Managing simple re-entrant flow lines Theoretical foundation and Experimental Results

Carlos F. Bispo Sridhar Tayur
GSIA, Carnegie Mellon University, Pittsburgh, PA 15213.

June 1997; Revised April 1998; February 2000; July 2000

Abstract

We address several aspects related to managing re-entrant lines in a unified manner – capacity allocation, inventory management and production control. Our approach to study these systems is through *simulation based optimization*. Simulation offers the flexibility to model the complexities adequately while the gradient computation (via *Infinitesimal Perturbation Analysis*) helps identify good solutions quickly. Our framework is a discrete time capacitated multiple product production-inventory system operating under a *base stock policy*. We analyze several different production and capacity allocation rules. We develop expressions for and validate the appropriate IPA derivatives. These derivatives can then be used in an optimization tool which enables the determination of optimal parameters for the several policies proposed. We also present a summary of insights obtained through an extensive computational study.

1 Introduction

A flow line is a manufacturing system where several products flow through the same sequence of operations. *Re-entrant flow lines* are flow lines where the same sequence is traversed several times (*levels*) to complete the products. In semiconductor manufacturing, wafers traverse flow lines several times to produce the different layers composing each circuit. The need to understand and manage the complexities of semiconductor manufacturing has motivated a large body of research on re-entrant lines. Our work presents a unified treatment of several managerial issues for a family of re-entrant flow lines, including capacity allocation, inventory management, and production control. Our framework is a discrete time, capacitated, multi-period, multi-product system operating under a *capacitated multi-echelon base stock policy*. Due to the intractability of analytic models, our approach is supported on simulation based optimization, which offers modeling flexibility and the ability to quickly obtain good solutions based on gradient estimates derived through *Infinitesimal Perturbation Analysis* (IPA). Our goal is to derive insights and establish a framework to study more complex re-entrant systems. The special case without re-entrance is applicable very widely in discrete-part manufacturing (in furniture industry and at Tier-1 automotive supplier plants for example) and thus our framework has broader applicability than the semi-conductor industry.

It is well known that base stock policies are optimal in a variety of settings for single machine and single product systems, and continue to be optimal for multiple machines in series and a single product in an uncapacitated setting; see [Tayur et al., 1999]. When there are capacity bounds, the optimal policy structure in a multiple stage setting is not easily characterized and can be complex: see [Speck and der Wall, 1991]. Like much of the past research, therefore, we concentrate on understanding implementable classes of policies – these are base stock policies

(for discrete time models like ours) or workload regulating and starvation avoidance (in continuous time models) – via experimental studies based on approximations and simulation. We validate a methodology to compute the optimal parameters within base stock policies. Our experimental study considers capacity allocation, production rules, and inventory levels simultaneously, and we provide an understanding into the performance of capacitated multi-product re-entrant flow lines.

An exhaustive review of related production planning and scheduling models can be found in [Uzsoy et al., 1992, Uzsoy et al., 1994]. These mainly use continuous time models with rules similar to a base stock policy, but do not consider the several inter-related issues simultaneously. We mention only the papers that are closely related to our work. Closed queuing models that address the control of the input of new material into the system are studied in [Wein, 1990, Harrison and Wein, 1990, Wein, 1988, Glassey and Resende, 1988]. Their input policies are conceptually of the base stock type for the amount of work in the system. In contrast, [Perkins and Kumar, 1989, Kumar and Seidman, 1990, Lu and Kumar, 1991, Lu et al., 1994] use open queuing models assuming no control over the input process, to address the scheduling problem of each server. Open queuing systems are implicitly operated under a base stock policy, since each server "wants to see" zero costumers in front of it. A third stream of research that includes [Kimemia and Gershwin, 1983, Akella et al., 1984, Akella and Kumar, 1986, Lou and Kager, 1989, van Ryzin et al., 1993, Bai and Gershwin, 1996] applies a hierarchical control framework, in the context of flow control. Their optimal policies are shown to be of the *hedging point* type for end-product surplus, which is also a version of base stock policy.

The rest of the paper is organized as follows. We introduce the basic recursion equations for the state variables, the production decisions, and performance measures in Section 2. Section 3 develops the recursion equations of their derivatives. In Section 4 we validate the IPA procedure; [Ho and Cao, 1991] is a basic reference and we use [Glasserman and Tayur, 1995] heavily. In Section 5 we present experimental data illustrating the main insights. We conclude in Section 6.

2 Model

Consider a production system composed of M machines, able to process P different products, each needing to cycle K times through all the machines. Moreover, we assume that each product unit at any given point imposes the same load on the visited machines (*uniform load*) and that the yield is perfect. Products which are visiting a given machine for the k-th time, $k = 1, 2, \ldots, K$, are said to be in *level* $(K - (k - 1))$, and each machine we term a *stage*. The

first machine to be encountered by any product is stage M and the last is stage 1. We denote the inventory of product p on stage m and level k at the beginning of period n as I_n^{kmp}. With our numbering, the end product inventory for any product at the beginning of period n is I_n^{11p}, $p = 1, 2, \ldots, P$, irrespective of the values of K and M.

In any period, each machine can process different parts belonging to different levels; the total production is limited by the machine's capacity. After being processed by a machine, parts are placed in intermediate buffers, where they wait their turn to be processed by next machine or until they are depleted by external demand once they complete all KM operations. We assume these buffers to have infinite capacity. Demand is satisfied out of the end product inventory or backlogged for future replenishment whenever demand exceeds the available inventory. Demands are assumed continuous (with density ϕ^p), independent across products, and i.i.d. for each product along time, possibly with a point mass at zero. Although semiconductor manufacturing in general, and wafer fabrication in particular possesses many other sources of uncertainty we have chosen to present a model stripped of those, given that managing reentrant systems with deterministic capacity, deterministic processing times, and perfect yield in a multiple product setting with random demand is not well understood yet.

In each period, the sequence of events is as follows. Demands occur, and then production decisions will be made upon observing the inventories. Cost is incurred as a function of the amount of inventory at the end of each period. We restrict our study to the following class of inventory control, capacity allocation, and production rules.

Inventory Control – Every level and stage operates on a capacitated base stock policy for echelon inventory. This means that given a particular product, the decision maker adds all inventory downstream from that level and stage to determine the echelon inventory. If the echelon inventory falls below the corresponding base stock value, the decision will be to produce the difference, provided there is enough capacity and (relevant) upstream inventory.

How much capacity is available depends on all products competing for capacity at any given machine and on the capacity allocation policy. We propose and analyze several production and capacity allocation rules.

Capacity Allocation – Each machine m, with $m = 1, \ldots, M$, has a fixed total capacity C^m. We can take this total capacity and define different degrees of capacity sharing. If we divide the capacity of each machine into $K \times P$ slots and assign each slot to a single product and level, we have what we call the *no sharing* mode (NS). If we define K slots and allow each of these to be shared by all products belonging to the same level, we have the *partial sharing* mode (PS). The

total sharing mode, (TS), is the extreme case where the capacity of each machine is accessible to any product and level. Note that the NS mode corresponds to a situation where we would be managing P different production systems, each with a single product and non re-entrant flow. We consider it here to compare with the other two sharing modes.

Production Rules – Whenever there is some degree of capacity sharing, it is necessary to establish a capacity management scheme, which defines how capacity is distributed among all products when there is a bound in capacity. To take care of this dynamic decision we propose three production rules: Linear Scaling (LSR), Priority (PR), and Equalize Shortfall (ESR).

The basic recursions governing a re-entrant flow line are discussed next.

2.1 Basic Recursions

The inventory dynamic equations are given by

$$I_{n+1}^{kmp} = I_n^{kmp} - P_n^{(km)^- p} + P_n^{kmp}, \tag{1}$$

where I_n^{kmp} is the inventory of product p, on level k and stage m at the beginning of period n; P_n^{kmp} is the production decision for period n; and $(km)^-$ designates the level and stage which produces out of inventory on level k and stage m. The external demand for product p during period n, d_n^p, is represented by $P_n^{(11)^- p}$.

Defining $E_n^{(11)^- p} = 0$, the echelon inventories at the beginning of period n are given by

$$E_n^{kmp} = I_n^{kmp} + E_n^{(km)^- p}. \tag{2}$$

Except for I_n^{11p}, every inventory variable is always non negative, given that the production decisions are always bounded by the available inventory. However, any echelon inventory variable may take negative values, as long as the value of I_n^{11p} is sufficiently negative. The shortfall is defined as

$$Y_n^{kmp} = z^{kmp} - E_n^{kmp}, \tag{3}$$

where z^{kmp} is the echelon base stock for product p, on level k, and stage m. The shortfall measures the distance to the target echelon inventory and is always non negative. The dynamic equations for the shortfall are

$$Y_{n+1}^{kmp} = Y_n^{kmp} + d_n^p - P_n^{kmp}. \tag{4}$$

The production decision is influenced by the production rule being used, which becomes active when the available capacity is exceeded. We define the amounts that would be produced should there be no capacity bounds as the *production net needs* for each product, level, and stage as:

$$f_n^{kmp} = \min\left\{z^{kmp} + d_n^p - E_n^{kmp}, I_n^{(km)^+p}\right\}, \tag{5}$$

where $(km)^+$ denotes the level and stage which provides inventory to production at level k and stage m. We assume infinite raw material, so $I_n^{(KM)^+p} = \infty$ for all n. Note that $z^{kmp}+d_n^p-E_n^{kmp}$ is always positive, since it is the shortfall plus one demand occurrence. We assume the system to start with inventories at their base stock values: $I_0^{kmp} = z^{kmp} - z^{(km)^-p}$ and the echelon inventories set according to (2). All other initial variables are set to zero.

For convenience, we define the control policy by an alternative set of control variables that are directly related to the multi-echelon base stock variables:

$$\Delta^{kmp} = z^{kmp} - z^{(km)^-p}, \tag{6}$$

with $z^{(11)^-p} = 0$. We refer to these as the Δ (or 'delta') variables.

2.2 The Production Decisions

2.2.1 Linear Scaling Rule

For the Linear Scaling Rule (LSR) with partial sharing of capacity (PS mode), the production decision for period n is defined as

$$P_n^{kmp} = f_n^{kmp} g_n^{km}, \tag{7}$$

with f_n^{kmp} given by (5) and g_n^{km}, expressing the way capacity is distributed, is given by

$$g_n^{km} \quad = \quad \min\left\{\frac{C^{km}}{\sum_p f_n^{kmp}}, 1\right\}, \tag{8}$$

where C^{km} is the amount of C^m assigned to level k. For the LSR in the TS mode, g_n^{km} is replaced in (7) by g_n^m, where

$$g_n^m \quad = \quad \min\left\{\frac{C^m}{\sum_p \sum_k f_n^{kmp}}, 1\right\}.$$

2.2.2 Priority Rule

Consider the PS mode. We assign capacity according to a static priority for the products. Let $p(i)$, for $i = 1, \ldots, P$, be the product that comes in the ith position on the priority list; that is, product $p(1)$ is the product with the highest priority and product $p(P)$ has the lowest priority. The production is

$$P_n^{kmp(i)} = \min\{f_n^{kmp(i)}, \max\{0, C^{km} - \sum_{j=1}^{i-1} P_n^{kmp(j)}\}\}. \tag{9}$$

In the TS mode, we assign capacity according to a priority for the products and levels. Let $k(i)$ and $p(i)$, for $i = 1, \ldots, K \times P$ be the level and product with the ith position on the priority list. The production decision is similar in structure to (9).

2.2.3 Equalize Shortfall Rule

Another way to dynamically allocate capacity is to *equalize the shortfall* for every product. The shortfall has been defined in (3). In contrast to the previous two production rules, the production decision is obtained iteratively. We first allocate capacity to the product with the highest shortfall, until it reaches the level of the product with the second highest shortfall. At this point, capacity will be assigned in equal parts to both products until their individual shortfalls equal the third highest shortfall, at which point these three products share capacity equally, and so forth. (Note that at any point it may be the case that the equalization cannot be accomplished because of insufficient feeding inventory or because capacity is exhausted.) See Appendix A for the procedure in the PS mode.

At the end of this procedure, P_n^{kmp} is the production decision and P'^{kmp}_n is its derivative with respect to the parameters being optimized. To handle the TS mode, only minor changes of this procedure are necessary; we skip the details.

2.3 Performance Measures

We concentrate on infinite horizon average cost and Type-1 service level. The average cost is calculated through the assignment of cost to inventories and backlogs in each period and averaging across periods. The service level relates to the fraction of times where a stock-out does not occur. Let h^{kmp} be the holding cost rate for level k, stage m, and product p. Let b^p be the back-logging cost rate for level 1, stage 1, and product p. The single period cost is defined as $J_n = \sum_{p=1}^{P} J_n^p$, with J_n^p given by

$$J_n^p = (I_n^{11p})^- b^p + (I_n^{11p})^+ h^{11p} + \sum_{m=2}^{M} I_n^{1mp} h^{1mp} + \sum_{k=2}^{K} \sum_{m=1}^{M} I_n^{kmp} h^{kmp}, \tag{10}$$

where $(X)^- = \max\{0, -X\}$ and $(X)^+ = \max\{0, X\}$. We consider the infinite horizon average cost

$$J_\infty = \lim_{N\to\infty} \mathbf{E}\left[\frac{1}{N}\sum_{n=1}^{N} J_n\right]. \tag{11}$$

As to the service level, let

$$V_N^p = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{I_n^{11p} \geq d_n^p \text{ or } d_n^p = 0\} \tag{12}$$

be the fraction of periods in which demand for product p is filled immediately. Let $\bar{V}_N = \frac{1}{P}\sum_{p=1}^{P} V_N^p$ and $\bar{v}_N = \mathbf{E}[\bar{V}_N]$. We study the infinite horizon version $\bar{v}_\infty = \lim_{N\to\infty} \bar{v}_N$.

3 Derivative Recursions

Here we only detail the process of taking derivatives with respect to the echelon base stock variables. We assume that for each p, $0 < z^{kmp} < z^{(km)^+ p}$ for all k, m. Let $z^* = z^{k^* m^* p^*}$ denote the variable with respect to which the derivatives are taken for some $k^* = 1, \ldots, K; m^* = 1, \ldots, M; p^* = 1, \ldots, P$.

3.1 Derivatives of the State Variables

For the state variables defined in Section 2.1 we start by differentiating the dynamic equation for the inventories, yielding

$$I'^{kmp}_{n+1} = I'^{kmp}_{n} - P'^{(km)^{-}p}_{n} + P'^{kmp}_{n}, \tag{13}$$

where $P'^{(11)^{-}p}_{n} = d'^{p}_{n} = 0$, since demands are independent of the control policy. The derivatives for equations (2) and (4) follow a similar principle. Expressions (3) and (5) deserve a particular attention because of their explicit dependence on the echelon base stock variables. We detail the specifics of the latter, since it possesses a more interesting structure, due to the existence of the $\min\{.\}$ operator. The derivative of the production net needs is given by

$$f'^{kmp}_{n} = \begin{cases} \mathbf{1}\{z^* = z^{kmp}\} - E'^{kmp}_{n} & \text{if } z^{kmp} + d^{p}_{n} - E^{kmp}_{n} < I^{(km)^{+}p}_{n} \\ I'^{(km)^{+}p}_{n} & \text{if } z^{kmp} + d^{p}_{n} - E^{kmp}_{n} > I^{(km)^{+}p}_{n} \\ 0 & \text{if } f^{kmp}_{n} = 0 \end{cases} \tag{14}$$

Note that $I'^{(KM)^{+}p}_{n} = 0$ because raw material is assumed to be infinite. The derivative for the initial inventories is given by $I'^{kmp}_{0} = \mathbf{1}\{z^* = z^{kmp}\} - \mathbf{1}\{z^* = z^{(km)^{-}p}\}$; note that $z^{(11)^{-}p} = 0$ is not a variable of the problem. For the initial echelon variables, the derivatives are trivially given by $E'^{kmp}_{0} = \mathbf{1}\{z^* = z^{kmp}\}$. The derivatives for the initial shortfall variables are set to zero.

3.2 Derivatives for the Production Decisions

We now consider the derivatives of the production decisions for each of the production rules proposed in Section 2.2. We detail only the PS mode for the LSR and the PR. The TS mode is a simple extension and the derivatives under ESR are detailed in Appendix A.

The derivative under LSR in the PS mode is

$$P'^{kmp}_{n} = f'^{kmp}_{n} g^{km}_{n} + f^{kmp}_{n} g'^{km}_{n}. \tag{15}$$

The derivative of g^{km}_{n} is

$$g'^{km}_{n} = \begin{cases} \frac{-C^{km} \sum_p f'^{kmp}_{n}}{(\sum_p f^{kmp}_{n})^2} & \text{if } \sum_p f^{kmp}_{n} > C^{km} \\ 0 & \text{if } \sum_p f^{kmp}_{n} < C^{km} \end{cases} \tag{16}$$

The derivatives obtained under PR in the PS mode are given by

$$P'^{kmp(i)}_{n} = \begin{cases} f'^{kmp(i)}_{n} & \text{if } 0 < f^{kmp}_{n} < C^{km} - \sum_{j=1}^{i-1} P^{kmp(j)}_{n} \\ -\sum_{j=1}^{k-1} P'^{kmp(j)}_{n} & \text{if } 0 < C^{km} - \sum_{j=1}^{i-1} P^{kmp(j)}_{n} < f^{kmp}_{n} \\ 0 & \text{if } C^{km} - \sum_{j=1}^{i-1} P^{kmp(j)}_{n} < 0 \end{cases} \tag{17}$$

3.3 Derivatives of the Performance Measures

To determine the derivatives of our infinite horizon performance measures, we need the derivatives of their single period and finite horizon counterparts. The cost derivative in period n is $J'_n = \sum_p J'^p_n$, where

$$\begin{aligned} J'^p_n = & -\mathbf{1}\{I^{11p}_n < 0\}(I')^{11p}_n b^p + \mathbf{1}\{I^{11p}_n > 0\}(I')^{11p}_n h^{11p} + \\ & + \sum_{m=2}^{M} I'^{1mp}_n h^{1mp} + \sum_{k=2}^{K}\sum_{m=1}^{M} I'^{kmp}_n h^{kmp}, \end{aligned} \tag{18}$$

Since $\bar{V}_N$ is not continuous, to obtain a differentiable representation, we replace the indicator function in (12) with a conditional expectation. This is explained in [Glasserman and Tayur, 1995] for the single product case. It is trivial to extend their method to the multiple product case in order to show that

$$P^{-1}\sum_{p=1}^{P} N^{-1}\sum_{n=1}^{N} \mathbf{1}\{I^{11p}_n > 0\}\phi^p_n(I^{11p}_n)(I'^{11p}_n) \tag{19}$$

is the derivative of $\bar{V}_N$ (see [Bispo, 1997] for details).

4 Validation of IPA Derivatives

We show that the system variables are smooth functions of the control parameters. This is a relatively easy task if it were not for the existence of min{.} and max{.} functions in some of the expressions. The derivative of such functions is well defined away from the points where ties occur. We show that either (1) ties occur with zero probability or (2) the derivative is the same for all the arguments. We only detail the validation under LSR in the PS mode for brevity.

A function ϕ mapping $S \in \mathbf{R}$ into $\mathbf{R}$ is *Lipschitz* if there exists a constant k_ϕ, called the modulus, for which $|\phi(x) - \phi(y)| \leq k_\phi |x - y|$. A random function is *Lipschitz with probability one* if there exists a random variable K that serves as a path-wise modulus. Our validation in the finite horizon setting will be based on Lemma 3.2 of [Glasserman and Tayur, 1995]:

Lemma 4.1 Let $\{X(s), s \in S\}$ be a random function with S an open subset of $\mathbf{R}$. Suppose that $\mathbf{E}[X(s)] < \infty$ for all $s \in S$. Suppose, further, that X is differentiable at $s_0 \in S$ with probability one, and that X is almost surely Lipschitz with modulus K_X satisfying $\mathbf{E}[K_X] < \infty$. Then $\mathbf{E}[X(s_0)]'$ exists and equals $\mathbf{E}[X'(s_0)]$.

Theorems 4.2 and 4.3 establish the main validation result for the state variables, decisions, and their derivatives with respect to the echelon base stock variables.

Theorem 4.2 If $\{d_n^p, n = 1, 2, \ldots, p = 1, 2, \ldots, P\}$ are independent and each d_n^p has a density on $(0, \infty)$, then the following hold:

- For $k = 1, \ldots, K, m = 1, \ldots, M, p = 1, \ldots, P$, and $n = 1, 2, \ldots$, each I_n^{kmp} as given by (1) is, w.p.o., differentiable at $(z^{111}, \ldots, z^{KMP})$ with respect to each z^{qrs}, $q = 1, \ldots, K, r = 1, \ldots, M$, and $s = 1, \ldots, P$. Moreover, these derivatives satisfy (13).
- P_n^{kmp} as given by (7), with f_n^{kmp} and g_n^{km} given by (5) and (8), respectively, is also differentiable w.p.o. and its derivative satisfies (15), with f'^{kmp}_n and g'^{km}_n given by (14) and (16), respectively.

Proof: The differentiability of the state variables relies on the structure of the recursive equations defining them. Due to the structure of (7), we only have to check if (5) and (8) are differentiable. The remaining equations are linear combinations of state variables and so do not pose any problem.

We start with expression (5). A tie between the two terms may induce non-differentiability. Since demands are continuous with only a point of mass at zero, ties occur with probability zero, except for the case where both terms are zero. The term $z^{kmp} + d_n^p - E_n^{kmp}$ equals zero if $d_n^p = 0$, in which case the echelon inventory reached its base stock level in the previous period. If this happens at some value for z^*, then w.p.o. it does so in a neighborhood of z^*, which implies that the derivative is zero. A similar reasoning is valid for the second term of (5), i.e., if the second term is zero, then it remains zero in a neighborhood of z^* w.p.o. Since both terms have zero derivative, the expression is differentiable and its derivative is a defined in (14).

Now consider (8). Again, a tie in the two terms may induce non-differentiability. The random variables f_n^{kmp} have two points of mass: one at zero and the other at $\Delta^{(km)^+p}$. This latter occurs when the net needs are bounded by feeding inventory, and this equals its local base stock level. Thus, the tie between $\sum_p f_n^{kmp}$ and C^{km} may occur with non-zero probability as in the following two cases.

1. The first case occurs when $C^{km} = C^{(km)^+}$. When this is the case, $\Pr\{\sum_p f_n^{kmp} = C^{km}\} \neq 0$, irrespective of the control parameters, because $\Pr\{I_n^{kmp} = 0\} \neq 0$. If in a period the feeding stage is bound by capacity at the same time stage (km) has zero inventory, in the next period it may be the case that $C^{km} = \sum_p f_n^{kmp}$. However, since a bound in capacity will occur w.p.o. in a neighborhood of z^*, the derivative will be zero for both terms and differentiability will be preserved.

2. The second case occurs when $C^{km} = \Delta^{(km)^+p}$ or a sum of some delta variables matches C^{km} exactly. If this ever happens we have a case of non equal derivatives on both terms because $\frac{d}{dz^{(km)^+p}} \sum_p f_n^{kmp} = 1$ and $\frac{d}{dz^{(km)^+p}} C^{km} = 0$. Given that the simulation is run with echelon base stock values resulting from iterations of an optimization procedure generating real values, the probability of a perfect tie at a given simulation run is zero, as in [Glasserman and Tayur, 1995].

Therefore, w.p.o., differentiability is preserved at each period.

□

Theorem 4.3 If, in addition to the conditions of Theorem 4.2, $\mathbf{E}[d_n^p] < \infty$ for all n, then $\mathbf{E}[I_n^{kmp}]'$, and $\mathbf{E}[P_n^{kmp}]'$ exist and equal $\mathbf{E}[I'^{kmp}_n]$, and $\mathbf{E}[P'^{kmp}_n]$.

Proof Sketch: We outline the logic of proof. To invoke Lemma 4.1, we have to show that with probability one the system variables are Lipschitz functions with integrable moduli. Since the state variables at time zero are linear on the base stock levels, they are Lipschitz. Since the operators min{.}, max{.}, addition, and multiplication preserve that property, it follows that each I_n^{kmp} and P_n^{kmp} is a composition of Lipschitz functions, and therefore is Lipschitz.

Since $\mathbf{E}[d_n^p] < \infty$ for all n, every I_n^{kmp} has finite expectation. Also, each P_n^{kmp} is integrable because it is bounded. Division (in our context) does not pose a problem as whenever the term $\sum_p f_n^{kmp}$ drops below C^{km}, $g_n^{km} = 1$.

The main thrust of the proof is to show that if some echelon base stock variable changes by a small amount δ, this induces bounded derivatives for the state variables. This can alternatively be done for the Δ variables and it turns out to be easier with these. Let us see what happens to the state variables on a sample path if only one of the Δ variables is infinitesimally disturbed. Each production decision may be bounded by capacity or not. Also, each production decision may be bounded by local inventory or not. The analysis has to be made for all possible combinations of these, for the specific case of the product, level, and stage which has been disturbed.

As long as the production decisions are not bounded by the disturbed inventory, with or without capacity bound, the only change is the disturbed inventory, which incurs a derivative of zero for all the other state variables and of one for the disturbed inventory. Things change for the first time when there is a bound in inventory due to the disturbed inventory. For this we have:

1. *No bound in capacity* – Since there is no simultaneous bound in capacity the whole disturbance moves down to next stage and level and stays associated with the same product. Therefore, the derivative incurred is zero for all the state variables except for the inventory fed by the originally disturbed inventory, for which the derivative is one.

2. *Bound in capacity* – Given the simultaneous bound in capacity, all production decisions for the level (PS mode) will be affected. Since the capacity is linearly scaled, this translates into fractions of the disturbance to be distributed among the inventories feeding the stage. Given the fact that the total production of the stage is bounded by capacity, the stage and level fed by this set of decisions will also be fractionally affected, but the sum of the disturbances adds up to zero. This means that all derivatives are either zero or have absolute values smaller than one.

Subsequent changes are similar: each time a production decision is solely bounded by disturbed inventories, the disturbance moves down the production line keeping the same magnitude. If it is simultaneously bounded by capacity, it redistributes itself in fractions among the stage and level for which the capacity bound occurs and some disturbance is provoked on the inventories fed in a way that the total added disturbance is zero. If the production decisions are solely bounded by capacity, the disturbances remain unchanged.

Whenever a single production decision of a disturbed inventory is solely bounded by demand we have two cases:

1. *This is the originally disturbed variable* – this has the effect of recovering the original

echelon base stock value, which means that the original disturbance returns to its original place, in terms of echelon inventory. Naturally if individual inventories for this product down the line are not yet reset to their original values the sum of disturbances for those will be zero.

2. *This is not the originally disturbed variable* – it moves the disturbance upwards to the feeding inventory, given that a bound by demand resets the whole echelon to its original value. The last sentence of case 1 above applies here as well.

Concluding, it can be shown that a singular disturbance propagates along the line only on levels and stages fed by it; the transference of disturbances does not increase from the original size; the transference of a mixture of disturbances can at most add up to the existing disturbances; when disturbances are moved upwards they will not go beyond the originally disturbed variable.

Given the line is finite in length there is a bound on the amount of forward propagation, which imposes a bound on the maximum disturbed amount for the finished goods inventory. This bound is proportional to δ. Therefore, $|I'^{kmp}_n|$ and $|P'^{kmp}_n|$ are bounded as functions of the echelon base stock variables in a finite horizon setting. Thus, Lemma 4.1 is applicable and the result follows.

□

Although technically more complex, it is also possible to establish the above result for the TS mode and for the other production rules using similar arguments. For the ESR, a couple of technical results are required to validate IPA and are presented in Appendix A. The validation for the finite horizon cost function follows because C_n is Lipschitz with modulus $K_I(\sum_{p=1}^{P} b^p + \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{p=1}^{P} h^{kmp})$, and K_I given by $\max_{k,m,p,n}\{|I'^{\ kmp}_{(z)_n}|\}$ is finite. Lemma 4.1 applies. For the infinite horizon, we first note (see [Bispo and Tayur, 1997]) that the necessary and sufficient condition for stability in the PS mode is

$$\mathbf{E}[\textstyle\sum_{p=1}^{P} d_0^p] < \min_{k,m}\{C^{km}\}, \tag{20}$$

and for the TS mode it is

$$K\mathbf{E}[\textstyle\sum_{p=1}^{P} d_0^p] < \min_{m}\{C^{m}\}. \tag{21}$$

We can show, as in [Glasserman and Tayur, 1995], that these conditions imply regeneration in finite time with probability one. Applying the same logic as in [Glasserman and Tayur, 1995], the validation of the derivatives of the cost and service level follow, and we have:

Theorem 4.4 Suppose $\{d_n^p, n = 1, 2, \ldots\}$ are i.i.d. for each p with finite expectation and that the adequate stability condition holds. Then $N^{-1}\sum_{n=1}^{N} C'_n \rightarrow c'_\infty$, with probability one, at almost every z^*. If, in addition, $\sup_x f^p(x) < \infty$, then $\bar{V}'_N \rightarrow \bar{v}'_\infty$ with probability one, at almost every z^*.

5 Experimental Study

A comprehensive study of re-entrant systems is an enormous task to perform, given the number of parameters to be taken into account: average demand, demand variance, holding costs, penalty costs, number of machines, number of levels, capacity of the machines, capacity allocation modes, and production rules. We have limited our computational study to one and two products (in Section 5.1 and 5.2 respectively) and assumed re-entrant structure on a single machine. Even in this simplified setting the range of parameters is very wide. All our experiments concern the infinite horizon average cost setting. Our task is somewhat simplified due to the following connection between optimal cost performance and achieved service level.

Theorem 5.1 For a production system, composed of any number of machines and levels, operated under a multi-echelon base stock policy, if $\{d_n^p, n = 1, 2, \ldots; p = 1, \ldots, P\}$ are independent and stationary, where each d_n^p is drawn from a density on $(0, \infty)$, the optimal base stock levels for the infinite horizon average cost measure for any production rule and any capacity sharing mode are such that

$$Pr(d_0^p \leq I^{11p}) = \frac{b^p}{b^p + h^{11p}} \quad \text{for all } p = 1, \ldots, P. \tag{22}$$

We call (22) as the *optimality condition*. We skip the proof as it follows along the lines of [Glasserman and Tayur, 1996].

Remark. There are cases where the optimization algorithm stops short of achieving a set of variables where the optimality condition is satisfied. These are the cases where the cost function is non differentiable at the optimum; we do see such situations in Section 5.1. Alternatively, as in [Anupindi and Tayur, 1998], one can optimize directly using service level constraints, which may be preferable in many cases.

Simulation details. The simulation length for each cost and gradient estimate is 20000 periods. This simulation length ensures the 95% confidence intervals are 2.5% wide relative to the central value. The comparison of relative performance for different production rules is made on a single

sample path; see [L'Ecuyer, 1994] for reasons to do so. We implemented a discrete step version of the *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* optimization algorithm ([Bertsekas, 1999]) to generate the successive values of the delta variables. (More specific details regarding Hessian updates and step sizes can be obtained from the authors.)

5.1 Single Product Setting

We first see how one should allocate a fixed total capacity to the different levels in the NS mode in Section 5.1.1. The results of [Glasserman and Tayur, 1995] show that capacity should be non-decreasing along the flow line; however, they do not study how a fixed amount to be allocated is distributed to the various levels. Next, in Section 5.1.2, we investigate the impact of holding costs and machine load on the relative performance of the production rules in both the NS and the TS modes. (There is no PS mode for single product systems.)

5.1.1 Capacity allocation to levels in NS mode

We take derivatives with respect to the capacity slots in order to determine their optimal values, constraining their sum to a fixed constant. The optimization is done simultaneously with respect to the base stock levels and capacity slots. One would expect the optimal allocation of capacity to level to depend on the holding costs along the production line. Surprisingly, in the majority of the cases, the optimal allocation is to divide capacity equally among the levels. We observed this in most of our experiments with different values of capacity, holding and penalty costs, number of levels, number of machines, and different demand distributions. A sample set of results is shown in Figure 1 for a system with $K = 2$ and $M = 1$. Keeping the values of $h^{111} = 10$ and $b^1 = 20$ constant, we varied h^{211} from 0 to 10. The sum $C^{21} + C^{11}$ was kept constant and equal to 25. For each case we computed the optimal cost for the optimal capacity allocation and the optimal cost for $C^{21} = C^{11} = 12.5$. (The second graph is a zoom of the graph on the left for low values of h^{211}/h^{111}.) To see why this should be the case intuitively, consider the situation when the penalty costs are high, and the large deviation approximation applies (see chapter 3 by Glasserman [1] in [Tayur et al., 1999]). Then, the optimal inventory levels (largely) depend on the bottleneck alone, and maximizing the capacity of the bottleneck is achieved by equal allocation of a fixed total capacity. This same conclusion carries through

[1] Very briefly: Under certain conditions, such as the one where the penalty costs are high, the tail distribution of the stochastic process that drives the inventory and backlog processes, can be approximated by an exponential distribution, whose parameters depend on capacity of the machines. In the limit, only the most stringent machine dominates the approximation.

for multiple products; we do not present the graphs here.

As the holding costs of early levels decrease to very low values, the optimal allocation of capacity changes. We notice that for values of h^{211} above 10% of h^{111}, the optimal capacity allocation is achieved by dividing C^1 into two exactly equal slots while below the 10% ratio, the optimal capacity allocation is achieved with $C^{11} > C^{21}$. It appears that the benefit of non-equal allocation to exploit holding cost differences gets washed away pretty quickly as it requires that the differential in holding costs be quite high. We also observed that the value of the penalty cost affects this ratio: The higher the value of b^1, the higher the value of h^{211}/h^{111} above which it is optimal to have $C^{21} = C^{11}$. As instances, take a case with $b^1 = 100$ where such solution is optimal for $h^{211}/h^{111} \geq 0.18$, and with $b^1 = 1000$ it is optimal for $h^{211}/h^{111} \geq 0.21$. This indicates that the benefit of exploiting the differential in holding costs is higher when the penalty cost is higher.

In the rest of our experiments, we allocate capacity to levels equally.

5.1.2 Performance of Production Rules in TS mode and Comparison between NS and TS modes

Before we can compare across the production rules, we need to find which is the best priority *within PR*. In the single product setting in the TS mode, among the $K!$ choices for static priority, the rule that gives priority to levels closer to demand always achieves the lowest cost. This is consistent with the LBFS (last buffer first serve) rule of [Lu and Kumar, 1991] and results of [Glassey and Resende, 1988, Lu and Kumar, 1991, Kumar and Kumar, 1994]. Consider a system with $K = 3$, $M = P = 1$, and 80% load to illustrate this fact. There are $3! = 6$ different priority assignments for the levels. Table 5.1.2 displays the optimal costs for each one of the priority assignments. The leftmost column lists the levels by decreasing order of their priority. Thus, 1-2-3 stands for priority to level 1, then to level 2, and finally to level 3. As shown in the table, the order 1-2-3 achieves a cost that is no higher than any other assignment. (In this example, 2-1-3 is also tied for first place; this is not always the case, although any sequence with stage 3 last in the priority is competitive with 1-2-3.)

Let us now compare the best of PR with the others. We thus compare the three production rules in the TS mode (LSR, ESR and PR) and the NS mode. We use the results obtained for a system with $K = 3$, $M = 1$, and $P = 1$ as the basis for discussion. We study the effect of different holding cost structures at different loads on relative performance.

Experimental Setup. We fixed the values of $h^{111} = 10$ and $b^1 = 20$. The other two holding

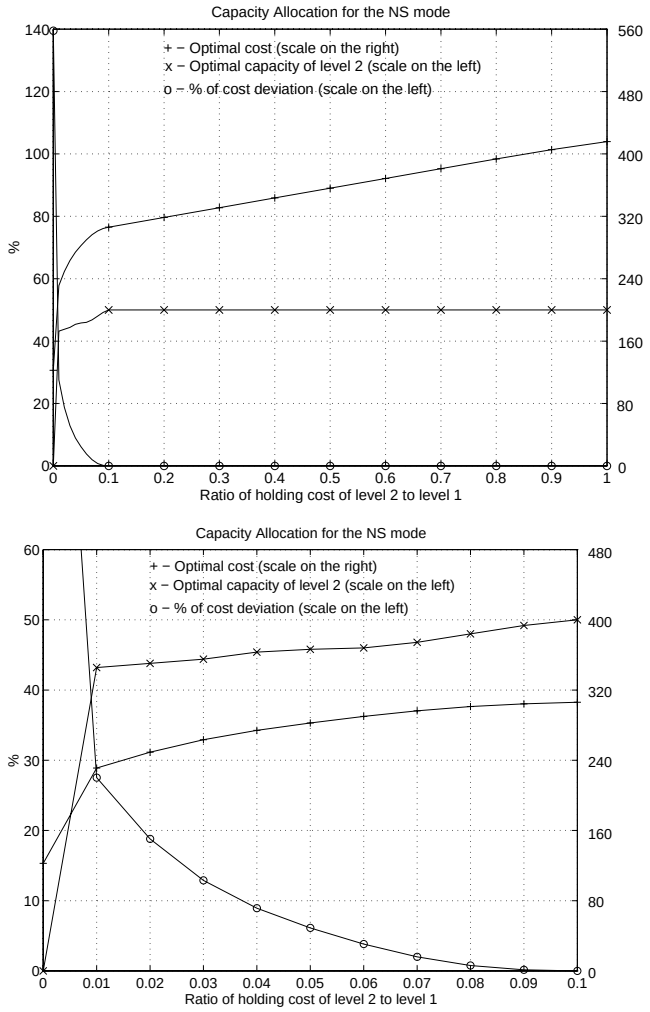

Figure 1: Capacity allocation as a function of holding costs. The upper figure shows that for a large range, the equal capacity allocation is optimal. The lower figure displays the small range when equal capacity allocation is not optimal. A total of 25 units of capacity are being allocated to two levels in a single stage, single product system.

Table 1: Optimal costs for alternative priority assignments.

Priority of levels	Optimal Cost
1 - 2 - 3	463.57
1 - 3 - 2	677.69
2 - 1 - 3	463.57
2 - 3 - 1	1110.14
3 - 1 - 2	701.75
3 - 2 - 1	1390.86

costs were varied from 0 to 10. With $[h^{311}, h^{211}, h^{111}, b^1]$ as notation, we study $[0, 0, 10, 20]$ to $[10, 10, 10, 20]$. On the x-axis of the figures, the entry $(4, 4)$ represents a system with $h^{311} = h^{211} = 4$, $h^{111} = 10$, and $b^1 = 20$, that is, $[4, 4, 10, 20]$. Between label $(2, 2)$ and label $(4, 4)$ lie labels $(2, 4), (2, 6), (2, 8)$, and $(2, 10)$ in that order, which correspond to the cost structures $[2, 4, 10, 20]$, $[2, 6, 10, 20]$, $[2, 8, 10, 20]$, and $[2, 10, 10, 20]$ respectively. We summarize the cost performance, the inventory behavior, and the connection with the Type-1 service level.

Cost Performance. In Figure 2 we compare all four possibilities in terms of cost for 90% load. (Different loads do not qualitatively change the relative performance of the rules; we do not display the plots here of our experiments conducted with 80% and 85 % loads.) It is easy to see that the Priority Rule outperforms the other two rules in the TS mode and also dominates the NS mode. The Equalize Shortfall Rule has only a very slight advantage relative to the NS mode, and converges to the same levels of performance as those of the Priority Rule as the intermediate holding costs increase.

It is worth noting that LSR performs terribly in the TS mode. The intuition is as follows. A closer inspection reveals that this is due to the way the scaling of production net needs is done. All levels except the entering level (level K) may be bound by feeding inventory. Level K is never bound by feeding inventory because this is assumed to be infinite. If there is a large shortfall, the production net needs of level K match the shortfall, but all other levels may be bound by inventory. Since the scaling (for capacity allocation) is done in terms of production net needs, it turns out that level K gets a higher share, thus affecting the lower levels. It is as if we are giving a higher priority to level K. As we saw in our study among the $K!$ choices for PR, this leads to the worst performances.

Inventory behavior. Figure 3 shows how the *optimal* Δ *levels* behave for the NS mode at 85% load. The reason we choose these plots, rather than the plots of the base stock variables, is

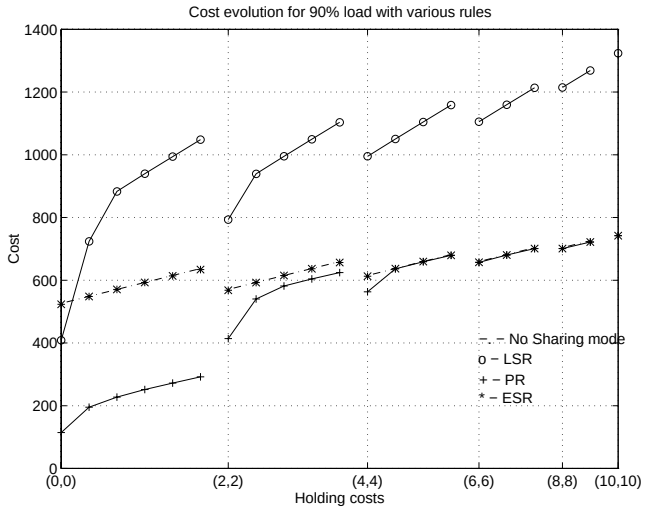

Figure 2: Optimal cost for 90% load comparing the four options: (1) NS mode; (2) LSR, (3) ESR and (4) PR in TS mode. LSR performs poorly, while PR performs well. ESR and NS are indistinguishable, and are competitive with PR at higher intermediate holding costs.

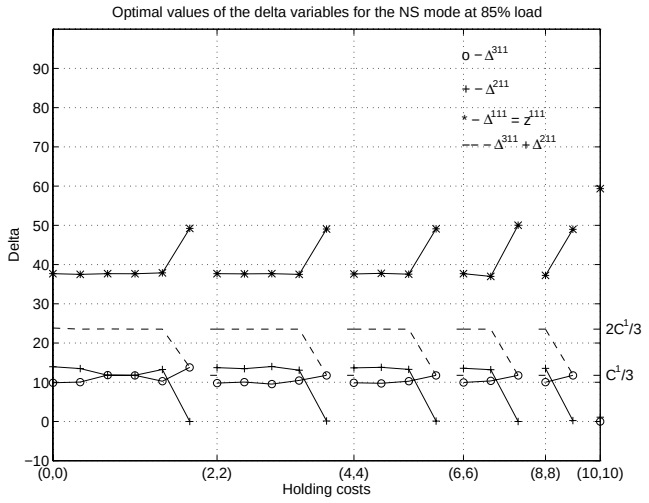

Figure 3: NS mode at 85 % load. We notice that inventory levels can be flat for a wide range of holding costs, and have a subtle relationship with capacity.

because they help make some of the structural properties of the solutions more evident.

Note the *almost constant* behavior of $z^{311} = \Delta^{311} + \Delta^{211} + \Delta^{111}$ across the different holding costs. Simply add the dashed lines with the starred lines. This also occurs for the three rules in the TS mode for a significant range of intermediate holding costs, namely very soon after moving away from zero for h^{311} and h^{211}. This seems to imply that the highest base stock is more sensitive to the terminal holding and penalty costs than it is to the intermediate holding costs, as long as these are not too small. For all the rules it is also evident that the different distribution of holding costs along the the system has the effect of just distributing the inventory on the different levels. Note in Figure 3 that $\Delta^{311} = z^{311} - z^{211}$ approaches zero when $h^{311} - h^{211}$ approaches zero and that $\Delta^{211} = z^{211} - z^{111}$ also approaches zero when $h^{211} - h^{111}$ approaches zero.

It is interesting to note that for some instances the sum $\Delta^{311} + \Delta^{211}$ is constant for a range of parameters The value achieved for those instances equals $\frac{2}{3}C^1$. In some other instances, the value Δ^{311} or Δ^{211} is constant and equals $\frac{1}{3}C^1$. This shows that the interaction between capacity and inventory levels is subtle, and in some sense, exact.

Connection with service level. Recall the *optimality condition*. Using a trivial gradient based optimization procedure most of the simulation runs converge to values where the condition is satisfied. Some times this simple optimization procedure fails to converge to values where the optimality condition holds. The failure to achieve this condition coincides with cases where either $\Delta^{311} + \Delta^{211} = \frac{2}{3}C^1$, $\Delta^{311} = \frac{1}{3}C^1$, $\Delta^{211} = \frac{1}{3}C^1$, or some $\Delta^{k11} = 0$ at the optimal. In these cases the cost function is not differentiable around the optimum, which calls for the utilization of more sophisticated optimization techniques; see [Lemaréchal, 1989].

5.2 Multiple Products

We now turn to systems processing multiple products. As in the single product case, we find that the best production rule for the TS mode always achieves better costs than the best production rule for the PS mode, although in some cases the difference may be very small. This reinforces the generally held belief that the larger the flexibility the better one can make use of the available resources. A notable fact is that LSR degrades its performance in TS mode relative to PS mode; this is intuitive because with a larger flexibility in capacity compared to PS mode, the LSR (implicitly allowing level K to have priority) is doing more damage.

First, in Section 5.2.1, we discuss the issue of priority assignment alternatives in the context of multiple products. In Section 5.2.2 we analyze the effect of holding costs, in (Section 5.2.3) the

effects of the penalty costs, and in (Section 5.2.4) the effects of demand variance.

5.2.1 Finding the best PR

We consider systems with a single machine processing two types of products, where each product is required to visit the machine three times before completion: $K = 3$, $M = 1$, $P = 2$. The presence of multiple products introduces additional options in the way priority can be assigned within the PR. There are a total of $(K \times P)! = 6! = 720$ different priority assignments. We restrict our attention to a logical subset of these.

Method 1: "Level first, Product second"

A situation where the priority for levels is $\{2, 1, 3\}$ and the priority of products is $\{2, 1\}$ signifies that, in the TS mode, production decisions are taken in the order: $\{P^{212}, P^{211}, P^{112}, P^{111}, P^{312}, P^{311}\}$. That is, we decide the production amounts level by level, according to the priority to levels and in each level we use the product priority. Comparing the optimal costs for the 12 different combinations (six different priority lists for each possible product priority), the order 1-2-3 always achieves the lower cost for each product priority (results not displayed here). This is in line with what we observed for single product.

Method 2: 'Product first, Level second'

Alternately, we may choose to prioritize primarily by product. That is, if priority to levels is $\{2, 1, 3\}$ and to products is $\{2, 1\}$ the production decisions may be taken by the order $\{P^{212}, P^{112}, P^{312}, P^{211}, P^{111}, P^{311}\}$. In Table 2 we present the comparison of this method with the previous one. (We only ran the systems for the best choice of priority for the levels based on previous observations.) We also investigated if changing demand variance and penalty costs produces any qualitative change to the above conclusions. We found that neither the penalty cost nor the demand variance affect the conclusion of this test. Thus, in what follows we use method 1 for PR.

Table 2: Comparison of method 1 and method 2.

Priority of levels	Priority of products	Method	Optimal Cost
1 - 2 - 3	1 - 2	1	771.41
1 - 2 - 3	1 - 2	2	786.38
1 - 2 - 3	2 - 1	1	752.62
1 - 2 - 3	2 - 1	2	783.04

5.2.2 Varying the holding cost structure for products

We study the impact of the holding costs on the performance of the several rules. This experiment is composed of two parts: (1) *same* holding cost for the different products; and (2) *different* holding costs for the products.

Experimental setup #1. We have $K = 3$, $M = 1$, and $P = 2$. The average demand for both products is fixed: $\mathbf{E}[d_0^1] = 8$ and $\mathbf{E}[d_0^2] = 12$. The total capacity of the single machine is fixed to an average load of 80%, that is, $C^1 = (3 \times 8 + 3 \times 12)/0.8 = 75$. The coefficient of variation for both products is fixed at 1. The costs h^{11p} and b^p were fixed at 10 and 20, respectively for $p = 1, 2$. We generated 21 different systems by changing the holding costs of level 3 and level 2, that is h^{31p} and h^{21p} for $p = 1, 2$. All the 21 systems have identical cost structure for both products, that is, $h^{311} = h^{312}$ and $h^{211} = h^{212}$. The cost structure of the first system is given by $[0, 0, 10, 20]$ for both products and the cost structure of system number 21 is $[10, 10, 10, 20]$. For each one of the 21 systems we obtained the optimal solution for all three production rules each with PS and TS. For the case of the PR, recalling the study on priorities, we only have to consider two choices (out of 120): either product 1 has priority or product 2 has priority. In all, therefore, we generated (4 for PS + 4 for TS =) 8 solutions per system and so obtaining (21(8)=) 168 solutions.

Experimental setup #2. The study is further subdivided into two subsets of experiments. In the first subset, the cost structure of product one was kept constant at $[2, 6, 10, 20]$, and we changed the cost structure of product two from $[0, 0, 10, 20]$ to $[10, 10, 10, 20]$ thus generating 21 different systems. In the second subset we exchanged the positions of product one and product two, generating another set of 21 different systems. For each one of the two subsets of experiments we generated 8 solutions as before, leading to 336 cases.

For the sake of brevity we do not display the plots. In qualitative terms, they are no different from the ones presented for the single product setting. We summarize the main findings.

Cost Performance in PS mode. In the PS mode, the change in holding costs does not affect the relative performance for the several rules. The LSR and the ESR achieve practically the same costs and perform better than any of the two tested priority assignments. Within PR, priority should be given to product 1 over product 2 to achieve the best performance. In general, all things being equal, priority should be given to products with the lower average demand within PR, which we explain intuitively later. Exceptions to this rule only occur when initial holding costs are close to zero.

Cost Performance in TS mode. There is no one best rule across all costs in this setting in contrast to the PS setting. For the TS mode also, priority should be given to the product with the lowest demand to achieve better performance. Priority to either product achieve practically the same best costs for situations where h^{31p} and h^{21p} are low. ESR performs best in any situation with higher holding costs. The advantage of the PR in the low holding cost cases is due to the build up of inventory since the costs are so low. Intuitively, the good performance of ESR is due to the following. Since the penalty for backlog is the same for both products, as are the terminal holding costs, it is as if the shortfall has the same price (or cost), and therefore trying to equalize it should be a good strategy.

These studies have shown that the average demand seems to be a determinant factor in deciding to which product we should give higher priority. However, there is a relationship with costs as well. Priority to product 2 is better than priority for product 1 only for low holding costs. For moderate to high costs, priority to product 1 outperforms priority to product 2, the lowest average demand has a strong effect. To intuitively see why this is so, consider again the large deviation approximation. It is easy to show that (if product [1] is the first in priority and product [2] is second priority sharing a capacity C), the required inventory levels (and costs) are asymptotically proportional to $\frac{\sigma^2_{[1]}}{2(C-\mu_{[1]})} ln \frac{b_{[1]}+h_{[1]}}{h_{[1]}}$ and $\frac{\sigma^2_{[1]}+\sigma^2_{[2]}}{2(C-\mu_{[1]}-\mu_{[2]})} ln \frac{b_{[2]}+h_{[2]}}{h_{[2]}}$ respectively for products [1] and [2]. (Note that h, b, μ, σ stand for holding, penalty, mean demand and standard deviation of demand respectively for the appropriate products.) This shows that if both products have the same holding and penalty costs, and the same coefficient of variation, then the product with a lower mean demand should be given a higher priority. As the costs change, and if the product with the higher mean demand also has higher penalty cost, the switch takes place in priority.

5.2.3 Varying the penalty costs

Experimental setup. The basic features remain the same as those of the two earlier studies. We ran two sets of experiments. In the first set we kept the cost structure for product one fixed at $[6, 8, 10, 20]$ and the cost structure of product two is $[2, 6, 10, b^2]$, with $b^2 \in [10, 50]$. For the second set of experiments we set the cost structure of product one at $[6, 8, 10, b^1]$, with $b^1 \in [10, 50]$ and kept the cost structure of product two fixed at $[2, 6, 10, 20]$. Each of the two sets comprises 21 different systems, leading to the generation of 336 different solutions as before. We summarize the results.

Set 1: Cost performance in PS mode. Figure 4 (upper) shows that both ESR and LSR are tied

for the first place (as before). It turns out that the ESR wins for high values of b^2 and loses to the LSR for low values. For the PR, the penalty cost variation introduces a more interesting behavior. There is a value of b^2 above which the best performance is achieved by PR when priority is given to product 2. (On the other hand, when b^2 is very low, priority to product 1 approaches the performance of the ESR and the LSR).

Set 1: Cost performance in TS mode. The ESR is still the winner for the TS mode for a wide range of values. Off those cases, the PR with priority given to product 1 achieves the best performance for low values of b^2. As in the PS mode there is a value for b^2 above which priority should be given to product 2. Observing the slope of the curves in the Figure 4 (lower), we can argue that eventually there will also be a value for b^2 above which the best production rule is the PR, with priority given to product 2. To confirm this, we ran a case with $b^2 = 200$. The optimal cost achieved with the ESR was 1204.4 and the PR, giving priority to product 2, achieved a cost of 1176.5.

In the second set of this study, where we changed the value of b^1, we observed the same qualitative features as in the first set. Thus, we omit the plots and a summary.

Concluding, in the PS mode, ESR works well. In the TS mode, ESR works best except when b^1 or b^2 is high: PR outperforms ESR in this case. Priority, in either modes, should be given to the product with lower mean demand (all other things being equal). However, if the penalty cost of the product with higher mean demand is significantly higher, then priority should be given to this product.

5.2.4 Varying the coefficient of variation (cv) for the demand

Experimental setup. The average demand for the products was kept the same as before in a $K = 3, P = 2$ and $M = 1$ system. We ran eight sets of experiments (differing in costs and product for which cv is changed). The cv range was between 0.1 to 1.0. Each of the eight sets has 10 different systems (one for each coefficient of variance). For each of the 80 settings, we generate 8 solutions (4 for PS and 4 for TS). The total is 640 solutions. The summary of the experiment is as follows. ESR performs the best in both PS and TS modes, unless the combination of high penalty cost, low mean demand and low variance in demand occurs, in which case PR with priority to this product wins. LSR is competitive in the PS mode, but falls substantially behind in the TS mode.

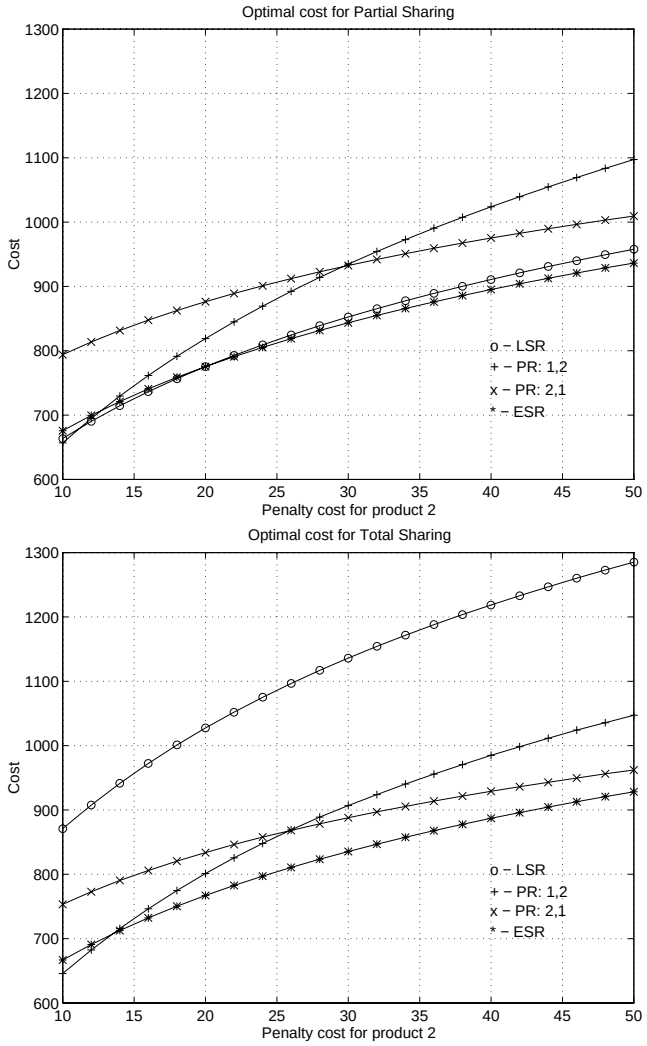

Figure 4: Optimal cost for the PS and TS mode as a function of the penalty cost for product 2. ESR performs well in both PS and TS modes. Note the switch in priorities on the products as a function of penalty cost, as explained by the large deviation approximation. LSR degrades in performance as we move from PS to TS mode.

6 Conclusions

This paper proposed a framework to manage re-entrant flow lines producing multiple products. It concentrated the analysis on a simple (and implementable) set of capacity management schemes and production rules as a first step towards understanding broader classes of systems. The re-entrant lines were modeled as discrete time capacitated multi-product production/inventory systems, operating under modified multi-echelon base stock policies – production decisions are constrained by available inventory and capacity. Several capacity sharing mechanisms were discussed and some production rules to manage capacity both from dynamic and static points of view were proposed.

Since these systems are too complex to handle analytically, the study used simulation-based optimization. After validating Infinitesimal Perturbation Analysis (IPA), a series of computational studies provided many insights on how to manage re-entrant systems. We briefly review some interesting findings. Equalize Shortfall Rule achieves the best performances across a wide range of parameters. The Priority Rule outperforms the Equalize Shortfall Rule when it is possible to unambiguously order the products with these three combined parameters, i.e., when the product with the lowest expected demand has lowest variance and highest penalty cost. The equivalence between penalty costs and service level estalished by Theorem 5.1 allows one to define a target service level instead of a penalty cost.

There are other issues that are worthy of investigation, such as systems with non-uniform loads and managing other sources of uncertainty. We are studying these at this time.

References

[Akella et al., 1984] Akella, R., Choong, Y. F., and Gershwin, S. B. (1984). Performance of Hierarchical Production Scheduling Policy. *IEEE Trans. on Components, Hybrids, and Manuf. Technol.*, CHMT-7(3):225–240.

[Akella and Kumar, 1986] Akella, R. and Kumar, P. R. (1986). Optimal Control of Production Rate in a Failure Prone Manufacturing System. *IEEE Trans. on Automatic Control*, AC-31(2):116–126.

[Anupindi and Tayur, 1998] Anupindi, R. and Tayur, S. (1998). Managing Stochastic Multi-Product Systems: Models, Measures and Analysis. *Operations Research*, 46(5):S98–S111.

[Bai and Gershwin, 1996] Bai, S. X. and Gershwin, S. B. (1996). Scheduling Manufacturing Systems with Work-In-Process Inventory Control: Reentrant Systems. *OR Spectrum*, 18(4):187–195.

[Bertsekas, 1999] Bertsekas, D. P. (1999). *Nonlinear Programming.* Athena Scientific, Massachusetts.

[Bispo, 1997] Bispo, C. F. G. (1997). *Re-Entrant Flow Lines.* PhD thesis, Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA.

[Bispo and Tayur, 1997] Bispo, C. F. G. and Tayur, S. (1997). Managing Simple Re-entrant Flow Lines II: Stability. *Working paper*.

[Glasserman and Tayur, 1995] Glasserman, P. and Tayur, S. (1995). Sensitivity Analysis for Base-Stock Levels in Multi-Echelon Production-Inventory Systems. *Management Science*, 41(2):263–281.

[Glasserman and Tayur, 1996] Glasserman, P. and Tayur, S. (1996). A Simple Approximation for a Multi-stage Capacitated Production-Inventory System. *Naval Research Logistics Quarterly*, 43(1):41–58.

[Glassey and Resende, 1988] Glassey, C. R. and Resende, M. G. C. (1988). Closed-Loop Job Release Control for VLSI Circuit Manufacturing. *IEEE Trans. on Semiconductor Manufacturing*, 1(1):36–46.

[Harrison and Wein, 1990] Harrison, J. M. and Wein, L. M. (1990). Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Closed Network. *Operations Research*, 38(6):1052–1064.

[Ho and Cao, 1991] Ho, Y.-C. and Cao, X. R. (1991). *Perturbation Analysis of Discrete Event Dynamic Systems*. Kluwer Academic Publishers.

[Kimemia and Gershwin, 1983] Kimemia, J. G. and Gershwin, S. B. (1983). An Algorithm for the Computer Control of a Flexible Manufacturing System. *IIE Transactions*, 15(4):353–362.

[Kumar and Seidman, 1990] Kumar, P. R. and Seidman, T. I. (1990). Dynamic Instabilities and Stabilization Methods in Distributed Real-Time Scheduling of Manufacturing Systems. *IEEE Trans. on Automatic Control*, 35(3):289–298.

[Kumar and Kumar, 1994] Kumar, S. and Kumar, P. R. (1994). Performance Bounds for Queueing Networks and Scheduling Policies. *IEEE Trans. on Automatic Control*, 39(8):1600–1611.

[L'Ecuyer, 1994] L'Ecuyer, P. (1994). Efficiency Improvement and Variance Reduction. In *1994 Winter Simulation Conference*, pages 122–132.

[Lemaréchal, 1989] Lemaréchal, C. (1989). Nondifferentiable optimization. In Nemhauser, G. L. et al., editor, *Hanbooks in OR and MS*, volume 1, pages 529–572. Elsevier Science Publishers.

[Lou and Kager, 1989] Lou, S. X. C. and Kager, P. W. (1989). A Robust Production Control Policy for VLSI Wafer Fabrication. *IEEE Trans. on Semiconductor Manufacturing*, 2(4):159–164.

[Lu et al., 1994] Lu, S. C. H., Ramaswamy, D., and Kumar, P. R. (1994). Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing Plants. *IEEE Trans. on Semiconductor Manufacturing*, 7(3):374–388.

[Lu and Kumar, 1991] Lu, S. H. and Kumar, P. R. (1991). Distributed Scheduling Based on Due-Dates and Buffer Priorities. *IEEE Trans. on Automatic Control*, 36(12):1406–1416.

[Perkins and Kumar, 1989] Perkins, J. R. and Kumar, P. R. (1989). Stable, Distributed, Real-Time Scheduling of Flexible Manufacturing/Assembly/Disassembly Systems. *IEEE Trans. on Automatic Control*, 34(2):139–148.

[Speck and der Wall, 1991] Speck, C. J. and der Wall, V. (1991). The Capacitated Multi-echelon Inventory System with Serial Structure: The average cost criterion. Technical Report COSOR 91-39, Dept. of Math. and Comp. Sci., Eindhoven Institute of Technology, Eindhoven, The Netherlands.

[Tayur et al., 1999] Tayur, S., Ganeshan, R., and Magazine, M. (1999). *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers.

[Uzsoy et al., 1992] Uzsoy, R., Lee, C.-Y., and Martin-Vega, L. A. (1992). A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part I: System Characteristics, Performance Evaluation and Production Planning. *IIE Transactions*, 24(4):47–60.

[Uzsoy et al., 1994] Uzsoy, R., Lee, C.-Y., and Martin-Vega, L. A. (1994). A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part II: Shop-Floor Control. *IIE Transactions*, 26(5):44–55.

[van Ryzin et al., 1993] van Ryzin, G., Lou, S. X. C., and Gershwin, S. B. (1993). Production Control for a Tandem Two-Machine System. *IIE Transactions*, 25(5):5–20.

[Wein, 1988] Wein, L. M. (1988). Scheduling Semiconductor Wafer Fabrication. *IEEE Trans. on Semiconductor Manufacturing*, 1(3):115–130.

[Wein, 1990] Wein, L. M. (1990). Optimal Control of a Two-Station Brownian Network. *Math. Opns. Res.*, 15:215–242.

A Equalize Shortfall Rule

Under partial sharing and uniform loads the following algorithm is applied for each $k = 1, \ldots, K$ and $m = 1, \ldots, M$.

Algorithm for ESR

Step 0. For all $p = 1, \ldots, P$ set $\underline{Y}^{kmp} = Y_n^{kmp} + d_n^p$, $\underline{Y'}^{kmp} = {Y'}_n^{kmp}$, $P_n^{kmp} = {P'}_n^{kmp} = 0$, $\underline{I}^{kmp} = I_n^{kmp}$, and $\underline{I'}^{kmp} = {I'}_n^{kmp}$.
Also, set $\underline{C}^{km} = C^{km}$, $\underline{C'}^{km} = C'^{km}$, and $j = P$.

Step 1. Order the products by decreasing value of their shortfall after demand is realized. Let $p(1), \ldots, p(j)$ denote that ordering, that is $\underline{Y}^{kmp(1)}$ is the maximum value and $\underline{Y}^{kmp(j)}$ is the minimum.
Set $l = 1$ and $\underline{Y}^{km(j+1)} = \underline{Y'}^{km(j+1)} = 0$.

Step 2. Let $H = \underline{Y}^{kmp(l)} - \underline{Y}^{kmp(l+1)}$. If $H \neq 0$, set $H' = \underline{Y'}^{kmp(l)} - \underline{Y'}^{kmp(l+1)}$ and go to Step 4. Otherwise, continue.

Step 3. If $l < j$, set $l = l + 1$ and go to Step 2. Otherwise, STOP.

Step 4. We have the first l products tied. Therefore the production decision and its derivative are updated as follows:

$$P_n^{kmp(i)} = P_n^{kmp(i)} + \underline{P}^{kmp(i)} \quad \text{for } i = 1, \ldots, l. \tag{23}$$

$${P'}_n^{kmp(i)} = {P'}_n^{kmp(i)} + \underline{P'}^{kmp(i)} \quad \text{for } i = 1, \ldots, l. \tag{24}$$

where

$$\underline{P}^{kmp} = \min\{H, \underline{I}^{(km)^+p}, \underline{C}^{km}/l\} \tag{25}$$

and

$$\underline{P'}^{kmp} = \begin{cases} H' & \text{if bound by the jump size} \\ \underline{I'}_n^{(km)^+p} & \text{if bound by inventory} \\ \underline{C'}^{km}/l & \text{if bound by capacity} \end{cases} \tag{26}$$

Step 5. Update the shortfalls, inventories, and available capacity.

$$\begin{aligned} \underline{Y}^{kmp(i)} &= \underline{Y}^{kmp(i)} - \underline{P}^{kmp(i)} \\ \underline{I}^{(km)^+p(i)} &= \underline{I}^{(km)^+p(i)} - \underline{P}^{kmp(i)} \quad \text{for } i = 1, \ldots, l, \\ \underline{C}^{km} &= \underline{C}^{km} - \sum_{i=1}^{l} \underline{P}^{kmp(i)} \end{aligned} \tag{27}$$

The derivatives are

$$\begin{aligned} \underline{Y}'^{kmp(i)} &= \underline{Y}'^{kmp(i)} - \underline{P}'^{kmp(i)} \\ \underline{I}'^{(km)^+p(i)}_n &= \underline{I}'^{(km)^+p(i)}_n - \underline{P}'^{kmp(i)} \quad \text{for } i = 1, \ldots, l. \\ \underline{C}'^{km} &= \underline{C}'^{km} - \sum_{i=1}^{l} \underline{P}'^{kmp(i)} \end{aligned} \tag{28}$$

Step 6. If $\underline{C}^{km} = 0$, STOP. The total production for level k and stage m is bound by capacity. Otherwise, continue.

Step 7. For each $i = 1, \ldots, l$, if $\underline{I}^{(km)^+p(i)} = 0$ remove product $p(i)$ from the list and set $j = j - 1$. If $j = 0$, STOP. The total production for level k and stage m does not use up all capacity. Otherwise, go to Step 1.

In order to validate the above procedure two results specific to the ESR are needed. Consider the PS mode. Recall that the final production decision for any period is computed iteratively. Each time a new amount is added its respective derivative has to be computed also as in (25) and (26). The derivative in (26) is only valid if the parameter l does not change with small changes of the control parameters. The following two results establish this.

Theorem A.1 If $\{d_n^p, n = 1, 2, \ldots, p = 1, 2, \ldots, P\}$ are independent and each d_n^p has a density on $(0, \infty)$, the ordering generated in Step 1 at the end of the first iteration remains unchanged with probability one, (w.p.o.), in a neighborhood of the base stock levels.

Proof: Assume first that for a given vector z there are no ties in the shortfall quantities after demand is realized at the beginning of period n. Under this assumption, there exists a $\delta = \min_i\{\underline{Y}^{kmp(i)} - \underline{Y}^{kmp(i+1)}\} > 0$. Therefore, there exists an $\epsilon > 0$ such that a change smaller than ϵ in any of the components of z will produce a change in shortfall variables which is smaller than δ. Thus, the ordering remains unchanged.

Now consider the case where at least two shortfall quantities are tied at period n. A tie in $\underline{Y}$ can only occur if they were equal at the end of period $n-1$ and demand in period n was zero for at least those two products. However, if at least two shortfall quantities were made equal in period $n-1$ for a given vector z, this means that with probability one they will also be made equal in period $n-1$ in a neighborhood of z because with probability one either (1) capacity was exhausted but there was some inventory not used up for at least those two products in period $n-1$ with vector z, or (2) capacity was not exhausted and the shortfalls were made zero on period $n-1$.

□

Theorem A.2 Under the same assumptions of Proposition A.1, at each iteration of the algorithm, the number of tied products, l, at the beginning of Step 4 remains unchanged w.p.o. in a neighborhood of z.

Proof: The proof proceeds by induction on the number of iterations of the above algorithm. Denote by $l(r)$ the number of products tied at the beginning of Step 4 during iteration r. Proposition A.1 establishes the result for the first iteration. Now assume that at some iteration $r-1$, $l(r-1)$ is invariant relative to sufficiently small changes in the base stock variables. We want to see what happens to $l(r)$.

From the $l(r-1)$ products, let us assume that some $\lambda(r-1) \leq l(r-1)$ remain tied upon application of Step 4. With probability one there are only two ways under which this can happen. (1) The $\lambda(r-1)$ have each enough inventory so that H can be assigned to them *and* adequate capacity is available to do so. In this situation, for a sufficiently small neighborhood of z, the inventory and the capacity available will still allow the same $\lambda(r-1)$ products to remain tied. (2) There is a bound in capacity (affecting all $\lambda(r-1)$ products). However, a bound in capacity will remain for a sufficiently small neighborhood of z, w.p.o. Also, $\lambda(r-1) < l(r-1)$ if and only if some products have their share reduced due to insufficient inventory, which again will remain so w.p.o. in a neighborhood of z. Thus, $\lambda(r-1)$ is invariant in a neighborhood of z.

The value of $l(r)$ depends on $\lambda(r-1)$ and on what happens in Step 4. If the production decision for the $\lambda(r-1)$ products is bound by H, then there is more capacity available to execute a new iteration. Also, a new iteration will take place if there are still products with non-zero shortfalls. In this case $l(r)$ will be the sum of $\lambda(r-1)$ with the products that were tied in second place before iteration $r-1$. This is because the shortfall of the $\lambda(r-1)$ products were brought down to the same levels as these. By similar arguments, the number of products tied for second place does not change for sufficiently small changes of any base stock variable.

There is also the possibility that the tie of the $\lambda(r-1)$ products occurred due to a bound in capacity. In this situation, if $\lambda(r-1) = l(r-1)$, there will be no more iterations, since the available capacity at the beginning of iteration $r-1$ will be exhausted during this iteration. It can also happen that $\lambda(r-1) < l(r-1)$, since some of the products may have been bounded by inventory. In this situation, there will be capacity available to perform at least one more iteration, and $l(r) = \lambda(r-1)$.

Therefore, the result follows.

□

Propositions A.1 and A.2 are valid for the TS mode without change. These two results establish that l in (26) is not a function of the base stock variables.