Performance Bounds for Queueing Networks and Scheduling Policies*†

Sunil Kumar and P. R. Kumar‡

Keywords: Control of Manufacturing Systems, Performance of Manufacturing Systems, Scheduling, Queueing Networks, Discrete Event Systems

Extended Abstract

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 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 little is known concerning performance, see [2].

In this paper we propose a technique for obtaining upper and lower bounds on performance. It is applicable to a broad, even non–classical, class of 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, one can obtain a supplementing set of linear 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 [3] and Meyn, see [2], where the square of the workload is studied.

The following Theorem illustrates the type of results that can be obtained. Consider an open re-entrant line where 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. 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 interarrival times are independent. Let $I(i) := \{j : b_i \text{ and } b_j \text{ share the same machine}\}$.

---

*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, USA. Phone number: (217)-333-7476. Fax: (217)-244-1653. Email: prkumar@decision.csl.uiuc.edu.

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

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

subject to the constraints,

$$\begin{aligned}
2\lambda \sum_{i \in I(1)} z_{i1} + 2\lambda - 2\mu_1 z_{11} &= 0 \\
2\mu_{j-1} z_{j-1,j} + 2\lambda - 2\mu_j z_{j,j} &= 0 \quad \text{for } j = 2, \ldots, L \\
\lambda \left(\sum_{j \in I(2)} z_{j2} - 1 \right) - \mu_1 (z_{12} - z_{11}) - \mu_2 z_{21} &= 0 \\
\lambda \sum_{i \in I(j)} z_{ij} - \mu_1 z_{1j} - \mu_j z j 1 + \mu_{j-1} z_{j-1,1} &= 0 \quad \text{for } j = 3, \ldots, L \\
\mu_{i-1} z_{i-1,i+1} - \mu_i z_{i,i+1} - \lambda + \mu_i z_{ii} - \mu_{i+1} z_{i+1,i} &= 0 \quad \text{for } i = 2, \ldots, L-1 \\
\mu_{i-1} z_{i-1,j} - \mu_i z_{i,j} + \mu_{j-1} z_{j-1,i} - \mu_j z_{ji} &= 0 \quad \text{for } i = 1, \ldots, L-2 \\
& \qquad \text{and } j \geq i+2,
\end{aligned}$$

$$\sum_{\{j | \sigma(j) = \sigma\}} z_{ji} \leq \sum_{j \in I(i)} z_{ji} \quad \text{for } i = 1, \ldots, L \text{ and } \sigma = 1, \ldots, S \text{ with } \sigma \neq \sigma(i),$$

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

2. Under the same conditions, an upper bound is obtained by replacing the "min" above by a "max."

3. Consider a buffer priority policy which provides pre-emptive resume priority according to a rank ordering of the buffers at each machine. Then one obtains upper and lower bounds, under the same conditions as in (i) and (ii) above, by appending the equality constraints:

$$z_{ij} = 0 \text{ if buffers } b_i \text{ and } b_j \text{ share the same machine, and } b_j \text{ receives priority over } b_i \,. \quad (1)$$

The method leading to the results above can be generalized considerably. We will illustrate the application of this method on several typical situations of interest in manufacturing systems. For an open re-entrant line modeling a semiconductor manufacturing plant, we plot a lower bound on the so called "actual-to-theoretical" ratio of delay to mean total processing time. We show that the Last Buffer First Serve (LBFS) policy of [4] is almost optimal in light traffic. In another example, 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, thus showing that LBFS dominates FBFS. For a closed re-entrant line modeling the so called "closed loop" release policy of Glassey and Resende [5], we bracket the performance of all the buffer priority scheduling policies [4], as well as the one

conjectured to be optimal in heavy traffic. For another closed queueing network, we show that the workload balancing policy suggested by the Brownian network analysis of Harrison and Wein [6] is almost optimal. Our bounds on throughput compare favorably with the simulation results reported there. For a manufacturing system with machine failures, we show how the performance changes with the mean time between failures. For a finite buffer system, we outline how one may bound the throughput. Finally for a GI/GI/1 queue, we obtain a better bound than Kingman's for a large range of utilization factors.

The full details may be found in [7].

Note: The method of Section 2.1 in [7] is the same as the "non–parametric method" of Section 4.2 of Bertsimas, Paschalidis and Tstisiklis [8]. Both were obtained simultaneously and independently. The idea of using a general "potential" function, i.e., several Lyapunov functions, see [9], was recognized by them in February 1992. We urge readers of our work to also read theirs, and future authors citing our work to also cite theirs.

References

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

[2] P. R. Kumar. Re-entrant lines. To appear in *Queueing Systems: Theory and Applications*, 1993.

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

[4] S. H. Lu and P. R. Kumar. Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control*, 36(12):1406–1416, December 1991.

[5] C. R. Glassey and M. G. C. Resende. Closed-loop job release control for VLSI circuit manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 1(1):36–46, February 1988.

[6] J. M. Harrison and L. M. Wein. Scheduling networks of queues: Heavy traffic analysis of a two-station closed network. *Operations Research*, 38(6):1052–1064, 1990.

[7] Sunil Kumar and P. R. Kumar. Performance bounds for queueing networks and scheduling policies. Technical report, Coordinated Science Laboratory, University of Illinois, Urbana, IL, 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] 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.