A Unified Approach to Linear Programming Bounds for Queueing Networks: Systems with Polyhedral Invariance of Transition Probabilities [1] [2]

James R. Morrison[3] and P. R. Kumar[4]

Abstract

We develop a framework for obtaining linear programming performance bounds in queueing networks. The structure allowing for the development of the bounds requires that the underlying Markov chain model possess translational invariance of its transition probabilities on polyhedra. Such a structure is exhibited by many systems of interest. The bounds are then obtained via a performance–to–performance duality.

Keywords: Wafer fabs, queueing networks, scheduling, performance evaluation, semiconductor manufacturing plants, polyhedral invariance.

1 Introduction

Outside of a limited class of queueing networks, few explicit equilibrium distribution solutions are known. In particular, for re–entrant line models, the baseline model for the consideration of semiconductor manufacturing plants, almost no explicit solutions are known. Hence, analytic performance bound techniques have been developed as an alternative to simulation. Linear programming (LP) performance bound techniques were developed in [1] and [6]. For fixed arrival rates, linear constraints which the model must satisfy were obtained for buffer priority policies and all non–idling polices. As a result, a linear program could be formulated to obtain analytic upper and lower bounds on the mean number of lots (synonymously, parts or customers) in the network. Subsequently, the first general LP stability test was developed in [4]. Again for buffer priority policies and all non–idling policies, [4] developed a linear program which obtained the coefficients of a quadratic Lyapunov function – thereby proving stability, provided the resulting Lyapunov function was copositive on the state space of the model. In [5], the duality of the upper bound performance LP of [6] to the stability LP of [4] was established. Further related work appeared in [3] and [2], where the goal of obtaining functional bounds (a bound parameterized by the system loading) was pursued.

In [7], tighter LP performance bounds were obtained through the average cost inequality by consideration of a piecewise quadratic Lyapunov function (termed the surrogate for the differential cost function). The approach identified regions of the state space in which the average cost inequality took an invariant form. Buffer priority policies and all non–idling policies remained the focus in [7].

Here, we recognize that, for stationary scheduling policies whose transition probabilities are translation invariant within polyhedral regions of the state space, the average cost inequality has an invariant form on polyhedrons (for particular choices of the surrogate of the differential cost function). If there are a finite number of such polyhedral regions, a finite LP may be developed to bound the mean number of lots in the re–entrant line. The key tool in obtaining a linear program is the use of a performance to performance duality (as opposed to the performance to stability duality of [5]). Thus, performance bounds for scheduling policies with translation invariant transition probabilities on polyhedrons may be obtained.

As an example of the applicability of this structure, one can consider fluctuation smoothing policies, which have an affine form of the index for when a buffer has priority, and bounds for the class of scheduling policies termed affine index policies may be developed. For reasons of space, we refer the reader to [8] for more details. These policies subsume buffer priority policies, the fluctuation smoothing for the mean of cycle time (FSMCT) policy, linear switch curve policies, and others. Even those are however still only special cases of what one may apply this approach to. One can consider more general policies, and one can even consider

[1] The research reported here has been supported in part by the National Science Foundation under Grant No. DMI-9743165, the Semiconductor Research Corporation under the Contract No. 97-FJ-489, EPRI and the USARP under subcontract to Cornell University under Contract Nos. 8333-04 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.

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

[3] Email:morrison@decision.csl.uiuc.edu

[4] University of Illinois, Coordinated Science Laboratory, 1308 West Main Street, Urbana, IL 61801. Email: prkumar@decision.csl.uiuc.edu. Tel: (217) 333-7476.

systems with machine failures and models more general than those mentioned here. Thus, this work provides a unified framework for the development of LP performance bound techniques, subsuming many previous developments.

In the next section, we detail the basic open re–entrant line model. Section 3 provides, as a starting point, the average cost inequality for a general Markov chain. The general polyhedral invariance principles and performance bound theorem are developed in Section 4. It is in this section that the performance to performance duality is exploited. With this general result in hand, we illustrate an application of this result for an affine index policy, and show how it induces a polyhedral invariance, thus allowing for an LP performance bound. Finally, in Section 6, we present concluding remarks.

2 System Description

The basic open re–entrant line is a queueing network consisting of S stations, labelled $\sigma_1, \ldots, \sigma_S$, at which lots receive service, and L buffers, labelled $b_1, \ldots, b_L$, at which lots await service. Buffer b_i is located at station $\sigma(i) \in \{\sigma_1, \ldots, \sigma_L\}$ and we use the notation $i \in \sigma$ to indicate that $\sigma(i) = \sigma$.

Lots are released into the network as a Poisson process of rate λ to buffer b_1. The lots traverse the network along a deterministic route, that is, after a lot in buffer b_i receives service from station $\sigma(i)$, the lot moves next to buffer b_{i+1}, unless $i = L$, in which case the lot exits the network. The service time for a lot in buffer b_i is assumed to be exponential of rate μ_i. We further assume that only one lot, from among those present in the buffers at a station, may be in service at any time. All service processes are assumed to be independent of the arrival process and one another, and all stochastic processes are right–continuous with left limits. Figure 1 provides an example of an open re–entrant line.

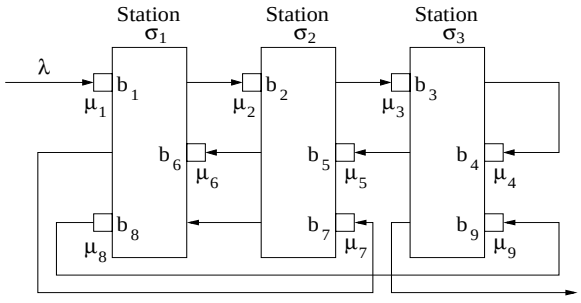

Figure 1: An open re–entrant line.

The state of the network is $x(t) = (x_1(t), \ldots, x_L(t))^T$, where $x_i(t)$ is the number of lots in buffer b_i at time t. As a station can serve at most one lot at a time, a scheduling policy is necessary to direct the stations in making their scheduling decisions. We restrict attention to stationary scheduling policies (that is, policies dependent only upon the system state $x(t)$ at time t). Let $w_i(t) = 1$ if buffer b_i is receiving service from station $\sigma(i)$ at time t and 0 otherwise. The vector $w(t) = (w_1(t), \ldots, w_L(t))^T$ is thus the scheduling decision at time t. As the scheduling policy is assumed stationary, we may instead write $w_i(x(t))$, without loss of generality. Thus, the queueing network model operating under a stationary scheduling policy is a controlled continuous–time discrete–state time–homogeneous Markov chain with state space $S \cup Z_+^L$.

To convert this model to an equivalent discrete–time Markovian model, we resort to uniformization. That is, we rescale time so that $1 = \lambda + \sum_{i=1}^{L} \mu_1$ and suppose that each buffer not receiving service from it's station has a virtual lot in service. We then sample the system state at times corresponding to arrivals and real or virtual service completions. Let $\tau_0 = 0$ and τ_n denote the n^{th} sampling time. Defining $x(n) := x(\tau_n)$ and $w(x(n)) := w(x(\tau_n))$, the sampled system is a controlled *discrete–time* discrete–state time–homogeneous Markov chain with state space $S \cup Z_+^L$.

We use $w(x)$ for the control as the policy is stationary. Let e_i be the unit vector with a 1 in the i^{th} position, and $e_{L+1} = 0$. The controlled transition probabilities are

$$\begin{aligned} p_{x,x+e_1} &= \lambda, \\ p_{x,x-e_i+e_{i+1}} &= \mu_i w_i(x), \forall i \in \{1, \ldots, L\}, \\ p_{x,x} &= 1 - \lambda - \sum_{i=1}^{L} \mu_i (1 - w_i(x)), \end{aligned}$$

where $p_{x,x}$ accounts for virtual service completions.

We note that our approach applies to systems more general than this model; however this is a good class of systems to illustrate the idea.

3 Average Cost Inequality

The average cost inequality is the starting point for our development. The following lemma recalls the average cost inequality.

Lemma 3.1: Average cost inequality performance bounds. *Let $c : S \to R$ be a cost function on the state space S of a discrete–time discrete–state time–homogeneous Markov chain with transition probability matrix $P = [p_{x,y}]$, whose evolution begins with some deterministic initial condition $x(0)$. Suppose $E|c(x(k))| < +\infty$ for every k, where $x(k)$ is the state*

of the Markov chain at time k. Let $W : S \to R$ be a function on S.

(i) *Suppose there exists a $J \in R$ and a function W with $W(x)$ bounded from below, satisfying*

$$J + W(x) \geq c(x) + \sum_{y \in S} p_{x,y} W(y), \tag{1}$$

for all $x \in S$, then

$$J \geq \limsup_{T \to +\infty} \frac{1}{T} \sum_{n=0}^{T-1} E[c(x(n))]. \tag{2}$$

(ii) *Suppose there exists a $J \in R$ and a function W with $W(x)$ bounded from above, satisfying*

$$J + W(x) \leq c(x) + \sum_{y \in S} p_{x,y} W(y), \tag{3}$$

for all $x \in S$, then

$$J \leq \liminf_{T \to +\infty} \frac{1}{T} \sum_{n=0}^{T-1} E[c(x(n))]. \tag{4}$$

4 Polyhedral Invariance and Linear Programming Performance Bounds

Here we suppose that the stationary scheduling policy of interest is translation invariant on polyhedral regions within the state space. That is, for a given policy described by $w(x)$, there exist M polyhedrons partitioning the state space S, denoted $P^1, \ldots, P^M$, on which $w(x)$ is constant. Let polyhedron m be described by a matrix A^m and a vector b^m, such that

$$P^m := \{x \in S : A^m x \geq b^m\}.$$

Thus, we require that there exist M polyhedrons as above with $w(x) = w^m$ for all $x \in P^m$. Figure 2 illustrates the idea.

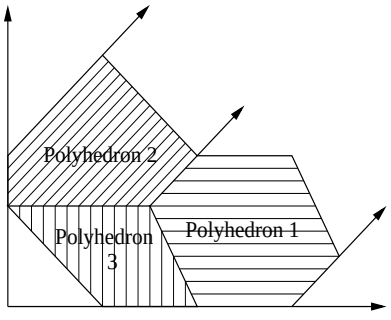

Figure 2: On polyhedron m, the control has constant value $w(x) = w^m$.

Recall that the cost function of interest is $c(x) =\mid x \mid= e^T x$, where e is the L length vector of 1's, so that J from the average cost inequality will be an upper or lower bound on the mean number of lots in the network. Consider the surrogate for the differential cost function $W(x) = p^T + (1/2)x^T Q x$, where p is an unknown vector and Q is an unknown symmetric matrix to be determined.

We next focus attention on the average cost inequality for an upper bound in polyhedron m, plug in the form for $W(x)$, and recall that $1 = \lambda + \sum_{i=1}^{L} \mu_i$. The goal is to find J, p, Q such that

$$\begin{aligned} J \quad \geq \quad & e^T x + \lambda[p^T e_1 + \frac{1}{2} e_1^T Q e_1 + e_1^T Q x] \\ & + \sum_{\{i : w_i^m = 1\}} \mu_i [p^T (e_{i+1} - e_i) \\ & \quad + \frac{1}{2}(e_{i+1} - e_i)^T Q (e_{i+1} - e_i) \\ & \quad + (e_{i+1} - e_i)^T Q x], \end{aligned}$$

for all $x \in P^m$. These inequalities may be written as

$$J \geq d^m + r^{m^T} x,$$

for all $x \in P^m$, where

$$\begin{aligned} d^m \quad := \quad & \lambda[p^T e_1 + \frac{1}{2} e_1^T Q e_1] \\ & + \sum_{\{i : w_i^m = 1\}} \mu_i [p^T (e_{i+1} - e_i) \\ & \quad + \frac{1}{2}(e_{i+1} - e_i)^T Q (e_{i+1} - e_i)], \end{aligned}$$

and

$$r^m := \lambda[e_1^T Q] + \sum_{\{i : w_i^m = 1\}} \mu_i [Q(e_{i+1} - e_i)],$$

and d^m, r^m are linearly dependent upon the unknowns p, Q of the surrogate for the differential cost function. Similarly treating each polyhedron P^m we see that the average cost inequality has a fixed form on polyhedrons.

Our ultimate goal is to obtain a p and Q so that J is as small as possible (the best upper bound). Thus we would like to formulate a program minimizing J subject to the average cost inequality constraints within each polyhedron (recall the polyhedrons are assumed to constitute a partition of the state space S). Unfortunately, r and x are both variable and the resulting program would be bi–linear. This difficulty is overcome by identification of a performance to performance duality and addressed in the following theorem.

Theorem 4.1: Linear programming performance bounds for policies which are translation invariant on polyhedrons. *Suppose that an open re–entrant line is operating under a stationary scheduling*

policy $w(x)$ for which there exist M polyhedrons, denoted $P^1, \ldots, P^M$, with $P^m := \{x \in S : A^m x \geq b^m\}$ for fixed given A^m, b^m, such that $w(x) = w^m$ for all $x \in P^m$. Further suppose that no P^m is empty, and that the P^m form a partition of the state space.

(i) *Let J, p, Q (symmetric), and y^m (for all $m = 1, \ldots, M$) be the decision variables in the linear program*

$$\text{Min } J$$

subject to

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

for all $m = 1, \ldots, M$, with value $\overline{V}$, where d^m and r^m are as defined previously (and are linear in the decision variables p and Q). If the function $W(x) = p^T x + (1/2)x^T Q x$ is bounded from below, then $\overline{V}$ is an upper bound on the long term average number of lots in the re–entrant line as in (2).

(ii) *Let J, p, Q (symmetric), and y^m (for all $m = 1, \ldots, M$) be the decision variables in the linear program*

$$\text{Max } J$$

subject to

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

for all $m = 1, \ldots, M$, with value $\underline{V}$, where d^m and r^m are as defined previously (and are linear in the decision variables p and Q). If the function $W(x) = p^T x + (1/2)x^T Q x$ is bounded from above, then $\underline{V}$ is a lower bound on the long term average number of lots in the re–entrant line as in (4).

Proof. (i) Consider the average cost inequality for an upper bound on the polyhedron P^m. The average cost inequality will hold if and only if J, d^m, and r^m are such that the linear program $\min_x -{r^m}^T x$ subject to $A^m x \geq b^m$ and $x \geq 0$ (componentwise) has value greater than or equal to $d^m - J$ (where we subsume the problem on the integer state space Z_+^L by the same problem on R_+^L). Denote this linear program as P with value VP. Associating dual variables y^m to the constraints of this linear program, it's dual, denoted as D, is

$$\text{Max } {b^m}^T y^m$$

subject to

$$\begin{aligned} {A^m}^T y^m &\leq -r^m \\ y^m &\geq 0 \text{ (componentwise).} \end{aligned}$$

Let VD denote the value of the dual linear program, if it exists.

If $VP \geq d^m - J$, so that P is bounded (and feasible by the assumption that P^m is not empty), then the dual D is bounded and feasible with value $VD \geq d^m - J$. Thus, there exists a vector y^m such that $y^m \geq 0$ componentwise, ${A^m}^T y^m \leq -r^m$, and ${b^m}^T y^m \geq d^m - J$.

From the other direction, if there exists a vector y^m such that $y^m \geq 0$ componentwise, ${A^m}^T y^m \leq -r^m$, and ${b^m}^T y^m \geq d^m - J$, then D is feasible and $VD \geq d^m - J$. Also, the value of D is finite, otherwise the primal P would be infeasible (but this is not the case by the assumption that P^m is non–empty). Hence, the linear program P is feasible with value $VP \geq d^m - J$.

This proves that the average cost inequality on the polyhedron P^m holds if and only if there exists a vector y^m such that $y^m \geq 0$ componentwise, ${A^m}^T y^m \leq -r^m$, and ${b^m}^T y^m \geq d^m - J$. Repeating the argument for each polyhedron P^m, for all $m = 1, \ldots, L$, ensures that the average cost inequality holds on the entire state space (recall that P^m form a partition of S) if the constraints of the linear program of Theorem 4.1 hold. The Max determines the best upper bound, if the resulting variables p and Q yield a function $W(x) = p^T x + (1/2)x^T Q X$ which is bounded from below. Notice that the resulting program is indeed a linear program – appealing to duality eliminates the bi–linearity.

(ii) The lower bound is proved in a similar fashion. ∎

We have demonstrated how to obtain a linear program to bound the network performance for policies which take a constant form on polyhedrons. The approach utilized a performance to performance duality to remove the bi–linear nature from the problem.

5 Linear Programming Performance Bounds for Affine Index Policies

One of our motivations has been the class of fluctuation smoothing policies. These policies have the following affine structure with some particular values of the coefficients which are not important for the applicability of this approach. Assign to buffer b_i an affine index

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

A station σ serves the non–empty buffer b_i, with $i \in \sigma$, whose policy index is greatest. In the event of a tie, the policy may assign priority to which ever buffer it chooses – we will obtain performance bounds independent of the choice made in the event of a tie. It will be

shown later that affine index policies have translation invariant transition probabilities in polyhedral regions of the state space. By choosing the coefficients appropriately, one can model the following policies:

Example 1: Buffer priority policies. A buffer priority policy is a stationary scheduling policy described by a vector $\theta := (\theta(1), \ldots, \theta(L))$ which is a permutation of the buffer labels (i.e., the buffer label of buffer b_i is i). A station σ works on the non–empty buffer b_i, $i \in \sigma$, if $\theta(i) < \theta(j)$ for all other non–empty buffers b_j where $j \in \sigma(i) = \sigma$. The affine index policy corresponding to a buffer priority policy assigns the policy index $\eta_i(x) = -\theta(i)$ to buffer b_i.

Example 3: Linear switch curve policy. The policy with $\eta_i(x) = m_i x_i$ is termed a linear switch curve policy.

One may note that in the fluctuation smoothing policy for the mean cycle time, the particular choice of the coefficients is as follows:

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

where $\overline{x}_j$ is an estimate of the mean number of lots in buffer b_j.

Now to apply Theorem 4.1, we must identify polyhedral regions (forming a partition) of the state space for which the transition probabilities under affine index policies are translation invariant. As mentioned briefly in Section 2, we will develop the bounds independent of how one chooses to proceed when two or more non–empty buffers at a station have the same policy index for a given state $x \in S$. This will result in the identification of polyhedral regions whose union is indeed S, but which may overlap.

As the policy may behave in a discontinuous manner when buffer b_i is empty or has lots, we first introduce a partition of S indexed by the vectors $\phi = (\phi_1, \ldots, \phi_L)^T$, where $\phi_i \in \{0, 1\}$. We use Φ to denote the set of all such vectors. Let $S^\phi =: \{x \in S : x_i = 0 \text{ if } \phi_i = 0 \text{ and } x_i = 1 \text{ if } \phi_i = 1\}$. Clearly, $\cup_{\{\phi \in \Phi\}} S^\phi = S$.

Let $C(\sigma) := \{i : i \in \sigma\}$ denote the constituency of station σ. Now, let $\omega = (\omega_1, \ldots, \omega_S)^T$ be a vector with each $\omega_s \in C(\sigma_s) \cup \{0\}$. Use Ω as the set of all such ω. Each element of ω will indicate that a particular buffer is receiving service (or, if $\omega_s = 0$, that no buffer is in service at σ_s so that $\sum_{\{i \in \sigma_s\}} x_i = 0$). Thus, we let

$$\begin{aligned} S^\phi_\omega := \{x \in S^\phi \quad : \quad & \eta_{\omega_s}(x) \geq \eta_i, \\ & \forall i \in \sigma_s, \text{ with } \phi_i = 1, \\ & \forall s \text{ with } \omega_s \neq 0\}. \end{aligned}$$

Let $\psi = (\phi, \omega)$ and $S^\psi = S^\phi_\omega$, and use Ψ to denote the set of all ψ for which S^ψ is *non–empty* (recall from the proof of Theorem 4.1 that it is essential that the polyhedrons are non–empty). Merely substituting the form for $\eta_i(x)$ demonstrates that these regions are polyhedrons as they may be described as

$$\begin{aligned} S^\phi_\omega := \{x \in S \quad : \quad & \sum_{j=1}^{L} m_j^{\omega_s} x_j - \sum_{j=1}^{L} m_j^i \geq k_i - k_{\omega_s}, \\ & \forall i \in \sigma_s, \text{ with } \phi_i = 1, \\ & \forall s \text{ with } \omega_s \neq 0, \\ & x_i = 0 \text{ if } \phi_i = 0, \\ & \text{ and } x_i \geq 1 \text{ if } \phi_i = 1\}. \end{aligned}$$

Though Theorem 4.1 assumes $A^\psi x \geq b^\psi$, the equality constraints above may be readily handled by simply allowing the corresponding dual variable (an element of the dual vector y^ψ) to be a free variable, that is not constrained to be positive.

It is not difficult to see that $\cup_{\{\psi \in \Psi\}} S^\psi = S$, and that the S^ψ may not be disjoint. Within each region S^ψ we seek to ensure the inequality for an upper bound (similarly for a lower bound)

$$\begin{aligned} J \quad \geq \quad & e^T x + \lambda [p^T e_1 + \frac{1}{2} e_1^T Q e_1 + e_1^T Q x] \\ & + \sum_{\{i \in \omega, i \neq 0\}} \mu_i [p^T (e_{i+1} - e_i) \\ & \quad + \frac{1}{2} (e_{i+1} - e_i)^T Q (e_{i+1} - e_i) \\ & \quad + (e_{i+1} - e_i)^T Q x], \end{aligned}$$

holds for all $x \in S^\psi$. If we can ensure this inequality for each $\psi \in \Psi$, the average cost inequality will hold for all $x \in S$. The issue of the overlap of the S^ψ accounts for the arbitrary tie breaking assumption in the event of two or more non–empty buffers achieving the maximum policy index at a station for a given state x. This is because we are ensuring that for each region of the state space where $\eta_i(x) = \eta_j(x)$ all possible choices of which buffer to serve are accounted for.

The inequalities above clearly have the form specified in Theorem 4.1, and requiring them to hold on each S^ψ (a polyhedron) we can apply Theorem 4.1 to obtain linear programming performance bounds.

Example 4: LP upper bound for a linear switch curve. Consider the network of Figure 3 operating under the linear switch curve policy given by the policy indices $\eta_i(x) = m_i x_i$. Suppose that $\lambda = 0.9$, $\mu_1 = \mu_4 = \mu_5 = \mu_6 = 4$, $\mu_2 = \mu_3 = 2$, $m_1 = m_3 = m_6 = 10$, and $m_2 = m_4 = m_5 = 1$. The maximum sustainable throughput rate for the network is $\lambda = 1$. Figure 4 gives the upper bound obtained via the approach of this paper and compares the value to the value one would obtain by extending the work of [6] to this scheduling policy.

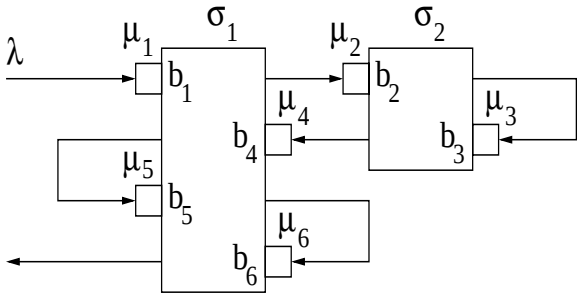

Figure 3: The open re–entrant line of Example 4.

Mean WIP	Upper Bound
Previous Approach	93.03
New Approach	41.74

Figure 4: Upper bounds for the network of Example 4.

6 Concluding Remarks

We have demonstrated how to obtain linear programming performance bounds for the class of networks exhibiting shift invariance of the transition probabilities on a finite set of polyhedra whose union is the state space. The results are more generally applicable than to the class of re-entrant line models considered here. Even for re-entrant lines, on can consider more general policies than have been illustrated here.

References

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

[2] 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*, 22(4):921–954, November 1997.

[3] 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*, 22(4):886–920, 1997.

[4] P. R. Kumar and S. P. Meyn. Stability of queueing networks and scheduling policies. *IEEE Transactions on Automatic Control*, 40(2):251–260, February 1995.

[5] 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*, 41(1):4–17, January 1996.

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

[7] James R. Morrison and P. R. Kumar. New linear program performance bounds for queueing networks. To appear in *Journal of Optimization Theory and Applications*, vol. 100, no. 3, pp. 575-597. Plenum Publishers, USA, March 1999.

[8] James R. Morrison, Queueing Network Analysis of Semiconductor Manufacturing Plants. Ph. D. Thesis, Department of Electrical and Computer Engineering, University of Illinois, September 2000.