PART DISPATCH IN RANDOM YIELD MULTISTAGE FLEXIBLE TEST SYSTEMS FOR PRINTED CIRCUIT BOARDS

RAM AKELLA
Carnegie Mellon University, Pittsburgh, Pennsylvania

S. RAJAGOPALAN
University of Southern California, Los Angeles, California

MEDINI R. SINGH
The University of Michigan, Ann Arbor, Michigan

(Received June 1987; revision received August 1990; accepted September 1990)

This paper concerns dynamic part dispatch decisions in electronic test systems with random yield. A discrete time, multiproduct, miltistage production system is used as a model for the test system with the objective to minimize the sum of inventory holding, backlogging, and overtime costs over a finite horizon. Exact results for such systems have been limited to either single-stage, multiple time period, or multistage, single time period problems with a single product. Here we develop two approximate policies: the linear decision rule, and the myopic resource allocation. The effectiveness of the two policies is evaluated through simulation under different operating conditions representative of those encountered in IBM and Tandem Computer facilities. The extensive computational study clearly demonstrates the overall superiority of the linear decision rule.

We consider a two-stage production system, shown in Figure 1, where various electronic components are tested. Each item requires testing at both stages. There are three inventories: an input inventory before the first stage, an in-process inventory between the stages, and a finished item inventory after the second stage. A random fraction of the items may not meet the required specifications at the tester stages. Good products from a tester stage go into the output inventory for the stage, while bad product is reworked and then returned to the input inventory for the stage. Both the supply of raw material to the system and the demand on it are subject to uncertainties. However, a higher level planning system ensures that the supply and demand are roughly matched over an appropriate time horizon. Demand that cannot be met from inventory is backordered until inventory becomes available.

The problem is to determine how much of each item to dispatch into each stage at the start of each period. The outputs from the tester stages are a function of the dispatch quantities and a random yield. The dispatch quantities themselves are constrained by the available capacity. The capacity constraint can, however, be violated at a cost. This overtime cost applies to each stage. Each stage has a nominal capacity level; if production exceeds this capacity level, then there is an overtime cost. There are neither setup times nor setup costs. The other two types of costs considered here are for holding inventory and backordering. The objective is to determine a dispatch policy that minimizes the long-term expected average costs.

Multistage systems have been examined in the literature, in the context of production planning and scheduling. Hax and Candea (1984), and Gershwin, Akella and Choong (1985) discuss a variety of models and their effectiveness. An important feature of actual manufacturing and assembly systems that is not quite captured in these models is the internal uncertainty that results from causes, such as random yield. This paper is a first step in bridging that gap. The dominant uncertainties vary from system to system. We focus here on an electronic assembly facility that produces printed circuit boards for mainframe computers. The objective of the entire facility is to respond to incoming orders for printed circuit boards and to meet production targets on schedule, through a

Subject classifications: Dynamic programming/optimal control, applications: linear decision rule. Industries, computers/electronics: PCB testing, semiconductor fabrication. Inventory/production, multi-item, echelon, stage: dispatch policies for uncertain yield systems.
Area of review: MANUFACTURING, PRODUCTION AND SCHEDULING.

0030-364X/92/4004-0776 $01.25

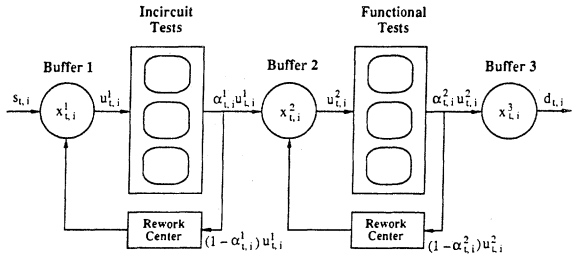

Figure 1. Test stages.

combination of component ordering policies, aggregate planning and detailed shop floor dispatch. We deal with the last issue in this paper.

Production control in the presence of random yield has attracted considerable interest. We discuss some of the work that is closely related to the present paper. The intent is not to survey the work in the area but to contrast what is being done here with what has appeared in the literature. Interested readers are referred to Yano and Lee (1989) for an excellent survey of lot sizing problems in the presence of random yields.

Previous research in the area has mainly concerned lot sizing decisions for a single product. Yano (1986) considers single-stage, finite and infinite horizon problems with linear costs, deterministic demands and independent, identically distributed yields. Under some restriction on the yield distribution, it is shown that the optimal production quantity is *multiplicative*, i.e., it is simply a multiple of the net demand for the period. Gerchak, Vickson and Parlar (1988) consider a similar model but they allow stationary random demands. They show that the optimal production quantity has neither a simple *order-up-to* or multiplicative structure nor is it myopic in nature. The lot size is a complicated function of system parameters not amenable to efficient computation.

The only multiproduct, multiple period model with random yield, to our knowledge, is by Karmarkar and Lin (1986). They present a single-stage model with linear cost structure reminiscent of classical LP-based production smoothing models. The solution approach comprises developing lower and upper bounds on the optimal solution. A good lower bound is obtained by using modified (or heuristic) Lagrangian relaxation that produces independent, single period subproblems. Three different procedures are presented to provide upper bounds. However, the only upper bound that produces small duality gaps is obtained through simulation. The more efficient procedures for deriving upper bounds that are presented are not very encouraging in terms of tightness of bounds. However, the lower bounds seem to be good and also suggest a heuristic procedure to directly obtain upper bounds.

Lee and Yano (1988) analyze a multistage (serial) system, similar to the one considered here, but with a single period and a single product. They show that the optimal target input is given by a unique critical number which can be computed efficiently in a sequential fashion. The multistage, multiproduct, multiperiod model presented here can be considered a generalization of the above models, where production resources have to be allocated in the presence of time varying demands for a portfolio of products. We allow nonstationary yield which can be correlated between the stages. These complexities are an essential part of the complex manufacturing environment considered here, and our effort is directed toward the development of efficient solution methods for resource allocation and dispatch decisions in such systems.

In the next section, we provide the background for the problem and discuss key features of the system that need to be incorporated in the model. A general formulation for the problem is given in Section 2 as a dynamic optimization problem subject to a set of linear constraints. This formulation encompasses all the complexities of the problem and allows general inventory and capacity related costs. Though intractable to solve in the proposed form, it forms the basis for various approximations in the following sections. When inventory and capacity related costs are quadratic or can be approximated by quadratic functions, the dynamic optimization problem can be solved very efficiently. We show in Section 3 that the optimal dispatch rule in this case is affine in available inventory as well as expected demand and supply; we call this policy the linear decision rule (LDR). Like the classical production smoothing models with quadratic costs (Holt et al. 1960), the LDR requires only expected values of demand and supplies. But unlike those models, expected values are not sufficient here for all uncertain quantities. For example, second-order moments (covariance terms) of yield distribution are needed. The stochastic dynamic programming formulation is quite general and captures many of the complexities of the model; yet the computational burden is very modest. The linear decision rule, however, is only heuristic due to many approximations made to arrive at the solution. To evaluate its performance against a reasonable alternative, we propose another solution scheme, myopic resource allocation (MRA), in Section 4. The MRA is based on decomposing the

problem by tester stage and time period, where the decomposed subproblems are solved efficiently using an adaptation of the resource allocation algorithm proposed by Luss and Gupta (1975) and Zipkin (1980). We compare the two dispatch policies in Section 5 using a comprehensive simulation study which mimics the cardline tester systems found in IBM and Tandem. The experiments, based on a linear holding, backlogging and overtime costs, clearly demonstrate the overall superiority of LDR under a variety of operating conditions. We expect that the LDR will perform even better under a more general setting because it explicitly takes account of nonstationarity and correlation of various parameters not considered in the computational experiments. A summary of results is provided in Section 6.

1. CARDLINE DESCRIPTION AND BACKGROUND OF PROBLEM

A typical cardline or electronic assembly facility is shown schematically in Figure 2. Our model is based on two facilities: IBM and Tandem. These systems have four stages:

1. the stocking warehouse, where raw components are received and stored;
2. the insertion stage, where the electronic components are inserted into printed circuit boards;
3. the soldering area, where cards that have had components inserted are wave soldered;
4. the test and rework area, where the cards are tested and reworked if necessary.

A detailed description of the system and the real-time part dispatch issues at the insertion stage, where uncertainties such as machine failures are explicitly modeled, can be found in Akella, Choong and Gershwin (1984), Akella and Kumar (1986), Akella, Singh and Krogh (1990), and Gershwin, Akella and Choong. Here we focus on the tester area and describe a dynamic part dispatch problem where uncertainties such as random yield play a key role.

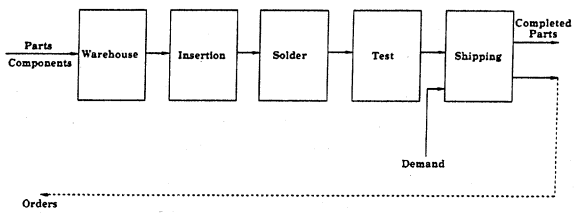

Figure 2. Cardline for electronic assembly.

Consider the following simplified representation of a tester area, with two substages, where in-circuit and functional tests are performed (Figure 1) to identify defective connections. The cause for faulty connections can be traced to the soldering stage, where an entire batch of cards may be affected, leading to correlated errors. We now describe some key features of these test systems that need to be incorporated in any model for determining dynamic dispatch policies.

1.1. Arrival and Departure Processes

Cards of different types arrive from the soldering area. Supply from the soldering area is based on the derived demands determined by a higher-level planning system. This system ensures that average inventories or backorders are bounded. Despite higher-level planning to coordinate arrivals with derived demand at each stage, there is some randomness due to uncertainties, such as machine failures, in the previous stages.

1.2. Random Yield

The main uncertainty that we focus on here is the random yield at each test stage. This results in stochastic workloads at each test stage, uncertainty in meeting the demand and increased, uncertain inventories at the buffers. We use a multiplicative yield model, where a random fraction of a batch that is released into the system is found to be defective. Occasionally, an entire batch of cards moving through the wave solder area is affected by fluctuations in belt speed and this results in batch correlated connection defects. The multiplicative yield model is especially suited for the high volume production with large batch sizes and correlated defects. Note that this model differs from the Bernoulli trial models, where each card defect is assumed to be independent of the others, the process is stationary, and batch correlations are ignored.

Product life cycles also affect the yield. During the product introduction phase, the σ/μ (standard deviation by mean) ratio is high. As the product matures, this ratio decreases to a relatively small value due to technological improvements and learning.

1.3. Rework

Here we assume that defective boards during a given day are sent to a separate rework station where they are reworked by the end of the same day. The model can easily be extended to allow any arbitrary but known rework times.

2. PROBLEM FORMULATION

We use a discrete time model to represent the dispatching problem outlined in the Introduction. This model assumes a finite horizon with N periods. First, we write the inventory balance equations for the test system shown in Figure 1. Let $s_{t,i}$ be the random arrival at stage 1 and $d_{t,i}$ be the random demand at stage 3 for card type i in period t. Both $s_{t,i}$ and $d_{t,i}$ are bounded and roughly matched by a higher-level planning system. The inventory level of card type i at buffer j (= 1, 2, 3) in period t is $x^j_{t,i}$. Let $u^j_{t,i}$ be the number of cards of type i tested during period t at stage j (= 1, 2), and $\alpha^j_{t,i}$ be the corresponding random yield term, which is the random fraction of good cards. The number of cards that need to be reworked and retested is given by $(1 - \alpha^j_{t,i})u^j_{t,i}$.

We can now write the *state* (or inventory-balance) *equations*:

$$x^1_{t+1,i} = x^i_{t,i} - u^1_{t,i} + (1 - \alpha^1_{t,i})u^1_{t,i} + s_{t,i}$$

$$x^2_{t+1,i} = x^2_{t,i} - u^2_{t,i} + (1 - \alpha^2_{t,i})u^2_{t,i} + \alpha^1_{t,i}u^1_{t,i}$$

$$x^3_{t+1,i} = x^3_{t,i} + \alpha^2_{t,i}u^2_{t,i} - d_{t,i}.$$

Combining these in vector form, we obtain

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{G}\mathbf{w}_t \tag{1}$$

where,

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{x}_{t,1} \\ \vdots \\ \mathbf{x}_{t,M} \end{bmatrix}; \mathbf{x}_{t,i} = \begin{bmatrix} x^1_{t,i} \\ x^2_{t,i} \\ x^3_{t,i} \end{bmatrix};$$

$$\mathbf{u}_t = \begin{bmatrix} \mathbf{u}_{t,1} \\ \vdots \\ \mathbf{u}_{t,M} \end{bmatrix}; \mathbf{u}_{t,i} = \begin{bmatrix} u^1_{t,i} \\ u^2_{t,i} \end{bmatrix};$$

$$\mathbf{w}_t = \begin{bmatrix} \mathbf{w}_{t,1} \\ \vdots \\ \mathbf{w}_{t,M} \end{bmatrix}; \mathbf{w}_{t,i} = \begin{bmatrix} s_{t,i} \\ d_{t,i} \end{bmatrix};$$

$$\mathbf{B}_t = \begin{bmatrix} \mathbf{B}_{t,1} & & \\ & \ddots & \\ & & \mathbf{B}_{t,M} \end{bmatrix};$$

$$\mathbf{B}_{t,i} = \begin{bmatrix} -\alpha^1_{t,i} & 0 \\ \alpha^1_{t,i} & -\alpha^2_{t,i} \\ 0 & \alpha^2_{t,i} \end{bmatrix};$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & & \\ & \ddots & \\ & & \mathbf{G}_M \end{bmatrix}; \mathbf{G}_i = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & -1 \end{bmatrix}.$$

We now represent the constraints on the time available for testing all the card types at the two stages. Let β^j_t be the amount of regular time available for processing at stage j in period t. Also, let τ^j_i represent the unit test time of card type i at stage j. Then the capacity constraints can be represented by

$$\sum_{i=1}^{M} \tau^j_i u^j_{t,i} \leq \beta^j_t,$$

or, in matrix form

$$\mathbf{T}\mathbf{u}_t \geq \boldsymbol{\beta}_t. \tag{2}$$

$(\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t)$ represents the overtime on which there are no limits.

Finally, we define the *objective function* as:

$$\underset{\substack{\mathbf{u}_t \geq 0 \\ t=1,2,\ldots,N-1}}{\text{minimize}} \; \underset{\mathbf{B}_t,\mathbf{w}_t}{E} \left\{ \sum_{t=1}^{N} g_t(\mathbf{x}_t) + \sum_{t=1}^{N-1} f_t(\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t) \right\}, \tag{3}$$

where $g(\mathbf{x}_t)$ represents inventory holding and backordering costs, and $f(\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t)$ represents the cost of overtime. The specific form of these functions depends on the manufacturing environment. The classical literature has assumed convex cost functions, in particular, linear (e.g., Karmarkar and Lin) and quadratic (e.g., Holt et al. (HMMS)). These functions often provide good approximations and have the merit of being analytically tractable. In the next two sections, we present two possible approaches to solve the above problem. The first approach assumes quadratic cost functions and models the coupling between the different stages explicitly. The second approach assumes linear cost functions and is based on decomposing the problem stagewise.

3. LINEAR DECISION RULE

We will assume that the cost functions $g(\mathbf{x}_t)$ and $f(\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t)$ are either quadratic or can be approximated closely by quadratic functions. Specifically, we assume that

$$g(\mathbf{x}_t) = \mathbf{x}_t'\mathbf{Q}_t\mathbf{x}_t \tag{4}$$

and

$$f(\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t) = (\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t)'\mathbf{R}_t(\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t), \tag{5}$$

where $\mathbf{Q}_t$ is a diagonal matrix representing the inventory carrying/backordering penalty coefficient for all the part-types at the three buffers in period t. Correspondingly, $\mathbf{R}_t$ is the diagonal matrix representing the overtime/undertime penalty coefficient for the resources at the two tester stages in period t. Note that

the inventory cost function (4) penalizes positive and negative inventories equally. In reality, backordering is much costlier than holding inventory. The following modifications could be used to overcome these limitations:

$$g(\mathbf{x}_t) = \mathbf{x}_t'\mathbf{Q}_t\mathbf{x}_t - \mathbf{C}_t\mathbf{x}_t \tag{6}$$

$$g(\mathbf{x}_t) = (\mathbf{x}_t - \mathbf{s}_t)\mathbf{Q}_t(\mathbf{x}_t - \mathbf{s}_t) \tag{7}$$

where nonnegative vector $\mathbf{C}_t$ is used to create the desired asymmetry in cost function and vector $\mathbf{s}_t$ specifies a desirable positive inventory at each buffer to be tracked by production. For the purpose of optimization, both modifications are equivalent as substituting $\mathbf{C}_t = 2\mathbf{s}_t\mathbf{Q}_t$ in (6) gives (7), except an inconsequential constant term $\mathbf{s}_t'\mathbf{Q}_t\mathbf{s}_t$. We will use quadratic form (6) in the remainder of this paper.

Backlogging is not permitted at buffers 1 and 2, but this constraint cannot be imposed directly in this formulation. When a situation arises such that the optimal input quantity is larger than the available inventory, the input quantity is truncated to the level of available inventory. The backlogging costs at buffers 1 and 2 can be interpreted as the implied cost of not having enough inventory to satisfy the input quantity and hence lowering the output at the downstream buffer.

The cost function (5) could also be modified to make overtime more expensive than undertime. Following the arguments above, we can use the modification

$$f(\mathbf{Tu}_t - \boldsymbol{\beta}_t) = (\mathbf{Tu}_t - \boldsymbol{\beta}_t)'\mathbf{R}_t(\mathbf{Tu}_t - \boldsymbol{\beta}_t) + \mathbf{F}_t(\mathbf{Tu}_t - \boldsymbol{\beta}_t), \tag{8}$$

where the positive vector $\mathbf{F}_t$ increases the penalty for overtime compared to that for underutilization of capacity. It is also required that the input quantity, $\mathbf{u}_t$, be nonnegative. However, we cannot incorporate this constraint explicitly into the present formulation. We shall impose this constraint heuristically by setting any negative production rate to zero.

The problem can now be stated as

$$\underset{\substack{\mathbf{u}_t \\ t=1,2,\ldots,N-1}}{\text{minimize}} \; \underset{\mathbf{B}_t,\mathbf{w}_t}{E} \left\{ \sum_{t=1}^{N} \mathbf{x}_t'\mathbf{Q}_t\mathbf{x}_t - \mathbf{C}_t\mathbf{x}_t + \sum_{t=1}^{N-1} (\mathbf{Tu}_t - \boldsymbol{\beta}_t)'\mathbf{R}_t(\mathbf{Tu}_t - \boldsymbol{\beta}_t) + \mathbf{F}_t(\mathbf{Tu}_t - \boldsymbol{\beta}_t) \right\} \tag{9}$$

subject to (1).

This is a variant of the classical linear quadratic control problem with random coefficient matrix (Bertsekas 1976). The above formulation is quite general in the sense that it allows multiple products, multiple periods and multiple stages of production; the demand for finished products and the supply of raw materials can be nonstationary and random; production yields can vary with time and be correlated among part types as well as between the stages; available capacity can vary from period to period and all costs can be nonstationary. Any delay between the production stages can also be incorporated simply by rewriting the state-equation (1). The following result (proved in the Appendix) gives the optimal production quantitities in terms of system parameters.

Theorem 1. *For a production system with dynamics described by* (1), *and the objective function given by* (9), *the optimal dynamic dispatch policy is given by*

$$\mathbf{u}_t^* = \mathbf{L}_t\mathbf{x}_t + \mathbf{M}_t, \tag{10}$$

where

$$\left.\begin{aligned} \mathbf{L}_t &= -\boldsymbol{\Gamma}_t^{-1}\bar{\mathbf{B}}_t'\mathbf{K}_{t+1} \\ \mathbf{M}_t &= \boldsymbol{\Gamma}_t^{-1}[\mathbf{T}'\mathbf{R}_t\boldsymbol{\beta}_t - \bar{\mathbf{B}}_t'\mathbf{K}_{t+1}\mathbf{G}\bar{\mathbf{w}}_t \\ &\qquad - \bar{\mathbf{B}}_t'\mathbf{P}_{t+1}' - \tfrac{1}{2}\mathbf{T}'\mathbf{F}_t'] \\ \boldsymbol{\Gamma}_t &= E\{\mathbf{B}_t'\mathbf{K}_{t+1}\mathbf{B}_t\} + \mathbf{T}'\mathbf{R}_t\mathbf{T} \\ \mathbf{K}_t &= \mathbf{K}_{t+1} - \mathbf{K}_{t+1}\bar{\mathbf{B}}_t\boldsymbol{\Gamma}_t^{-1}\bar{\mathbf{B}}_t'\mathbf{K}_{t+1} + \mathbf{Q}_t \\ \mathbf{P}_t &= \mathbf{P}_{t+1} + [\mathbf{G}\bar{\mathbf{w}}_t + \bar{\mathbf{B}}_t\mathbf{M}_t]\mathbf{K}_{t+1} - \tfrac{1}{2}\mathbf{C}_t \\ \mathbf{K}_N &= \mathbf{Q}_N;\ \mathbf{P}_N = -\tfrac{1}{2}\mathbf{C}_N. \end{aligned}\right\} \tag{11}$$

The optimal policy for each part type is affine in the inventory levels. Also, observe from the form of $\mathbf{M}_t$ that the feedback policy is also affine in a linear combination of the expected demand and supply. As in classical production smoothing models with quadratic costs (HMMS, p. 123) the linear decision rule requires only expected values (shown with overbars) of future demands and supplies, all other distributional information about these quantities are irrelevant. But, unlike those models, expected value alone is not sufficient for all uncertain quantities. For example, second-order moments (covariance terms) of yield distribution are needed, as can be seen from the presence of the $E\{\mathbf{B}_t\mathbf{K}_{t+1}\mathbf{B}_t\}$ term in $\boldsymbol{\Gamma}_t$. The certainty equivalence obviously does not hold for this model and this is mainly because of the random coefficient matrix $\mathbf{B}_t$ in the state equation. The alternate objective functional forms (7 and 8) can be incorporated in Theorem 1 simply by substituting $(\mathbf{x}_t - \mathbf{s}_t)$ and

$(\beta_t - \gamma_t)$ in place of $\mathbf{x}_t$ and β_t, respectively; all the observations made above remain valid.

For such a general formulation, it is interesting that the decision rule is simple, and essentially linear. Furthermore, the calculation of affine constants $\mathbf{L}_t$ and $\mathbf{M}_t$ can be done recurisvely, overwriting the intermediate matrices $\mathbf{K}_t$ and $\mathbf{P}_t$ at each recursion. Unlike the LP formulation of similar planning problems, where the size of the coefficient matrix increases proportionately with the number of periods, matrices $\mathbf{K}_t$ and $\mathbf{P}_t$ are independent of the number of periods. The recursive calculation involves simple matrix operations at each step. As a result, the computational burden for the proposed decision rule turns out to be far less than that needed for even a deterministic LP solution.

We point out, however, that this simplicity is closely linked to the quadratic form of the objective function. When the actual costs are not quadratic, the method can still be used by choosing quadratic cost functions that yield a good linear decision rule for the original cost environment.

We now discuss the issue of choosing the quadratic cost coefficients. Schneeweiss (1971, 1974) developed a two-stage procedure for choosing the LDR parameters optimally for production smoothing problems with nonquadratic costs and Gaussian demands. Using Wiener filtering theory, the stationary probability distribution in the inventory-production space is first derived as a function of quadratic cost parameters. The optimal decision rule parameters are then computed so that the expected cost resulting from the probability distribution, given the firm's actual cost function, is minimized. The Wiener filtering procedure is intimately related to the certainty equivalence principle, which does not hold in the proposed model. We present another two-stage iterative procedure where, given the probability distributions of inventories and overtime/undertime for each period, the quadratic cost parameters are fitted such that the weighted least square deviation from the firm's actual cost is minimized, where weights correspond to the probabilities of being in various inventory and overtime states (see the Appendix). A linear decision rule is then computed using these cost parameters. The LDR is used, in turn, to generate, by repeated simulation runs, an updated distribution of inventories and overtime/undertime, which is then used to achieve a better quadratic fit.

While we do not prove the convergence of the proposed method (Schneeweiss's iterative method also suffers from the same limitation), computational experience shows that the fitted cost parameters converge to a narrow range within a few iterations, provided that we start with a good initial guess of probable inventories and overtimes. It turns out that the LDR parameters $\mathbf{L}_t$ and $\mathbf{M}_t$ are not very sensitive to the quadratic cost parameters, and the iterative process can be terminated whenever improvement in expected total cost due to a new fit becomes insignificant. The proposed procedure for finding quadratic cost function has another advantage. Notice that the distribution of inventories and overtime/undertime will depend on the distribution of demand and supply. The fitted quadratic cost parameters, as a result, will depend on the demand and supply distribution. The linear decision rule, which does not require any higher order moments of the demand and supply distributions, is now dependent on them indirectly through the quadratic cost parameters. We believe that this further enhances the performance of the LDR.

As a final point, we note that the applicability of the linear decision rule is not limited to problems with quadratic costs alone. For example, the optimality of the linear policy is established in Yano for a single-stage production system with variable yield, linear costs and deterministic demands for both finite and infinite horizon problems. For many other dynamic optimization problems with linear state equations and nonquadratic costs, Schneeweiss (1971, 1974) shows that linear policies can be a good approximation.

4. MYOPIC RESOURCE ALLOCATION

We present an alternative approach to solve the problem using a capacitated newsboy model with random yields and deterministic demands. The method is based on decomposing the problem by tester stage and time period. The decomposed subproblems consist of allocating the capacity, including overtime, to the various part types in a newsboy fashion. This problem is solved efficiently using a variant of the resource allocation approach proposed by Luss and Gupta (1975) and Zipkin (1980).

For any period, given beginning inventories, the optimal dispatch quantities are determined in the following fashion: Starting with the final stage, the dispatch quantities are computed sequentially for each stage using corresponding costs and the net demand (demand *less* inventory). Barring the final stage, where actual demands occur, for all other stages the dispatch quantity at the following stage is used as the demand for the previous stage. Assume that

yields are not correlated between the stages or part types; inventory costs are linear and separable for part types and capacity violation is penalized by a linear overtime cost.

The cost of backlogging at the intermediate buffer is the price one pays for not being able to feed the following stage when required by the optimal decision. In the worst case, a unit short at the intermediate buffer may result in a unit backlogged at the final stage. In this case, the cost of backlogging at the intermediate buffer will be exactly equal to the backlogging cost at the final buffer. It is possible that the cost of backlogging at the intermediate buffer is less than that at the final buffer due to the benefits derived from the alternative use of capacity at stage 2 freed by insufficient supply at the intermediate buffer. However, the value of a unit of capacity can vary from zero to overtime cost, depending upon the marginal value of capacity for other part types. As a result, the backlogging cost at the intermediate buffer for any part type is a complex function of stock levels and yield distributions of all part types. To preserve the separable structure of the cost function, a property critical for efficient solution of the decomposed problems, we have taken the backlogging cost at the intermediate buffer to be the same as that at the final buffer.

Consider a tester stage j which, according to Figure 2, draws the components from buffer j and after testing puts the good output into buffer $j+1$. Let d_i^j be the net demand and $f_i^j(.)$ the probability density function of yield distribution for part type i with corresponding upper and lower limits UL_i and LL_i, respectively. The per unit inventory holding and backordering costs at the output buffer, $j+1$, are h_i^{j+1} and b_i^{j+1}, respectively for part type i. Choosing u_i^j as the input quantity implies that a backordering cost $b_i^{j+1}(d_i^j - \alpha_i^j u_i^j)$ may incur if the good output $\alpha_i^j u_i^j$ turns out to be less than the net demand; otherwise, an extra holding cost $(h_i^{j+1} - h_i^j)(\alpha_i^j u_i^j - d_i^j)$ may incur due to surplus production. Due to value added at tester stages, holding costs at consecutive buffers satisfy the relationship $h_i^{j+1} \geq h_i^j$. The sum of backordering and excess holding costs for part type i can be represented by a convex cost function

$$L_i^j(u_i^j) = \int_{\alpha_i^j = LL_i}^{\alpha_i^j = (d_i^j/u_i^j)} b_i^{j+1}(d_i^j - \alpha_i^j u_i^j) f_i^j(\alpha_i^j)\, d\alpha_i^j + \int_{\alpha_i^j = (d_i^j/u_i^j)}^{\alpha_i^j = UL_i} (h_i^{j+1} - h_i^j)(\alpha_i^j u_i^j - d_i^j) f_i^j(\alpha_i^j)\, d\alpha_i^j.$$

The optimization problem for stage j can now be stated as

$$\text{minimize} \sum_{i=1}^{M} L_i^j(u_i^j) + r^j y^j$$

$$\text{subject to} \sum_{i=1}^{M} \tau_i^j u_i^j \leq \beta^j + y^j$$

$$y^j \geq 0; \quad u_i^j \geq 0 \; i = 1, \ldots, M,$$

where y^j is the amount of overtime and r^j the unit cost of overtime at stage j. In what follows, we suppress superscript j for notational simplicity. The optimization problem described above has a derivative separable objective function with a single resource constraint. The expected marginal decrease in cost, $-(1/\tau_i)(dL_i/du_i)$, is a nonincreasing function of total machine time assigned to a part type. Starting with a maximum value of $(b_i^{j+1}\bar{\alpha}_i)/\tau_i$ the expected marginal decrease in cost remains constant until the allocated machine time exceeds $\tau_i d_i / UL_i$, after which it starts decreasing. The optimal allocation is such that the marginal cost of capacity equals the expected marginal benefit derived from an extra unit of capacity; those part types that cannot afford this price do not get produced. If overtime is used in the optimal solution, then the marginal cost of capacity also equals the unit cost of overtime and it is easy to identify which part types do not get produced in the optimal solution. When overtime is not used in the optimal solution, a simple ranking procedure due to Luss and Gupta (1975), and Zipkin (1980) can be used to identify which part types do not get produced. The following algorithm is used to obtain an optimal solution.

Algorithm 1

Step 1. For $i = 1, \ldots, N$ find $u_i^{(1)}$ such that

$$\left.\frac{dL_i}{du_i}\right|_{u_i^{(1)}} = 0.$$

If $\sum_{i=1}^{M} \tau_i u_i^{(1)} < \beta$, then $u_i^* = u_i^{(1)}$ for all i. Stop.

Step 2. Assign each part type a unique rank $[i]$ such that

$$-\frac{1}{\tau_{[i]}}\left.\frac{dL_{[i]}}{du_{[i]}}\right|_{u_{[i]}=0} \geq -\frac{1}{\tau_{[i+1]}}\left.\frac{dL_{[i+1]}}{du_{[i+1]}}\right|_{u_{[i+1]}=0}.$$

Break the tie arbitrarily. Store the transformation $i \rightarrow [i]$. Let j be the largest index such that

$$-\frac{1}{\tau_{[j]}}\left.\frac{dL_{[j]}}{du_{[j]}}\right|_{u_{[j]}=0} \geq r, \; 1 \leq j \leq M.$$

For $i = 1, 2, \ldots, j$, find $u_{[i]}^{(2)}$ such that

$$-\frac{1}{\tau_{[i]}}\frac{dL_{[i]}}{du_{[i]}}\bigg|_{u_{[i]}^{(2)}} = r.$$

Set $u_{[i]}^{(2)} = 0$ for $i = j + 1, \ldots, M$. If $\sum_{i=1}^{M} \tau_{[i]} u_{[i]}^{(2)} - \beta > 0$, then $u_{[i]}^{*} = u_{[i]}^{(2)}$ for all i. Transform solution $u_{[i]}^{*} \rightarrow u_i^{*}$ and stop.

Step 3. a. Set $k = j$.

b. Compute $\lambda(k)$ and $u_{[i]}^{(3)}$, $i = 1, \ldots, N$ by simultaneous solution of the following system of equations

$$-\frac{1}{\tau_{[i]}}\frac{dL_{[i]}}{du_{[i]}}\bigg|_{u_{[i]}^{(3)}} = \lambda(k), \quad i = 1, \ldots, k \tag{12}$$

$$u_{[i]}^{(3)} = 0, \quad i = k + 1, \ldots, M \tag{13}$$

$$\sum_{i=1}^{M} \tau_{[i]} u_{[i]}^{(3)} = \beta. \tag{14}$$

c. If $-\frac{1}{\tau_{[k+1]}}\frac{dL_{[k+1]}}{du_{[k+1]}}\Big|_{u_{[k+1]}=0} \leqslant \lambda(k)$,

then $u_{[i]}^{*} = u_{[i]}^{(3)}$ for all i. Transform solution $u_{[i]}^{*} \rightarrow u_i^{*}$ and stop.

d. Set $k = k + 1$ and return to Step 6.

The proof that the above algorithm yields the optimal solution is given in the Appendix. The algorithm exploits the convex, separable nature of the cost function and the single resource constraint to obtain a simple ranking of products, which is then used to obtain the optimal solution efficiently. Note that if the optimal solution is such that either capacity is underutilized, or overtime is used, then the algorithm terminates at either Step 1 or Step 2, respectively. Only when machine time is scarce and the overtime is prohibitively expensive that it goes to Step 3, which is essentially Zipkin's algorithm. The above algorithm can be extended to convex inventory costs. However, this will increase the computational effort in solving the simultaneous equations (12–14). Note also that cost functions $L_i(u_i)$ are not strictly convex, which implies that for the same marginal cost of capacity (λ), a number of solutions (u_i) may exist. However, this does not cause any problem as long as Steps 2 and 3 above are interpreted suitably. For example, corresponding to $\lambda(k)$ a set of $u_{[i]}^{(3)}$ may satisfy (12) in Step 3b and all such solutions should be considered as candidates for simultaneous solution of the set of equations (12)–(14).

5. COMPUTATIONAL RESULTS

In this section, we report the results of a computational study performed to assess the effectiveness of the two policies: the linear decision rule (LDR) and the myopic resource allocation (MRA), discussed in previous sections. Different assumptions were made for the development of the two dispatch policies and our goal is to illustrate their performance under various operating conditions. Since the manufacturing problem addressed here is too complex to be solved optimally, we use simulation as a benchmark to evaluate the approximations. This approach is common for problems for which exact results are not known; for example, see Bitran and Tirupati (1988) and many references therein.

We consider a production environment that is representative of the cardline tester systems found in IBM and Tandem. However, the system details have been simplified and parameter values disguised for the study. We first discuss the details of the experiment and then the results.

Each of the problem sets we consider has four part types. For each problem a horizon of 10 time periods was considered. Short-term dispatch decisions are typically based upon a horizon of approximately this length; see, for example, Graves (1982). The relative performance of the two decision rules is not very sensitive to the horizon length due to a careful choice of initial inventories with which we start all simulation runs.

5.1. Initial Inventories ($x_{1,l}^{l}$)

At buffers 1 and 2, due to the same-day rework policy, there is always a certain residual inventory. We assume that the initial inventories at buffers 1 and 2 are equal to average expected residual inventories. At buffer 3, they are assumed to be zero.

5.2. Yield Values ($\alpha_{t,l}^{l}$)

Yields for all part types were taken to be stationary and uncorrelated. They were generated from uniform distributions with averages shown in Table I. To study the effect of yield variance, two ranges of the σ/μ ratio were considered: *Low* (0.03 – 0.1), and *High* (0.2 – 0.4).

Table I
Average Yield and Test Times

	Expected Yield at Test Stage		Test Time at Test Stage	
Part Type	1	2	1	2
1	0.550	0.450	0.160	0.284
2	0.700	0.550	0.228	0.226
3	0.525	0.525	0.258	0.327
4	0.400	0.525	0.232	0.310

5.3. Demand and Supply ($d_{t,i}$ and $s_{t,i}$)

As discussed, demand is either deterministic (based on master production schedules) or has a low noise level at the time scale considered. In this computational study, we assume demand to be deterministic and mildly fluctuating over time. Demand values were generated randomly from the range 200–300.

Supply to the tester stage was generated such that it was reasonably well-matched with the demand. Supplying more than what is needed will incur an unnecessary holding cost; similarly, an acute shortage of supply will drive the backlogging costs up no matter what policy is chosen. In both cases, the total cost will be inflated and the savings due to the better policy devalued.

5.4. Test Times and Tester Capacities (τ_i^j and β_t^j)

Test times are of the order of a fraction of a minute. The specific values used in the computational study are listed in Table I. Based on these test times, the mean yield values at the two stages, and the demands, we generated three different scenarios for capacities:

i. Matched Capacity. Capacity over 10 periods roughly matched to the demand at both stages.

ii. Surplus Capacity. Capacity available 20% greater than in i at both stages.

iii. Inadequate Capacity. Capacity available 20% less than in i at both stages.

The capacity was assumed to be constant over time.

5.5. Performance Measure

The exact functional forms of inventory and capacity related costs are difficult to establish and they change with the manufacturing environment. While the LDR can be used for any convex cost function subject to an accurate approximation by quadratic form in the region of interest, the MRA was developed using linear inventory and overtime cost functions. Hence, we chose to compare the performance of the two dispatch rules based on linear inventory and overtime costs. This will also put to test how well LDR performs in extreme circumstances because achieving an accurate quadratic fix to linear costs is usually more difficult than fitting quadratic forms to convex functions. The final comparison is based on the *actual* total cost incurred by the two policies regardless of how they were developed.

5.6. Costs

To test the performance of the dispatch policies extensively, several cost scenarios were considered and are described below.

Holding Costs. Holding costs were based on the value of the printed circuit boards and a 30% annual holding charge. They ranged from 0.2 to 0.4 per unit per period and were taken to be the same at the three buffers. They were used as base costs compared to which other costs were defined on a relative scale.

Backlogging Costs. Backlogging cost is incurred only at buffer 3. Three ratios of backlogging cost/holding cost(b/h) were considered: 2, 5, and 10. They provide various levels of cost asymmetries against which performance of LDR can be judged.

Overtime Costs. The overtime cost is expressed here as cost per unit time of extra capacity and is assumed to be the same at the two test stages. A judicious tradeoff between overtime and backlogging costs is central to a good dispatch policy and the relative cost of overtime compared to backlogging plays a key role in this process. An inexpensive overtime cost can mitigate the effect of bad decisions by employing overtime capacity whenever needed, without much penalty. A very expensive overtime, on the other hand, may not serve any purpose since it might become cheaper to backlog than produce using overtime. For any part type, if the maximum expected marginal decrease in cost due to use of a unit of capacity, $b\bar{\alpha}/\tau$, is less than the overtime cost, it will never be produced using the overtime capacity. However, if the quantity $b\bar{\alpha}/\tau$ for a part type is greater than the cost of overtime, it will qualify for production using regular time, and it may even qualify for overtime if, by optimal allocation of regular capacity, the marginal cost of capacity turns out to be greater than the cost of overtime. To examine the various possibilities of overtime-backlogging tradeoffs, we consider three scenarios for overtime cost for each case of b/h:

Prohibitively Expensive Overtime: The overtime cost is greater than the maximum of $b\bar{\alpha}/\tau$ over all part types at both the stages. This is equivalent to a hard capacity constraint since use of overtime is never beneficial. However, due to quadratic approximation, LDR may use some amount of overtime and pay a high penalty.

Inexpensive Overtime: The overtime cost is significantly less than the minimum of $b\bar{\alpha}/\tau$ over all part

types at both the stages. A liberal use of overtime can be expected in this case both to avoid the backlogging as well as to fill any outstanding order due to erroneous allocation in the past. Since overtime can be readily used to bail out from any backlog, the effect of a bad decision does not propagate beyond one period.

Moderate Overtime Cost: The overtime cost, in this case, is chosen such that it is within the range over which $b\bar{\alpha}/\tau$ lies for different part types. Overtime is used sparingly in this case and tradeoff between backlogging and overtime costs becomes critical.

5.7. Experimental Details

A total of 54 problem sets were constructed based on three cases of available capacity, two types of yield variability, three levels of backlogging-to-holding cost asymmetry, and three scenarios for overtime cost. Each problem set consisted of 10 problems, each of which was constructed using a different speed for the generation of the demand sequences. For each problem, the simulation run consisted of a pilot run, and a main study, as described below.

Pilot Run. The purpose of the pilot run was to determine the quadratic cost parameters **Q**, **C**, **R**, and **F**. To generate an initial value of these parameters, some knowledge about the range of inventory/backorder and overtime values is needed so that a good fit to actual cost can be obtained using the weighted least square method described in the Appendix. To provide this knowledge, a number of simulation runs were made using the MRA. Once initial estimates of quadratic cost parameters were available, they were used to generate the linear decision rule, which were then used in simulation runs to update the knowledge about the range of inventory/backorder and overtime values. This process was used to successively improve the fit and in our experience 2 to 3 iterations were sufficient to achieve a good fit.

Main Study. The main study consisted of 20 simulations using different sample paths of yield realizations for each of the 540 problems in 54 categories. However, both the decision rules used identical sample paths for comparison. A period-by-period account of holding and backlogging costs incurred by each part type and overtime costs incurred at each production stage was maintained to understand the behavior of the two policies.

5.8. Results

A comparative summary of results is provided in Tables II–IV, which are classified by the backlogging-to-holding cost ratios. The entries indicate the average total cost for the MRA as *a percentage excess* over the average total cost for the LDR. For example, entry 21.1 means that the average total cost for the MRA was found to be 1.211 times the corresponding average total cost for the LDR. The absence of any negative terms in the three tables indicates the overall superiority of LDR over MRA. However, the relative cost

Table II
Relative Cost for MRA (Percentage Excess Over LDR), $b/h = 2$

	Inadequate Capacity		Matched Capacity		Surplus Capacity	
Overtime	Low Yield Variance	High Yield Variance	Low Yield Variance	High Yield Variance	Low Yield Variance	High Yield Variance
Prohibitively Expensive	21.1	49.8	19.4	33.8	5.1	10.1
Moderate	30.5	51.2	22.7	50.0	5.1	10.1
Inexpensive	4.4	14.5	1.9	13.9	0.7	9.2

Table III
Relative Cost for MRA (Percentage Excess over LDR), $b/h = 5$

	Inadequate Capacity		Matched Capacity		Surplus Capacity	
Overtime	Low Yield Variance	High Yield Variance	Low Yield Variance	High Yield Variance	Low Yield Variance	High Yield Variance
Prohibitively Expensive	19.8	35.8	16.8	30.8	4.1	9.1
Moderate	26.2	43.3	18.0	35.9	4.1	9.1
Inexpensive	3.3	11.9	1.8	11.8	0.5	3.3

Table IV
Relative Cost for MRA (Percentage Excess Over LDR), $b/h = 10$

	Inadequate Capacity		Matched Capacity		Surplus Capacity	
Overtime	Low Yield Variance	High Yield Variance	Low Yield Variance	High Yield Variance	Low Yield Variance	High Yield Variance
Prohibitively Expensive	19.1	27.3	14.1	23.2	3.4	8.0
Moderate	24.7	33.8	14.7	23.2	3.4	8.0
Inexpensive	2.4	11.0	1.7	10.3	0.4	1.4

of MRA compared to LDR varies from an insignificant difference of less than 1% to a substantial deviation of more than 50%. An understanding of the sensitivity of their relative performance to various operating parameters is in order.

Effect of Yield Variance. As the uncertainty in yield increases, both policies pay a higher backlogging cost due to uncertainty in stock at the intermediate buffer. But the MRA, which does not take into account yield uncertainty at stage 1 while making dispatch decisions at stage 2, pays a heavier toll than LDR which takes a coordinated dispatch decision for both the stages. As a result, the relative performance of MRA compared to LDR becomes worse as yield variance increases. This effect would be further accentuated if more than two test stages were involved.

Effect of Available Capacity. The LDR utilizes the scarce capacity much more efficiently than the MRA and this results in a significant difference in the total cost of the two policies, particularly when overtime is not too cheap. A close look at the period-to-period simulation results revealed that MRA occasionally left a portion of the capacity unused even when capacity was inadequate. This happens because the MRA allocates capacities sequentially at the two test stages which may result in an unproductive use of capacity. To illustrate this point, consider a part type which offers a relatively large expected marginal decrease in the total cost $b\bar{\alpha}/\tau$ at the second stage compared to other part types. As a result, it gets a significant amount of capacity at the second test stage. The demand for this part type, among others, is placed on the first test stage. Consider further that this particular part type has a very small $b\bar{\alpha}/\tau$ at the first test stage compared to the other part types. As a result, it does not get any capacity allocated at the first stage. The net result is starvation of the second stage due to an uncoordinated allocation of capacities which, in turn, leads to a high backlogging cost. The scarcer the capacity, the higher the price paid for the starvation. When surplus capacity is available, the difference between the two policies is minimal.

Effect of Overtime Cost. As the cost of overtime increases, the percentage excess total cost of MRA over LDR first increases and then gradually decreases. This can be explained as follows. An inexpensive overtime belittles the advantages of the coordinated capacity allocation accomplished by the LDR. As the overtime cost increases, the need to use less overtime and make coordinated capacity allocation decisions to reduce backlogging costs becomes important. As mentioned in the previous paragraph, MRA is less effective than LDR in this respect. We observed in a sequence of simulation experiments (not reported here) with gradually increasing overtime costs that initially MRA used overtime liberally. Subsequently, the use of overtime decreased and backlogging costs increased dramatically.

Unfortunately, as the per unit cost of overtime increases, the performance of LDR downgrades gradually due to decreasing quality of quadratic fit to the undertime-overtime curve. This is especially true here because we assumed a zero undertime, linear overtime cost resulting in a ramp-like curve difficult to fit accurately by quadratic forms particularly for high ramp angles.

When capacity is in excess, the quality of fit is not an inssue as overtime is never used. As a result, after an initial increase the percentage excess total cost of MRA over LDR saturates to a constant value. The mild difference in performance for the two policies in this case is mainly due to the way they tradeoff inventory and backlogging costs: the LDR uses a smooth production plan taking into account the mild variation in period-to-period demands and supplies, the MRA computes dispatch quantity based on a single period's requirement alone. As a result, the MRA runs into more frequent raw material unavailability at the intermediate buffer than the LDR, particularly when yield variance is high.

Effect of Backlogging-to-Holding Cost Ratio. LDR performance degrades gradually as the backlogging to holding cost ratio increases. This is mainly due to the decreasing quality of quadratic fit to the inventory holding-backlogging cost curve. As a matter of fact, had we not incorporated modification (6) to the quadratic cost functional, the LDR would have performed even worse with the increase in the cost ratio.

6. CONCLUSIONS

We have modeled a flexible test system for printed circuit boards as a multiproduct, multistage production system with random yield. Two policies, LDR and MRA, for part dispatch have been developed by considering two different sets of approximations. The LDR assumes that inventory and capacity related costs have (or can be approximated by) quadratic forms. It allows efficient computation of decision rules in very general problem settings—costs are allowed to be nonstationary, yields can be time-varying and correlated among part types or between the stages, demand and raw material supply can be nonstationary and stochastic, and available capacity may vary with time. It needs only the first one or two moments of uncertain quantities, which are easier to obtain economically compared to complete distributional information required by MRA. The myopic resource allocation, on the other hand, makes dispatch decisions in a myopic capacitated newsboy fashion. Due to its myopic nature, it is not suited for nonstationary situations. However, the policy lends itself to intuitive economic interpretation and can be computed efficiently using a recently developed resource allocation algorithm.

We also performed extensive computational studies to assess the performance of the two policies under various operating conditions. The experiments, based on a linear holding, backlogging and overtime costs, clearly demonstrate overall superiority of LDR. When capacity is scarce and overtime not too cheap, the average total cost for the MRA compared to LDR is quite high. If capacity is in excess, both decision rules give similar performance, except when yield variance is high in which case LDR is again better. Availability of inexpensive overtime reduces the gap between them. We point out that LDR is expected to perform even better under a more general setting because it explicitly takes into account the nonstationarity and correlation of various parameters not considered in the computational experiments.

APPENDIX

Proof of Theorem 1

Rewriting the objective function (9) in the dynamic programming recursive form, we get

$$J_t(\mathbf{x}_t) = \min_{\mathbf{u}_t} E_{\mathbf{B}_t,\mathbf{w}_t} \{\mathbf{x}_t'\mathbf{Q}_t\mathbf{x}_t - \mathbf{C}_t\mathbf{x}_t + (\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t)'\mathbf{R}_t(\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t) + \mathbf{F}_t(\mathbf{T}\mathbf{u}_t - \boldsymbol{\beta}_t) + J_{t+1}(\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{G}\mathbf{w}_t)\} \quad (A.1)$$

$$J_N(\mathbf{x}_N) = \mathbf{x}_N'\mathbf{Q}_N\mathbf{x}_N - \mathbf{C}_N\mathbf{x}_N.$$

Let us rewrite the last equations as

$$J_N(\mathbf{x}_N) = \mathbf{x}_N'\mathbf{K}_N\mathbf{x}_N + 2\mathbf{P}_N\mathbf{x}_N \quad (A.2)$$

by setting $\mathbf{Q}_N = \mathbf{K}_N$ and $\mathbf{C}_N = -2\mathbf{P}_N$ without loss of generality. Starting from the last period, we can compute the optimal policy recursively. For $t = N - 1$, the cost-to-go $J_{N-1}(\mathbf{x}_{N-1})$ is obtained by first substituting (A.2) in (A.1) and then using the inventory balance equation (1) to represent $\mathbf{x}_N$ as a function of $\mathbf{x}_{N-1}$ and decision $\mathbf{u}_{N-1}$. We obtain

$$\begin{aligned} J_{N-1}(\mathbf{x}_{N-1}) &= \min_{\mathbf{u}_{N-1}} E_{\mathbf{B}_{N-1},\mathbf{w}_{N-1}} \{\mathbf{x}_{N-1}'\mathbf{Q}_{N-1}\mathbf{x}_{N-1} - \mathbf{C}_{N-1}\mathbf{x}_{N-1} \\ &\quad + (\mathbf{T}\mathbf{u}_{N-1} - \boldsymbol{\beta}_{N-1})'\mathbf{R}_{N-1} \cdot (\mathbf{T}\mathbf{u}_{N-1} - \boldsymbol{\beta}_{N-1}) + \mathbf{F}_{N-1} \cdot (\mathbf{T}\mathbf{u}_{N-1} - \boldsymbol{\beta}_{N-1}) + (\mathbf{x}_{N-1} \\ &\quad + \mathbf{B}_{N-1}\mathbf{u}_{N-1} + \mathbf{G}\mathbf{w}_{N-1})' \cdot \mathbf{K}_N(\mathbf{x}_{N-1} + \mathbf{B}_{N-1}\mathbf{u}_{N-1} \\ &\quad + \mathbf{G}\mathbf{w}_{N-1}) + 2\mathbf{P}_N(\mathbf{x}_{N-1} \\ &\quad + \mathbf{B}_{N-1}\mathbf{u}_{N-1} + \mathbf{G}\mathbf{w}_{N-1})\}. \end{aligned} \quad (A.3)$$

Substituting $\boldsymbol{\Gamma}_{N-1}$ for $E\{\mathbf{B}_{N-1}'\mathbf{K}_N\mathbf{B}_{N-1}\} + \mathbf{T}'\mathbf{R}_{N-1}\mathbf{T}$ and collecting terms that involve $\mathbf{u}_{N-1}$, $\mathbf{x}_{N-1}$, and those without them, the above equation can be rewritten as

$$\begin{aligned} J_{N-1}(\mathbf{x}_{N-1}) &= \mathbf{x}_{N-1}'(\mathbf{Q}_{N-1} + \mathbf{K}_N)\mathbf{x}_{N-1} \\ &\quad + (2\mathbf{P}_N + \bar{\mathbf{w}}_{N-1}'\mathbf{G}'\mathbf{K}_N - \mathbf{C}_{N-1})\mathbf{x}_{N-1} \\ &\quad + \min_{\mathbf{u}_{N-1}}\{\mathbf{u}_{N-1}'\boldsymbol{\Gamma}_{N-1}\mathbf{u}_{N-1} + 2\mathbf{u}_{N-1}'\bar{\mathbf{B}}_{N-1}'\mathbf{K}_N\mathbf{x}_{N-1} \\ &\qquad - 2\mathbf{u}_{N-1}'(\mathbf{T}'\mathbf{R}_{N-1}\boldsymbol{\beta}_{N-1} - \bar{\mathbf{B}}_{N-1}'\mathbf{K}_N\mathbf{G}\bar{\mathbf{w}}_{N-1} \\ &\qquad - \bar{\mathbf{B}}_{N-1}'\mathbf{P}_N' - \tfrac{1}{2}\mathbf{T}'\mathbf{F}_{N-1}')\} \\ &\quad + \text{terms free of } \mathbf{x}_{N-1} \text{ and } \mathbf{u}_{N-1}. \end{aligned} \quad (A.4)$$

By differentiating with respect to $\mathbf{u}_{N-1}$ and setting the derivative equal to zero, we get

$$\boldsymbol{\Gamma}_{N-1}\mathbf{u}^*_{N-1} = -\bar{\mathbf{B}}'_{N-1}\mathbf{K}_N\mathbf{x}_{N-1} + (\mathbf{T}'\mathbf{R}_{N-1}\boldsymbol{\beta}_{N-1} - \bar{\mathbf{B}}'_{N-1}\mathbf{K}_N\mathbf{G}\bar{\mathbf{w}}_{N-1} - \bar{\mathbf{B}}'_{N-1}\mathbf{P}'_N - \tfrac{1}{2}\mathbf{T}'\mathbf{F}'_{N-1}),$$

which yields the optimal production vector for period $N-1$,

$$\mathbf{u}^*_{N-1} = \mathbf{L}_{N-1} + \mathbf{M}_{N-1},$$

where, $\mathbf{L}_{N-1} = -\boldsymbol{\Gamma}^{-1}_{N-1}\bar{\mathbf{B}}'_{N-1}\mathbf{K}_N$, and

$$\mathbf{M}_{N-1} = \boldsymbol{\Gamma}^{-1}_{N-1}(\mathbf{T}'\mathbf{R}_{N-1}\boldsymbol{\beta}_{N-1} - \bar{\mathbf{B}}'_{N-1}\mathbf{K}_N\mathbf{G}\bar{\mathbf{w}}_{N-1} - \bar{\mathbf{B}}'_{N-1}\mathbf{P}'_N - \tfrac{1}{2}\mathbf{T}'\mathbf{F}'_{N-1}).$$

Similarly, to obtain the optimal production decision for period $N-2$, we first compute $J_{N-1}(\mathbf{x}_{N-1})$. By substituting $\mathbf{u}^*_{N-}$ in (A.4) we get

$$J_{N-1}(\mathbf{x}_{N-1}) = \mathbf{x}'_{N-1}\mathbf{K}_{N-1}\mathbf{x}_{N-1} + 2\mathbf{P}_{N-1}\mathbf{x}_{N-1} + \text{terms free of } \mathbf{x}_{N-1}, \tag{A.5}$$

where the matrices $\mathbf{K}_{N-1}$ and $\mathbf{P}_{N-1}$ are obtained by straightforward algebra and are given by

$$\mathbf{K}_{N-1} = \mathbf{K}_N - \mathbf{K}_N\bar{\mathbf{B}}_{N-1}\boldsymbol{\Gamma}^{-1}_{N-1}\bar{\mathbf{B}}'_{N-1}\mathbf{K}_N + \mathbf{Q}_{N-1}$$

$$\mathbf{P}_{N-1} = \mathbf{P}_N + [\mathbf{G}\bar{\mathbf{w}}_{N-1} + \bar{\mathbf{B}}_{N-1}\mathbf{M}_{N-1}]'\mathbf{K}_N - \tfrac{1}{2}\mathbf{C}_{N-1}.$$

The constant term in (A.5) will not affect the computation of $\mathbf{u}^*_{N-2}$ and can be dropped. Note the similarity between (A.2) and (A.5), which is of significance because the cost-to-go, $J_{N-2}(\mathbf{x}_{N-2})$, is obtained by using (A.1) for $t = N-2$ and substituting $J_{N-1}(\mathbf{x}_{N-1})$ from (A.5). This results in an expression identical to (A.3) except all subscripts are shifted by one period. Repeating the steps followed above gives

$$\mathbf{u}^*_{N-2} = \mathbf{L}_{N-2}\mathbf{x}_{N-2} + \mathbf{M}_{N-2}, \text{ and}$$

$$J_{N-2}(\mathbf{x}_{N-2}) = \mathbf{x}'_{N-2}\mathbf{K}_{N-2}\mathbf{x}_{N-2} + 2\mathbf{P}_{N-2}\mathbf{x}_{N-2} + \text{terms free of } \mathbf{x}_{N-2},$$

where matrices $\mathbf{L}_{N-2}$, $\mathbf{N}_{N-2}$, $\mathbf{K}_{N-2}$, and $\mathbf{P}_{N-2}$ are identical to those for $N-1$. The same argument can be repeated for $t = N-3, N-2, \ldots, 1$ to obtain the recursive equations for each period.

Fitting A Quadratic Function

Here we describe how to fit a quadratic form to a nonquadratic cost function, such that the weighted least squares deviation from the actual curve is minimized. Suppose that the inventory costs for part type i is given by a function, $g_i(x_i)$. For example, if the actual costs were linear, then $g_i(x_i) = h_i\max(x_i, 0) - b_i\max(-x_i, 0)$. Suppose that x_i^k, $k = 1, 2, \ldots I$ are the inventory realizations generated by repeated simulation runs using the best available policy. We would like to achieve a better fit in the region where x_i^k's fall more frequently. The objective of the weighted least square fit can be given by

$$\underset{Q_i,C_i}{\text{minimize}} \sum_{k=1}^{I} \{(Q_i(x_i^k)^2 - C_i x_i^k) - g_i(x_i^k)\}^2.$$

Taking partial derivatives with respect to Q_i and C_i, and setting them to zero gives the following set of equations

$$\begin{bmatrix} \sum_{k=1}^{I}(x_i^k)^4 & -\sum_{k=1}^{I}(x_i^k)^3 \\ \sum_{k=1}^{I}(x_i^k)^3 & -\sum_{k=1}^{I}(x_i^k)^2 \end{bmatrix} \begin{bmatrix} Q_i \\ C_i \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{I}(x_i^k)^2 g_i(x_i^k) \\ \sum_{k=1}^{I} x_i^k g_i(x_i^k) \end{bmatrix}$$

which can be solved to obtain the best fit.

Proof That Algorithm 1 Finds the Optimal Solution

The Kuhn-Tucker conditions for the problem stipulate that there exists a number $\lambda^* \geq 0$ such that $(\mathbf{u}^*, y^*, \lambda^*)$ satisfies

$$-\frac{1}{\tau_i}\frac{dL_i}{du_i} \leq \lambda \quad \text{for all } i \tag{A.6}$$

$$u_i > 0 \text{ implies } -\frac{1}{\tau_i}\frac{dL_i}{du_i} = \lambda \text{ or, equivalently,}$$

$$-\frac{1}{\tau_i}\frac{dL_i}{du_i} < \lambda \text{ implies } u_i = 0. \tag{A.7}$$

$$r \geq \lambda \tag{A.8}$$

$$y > 0 \text{ implies } r = \lambda \text{ or, equivalently,}$$

$$r > \lambda \text{ implies } y = 0 \tag{A.9}$$

$$\lambda\left(\sum_{i=1}^{M} \tau_i u_i - y - \beta\right) = 0. \tag{A.10}$$

Properties (A.6) and (A.7) together with the fact that $-(dL_i/du_i)$ are decreasing functions of u_i imply that

$$u_i^* > 0 \text{ iff } -\frac{1}{\tau_i}\frac{dL_i}{du_i}\bigg|_{u_i=0} > \lambda^*.$$

If the variables have been arranged such that

$$-\frac{1}{\tau_i}\frac{dL_i}{du_i}\bigg|_{u_i=0} \geq -\frac{1}{\tau_{i+1}}\frac{dL_{i+1}}{du_{i+1}}\bigg|_{u_{i+1}=0}$$

then the optimal solution, $\mathbf{u}^*$, is such that

$$\begin{cases} u_i^* > 0; & -\dfrac{1}{\tau_i}\dfrac{dL_i}{du_i}\bigg|_{u_i^*} = \lambda \quad i = 1, 2, \ldots, k^* \\ u_i^* = 0; & -\dfrac{1}{\tau_i}\dfrac{dL_i}{du_i}\bigg|_{u_i^*=0} < \lambda \quad i = k^*+1, \ldots, M. \end{cases} \tag{A.11}$$

If $\sum_{i=1}^{M} \tau_i u_i^* < \beta + y$, then $\lambda^* = 0$ from (A.10) which, in turn, implies that $y^* = 0$ (from A.9). Hence, the optimal solution is such that either: i) $\sum_{i=1}^{M} \tau_i u_i^* < \beta$, or ii) $\sum_{i=1}^{M} \tau_i u_i^* = \beta + y^*$. Case i can be tested easily by solving the problem without the resource constraint and then verifying that the unconstrained optimal production rates do not violate the regular time capacity (Step 1 of the algorithm). Case ii can be further subdivided into iia $y^* > 0$ which implies $\lambda^* = r$, and iib $y^* = 0$ in which case, $\lambda^* < 0$ and $\sum_{i=1}^{M} \tau_i u_i^* = \beta$. To check if Case iia holds, algorithm 1 uses property (A.11) to find the production quantities and then check if $y^* = \sum_{i=1}^{M} \tau_i u_i^* - \beta$ is greater than zero (Step 2). If this is not the case, then, by enumeration, Case iib must hold, which is exactly the problem dealt with in Zipkin. Step 3 is identical to that of Zipkin and proof of optimality for this step can be found there.

ACKNOWLEDGMENT

We are greatly indebted to the referees for many helpful comments which led to significant enhancement of the paper. The myopic resource allocation policy is based on the incremental greedy algorithm suggested by one of the referees. We also acknowledge the support for this research by IBM. In particular, we thank Chacko Abraham, Billy Crowder and Ranga Jayaraman of the Manufacturing Research group for having been extremely supportive over an extended period of time. We also thank Mike Nagem of IBM Charlotte, and Jack Cundari of Tandem Computers, Inc., and their groups, for their support and collaboration. This paper has benefited from the discussions with Steve Graves, Harry Groenevelt, Robin Roundy and Jie Sun.

REFERENCES

AKELLA, R., Y. CHOONG AND S. B. GERSHWIN. 1984. Performance of a Hierarchical Production Scheduling Policy. *IEEE Trans. Comput. Hybrids Manuf. Tech.* **CHMT-7,** 225–240.

AKELLA, R., AND P. R. KUMAR. 1986. Optimal Control of Production Rate in a Failure Prone Manufacturing System. *IEEE Trans. Autom. Control* **AC-31,** 116–126.

AKELLA, R., M. R. SINGH AND B. KROGH. 1990. Efficient Computation of Coordinating Controls in Hierarchical Structures for Failure-Prone Multi-Cell Flexible Assembly Systems. *IEEE Trans. Robotics Autom.* **6,** 659–672.

BERTSEKAS, D. 1976. *Dynamic Programming and Stochastic Control.* Academic Press, New York.

BITRAN, G. R., AND D. TIRUPATI, 1988. Mutiproduct Queueing Networks With Deterministic Routing. *Mgmt. Sci.* **34,** 75–100.

GERCHAK, Y., R. G. VICKSON AND M. PARLAR. 1988. Periodic Review Production Models With Variable Yields and Uncertain Demands. *IIE Trans.* **20,** 144–150.

GERSHWIN, S. B., R. AKELLA AND Y. CHOONG. 1985. Short-Term Production Scheduling of an Automated Manufacturing Facility. *IBM J. R&D* **29,** 392–400.

GRAVES, S. C. 1982. Using Lagrangean Techniques to Solve Hierarchical Production Planning Problems. *Mgmt. Sci.* **28,** 260–275.

HAX, A. C., AND C. CANDEA. 1984. *Production and Inventory Management.* Prentice-Hall, New York.

HOLT, C., F. MODIGLIANI, J. F. MUTH AND H. A. SIMON. 1960. *Planning, Production, Inventories, and Work Force.* Prentice-Hall, Englewood Cliffs, N.J.

KARMARKAR, U. S., AND S. C. LIN. 1986. Production Planning With Uncertain Yields and Demands. Working Paper Series No. QM 86-32, William E. Simon Graduate School of Business, University of Rochester, Rochester, N.Y.

LEE, H. L., AND C. A. YANO. 1988. Production Control in Multistage Systems With Variable Yield Losses. *Opns. Res.* **36,** 269–278.

LUSS, H., AND S. GUPTA. 1975. Allocation of Effort Resources Among Competing Activities. *Opns. Res.* **23,** 360–366.

SCHNEEWEISS, C. A. 1971. Smoothing Production by Inventory—An Application of the Wiener Filtering Theory. *Mgmt. Sci.* **17,** 472–483.

SCHNEEWEISS, C. A. 1974. Optimal Production Smoothing and Safety Inventory. *Mgmt. Sci.* **20,** 1122–1130.

YANO, C. A. 1986. Optimal Finite and Infinite Horizon Policies for a Single Stage Production System With Variable Yields. Technical Report 86-32, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Mich.

YANO, C. A., AND H. L. LEE. 1989. Lot-Sizing With Random Yields: A Review. Technical Report 89-16, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Mich.

ZIPKIN, P. H. 1980. Simple Ranking Methods for Allocation of One Resource. *Mgmt. Sci.* **26,** 34–43.