j^i are given constants. A station σ must work on a customer in its constituent buffers whose index is greatest. Ties may be settled arbitrarily.

Example 3.2: Buffer priority policies. *Let $\theta = \{\theta(1), \ldots, \theta(L)\}$ be a permutation of the buffer indices $\{1, \ldots, L\}$, and set $\eta_i(x) := -\theta(i)$.* □

Example 3.3: Fluctuation smoothing policy for mean cycle time (FSMCT) [24].

$$\eta_i(x) = \sum_{j=i}^{L} \overline{x_j} - \sum_{j=i}^{L} x_j,$$

where $\overline{x_j}$ is an estimate of the mean number of customers in buffer b_j. □

Example 3.4: Linear switch curve policies. $\eta_i(x) = m_i x_i$. □

3.3 Linear programming performance bounds

The key issue is to identify polyhedral regions of the state space for which the average cost inequality has a single form. As a station can work on a customer in a nonempty buffer, we first partition the state space into regions where specific buffers have at least one customer. For $x \in \mathcal{S}$, define $\phi(x) = (\phi_1(x0), \ldots, \phi_L(x))^T$ by $\phi_i(x) = 0$ if $x_i = 0$, and $\phi_i(x) = 1$, if $x_i \geq 1$. Denote by Φ the set of all such vectors. Now, define $\mathcal{X}^\phi \subseteq \mathcal{Z}_+^L$ by $\mathcal{X}^\phi := \{x \in \mathcal{Z}_+^L : \phi(x) = \phi\}$. Note that $\cup_{\phi \in \Phi} \mathcal{X}^\phi = \mathcal{S}$. Now we also partition the state space according to the buffers receiving service. Let Ω denote the set of all vectors $\omega = (\omega_1, \ldots, \omega_E)$ with $\omega_j \in \{0\} \cup \{i : \sigma(i) = \sigma_j\}$. $\omega_j = \ell$ will denote that buffer b_ℓ is receiving service, while $\omega_j = 0$ denotes that no buffer is in service at σ_j. Now define the composite index $\psi := (\phi, \omega)$ and set $\mathcal{S}^\psi := \{x \in S^\phi : u_{\omega_j}(x) = 1, \forall j \text{ with } \omega_j \neq 0, u_{\omega_j}(x) = 0, \forall j \text{ with } \omega_j = 0\}$. Let Ψ denote the set of ψ for which $\mathcal{S}^\psi$ is nonempty. Finally, we define the elements of the partition for $\psi = (\phi, \omega) \in \Psi$ by,

$$\overline{\mathcal{S}}^\psi := \{x \in S^\phi : \eta_{\omega_j}(x) \geq \eta_\ell(x), \forall \ell \in \sigma_j \text{ with } \phi_\ell = 1, \forall j \text{ with } \omega_j \neq 0\}.$$

Note that $\cup_{\psi \in \Psi} \overline{\mathcal{S}^\psi} = \mathcal{S}$. Clearly, $\mathcal{S}^\psi \subseteq \overline{\mathcal{S}^\psi}$, so that if we ensure that the average cost inequality holds on the set $\overline{\mathcal{S}^\psi}$ with ω dictating which buffers are working, then it will also hold on $\mathcal{S}^\psi$. We will thus ensure that the average cost inequality will hold no matter how ties for the maximum affine index are broken.

Let $I(\omega) := \{i \in \{1, \ldots, L\} : i = \omega_j, \text{ for some } j \in \{1, \ldots, E\}\}$ denote the indices of the buffers in service, and define for $x \in S^\psi$, $f^\psi(e_1) := \lambda$, $f^\psi(e_{i+1} - e_i) := \mu_i$, for all $i \in I(\omega)$, $f^\psi((0, \ldots, 0)^T) := 1 - \lambda - \sum_{i \in I(\omega)} \mu_i$, and $f^\psi(z) := 0$ for all other z. Now note that in each region $\mathcal{S}^\psi$ the transition probabilities exhibit translational invariance. Clearly, the regions $\overline{\mathcal{S}^\psi}$ are polyhedra, as substitution of the form of the affine index reveals: $\overline{S}^\psi = \{x \in S : x_i = 0 \text{ if } \phi_i = 0, x_i \geq 1 \text{ if } \phi_i = 1, \sum_{n=1}^L m_n^{\omega_j} x_n - \sum_{n=1}^L m_n^\ell x_n \geq k_\ell - k_{\omega_j}, \forall \ell \in \sigma_j \text{ with } \phi_\ell = 1, \text{ and } \forall j \text{ with } \omega_j \neq 0\}$. This is of the form $\overline{\mathcal{S}}^\psi = \{x \in S : A^\psi x \geq b^\psi\}$. We associate the vector of dual variables y^ψ with the inequality constraints of $\overline{\mathcal{S}^\psi}$.

We now turn our attention to the form of the average cost inequality in the region

$\overline{\mathcal{S}^\psi}$, assuming that ω is the list of working buffers for all $x \in \overline{\mathcal{S}^\psi}$. Recalling $W(x) = p^T x + (1/2)x^T Q x$, we have

$$\begin{aligned} J &\leq \lambda[p^T e_1 + \frac{1}{2} e_1^T Q e_1] + \sum_{i \in I(\omega)} \mu_i [p^T (e_{i+1} - e_i) + \frac{1}{2}(e_{i+1} - e_i)^T Q (e_{i+1} - e_i)] \\ &\quad + [e + \lambda Q e_1 + \sum_{\{i \in I(\omega)\}} \mu_i Q (e_{i+1} - e_i)]^T x \\ &=: d^\psi + r^{\psi^T} x, \text{ for all } x \in \overline{\mathcal{S}^\psi}. \end{aligned} \tag{12}$$

The fixed linear form is evident, and we note that d^ψ and r^ψ are linear in p and Q (symmetric), respectively. Application of Theorem 2.2 yields LP performance bounds.

Theorem 3.5. LP performance bounds for open re–entrant lines under affine index policies.

(i) *Let (J, p, Q (symmetric), and vectors $y^\psi, \forall \psi \in \Psi$) be the decision variables in the following linear program $\underline{T}$:*

$$Max\ J$$

subject to

$$\begin{aligned} A^{\psi^T} y^\psi &\leq r^\psi, \\ b^{\psi^T} y^\psi &\geq J - d^\psi, \\ y^\psi &\geq 0, \end{aligned}$$

for all $\psi \in \Psi$. If the $W(x)$ resulting from a feasible solution (J, p, Q (symmetric), and $y^\psi, \forall \psi \in \Psi$) is bounded above, then J is a lower bound on the average cost as in (2).

(ii) *Let (J, p, Q (symmetric), and vectors $y^\psi, \forall \psi \in \Psi$) be the decision variables in the following linear program $\overline{T}$:*

$$Min\ J$$

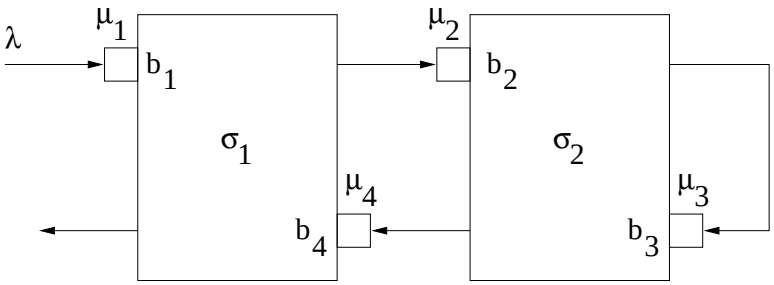

Figure 3: An open re–entrant line under the FSMCT policy.

subject to

$$\begin{aligned} A^{\psi^T} y^\psi &\leq -r^\psi, \\ b^{\psi^T} y^\psi &\geq d^\psi - J, \\ y^\psi &\geq 0, \end{aligned}$$

for all $\psi \in \Psi$. If the $W(x)$ resulting from a feasible solution (J, p, Q (symmetric), and $y^\psi, \forall \psi \in \Psi$) is bounded below, then J is an upper bound on the average cost as in (3).

Example 3.6: Polyhedral translational invariance for a re–entrant line operating under the FSMCT policy. *Consider the network of Figure 3. There are two stations σ_1 and σ_2 and four buffers. We will demonstrate the linear constraints for $\overline{S}^\psi$ where $\phi = (1,1,1,1)^T$ and $\omega = (1,3)^T$ The $\psi = (\phi, \omega)$ index indicates $x_i \geq 1, \forall i$, and $\eta_1(x) \geq \eta_4(x)$, $\eta_3(x) \geq \eta_2(x)$. Let $k_i := \sum_{j=i}^{L} \overline{x}_j$, where $\overline{x}_j \geq 0$. Note that $k_{i+1} \leq k_i, \forall i$. The constraints characterizing this region are given by*

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \geq \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ k_4 - k_1 \\ k_2 - k_3 \end{pmatrix}.$$

□

Example 3.7: Reduced complexity performance bounds for linear switching curves. *Consider a re–entrant line operating under an affine index policy given by $\eta_i(x) =$*

$m_i x_i, \forall i$, *with* $m_i > 0, \forall i$. *Let* $\psi = (\phi, \omega)$ *be as before and define*

$$\overline{S}^{\psi} := \{x \in S \ : \ x_i \geq 0 \text{ if } \phi_i = 1, x_i = 0 \text{ if } \phi_i = 0, \forall i,$$
$$m_{\omega_j} x_{\omega_j} - m_n x_n \geq 0, \forall j \text{ with } \omega_j \neq 0, \forall n \in \sigma_j \text{ with } \phi_n = 1\}.$$

These sets are supersets of the regions defined in the development of the affine index policy LP bound theorem, since x_i *is only required to be nonnegative rather than* ≥ 1 *when* $\phi_i = 1$, *and we still have* $\cup_{\{\psi \in \Psi\}} \overline{S}^{\psi} = S$.

If we require the average cost inequality of (12) to hold for all $x \in \overline{S}^{\psi}$, *for all* $\psi \in \Psi$, *we will obtain a performance bound as in Theorem 3.5. The utility of considering these supersets is that we are then able to eliminate certain* ψ*'s from consideration, as the constraints resulting from the average cost inequality on those* $S^{(\phi,\omega)}$ *are implied by another* (ϕ', ω). *Let* $\gamma(\sigma) := (\gamma_1(\sigma), \ldots, \gamma_L(\sigma))^T$ *be a vector (depending upon the station argument* σ*) with* $\gamma_i(\sigma) = 1$ *if* $i \in \sigma$, *and* 0 *otherwise, (i.e.,* $\gamma_i(\sigma) = I_{\{i \in \sigma\}}$*). One can see that only those* ϕ*'s which may be constructed as a sum* $\phi = \sum_\sigma w_\sigma \gamma(\sigma)$, *where* $w_\sigma \in \{0, 1\}$, *need to be considered in constructing the linear program. Further, one can see that an explicit form for* J *in terms of* Q *can be obtained by setting*

$$p_i = \sum_{j=i+1}^{L} q_{j,j} + (1/2) q_{i,i} - \sum_{j=i}^{L-1} q_{j,j+1}.$$

The result is that every inequality in which J *appears is satisfied with equality such that*

$$J = \lambda \sum_{i=1}^{L} q_{i,i} - \lambda \sum_{i=1}^{L-1} q_{i,i+1}.$$

This corresponds to the cost function of the duals of the linear programs of [14], and thus suggests a connection which may be worth pursuing.

For the example in Figure 4, let $\mu_1 = \mu_4 = \mu_5 = \mu_6 = 4$ *and* $\mu_2 = \mu_3 = 2$, *and consider the linear switch curve policy described by* $m_1 = 10$, $m_2 = 1$, $m_3 = 10$, $m_4 = 1$, $m_5 = 1$, *and* $m_6 = 10$. *The maximum sustainable throughput rate is* $\lambda^* = 1$. *Upper bounds for the mean number of lots (or WIP) are given in Figure 5. Note the superiority of our new upper bound.*

□

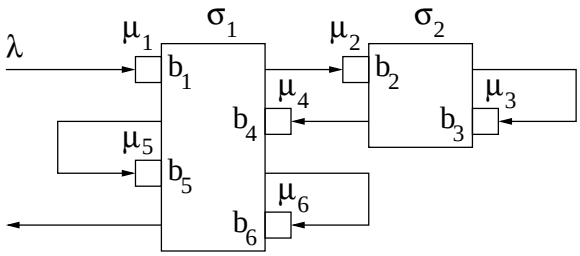

Figure 4: An open re–entrant line under a linear switch curve policy.

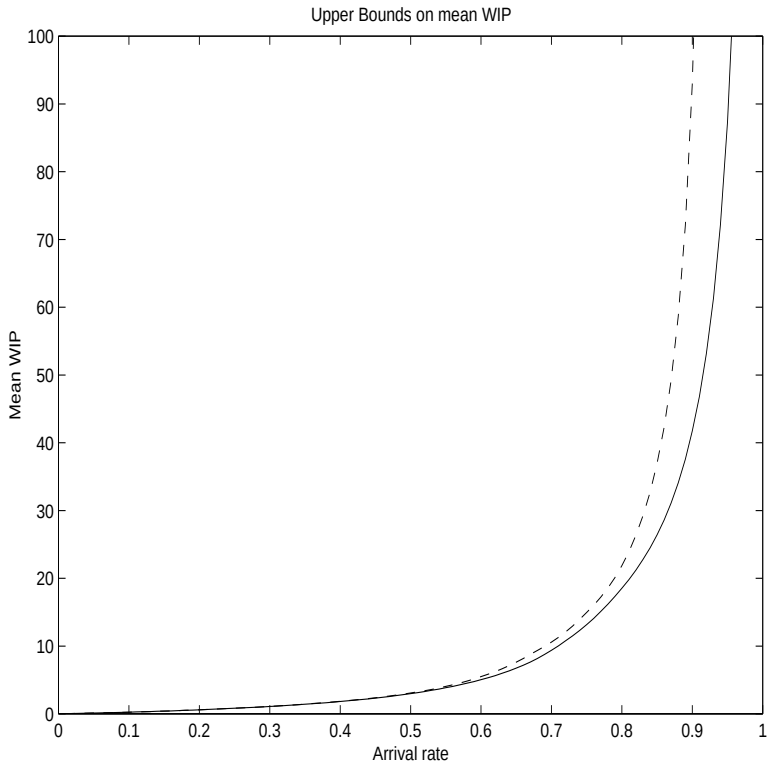

Figure 5: Upper bounds on mean WIP for Example 3.7. The solid curve denotes the upper bound obtained through the reduced complexity approach given here. The dashed curve denotes the upper bound obtained through the approach of [14] by incorporating linear constraints arising from the linear switch curve policy.

3.4 Improved bounds via piecewise quadratic surrogates of the differential cost function

The bounds obtained in Section 3.3 can be improved by allowing piecewise quadratic forms for the surrogate of the differential cost function. Piecewise linear functions have been employed for stability analysis in [34] and in performance analysis [20]. For a proposal for constructing a control law based on a connection between the differential cost function and the piecewise quadratic fluid total cost viewed as an $o(|x|^2)$ approximation, and other related references, we refer the reader to [31]. For performance bounds, piecewise quadratic functions allow for tighter bounds than attainable with a fixed quadratic function, as shown in [18, 21]. Such piecewise quadratic functions dovetail very naturally with a polyhedral decomposition, allowing us to separate the resulting quadratic form obtained for the average cost inequality into quadratic and linear terms, as we demonstrate below.

Consider an open re–entrant line operating under an affine index policy. With $\phi(x) : \mathcal{Z}_+^L \to \{0,1\}^L$ defined as earlier, let ϕ denote an element of $\{0,1\}^L$, and use Φ to denote the set of all ϕ's for which $\mathcal{X}^\phi := \{x \in S : \phi(x) = \phi\}$. Let the function $\xi(x) : \mathcal{Z}_+^L \to \{0,1,2\}^L$ be defined as $\xi_i(x) = 0$, if $x_i = 0$, $\xi_i(x) = 1$, if $x_i = 1$, $\xi_i(x) = 2$, if $x_i \geq 2$. Let ξ denote an element in the range of the function $\xi(\cdot)$, so that $\xi = (\xi_1, \ldots, \xi_L)^T$ is a vector with $\xi_i \in \{0,1,2\}$. Define $\mathcal{Y}^\xi := \{x \in S : \xi(x) = \xi\}$. Let $W^\phi(x) = c^\phi + p^{\phi^T} x + \frac{1}{2} x^T Q^\phi x$, where Q is a symmetric matrix, and define the surrogate of the differential cost function as the composite piecewise quadratic function

$$W(x) := W^\phi(x), \qquad \forall x \in \mathcal{X}^\phi.$$

The goal here is to identify polyhedral regions of the state space for which the average cost inequality has a fixed *quadratic* form.

As before, let $\omega = (\omega_1, \ldots, \omega_E)$ be a vector with $\omega_j \in \{0\} \cup \{i : \sigma(i) = \sigma_j\}$. Let $\psi = (\xi, \omega)$ and use S^ψ to denote the region of the state space in which $x \in \mathcal{Y}^\xi$ and the buffers indicated by ω are in service (or no buffer from a particular station σ_j is in service if $\omega_j = 0$). We let

Ψ denote the set of $\psi = (\xi, \omega)$ for which S^ψ is nonempty. Define the following subset of S:

$$\begin{aligned}\overline{S}^\psi := \{x \in S \quad : \quad & x_i = 0 \text{ if } \xi_i = 0, x_i = 1 \text{ if } \xi_i = 1, x_i \geq 2 \text{ if } \xi_i = 2, \\ & \sum_{n=1}^{L} m_n^{\omega_j} x_n - \sum_{n=1}^{L} m_n^{\ell} x_n \geq k_\ell - k_{\omega_j}, \\ & \forall \ell \in \sigma_j \text{ with } \xi_i \geq 1, \forall j \text{ with } \omega_j \neq 0\},\end{aligned}$$

so that $S^\psi \subseteq \overline{S}^\psi$. If we can ensure that the average cost inequality holds for this set with the working buffers given by ω, then the average cost inequality will hold on S^ψ.

Notice that each $x \in \overline{S}^\psi$ can be written as $x = z + \xi$, where z is an element of

$$\begin{aligned}\overline{Z}^\psi := \{z \in \mathcal{Z}_+^L \quad : \quad & z_i = 0 \text{ if } \xi_i = 0, z_i = 0 \text{ if } \xi_i = 1, z_i \geq 0 \text{ if } \xi_i = 2, \\ & \sum_{n=1}^{L} (m_n^{\omega_j} - m_n^{\ell}) z_n \geq k_\ell - k_{\omega_j} + \sum_{n=1}^{L} (m_n^{\ell} - m_n^{\omega_j}) \xi_n, \\ & \forall \ell \in \sigma_j \text{ with } \xi_i \geq 1, \forall j \text{ with } \omega_j \neq 0\}.\end{aligned}$$

Let the matrices A_{in}^ψ and A_{eq}^ψ and the vectors b_{in}^ψ and b_{eq}^ψ characterize the set $\overline{Z}^\psi$, in the same manner as in the previous section. Using this substitution and the function $\xi(x)$, one can see, as in [18] and [21], that there is a constant d^ψ, a vector r^ψ and a symmetric matrix M^ψ, all linearly dependent upon the variables c^ϕ, p^ϕ, and Q^ϕ, such that the average cost inequality for a lower bound on $\overline{S}^\psi$ will hold if one ensures the following inequality:

$$J \leq d^\psi + r^{\psi^T} z + \frac{1}{2} z^T M^\psi z,$$

for all $z \in \overline{Z}^\psi$. Hence, the average cost inequality has a fixed form in the region. This argument can be repeated for each $\psi \in \Psi$.

In order to ensure that these constraints hold, we first require that the quadratic terms hold and then apply the theory of Section 2.2 to the remaining linear inequality on the polyhedron $\overline{Z}^\psi$. We thus have the following theorem, which provides tighter bounds though at the cost of increased computational complexity.

Theorem 3.8: Improved LP performance bounds via a piecewise quadratic surrogate for the differential cost function for re–entrant lines under affine index policies.

(i) *Let $(J, c^\phi, p^\phi, Q^\phi$ (symmetric) $\forall\phi$ and vectors $y_{in}^\psi, y_{eq}^\psi, \forall\psi \in \Psi)$ be the decision variables in the following linear program $\underline{T}$:*

$$Max\ J$$

subject to

$$\begin{aligned} A_{in}^{\psi^T} y_{in}^\psi + A_{eq}^{\psi^T} y_{eq}^\psi &\leq r^\psi, \\ b_{in}^{\psi^T} y_{in}^\psi + b_{eq}^{\psi^T} y_{eq}^\psi &\geq J - d^\psi, \\ M^\psi &\geq 0, \\ y_{in}^\psi &\geq 0, \end{aligned}$$

for all $\psi \in \Psi$. If the $W(x)$ resulting from a feasible solution $(J, c^\phi, p^\phi, Q^\phi$ (symmetric) $\forall\phi$ and vectors $y_{in}^\psi, y_{eq}^\psi, \forall\psi \in \Psi)$ is bounded above, then J is a lower bound on the average cost as in (2).

(ii) *Let $(J, c^\phi, p^\phi, Q^\phi$ (symmetric) $\forall\phi$ and vectors $y_{in}^\psi, y_{eq}^\psi, \forall\psi \in \Psi)$ be the decision variables in the following linear program $\overline{T}$:*

$$Min\ J$$

subject to

$$\begin{aligned} A_{in}^{\psi^T} y_{in}^\psi + A_{eq}^{\psi^T} y_{eq}^\psi &\leq -r^\psi, \\ b_{in}^{\psi^T} y_{in}^\psi + b_{eq}^{\psi^T} y_{eq}^\psi &\geq d^\psi - J, \\ M^\psi &\leq 0, \\ y_{in}^\psi &\geq 0, \end{aligned}$$

for all $\psi \in \Psi$. If the $W(x)$ resulting from a feasible solution $(J, c^\phi, p^\phi, Q^\phi$ (symmetric) $\forall\phi$ and vectors $y_{in}^\psi, y_{eq}^\psi, \forall\psi \in \Psi)$ is bounded below, then J is an upper bound on the average cost as in (3).

For the class of buffer priority policies, Theorem 3.8 is equivalent to the LPs of [18] for open re–entrant lines and to the LPs of [21] for closed re–entrant lines.

To obtain bounds for the class of all non–idling policies one needs to ensure that the average cost inequality holds at each state in the state space no matter which buffer is chosen for service. Thus one must construct an LP which provides that the average cost inequalities hold no matter which buffer receives service. The result is more constraints. The bounds for the class of all non–idling policies are naturally looser than a bound for a specific scheduling policy.

If one can deduce regions of the state space which are transient, then those regions need not be included when verifying the average cost inequality. The bounds are thereby improved (fewer linear constraints result) and the complexity is reduced. See [18] and [21] for some ways to identify transient regions of the state space. Also, [21] suggests methods for reducing the computational complexity of LP performance bounds. The simplifying assumption that the service time distributions are exponential and arrivals are Poisson may be relaxed via the method of stages [33], see [14].

4 Bounds for unreliable manufacturing systems

Consider a system as in [3] or [2] with the additional complications that the demand for finished goods as well as the production process are themselves both random. As shown in Figure 6, there is a machine which can either be in an "Up" state or in a "Down" state, wandering back and forth between the two as a continuous time Markov chain with rates q_0 and q_1. When the machine is up it can produce goods as a Poisson process of rate μ, with the proviso that if we wish we can idle it, which means no goods are produced. When the machine is down, no goods are produced. The demand for finished goods is a Poisson process of rate λ. Let $x(t)$ denote the inventory of finished goods, defined as the cumulative production up to time t minus the cumulative demand, which indicates a backlog when it is negative, and a surplus when it is positive. Positive inventory incurs a holding cost of c_+

per unit time, while negative inventory incurs a backlog cost of c_- per unit time. As in [1], we consider a "hedging point" policy with N as the "produce up to" level:

If $(x(t) \leq N$ and machine is up), then produce at rate μ
$(x(t) > N$ or machine is down), then there is no production.

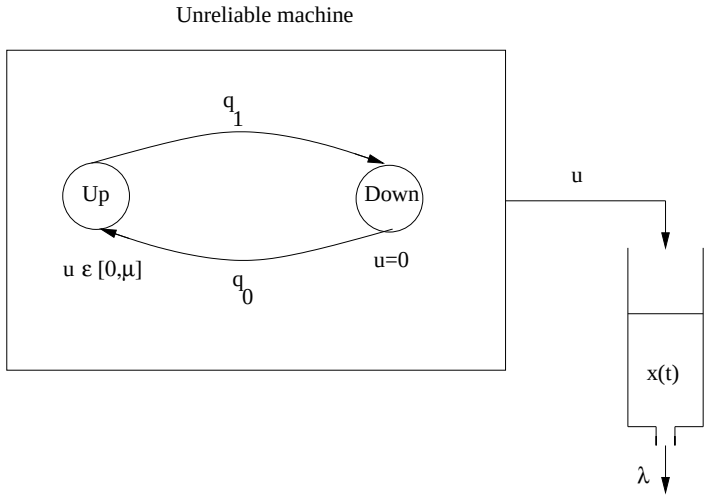

Figure 6: An unreliable machine

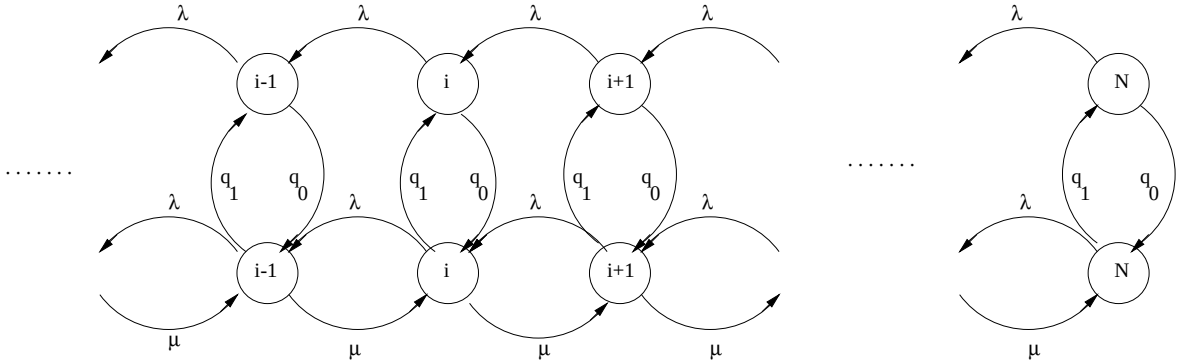

Figure 7: The Markov chain modeling the unreliable manufacturing system under a hedging point policy.

The cost of any such policy can be explicitly solved by analytically computing the steady state distribution of the Markov chain shown in Figure 7. The necessary condition for stability, i.e., finite mean inventory level, is $\frac{q_0}{q_0+q_1} \cdot \mu > \lambda$, as can be seen from comparing the mean production capability and demand rates. Denote by

$$\rho := \frac{\lambda(q_0 + q_1)}{\mu q_0}$$

the load on the system. Let

$$s_{\pm} \;=\; \frac{1 \pm \sqrt{1 - 4\mu(q_0+\lambda)}}{2}.$$

If $\pi_0(i)$ = Prob (System is "Down" and inventory level is i), and $\pi_1(i)$ = Prob (System is "Up" and inventory level is i), then it can be computed that

$$\begin{bmatrix} \pi_0(i) \\ \pi_1(i) \end{bmatrix} \;=\; \left(\frac{\lambda}{\mu(q_0+\lambda)}\right)^{N-i} \begin{bmatrix} \frac{s_- - 1}{s_- - s_+} \cdot s_+^{N-i} + \frac{1-s_+}{s_- - s_+} \cdot s_-^{N-i} \\ \frac{s_+ - \mu - q_1}{q_1} \cdot \frac{s_- - 1}{s_- - s_+} \cdot s_+^{N-1} + \frac{s_- - \mu - q_1}{q_1} \cdot \frac{1-s_+}{s_- - s_+} \cdot s_-^{N-i} \end{bmatrix}.$$

From this, the average cost J^* can be computed as

$$\begin{aligned} J^* \;=\; & \left\{ \frac{s_+ - \mu}{q_1} \cdot \frac{s_- - 1}{s_- - s_+} \left[\frac{N + \theta_+^{N+1}}{1-\theta_+} - \frac{c_+(\theta_+ - \theta_+^{N+2})}{(1+\theta_+)^2} + \frac{c_- \theta_+^{N+1}}{(1-\theta_+)^2} \right] \right. \\ & \left. + \frac{s_- - \mu}{q_1} \cdot \frac{1-s_+}{s_- - s_+} \left[\frac{c_+(N + \theta_-^{N+1})}{1-\theta_-} - \frac{c_+(\theta_- - \theta_-^{N+2})}{(1-\theta_-)^2} + \frac{c_- \theta_-^{N+1}}{(1-\theta_-)^2} \right] \right\} \cdot \\ & \cdot \left\{ \frac{s_+ - \mu}{q_1} \cdot \frac{s_- - 1}{s_- - s_+} \cdot \frac{1}{1-\theta_+} + \frac{s_- - \mu}{q_1} \cdot \frac{1-s_+}{s_- - s_+} \cdot \frac{1}{1-\theta_-} \right\}^{-1}, \end{aligned}$$

where

$$\theta_{\pm} \;:=\; \frac{\lambda s_{\pm}}{\mu(q_0+\lambda)}.$$

Since the explicit solution is known, this example allows us to compare the performance of our bounds. Figure 8 presents the results for $q_0 = 9$, $q_1 = 1$, and $c_+ = c_- = 1$, as a function of the system loading ρ. The optimal value of the hedging point is $N = 6$ at $\rho = 0.9$.

For $N = 0$, the upper bound is exact. As N is increased, the exact solution initially hugs the lower bound for small system loading ρ, while for large ρ it hugs the upper bound. When N is very large, it is the lower bound which is nearly exact for $\rho = 0, 0.1, 0.2, \ldots, 0.9$, the range for which computations were done). It should be noted that for $N = 0$, the exact cost increases with system loading which is captured by the upper bound. However, when N is very large, say $N = 100$, the exact cost decreases as the system loading is increased since machine failures actually decrease the buffer size from the high value of 100 until ρ is fairly large, which is in fact captured by the lower bound.

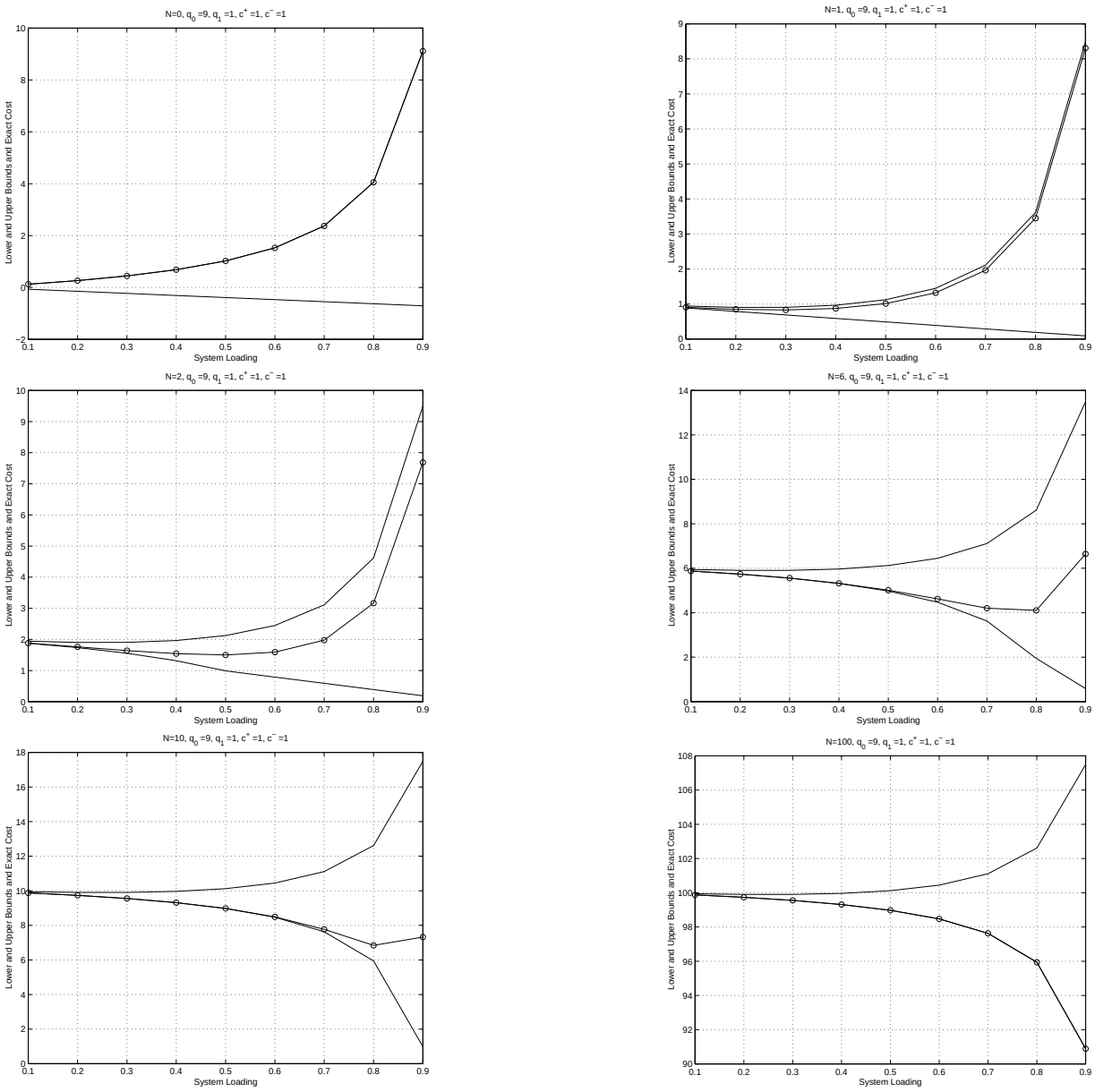

Figure 8: A comparison of the lower and upper bounds in relation to the exact value.

5 Bounds for manufacturing systems with batch tools

To model features important to semiconductor manufacturing plants, we need to consider stations which may feature batch or serial tools. A serial station (or serial tool) can process a single lot at a time as in the basic re–entrant line model. However, a batch tool σ, can

process up to M_σ^{max} lots at a time. Thus a serial tool σ is equivalently a batch tool with $M_\sigma^{max} = 1$, and so we will use $M_\sigma^{max} = 1$ to denote a serial station σ. All lots loaded into a batch tool are served concurrently with a common exponentially distributed service time of rate μ_i, depending upon the buffer b_i from which they are loaded, and independent of the number of lots loaded. We allow lots to be loaded from *different* buffers, as can happen in some wafer fabs. Upon completion of service, the lots are released to their next appropriate stage of processing.

For a batch tool, the critical scheduling decision is to choose how many lots to load into the tool. One common policy is to choose a minimum batch size M_σ^{min}, and to idle the tool until that many lots are present at the tool. At that time, all the lots available (at least M_σ^{min}) are loaded into the tool for processing. If less than M_σ^{min} lots are available for processing, the station σ is allowed to remain idle. Again, for a serial tool σ, we let $M_\sigma^{min} = 1$.

The state space of this model (employing uniformization as before with $\lambda + \sum_{i=1}^{L} \mu_i = 1$), may be considered as the vector $x = (x_1, \ldots, x_L)^T$, where x_i is the number of lots in buffer b_i. Again, we allow the stationary scheduling policy to be an affine index policy assigning to each buffer an index $\eta_i(x) = k_i + \sum_{j=1}^{L} m_j^i x_j$. The implementation of such a policy is clear when the batch size is one. However, for batch sizes greater than two, there is the additional issue of how to decide when multiple lots are to be loaded.

Focusing attention on batch tool scheduling, to implement the affine index policy, one chooses from among those buffers with at least the minimum number of lots. (If lots may be loaded concurrently from distinct buffers, then one needs to consider the sum of the lots in those buffers vying for contention with the minimum lot requirement M_σ^{min} of the batch loading rule). If a buffer b_i which may be processed concurrently with another buffer is scheduled to receive service (i.e., $\eta_i \geq \eta_k$, for all k with $x_k \geq M_{\sigma(i)}^{min}$, or k is such that $k \in C_k := \{l : \mu_l = \mu_k, \sigma(l) = \sigma(k) = \sigma(i)\}$ and $\sum_{l \in C_k} x_l \geq M_{\sigma(i)}^{min}$), we assume that as many lots as possible are loaded into σ from b_i. If the station can still process more lots, as many lots as possible are loaded from the next buffer $b_j \in C_i$ with the highest affine policy index

from among all buffers $b_k \in C_i$. Lots are loaded in this manner until station σ is fully loaded or no lots remain in the buffers in C_i. As before, ties may be broken arbitrarily, or at the dictate of some priority policy.

To identify polyhedral regions of the state space for which the transition probabilities are translation invariant, we proceed in a manner similar to that of Section 3.3 for the basic open re–entrant line. Let $\phi = (\phi_1, \ldots, \phi_L)^T$ be a vector with $\phi_i \in \{0, 1, \ldots, M^{max}_{\sigma(i)}\}$. Use Φ to denote the set of all such vectors. Let $n = (n_1, \ldots, n_L)^T$ be a vector with $n_i \in \{0, 1, \ldots, M^{max}_{\sigma(i)}\}$ for all $i \in \{1, \ldots, L\}$, $M^{min}_{\sigma} \leq \sum_{i \in \sigma} n_i \leq M^{max}_{\sigma}$ for all σ with $\sum_{i \in \sigma} n_i \neq 0$, and if $n_i \neq 0$ then $n_j = 0$ for all $j \in \sigma(i)$ with $j \notin C_i$. Use $\mathcal{N}$ to denote the set of all such vectors. We interpret n as a list of the number of lots receiving service from each buffer. As there may be multiple buffers receiving service at a station with different values of the affine index η_i, we need to distinguish between states for which the same n results but different buffers achieve the maximum η. To do this, we need to explicitly consider the case where multiple buffers are in service at a batch station and a particular ordering of the buffers results from their affine indices.

Recall that $C_i := \{k : \mu_i = \mu_k, \sigma(i) = \sigma(k)\}$ is the set of buffers in the same concurrency class as buffer b_i (it may contain only the buffer index itself, i.e., $C_i = \{i\}$). For a given $\phi \in \Phi$ and $n \in N$, let $\rho(n, \sigma) = (\rho_1(n, \sigma), \ldots, \rho_{l(n,\sigma)}(n, \sigma))$ be a permutation of the buffer subscripts (i.e., i for buffer b_i) of the buffers receiving service at σ, where $l(n, \sigma)$ is the number of buffers concurrently receiving processing from σ (if no buffers are in service, or if only one buffer is receiving service, the permutation is empty or has only one element, respectively). We interpret $\rho(n, \sigma)$ as the priority ordering of the buffers receiving service at σ according to the affine index policy. Let $\mathcal{T}(n, \sigma)$ be the set of all such permutations for a given n and σ. Given $\phi \in \Phi$ and $n \in N$, for each permutation $\rho(n, \sigma) \in \mathcal{T}(n, \sigma)$ for each σ, let $\psi = (\phi, n, \rho(n, \sigma_1), \ldots, \rho(n, \sigma_E))$ and define

$$\begin{aligned} \overline{S}_1^{\psi} := \{x \in S \ : \ & x_i \geq M^{max}_{\sigma} \text{ if } \phi_i = M^{max}_{\sigma}, x_i = \phi_i \text{ otherwise}, \\ & x_i = n_i \text{ if } M^{min}_{\sigma(i)} \leq \sum_{j \in C_i} n_j < M^{max}_{\sigma(i)}, \end{aligned}$$

$$x_i \geq M^{max}_{\sigma(i)}, \forall i \text{ with } n_i = M^{max}_{\sigma(i)},$$
$$x_{\rho_i(n,\sigma)} = n_{\rho_i(n,\sigma)}, i = 1, \ldots, l(n,\sigma) - 1, \text{ and}$$
$$x_{\rho_{l(n,\sigma)}(n,\sigma)} \geq M^{max}_{\sigma} - \sum_{i=1}^{l(n,\sigma)-1} n_{\rho_i(n,\sigma)},$$
$$\forall \sigma \text{ with } \sum_{i \in \sigma} n_i = M^{max}_{\sigma(i)} \text{ and } l(n,\sigma(i)) \geq 2\}$$

and

$$\overline{S}^{\psi}_2 := \{x \in S \ : \ \eta_{\rho_1(n,\sigma)}(x) \geq \eta_j(x), \forall j \in \sigma \text{ with } \sum_{k \in C_j} \phi_k \geq M^{min}_{\sigma},$$
$$\eta_{\rho_i(n,\sigma)}(x) \geq \eta_{\rho_{i+1}(n,\sigma)}(x), i = 2, \ldots, l(n,\sigma) - 1,$$
$$\forall \sigma \text{ with } \sum_{k \in \sigma} n_k \neq 0\}.$$

Let $\overline{S}^{\psi} := \overline{S}^{\psi}_1 \cap \overline{S}^{\psi}_2$ denote their intersection. These regions are polyhedra described by linear constraints. Let Ψ denote the set of ψ for which $\overline{S}^{\psi}$ is nonempty.

Letting $c(x) = |x| = \sum_{i=1}^{L} x_i$ and $W(x) = p^T x + (1/2) x^T Q x$, with Q symmetric, the average cost inequalities for a lower bound for all $x \in \overline{S}^{\psi}$ take the form

$$\begin{aligned} J \ \leq \ & \lambda[p^T e_1 + (1/2) e_1^T Q e_1] \\ & + \sum_{\{i \text{ with } n_i \neq 0\}} \mu_i [n_i p^T (e_{i+1} - e_i) + (1/2) n_i^2 (e_{i+1} - e_i)^T Q (e_{i+1} - e_i)] \\ & + [e + \lambda Q e_i + \sum_{\{i \text{ with } n_i \neq 0\}} \mu_i n_i Q (e_{i+1} - e_i)]^T x. \end{aligned}$$

Example 4.1: Batch tools. *Consider the network of Figure 9. Station σ_1 is a serial station, that is, $M^{min}_{\sigma_1} = 1 = M^{max}_{\sigma_1}$, and station σ_2 is a batch station with $M^{min}_{\sigma_2} = 3$ and $M^{max}_{\sigma_2} = 5$, as indicated by the wafer–like circles. Here we assume that the stations are operating under the last buffer first serve (LBFS) buffer priority policy, that is, $\eta_1 = 1$ and $\eta_3 = 3$ at serial station σ_1 (so that buffer b_3 has priority over buffer b_1) and $\eta_2 = 2$ and $\eta_4 = 4$ at batch station σ_2 (hence buffer b_4 is given priority). For $\mu_1 = \mu_3 = 10$ and $\mu_2 = \mu_4 = 2$, the maximum possible throughput obtainable is $\lambda = 5$. The bounds obtained given in Figure 10.*

Note that the upper bound for $M^{min}_{\sigma_2} = 3$ is better than that for $M^{min}_{\sigma_2} = 5$ in light traffic, but worse in heavy traffic. This behavior occurs because at low loading it can be advantageous

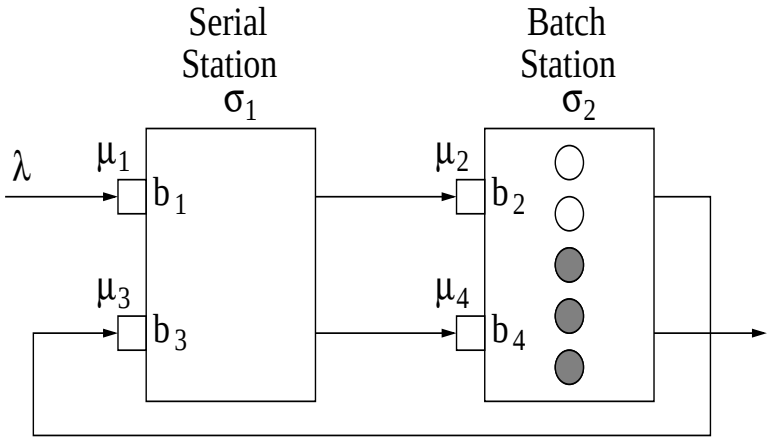

Figure 9: A re–entrant line with serial and batch servers.

to run partial loads at a batch tool. However at high loading, one needs to utilize the capacity more fully, hence higher batch sizes are necessitated.

6 Bounds for systems with setups

Returning to the open re–entrant line model of Section 3.1, we incorporate a station feature known as setups. Each station in a real plant may have multiple operating modes, possibly characterized by a tool head or a calibration. Prior to beginning service on a lot in a particular buffer, the station may have to change the tool head or calibration. To model this phenomenon, suppose that each station σ has $M(\sigma)$ possible setup modes, labeled $\{1, \ldots, M(\sigma)\}$. We assume that the change from one setup mode i to another setup mode j requires an exponentially distributed service time of rate $\mu_{i,j}^{\sigma}$, during which time the station is forced to idle (from the lot's perspective). We use $v_{i,j}^{\sigma}(t) = 1$ if the station is undergoing setup from mode i to j at time t, and 0 otherwise. The setup times are independent of each other, the arrival process, and the service time processes. Again using uniformization and letting $1 = \lambda + \sum_{i=1}^{L} \mu_i + \sum_{\sigma} \sum_{i,j \in \{1,\ldots,M(\sigma)\}} \mu_{i,j}^{\sigma}$, we obtain a controlled Markov chain, assuming that the scheduling policy is stationary (that is, it depends only upon the number of lots in the buffers and the station setup modes at time t).

The state for the resulting controlled discrete–time time–homogeneous Markov chain is

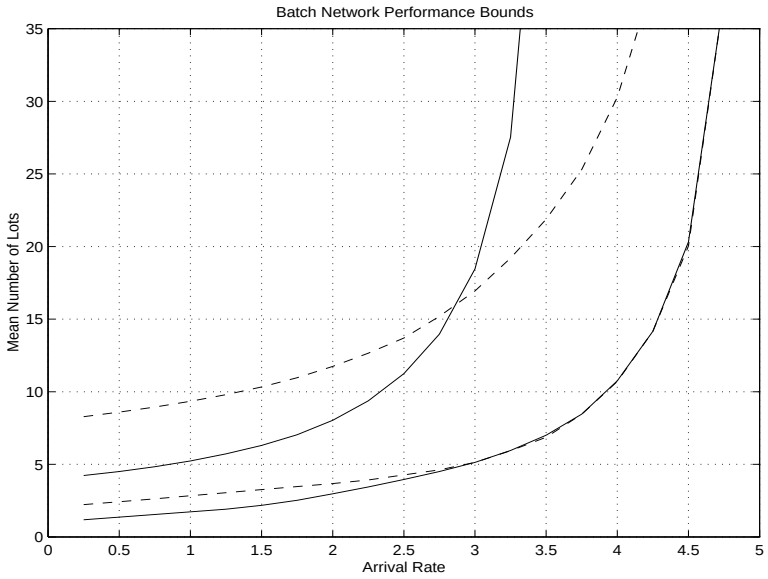

Figure 10: Performance bounds for the network of Example 4.1. The solid curves are the bounds for when a partial load is allowed, $M_{\sigma_2}^{min} = 3$. The dashed curves are bounds for when a full load is required, $M_{\sigma_2}^{min} = 5$.

$x = (y, z) = (y_1, \ldots, y_L, z_1, \ldots, z_E)$, where y_i is the number of lots in buffer b_i and z_k is the setup mode of station σ_k. We use $\text{Buff}(\sigma, k) \subseteq \{1, \ldots, L\}$ to denote the set of buffer indices which are enabled to receive service from station σ when that station is in setup mode $k \in \{1, \ldots, M(\sigma)\}$. We assume that these sets are disjoint and that for each station σ, $\cup_{k \in \{1,\ldots,M(\sigma)\}} \text{Buff}(\sigma, k) = \{i : i \in \sigma\}$. We let Δ denote the set of all possible vectors $z = (z_1, \ldots, z_E)$, and use $\delta \in \Delta$ to denote such a vector. Later, we will use the notation $\delta(i) = k$ to indicate that $i \in \text{Buff}(\sigma(i), k)$. In other words, the setup mode enabling buffer b_i to receive service from station $\sigma(i)$ is denoted by mode k, i.e., $\delta(i) = k$. Consider the function

$$W^z(y) = c^z + {p^z}^T x + (1/2) y^T Q^z y,$$

for all $y \in \mathcal{Z}_+^L$ and for each $z = (z_1, \ldots, z_E)$ where $z_k \in \{1, \ldots, M(\sigma_k)\}$. The composite $W(x) = W^z(y)$ for $x = (y, z)$ shall serve as the surrogate for the differential cost function.

Now generalize the affine index policies slightly to allow different indices to apply de-

pending upon the setup state z of the stations. That is,

$$\eta_i^z(y) := k_i^z + \sum_{j=1}^{L} m_j^{z,i} y_j$$

is the index for each buffer when the setup state of the stations is z.

Similar to the preceding, let $S^\phi := \{y \in \mathcal{Z}_+^L : y_i = 0 \text{ if } \phi_i = 0 \text{ and } y_i \geq 1 \text{ if } \phi_i = 1\})$, for each $\phi \in \Phi$. For each $\delta \in \Delta$, let

$$S^{\phi,\delta} := \{x \in S : y \in S^\phi, z = \delta\}.$$

Let $\omega_j \in \{0\} \cup \{i : \sigma(i) = \sigma_j\}$, for $j = 1, \ldots, E$, and $\omega = (\omega_1, \ldots, \omega_E)$. Use Ω to denote the set of all such ω's. For each $\phi \in \Phi$, $\delta \in \Delta$, and $\omega \in \Omega$, let $\psi = (\phi, \delta, \omega)$ and define

$$\begin{aligned} \overline{S}^\psi = \{x \in S^{\phi,\delta} \quad : \quad & \eta_{\omega_j}^\delta(y) \geq \eta_i^\delta(y), \\ & \forall i \in \sigma_j, i \neq j, \text{ with } \phi_i = 1, \forall j \text{ with } \omega_j(\phi) \neq 0\}. \end{aligned}$$

We suppose that for $x \in S^\psi$, if $\omega_j \neq 0$ and $\omega_j \in \text{Buff}(\sigma_j, \delta_j)$, then $u_{\omega_j}(x) = 1$ (recall that $u(x)$ is the vector of indicators of buffers receiving service), and station σ_j is processing a lot from buffer b_{ω_j}. Otherwise, the station σ_j is undergoing setup from $\delta_j \in \{1, \ldots, M(\sigma_j)\}$ to $\delta(\omega_j)$.

Use Ψ to denote the set of all indices ψ yielding nonempty $\overline{S}^\psi$. We then have that $\cup_{\psi \in \Psi} \overline{S}^\psi = S$ and the regions are polyhedra described by linear constraints. Now, use the notation $\text{Stat}(i)$ to indicate the station index of the station serving buffer b_i; that is, if $\sigma(i) = \sigma_j$, then $\text{Stat}(i) = j$. Also, let δ^i be the vector δ with the mode of station $\sigma(i)$ replaced by $\delta(i)$. Let $I(\omega) := \{i \in \{1, \ldots, L\} : i = \omega_j \text{ for some } j \in \{1, \ldots, E\}\}$. To obtain a lower bound, we seek to ensure the following inequality on the region $\overline{S}^\psi$:

$$\begin{aligned} J \leq c^T y \quad + \quad & \lambda[p^{\delta^T} e_1 + (1/2) e_1^T Q^\delta e_1 + y^T Q^\delta e_1] \\ + \quad & \sum_{\substack{i \in I(\omega) \\ \{_{i \in} \text{Buff}_{(\sigma(i),\delta}} \text{Stat}_{(i)}{}^{)}\}}} \mu_i [p^{\delta^T}(e_{i+1} - e_i) + (1/2)(e_{i+1} - e_i)^T Q^\delta (e_{i+1} - e_i) \\ & + y^T Q^\delta (e_{i+1} - e_i)] \end{aligned}$$

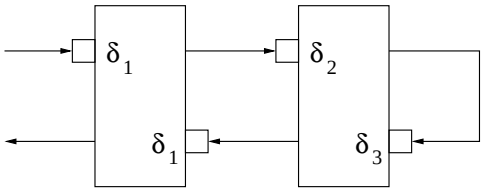

Figure 11: A network with setups.

$$+ \sum_{\substack{i \in I(\omega \\ \{_{i \notin} \text{Buff}_{(\sigma(i),\delta_{\text{Stat}_{(i)}})}\}}} \mu^{\sigma(i)}_{\delta_{\text{Stat}_{(i)}},\delta(i)} [c^{\delta^i} - c^{\delta} + p^{\delta^i T} y - p^{\delta T} y + (1/2) y^T Q^{\delta^i} y - (1/2) y^T Q^{\delta} y],$$

for all $x \in \overline{S}^{\psi}$. We thus see that the average cost inequalities for the proposed "partition" have a fixed quadratic form and we may apply the approach of Section 2 to obtain performance bounds.

Example 4.2. Bounds on systems with setups. *Consider the network of Figure 11, where station σ_1 has one setup state and station σ_2 has two setup states. Thus, the possible setup states for the network are $(1,2)$ and $(1,3)$. The setup changeover transition rates are $\mu^{\sigma_2}_{2,3} = 1.0$ and $\mu^{\sigma_2}_{3,2} = 2.0$. Let the arrival rate be $\lambda = 1$ and set $\mu_i = 3$, for all i. The scheduling policy under consideration is $\eta^{(1,2)}_i = x_i$ and $\eta^{(1,3)}_i = 2x_i$, for all i. With these choices we obtain the performance bound shown in Figure 12.* □

	Mean WIP
Upper Bound	93.03
Lower Bound	41.74

Figure 12: Performance bounds for a network with setups.

7 Concluding remarks

We have shown how to exploit the property of translational invariance of the transition probabilities on polyhedra to obtain performance bounds for Markov chains. This general

framework allows us to model a variety of manufacturing systems featuring complexities such as unreliable machines, set-ups, batch tools, or operating under more general classes of scheduling policies. It is hoped that methodologies such as these will take their place alongside the more established simulation modeling in the analysis and design of manufacturing systems.

References

[1] 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.

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

[3] 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.

[4] P. H. Algoet, "Flow balance for the steady state distribution of a flexible manufacturing system," *IEEE Transactions on Automatic Control*, vol. AC-34, pp. 917–921, August 1989.

[5] A. Sharifnia, "Production control of a manufacturing system with multiple machine states," *IEEE Transactions on Automatic Control*, vol. AC-33, pp. 620–625, July 1988.

[6] J. R. Perkins and R. Srikant, "Scheduling multiple part-types in an unreliable single machine manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. AC-42, pp. 364–377, March 1997.

[7] C. Shu and J. R. Perkins, "Optimal PHP production of multiple part types on a failure prone machine with quadratiic buffer costs," *IEEE Transactions on Automatic Control*, vol. 46, pp. 541–549, April 2001.

[8] J. R. Perkins and R. Srikant, "Failure-prone production systems with uncertain demand," *IEEE Transactions on Automatic Control*, vol. 46, pp. 441–449, March 2001.

[9] G. Liberopoulos and M. Caramanis, "Production control of manufacturing systems with production rate-dependent failure rates," *IEEE Transactions on Automatic Control*, vol. 39, pp. 889–895, April 1994.

[10] M. H. Veatch and M. C. Caramanis, "Optimal manufacturing flow controllers: Zero-inventory policies and control switching sets," *IEEE Transactions on Automatic Control*, vol. 44, pp. 914–921, May 1999.

[11] S. P. Sethi, Q. Zhang, and X. Y. Zhou, "Hierarchical production controls in a stochastic two machine flowshop with finite internal buffer," *IEEE Transactions on Robotics and Automation*, vol. 13, pp. 1–13, February 1997.

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

[13] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, "Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance," *Annals of Applied Probability*, vol. 4, pp. 43–75, 1994.

[14] S. Kumar and P. R. Kumar, "Performance bounds for queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. AC-39, pp. 1600–1611, August 1994.

[15] P. R. Kumar and S. P. Meyn, "Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies," *IEEE Transactions on Automatic Control*, vol. 41, pp. 4–17, January 1996.

[16] H. Jin, J. Ou, and P. R. Kumar, "The throughput of irreducible closed markovian queueing networks: Functional bounds, asymptotic loss, efficiency, and the harrison-wein conjectures," *Mathematics of Operations Research*, vol. 22, no. 4, pp. 886–920, 1997.

[17] C. Humes, Jr., J. Ou, and P. R. Kumar, "The delay of open markovian queueing networks: Uniform functional bounds, heavy traffic pole multiplicities, and stability," *Mathematics of Operations Research*, vol. 22, pp. 921–954, November 1997.

[18] J. R. Morrison and P. R. Kumar, "New linear program performance bounds for queueing networks," *Journal of Optimization Theory and Applications*, vol. 100, pp. 575–697, March 1999. Plenum Publishers.

[19] D. Bertsimas and J. Nino-Mora, "Optimization of multi-class queueing networks with changeover times via the achievable region approach: Part ii, the multiple state case," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 331–361, 1999.

[20] D. Bertsimas, D. Gamarnik, and J. N. Tsitsiklis, "Performance of multiclass queueing networks via piecewise linear Lyapunov functions." Preprint, 2000.

[21] J. R. Morrison and P. R. Kummar, "New linear program performance bounds for closed queueing networks," *Discrete Event Dynamic Systems: Theory and Applications*, vol. 11, pp. 291–317, October 2001.

[22] A. Hordijk and A. Shwartz, "Performance bounds for queues via generating functions," *IEEE Transactions on Automatic Control*, vol. 46, no. 1, pp. 137–142, 2001.

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

[24] S. C. H. Lu, D. Ramaswamy, and P. R. Kumar, "Efficient scheduling policies to reduce mean and variance of cycle–time in semiconductor manufacturing plants," *IEEE Transactions on Semiconductor Manufacturing*, vol. 7, pp. 374–385, August 1994.

[25] F. C. Kelly and C. N. Laws, "Dynamic routing in open queueing networks: Brownian models, cut constraints, and resource pooling," *Queueing Systems*, vol. 13, pp. 47–86, 1993.

[26] J. M. Harrison, "The BIGSTEP approach to flow management in stochastic processing networks," in *Stochastic Networks Theory and Applications* (F. P. Kelly, S. Zachary, and I. Ziedins, eds.), pp. 57–89, Oxford, UK: Clarendon Press, 1996.

[27] H. Chen and A. Mandelbaum, "Discrete flow networks: Bottleneck analysis and fluid approximations," *Math. Oper. Res.*, vol. 16, pp. 408–446, 1991.

[28] J. R. Perkins and P. R. Kumar, "Buffer level control of unreliable manufacturing systems," in *Proceedings of the 1993 American Control Conference*, pp. 826–827, June 1992.

[29] J. R. Perkins and P. R. Kumar, "Optimal control of pull manufacturing systems," *IEEE Transactions on Automatic Control*, vol. 40, pp. 2040–2051, December 1995.

[30] S. P. Meyn, "Stability and optimization of multiclass queueing networks and their fluid models," *American Mathematical Society: Lectures in Applied Mathematics*, vol. 33, pp. 175–199, 1997.

[31] S. P. Meyn, "Feedback regulator for sequencing and routing in multiclass queueing networks." To appear in *SIAM Journal of Control and Optimization*, 2000.

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

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

[34] D. Down and S. Meyn, "Piecewise linear test functions for stability and instability of queueing networks," *Queueing Systems*, vol. 27, pp. 205–226, 1997.