Design and Operation of Manufacturing Systems — Control- and System-Theoretical Models and Issues [1]

Stanley B. Gershwin
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307
gershwin@mit.edu
http://web.mit.edu/manuf-sys/www

1. Introduction

The design and operation of manufacturing systems are of great economic importance. In spite of the considerable literature and long history of manufacturing, much research remains to be done. This paper combines earlier work on the hedging point method for scheduling and the decomposition method for performance prediction and design to propose a unified approach for design and operation. It includes algorithms for selecting buffer sizes and policy parameters.

2. Design and Operation of Manufacturing Systems

2.1. Issues and Limitations

We focus on *quantity*-related issues: how much is made and how long it takes to make it; disruptions due to failures and similar events; and inventory. We do not consider quality, product design, and the aspects of process design that require an understanding of the physics of processes. We simplify the problem by not treating setups or batches.

Measures of factory performance include *production rate* (the rate at which material is produced), *inventory levels* (the amount of material in the system), *lead time* (the time between the arrival of an order and the delivery of a product), and *delivery reliability* (the fraction of time that deliveries occur when promised). The means and variances of these measures are of interest.

2.2. Dynamic Programming

2.2.1 Formulation Many authors formulate the problem of real-time scheduling of manufacturing systems as a dynamic programming problem [1], [2], [15], [17], [20], [22]. However, the curse of dimensionality drastically limits the size of problems that can be solved exactly; and there is no guarantee that an optimal solution is easy to implement. Finally, well-designed suboptimal solutions may be very nearly as effective as optimal solutions.

Figure 1 represents a tandem production system. Type q ($q = 1, ..., n$) material travels from an upstream inventory to Machine M_1 to Buffer B_{1q}, and so forth until Machine M_k. The machines are flexible. No cost is incurred or time spent in changing a machine from operating on one part to another.

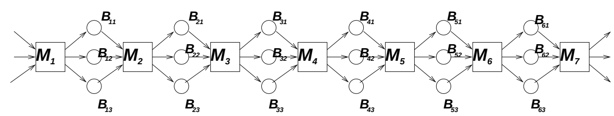

Figure 1: Production line with $k = 7$ stages and $n = 3$ part types

The machines are unreliable. The fraction of time that M_i is operational is e_i and μ_{iq} is the maximum rate at which Machine M_i can do operations on type q parts. Assume that the demand for Type q parts is constant at rate $d_q > 0$, and that $\sum_{q=1}^{n} d_q/\mu_{iq} < e_i$ for all i. This is a statement of *feasibility*.

Machine M_i's repair state is α_i. When M_i is operational, $\alpha_i = 1$. Otherwise, $\alpha_i = 0$. The production rates $u_i(t) = (u_{i1}(t), ...u_{in}(t))$ of Machine M_i at time t are control variables. They satisfy

$$\begin{aligned} &\text{if } \alpha_i = 0, \quad u_{iq} = 0 \text{ for } q = 1, ..., n; \\ &\text{if } \alpha_i = 1, \quad \textstyle\sum_{q=1}^{n} \frac{u_{iq}}{\mu_{iq}} \leq 1; u_{iq} \geq 0. \end{aligned} \tag{1}$$

Let $b_{iq}(t)$ be the level of Buffer B_{iq} at time t. Then $b_i = (b_{i1}, ...b_{in})$ satisfies $db_i/dt = u_i - u_{i+1}$. Let N_{iq} be the *size* of Buffer B_{iq}. Then

$$0 \leq b_{iq} \leq N_{iq} \tag{2}$$

Let x_k, an n-vector, satisfy

$$\frac{dx_k}{dt} = u_k - d. \tag{3}$$

This is the *system surplus*. It is the vector of cumulative differences between production and demand. Each component can be positive or negative. The *local surplus* n-vector $x_i = \sum_{j=i}^{k-1} b_j + x_k$ satisfies

$$x_i - x_{i+1} = b_i, \tag{4}$$

[1] Presented at the *Automatic Controls Conference, Albuquerque, New Mexico, June, 1997*

$$\frac{dx_i}{dt} = u_i - d. \tag{5}$$

The system states are $(\alpha_1, ..., \alpha_k)$ and $(x_1, ..., x_{k-1}, x_k)$. A natural problem formulation is

$$\min E \int_0^T g(x_1(s), ..., x_k(s))ds \tag{6}$$

subject to (1), (2), (3), (4), (5), and the dynamics of α

in which g is a function that penalizes surplus and backlog, and T is a long time.

2.2.2 $k = 1, n = 1$ For a single-stage, single-part-type system in which g was a generalized absolute value (8), Akella and Kumar [1] (for a discounted version of this problem) and Bielecki and Kumar [2] showed that the solution of this problem was in the form of a *hedging point policy*. There is some number Z such that the optimal control law is (when $\alpha = 1$)

$$\begin{array}{ll} \text{if } x > Z & u = 0 \\ \text{if } x = Z & u = d \\ \text{if } x < Z & u = \mu \end{array} \tag{7}$$

2.2.3 $k = 1, n \geq 2$ Srivatsan [22] studied a one-machine, two-part generalization of the Bielecki-Kumar [2] problem in which the generalized absolute value function cost is $g = g_1 + g_2$, and

$$g_q = \begin{cases} -g_q^- x_q & \text{if } x_q < 0 \\ g_q^+ x_q & \text{if } x_q \geq 0 \end{cases} \tag{8}$$

where g_q^- and g_q^+ are non-negative numbers. He was only able to obtain a complete solution when the hedging point (Z_1, Z_2) was 0. The policy is generalized in [17].

2.2.4 $k \geq 2, n = 1$ Van Ryzin [24], and colleagues [25] studied two-stage, single-part systems with discounted cost numerically. In [18] the optimal production control policy is approximated by a policy of the form

$$\begin{array}{l} \text{For } i = 1, 2, \text{ if } \alpha_i = 1, \\ \quad \begin{array}{ll} \text{if } x_i > Z_i & u_i = 0; \\ \text{if } x_i = Z_i & u_i = d; \\ \text{if } x_i < Z_i & \text{choose } u_i \text{ as large as possible.} \end{array} \end{array} \tag{9}$$

"As large as possible" means that if $x_1 - x_2 = N$, then $u_1 = \min(\mu_1, u_2)$, and that if $x_1 - x_2 = 0$, then $u_2 = \min(u_1, \mu_2)$.

For longer lines, the buffer level is $b_i = x_i - x_{i+1}$, so $0 \leq Z_i - Z_{i+1} \leq N_i$. We propose a policy which is the same as (9), but "as large as possible" means that if $x_i - x_{i+1} = N_i$, then $u_i = \min(\mu_i, u_{i+1})$, and that if $x_i - x_{i+1} = 0$, then $u_{i+1} = \min(u_i, \mu_{i+1})$.

2.2.5 $k \geq 2, n \geq 2$ It is conjectured in [18] that a good policy for lines with $k \geq 2$ and $n \geq 2$ would be to combine (9) and the policy of Section 2.2.3.

2.3. Other Configurations, Other Policies

2.3.1 General Network Configuration Consider a generalization of the production line in which Part Type q follows a sequence of $K(q)$ steps. It visits Machines $M_{S(1,q)}, M_{S(2,q)}, ..., M_{S(K(q),q)}$ in sequence where $S(s, q) \in \{1, ..., k\}$. After receiving its sth operation at Machine $M_{S(s,q)}$, the part goes to Buffer B_{sq}, and then to Machine $M_{S(s+1,q)}$ for its $s+1$st operation. An important example is semiconductor manufacturing.

It is important to distinguish the same part type at different stages. The buffers are *homogeneous* — each buffer holds parts that are indistinguishable, except possibly according to arrival time, due date, etc.

The feasibility requirement becomes

$$\sum_{\{s,q|S(s,q)=i\}} \frac{d_q}{\mu_{sq}} < e_i \text{ for all } i \tag{10}$$

in which the sum is taken over all parts at all production stages that visit Machine M_i.

2.3.2 Generalized Hedging Point Policy We conjecture that the following policy works well:

0. Before executing the policy, rank order all the products at each machine. One possible ranking is described below.

For each machine $M_i, i = 1, ..., k$, if $\alpha_i = 0$, then $u_{sq} = 0$ for all sq such that $S(s, q) = i$. If $\alpha_i = 1$, divide the parts into four sets. Initially, they are:

- $\mathcal{S}_1$: the set of all parts for which $x_{sq} > Z_{sq}$;
- $\mathcal{S}_2$: the set of parts for which $x_{sq} = Z_{sq}$ and for which there are no higher ranking parts such that $x_{sq} < Z_{sq}$;
- $\mathcal{S}_3$: the highest ranking part for which $x_{sq} < Z_{sq}$;
- $\mathcal{S}_4$: all other parts.

Then

1. For each part sq in $\mathcal{S}_1$, u_{sq}=0.
2. For each part sq in $\mathcal{S}_2$, $u_{sq} = d_{sq}$
3. If $\mathcal{S}_3$ and $\mathcal{S}_4$ are not empty, allocate as much of the remaining capacity as possible to the part type in $\mathcal{S}_3$, subject to the limitations discussed just below (9). If the part type in $\mathcal{S}_3$ is called s^*q^*, define

$$\mathcal{U}_{s^*q^*} = \mu_{s^*q^*} \left(1 - \sum \frac{d_q}{\mu_{sq}}\right) \tag{11}$$

in which the sum is over all part types in $\mathcal{S}_2$. Then

$$u_{s^*q^*} = \begin{cases} \min(\mathcal{U}_{s^*q^*}, u_{s^*-1,q^*}) & \text{if } b_{s^*-1,q^*} = 0; \\ \min(\mathcal{U}_{s^*q^*}, u_{s^*+1,q^*}) & \text{if } b_{s^*q^*} = N_{s^*q^*}; \\ \mathcal{U}_{s^*q^*} & \text{otherwise.} \end{cases} \tag{12}$$

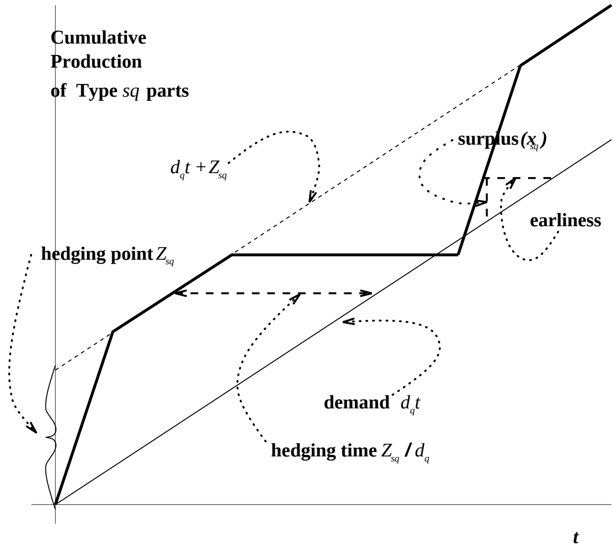

Figure 2: Earliness and lateness

4. • If $\sum_{sq \in \mathcal{S}_2 \cup \mathcal{S}_3} \frac{u_{sq}}{\mu_{sq}} = 1$, then $u_{sq} = 0 \,\forall\, sq \in \mathcal{S}_4$, and *stop.*

- If $\mathcal{S}_4$ is empty, *stop.*
- Otherwise,
  (a) Move the part from $\mathcal{S}_3$ into $\mathcal{S}_2$. Do not change its production rate u_{sq}.
  (b) Move the highest ranking part from $\mathcal{S}_4$ into $\mathcal{S}_3$. If there are more than one part with the same, highest, ranking, select one whose surplus is furthest behind its hedging point.
  (c) Go to Step 3.

Stop means that all the capacity of M_i has been allocated.

The ranking that follows from (8) is according to decreasing values of $g_{sq}^{-}\mu_{sq}$, the "$c\mu$ rule."

In Figure 2, the vertical axis represents cumulative demand for type q ($d_q t$) and the cumulative production of Type q parts at stage s (the heavy broken line).

2.3.3 Alternate Statement Let $P_{sq}(t) =$ the total cumulative production of Type q parts at stage s of its process at time t. Then

$$P_{sq}(t) = \int_0^t u_{sq}(r)dr = d_q t + x_{sq}(t) \leq d_q t + Z_{sq}. \quad (13)$$

Define

$$\mathcal{E}_{sq}(t) = \frac{P_{sq}(t)}{d_q} - t \leq \frac{Z_{sq}}{d_q}. \quad (14)$$

We can interpret $P_{sq}(t)/d_q$ as the time at which we *should* have produced a total of P_{sq} units. Therefore $\mathcal{E}_{sq}(t)$ represents the *earliness* of Part q at stage s at time t. It should be less than or equal to the *hedging time*, Z_{sq}/d_q.

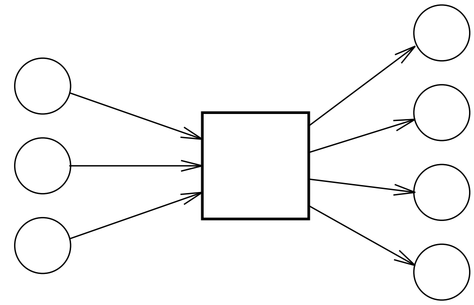

Figure 3: Assembly/Disassembly Machine

The new statement is the same that of Section 2.3.2 except that we replace $x_{sq} > Z_{sq}$ with $\mathcal{E}_{sq}(t) > Z_{sq}/d_q$; $x_{sq} = Z_{sq}$ with $\mathcal{E}_{sq}(t) = Z_{sq}/d_q$; and $x_{sq} < Z_{sq}$ with $\mathcal{E}_{sq}(t) < Z_{sq}/d_q$. This will keep $\mathcal{E}_{sq}(t)$ near Z_{sq}/d_q and $(P_{sq} - Z_{sq})/d_q$ near t.

2.3.4 Least Slack and Hedging Point At Machine M_i, ζ_{sq} is the expected amount of work (in time units) that remains to be done on Part Type sq (where $S(s,q) = i$) before it leaves the system. In the *least slack* policy [21], each individual part is assumed to have a due date. The slack for a Type q part at Machine i is the difference between its due date and ζ_{sq}. At each stage, the part with the least slack is selected next. Lu and Kumar show that this policy works very well.

Suppose the type sq parts have serial numbers starting at 0. The due date of the part with serial number P_{sq} is P_{sq}/d_q. According to the policy of Section 2.3.3, we want that part to obtain its sth operation (at Machine M_i, where $i = S(s,q)$) at time $(P_{sq} - Z_{sq})/d_q$ which is Z_{sq}/d_q time units earlier than its due date. Consequently, the hedging point and least slack policies are similar if $Z_{sq}/d_q = \zeta_{sq}$. and we can interpret $(P_{sq} - Z_{sq})/d_q$ as the slack time of Part Type q at stage s [3].

However, least slack does not distinguish between parts that are earlier than the slack time and those that are not. It considers only slack time in ranking the parts, and no other criterion.

3. Token-Based Control Policies

3.1. Definitions

In some scheduling systems, operations are performed only when both parts and *tokens* are present. Kanban, [4], [5], [11], [13] and Buzacott and Shanthikumar's PAC system [7], [8] are examples.

Figure 3 shows an *assembly/disassembly* machine. When it does an operation, it takes one part from each of the upstream buffers and places one part into each of the downstream buffers, which are homogeneous. It can only do an operation if there is at least one part in each upstream buffer and one space in each downstream buffer. A/D systems have been studied in [12], [16], [17].

Let some of the buffers in Figure 3 hold tokens rather

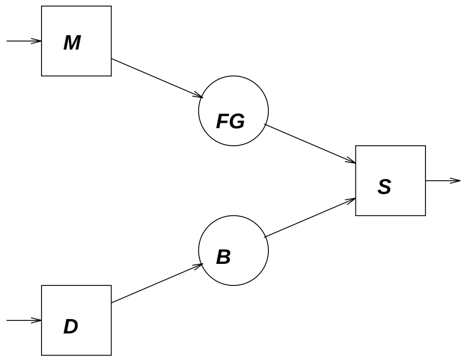

Figure 4: Three-Machine Assembly System

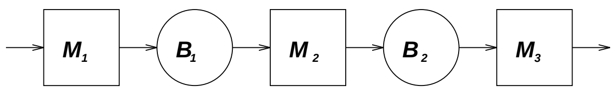

Figure 5: Three-Machine Production Line

than parts. The operation is only *authorized* if there is a token in each upstream token buffer and a space in each finite downstream token buffer. Therefore, a real-time scheduling policy is defined by the **topology** of the token network; the **sizes** of the token buffers; and the **sizes** of the part buffers.

3.2. Equivalence

Figure 4 shows a three-machine assembly system in which Machine S takes parts from Buffers FG and B, assembles them, and sends them on to customers. Buffers FG and B get their parts from Machines M and D. Assume that Machine S is infinitely fast and perfectly reliable. Material cannot be in both buffers simultaneously. Assume that Buffer B is infinite; the size of Buffer FG is Z, which is finite; and Machine D is perfectly reliable and operates at speed $\mu_D = d$.

The behavior of this system is shown in [4] to be *identical* to the system of Section 2.2.2. Machine D represents the *demand* process. When the sales department confirms an order, a token is generated by D and is put into B. If FG (the *finished goods* buffer) is not empty, Machine S (*synchronization* of demand and production) instantaneously assembles the order and a part and sends it out of the production area to the customer, to the shipping area, or elsewhere.

3.3. Larger Systems: $k \geq 2, n = 1$

Suppose the three-machine production line in Figure 5 is controlled using the methods of Section 2.2.4. Then its behavior is exactly the same as the assembly/disassembly system of Figure 6. The control structure of Figure 4 is replicated for each machine M_i. When M_i performs an operation, it removes a part from Buffer B_{i-1}, it adds a part to Buffer B_i, and it creates a token and adds it to

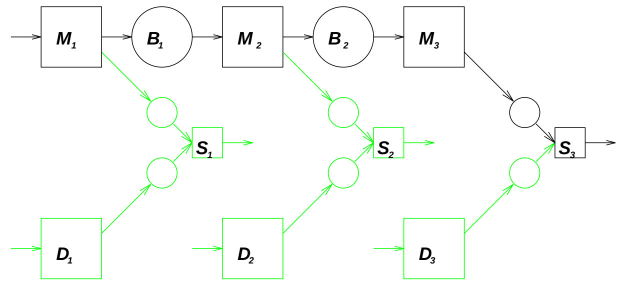

Figure 6: Controlled Three-Machine Production Line as a Tree-Structured Assembly System

the upper buffer of the new structure. If Buffer B_{i-1} is empty, or if Buffer B_i or the new buffer is full, an operation cannot take place. Synchronization Machines S_1 and S_2 require demands in order to prevent the new upper buffers from filling up. Synchronization Machine S_3 regulates the delivery of completed items to customers.

The upper buffer in each control structure, which has size Z_i, performs the function of the hedging point; the lower buffer represents a local backlog.

4. Decomposition Methods

Decomposition is an approximation technique for evaluating the performance of systems like that of Figure 6. It was developed for kanban-like systems in [14] and improved in [9]. It was extended to assembly/disassembly systems in [12] and [16]. It is reviewed in [10] and discussed in detail in [17] and [6]. Recent work [19] has extended it to systems with yield losses and rework. Experience shows it to be rapid and accurate. Systems like the one in Figure 6 can be analyzed by decomposition. This is a conceptual extension of earlier work because none of the earlier authors contemplated tokens in the network.

5. Optimization

Optimization methods have been built on decomposition and they are also very fast [23]. Therefore we can efficiently design real-time scheduling policies and manufacturing systems simultaneously. This is important because it leads to faster designs with better performance.

6. Conclusion

6.1. Summary

This paper describes a connection between dynamic programming-based (hedging point) and other scheduling policies for manufacturing systems. It shows that single-part type systems operated according to hedging point can be treated as token-based systems. Thus, these systems can be analyzed and optimized by means of decomposition methods. In particular, the paper suggests a method for choosing hedging points for multiple-stage systems.

6.2. Future Research

Decomposition methods for token-based control schemes are needed. Extensions of existing methods include non-acyclic networks with multiple loops, multiple-part-type systems, and reentrant systems. Optimization methods for these kinds of systems are also needed.

References

[1] 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, No. 2, pp. 116-126, February, 1986.

[2] T. Bielecki and P. R. Kumar, "Optimality of Zero-Inventory Policies for Unreliable Manufacturing Systems," *Operations Research,* Volume 36, No. 4, July-August, 1988, pp. 532-541.

[3] A. M. Bonvik, private communication.

[4] A. M. Bonvik, Performance Analysis of Manufacturing Systems Under Hybrid Control Policies, Ph.D. Thesis, Massachusetts Institute of Technology, Operations Research Center, June, 1996.

[5] A. M. Bonvik, C. E. Couch, and S. B. Gershwin, "A Comparison of Production-Line Control Mechanisms", *International Journal of Production Research*, to appear.

[6] M. H. Burman, "New Results in Flow Line Analysis", Ph.D. Thesis, Massachusetts Institute of Technology, Operations Research Center, June, 1995.

[7] J. A. Buzacott and J. G. Shanthikumar, "Design of Manufacturing Systems Using Queuing Models," *Queueing Systems Theory and Applications*, Volume 12, No. 1-2, December, 1992, pp. 135-213.

[8] J. A. Buzacott and J. G. Shanthikumar, *Stochastic Models of Manufacturing Systems*, Prentice Hall, 1993.

[9] Y. Dallery, R. David, and X.-L. Xie, "An Efficient Algorithm for Analysis of Transfer Lines with Unreliable Machines and Finite Buffers," *IIE Transactions,* Vol. 20, No. 3, pp. 280-283, September, 1988.

[10] Y. Dallery and S. B. Gershwin, "Manufacturing Flow Line Systems: A Review of Models and Analytical Results," *Queueing Systems Theory and Applications*, Volume 12, December, 1992, pp. 3-94.

[11] Y. Dallery and G. Liberopoulos, "Extended Kanban Control system: A New Kanban-Type Pull Control Mechanism for Multi-Stage Manufacturing Systems," Lab. MASI, Univ. P. et M. Curie, Paris, Dec., 1995.

[12] M. Di Mascolo, R. David, and Y. Dallery, "Modeling and Analysis of Assembly Systems with Unreliable Machines and Finite Buffers," *IIE Transactions*, Vol. 23, No. 4, December 1991, pp. 315-331.

[13] Y. Frein, M. Di Mascolo and Y. Dallery, "On the design of generalized kanban control systems", *International Journal of Operations and Production Management*, Vol. 15, No. 9, pp. 158-184.

[14] S. B. Gershwin, "An Efficient Decomposition Method for the Approximate Evaluation of Tandem Queues with Finite Storage Space and Blocking," *Operations Research*, pp. 291-305, March-April, 1987.

[15] S. B. Gershwin, "Hierarchical Flow Control: A Framework for Scheduling and Planning Discrete Events in Manufacturing Systems," Proceedings of the IEEE, Vol. 77, No. 1, pp. 195-209, January, 1989. Reprinted in *Discrete Event Dynamic Systems*, edited by Yu-Chi Ho, IEEE Computer Society Press, 1992.

[16] S. B. Gershwin, "Assembly/Disassembly Systems: An Efficient Decomposition Algorithm for Tree-Structured Networks," *IIE Transactions*, Vol. 23, No. 4, December 1991, pp. 302-314.

[17] S. B. Gershwin, *Manufacturing Systems Engineering,* Prentice-Hall, 1994.

[18] S. B. Gershwin, "Design and Operation of Manufacturing Systems — A Controls-Based Synthesis," *International Conference on Computer Integrated Manufacturing and Automation Technology* (CIMAT'96) at the Institut National Polytechnique de Grenoble, June, 1996.

[19] S. Helber, private communication.

[20] J. G. Kimemia and S. B. Gershwin, "An Algorithm for the Computer Control of Production in a Flexible Manufacturing Systems," *IIE Transactions*, Vol. 15, No. 4, pp. 353-362, December, 1983. Reprinted in *Modeling and Control of Automated Manufacturing Systems* by Alan A. Desrochers, IEEE Computer Society Press Tutorial, 1990.

[21] S. H. Lu and P. R. Kumar, "Distributed Scheduling Based on Due Dates and Buffer Prioritization," *IEEE Trans. on Automatic Control*, Vol. 36, December, 1991.

[22] N. Srivatsan, "Synthesis of Optimal Policies for Stochastic Manufacturing Systems," Ph.D. Thesis, Massachusetts Institute of Technology, Operations Research Center, September, 1993.

[23] J. E. Schor, "Efficient Algorithms for Buffer Allocation," MIT EECS Master's Thesis, May, 1995.

[24] G. J. Van Ryzin, "Control of Manufacturing Systems with Delay," MIT EECS Master's thesis, June, 1987.

[25] G. J. Van Ryzin, X.-C. Lou, and S. B. Gershwin, "Production Control for a Tandem Two-Machine System," *IIE Trans.*, Volume 25, No. 5, pp. 5-20, Sept., 1993.