Quantity Flexibility Contracts and Supply Chain Performance

A. A. Tsay • W. S. Lovejoy
Department of Operations & Management Information Systems, Leavey School of Business, Santa Clara University, Santa Clara, California 95053-0382
School of Business Administration, University of Michigan, Ann Arbor, Michigan 48109-1234

The Quantity Flexibility (QF) contract is a method for coordinating materials and information flows in supply chains operating under rolling-horizon planning. It stipulates a maximum percentage revision each element of the period-by-period replenishment schedule is allowed per planning iteration. The supplier is obligated to cover any requests that remain within the upside limits. The bounds on reductions are a form of minimum purchase commitment which discourages the customer from overstating its needs. While QF contracts are being implemented in industrial practice, the academic literature has thus far had little guidance to offer a firm interested in structuring its supply relationships in this way. This paper seeks to address this need, by developing rigorous conclusions about the behavioral consequences of QF contracts, and hence about the implications for the performance and design of supply chains with linkages possessing this structure. Issues explored include the impact of system flexibility on inventory characteristics and the patterns by which forecast and order variability propagate along the supply chain. The ultimate goal is to provide insights as to where to position flexibility for the greatest benefit, and how much to pay for it.
(*Supply Chain Management*; *Supply Contracts*; *Quantity Flexibility*; *Forecast Revision*; *Materials Planning*; *Bullwhip Effect*)

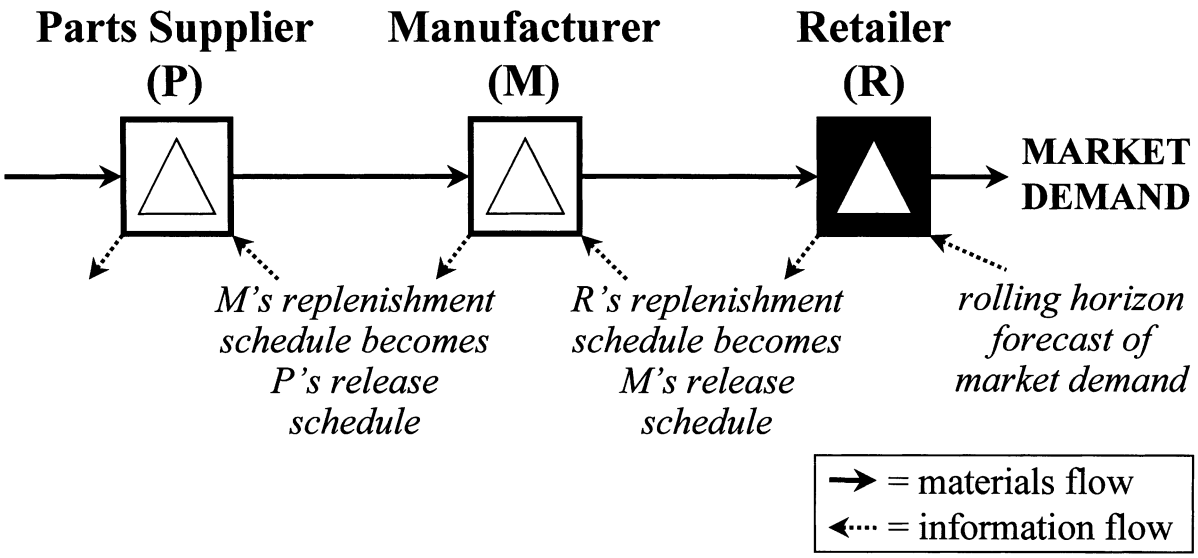
1. Introduction

Many modern supply chains operate under decentralized control for a variety of reasons. For example, outsourcing of various aspects of production is currently a popular business model in many industries (cf. Farlow et al. 1995, Iyer and Bergen 1997), which automatically distributes decision-making authority. Even for highly vertically integrated firms, today's characteristically global business environments often result in multiple sites worldwide working together to deliver product, while reporting to different organizational functions or units within the corporation. Operational control of these sites may be intentionally decentralized for informational or incentive considerations. However, decentralization is not without risks. For expository purposes, we describe some of these in the context of the single-product, serial supply chain depicted in Figure 1. Each node represents an independently managed organization, and each pair of consecutive nodes is a distinct supplier-buyer relationship.

To reconcile manufacturing/procurement time-lags with a need for timely response, agents within such supply chains often commit resources to production quantities based on forecasted, rather than realized demand. A period-by-period replenishment schedule (e.g., six months' worth of monthly volume estimates) is a common format by which many firms communicate information about future purchases to their supply partners. Rolling horizon updating is a standard operational means of incorporating new information as it accrues over time. For example, each period the

1523-4614/99/0102/0089$05.00

Figure 1 Decentralized Supply Chain



retailer creates a forecast of the uncertain and potentially non-stationary market demand e.g., [100, 120, 110, . . .] where the 100 denotes the current period's demand, 120 is an estimate of the next period's demand, and so on. Based on this, the retailer provides to the manufacturer a schedule of desired replenishments, e.g., [50, 150, 90, . . .], where the numbers may differ from the market forecast due to whatever inventory policy the retailer may use, and any stock carried over from the previous period. The manufacturer treats this schedule as its own "demand forecast" and in turn creates a replenishment schedule for the parts supplier to fill, and so on. This information flow is represented by the dotted lines in Figure 1. We assume that each party knows only the schedule provided by its immediate customer, and is only concerned with its own cost performance.

Such estimates are intended to assist an upstream supplier's capacity and materials planning. However, buyers commonly view them as a courtesy only, and indeed craft the supply contracts to preserve this position. To some buyers this presents an opportunity to inflate these figures as a form of insurance, only to later disavow any undesired product (cf. Lee et al. 1997). A careful supplier must then deflate the numbers to avoid over-capacity and inventory. This game of mutual deception may be individually rational given the circumstances, but increases the uncertainties and costs in the system (cf. Magee and Boodman 1967, Lovejoy 1998).

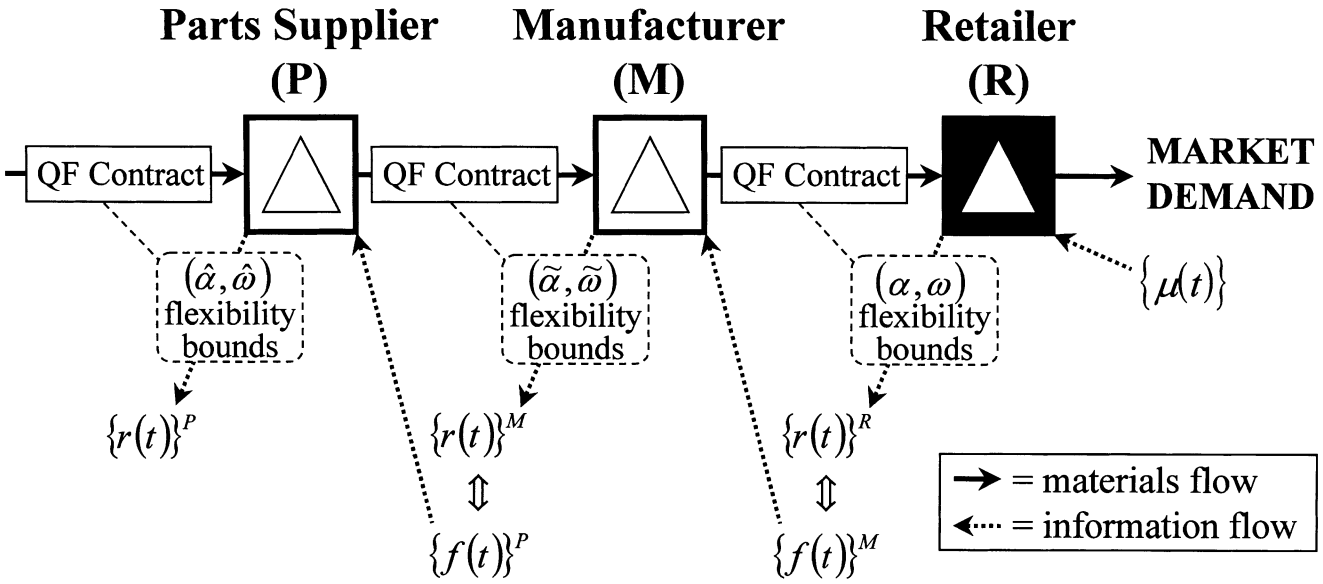
Various remedies to this well-known inefficiency have been attempted, a number of which are noted in §2. One approach that has become popular in many industries is the Quantity Flexibility (QF) contract, which attaches a degree of commitment to the forecasts by installing constraints on the buyer's ability to revise them over time. The extent of revision flexibility is defined in percentages that vary as a function of the number of periods away from delivery. This is made concrete in Figure 2.

Since individual nodes share common structure and we may wish to consider chains of considerable length, we use common variable names for node attributes wherever possible, and associate them with specific parties via superscripts (P, M, and R in the example in Figure 2).

At each time period, indexed by t, the period-by-period stochastic market demand is described by $\{\mu(t)\} = [\mu_0(t), \mu_1(t), \mu_2(t), \ldots]$, where

$$\begin{aligned} \mu_0(t) &= \text{actual market demand occurring in period } t \\ \mu_j(t) &= \text{estimate of period } (t + j) \text{ demand,} \\ &\quad \text{for each } j \geq 1. \end{aligned} \tag{1}$$

Figure 2 Decentralized Supply Chain with QF Contracts



The statistical structure of this process is known to the retailer, who incorporates it into supply planning. The retailer in turn provides the manufacturer with a *replenishment schedule* vector $\{r(t)\}^R = [r_0(t), r_1(t), r_2(t), \ldots]^R$, where

$$r_0(t) = \text{actual purchase made in period } t \quad (2)$$

$$r_j(t) = \text{estimate of purchase to be made in period } (t + j), \text{ for each } j \geq 1.$$

This becomes the upstream supplier's *release schedule* vector, denoted $\{f(t)\}^M = [f_0(t), f_1(t), f_2(t), \ldots]^M$, where

$$f_0(t) = \text{quantity sold in period } t \quad (3)$$

$$f_j(t) = \text{estimate of quantity to be sold in period } (t + j), \text{ for each } j \geq 1.$$

Thus far we have simply formalized the information flow described in Figure 1. Next, we consider the QF contract between each pair of nodes. The manufacturer-retailer QF contract is parametrized by (α, ω), where $\alpha = [\alpha_1, \alpha_2, \ldots]$ and $\omega = [\omega_1, \omega_2, \ldots]$. This places bounds on how the retailer may revise $\{r(t)\}^R$ going forward in time. Specifically, for each t and $j \geq 1$:

$$[1 - \omega_j]r_j(t) \leq r_{j-1}(t + 1) \leq [1 + \alpha_j]r_j(t). \quad (4)$$

That is, the estimate for future period $(t + j)$ cannot be revised upward by a fraction of more than α_j or downward by more than ω_j. Contingent on this, the contract stipulates that the retailer's eventual orders will all be filled with certainty.[1]

[1]It is natural to expect that any reasonable flexibility agreement should be such that the interval bounding a given future period's purchase becomes progressively smaller as that period approaches. Although not readily apparent from Equation (4), the QF arrangement has this feature. For instance, according to Equation (4), in planning for period $(t + 2)$ the retailer's period t estimate $r_2(t)$ constrains the period $(t + 1)$ estimate by

$$[1 - \omega_2]r_2(t) \leq r_1(t + 1) \leq [1 + \alpha_2]r_2(t).$$

In turn, by another application of Equation (4), $r_1(t + 1)$ is known to constrain the eventual purchase $r_0(t + 2)$ by

$$[1 - \omega_1]r_1(t + 1) \leq r_0(t + 2) \leq [1 + \alpha_1]r_1(t + 1).$$

Together these define from the period t perspective the window within which the eventual purchase must fall:

$$[1 - \omega_1][1 - \omega_2]r_2(t) \leq r_0(t + 2) \leq [1 + \alpha_1][1 + \alpha_2]r_2(t).$$

Hence, the window bounding the actual purchase evolves from $[(1 - \omega_1)(1 - \omega_2)r_2(t), (1 + \alpha_1)(1 + \alpha_2)r_2(t)]$ to $[(1 - \omega_1)r_1(t + 1), (1 + \alpha_1)r_1(t + 1)]$. Assuming Equation (4) is observed, the latter window

Because $\{f(t)\}^M = \{r(t)\}^R$, Equation (4) means the manufacturer can be sure that revisions to estimates of its "demand" will obey

$$[1 - \omega_j]f_j(t) \leq f_{j-1}(t - 1) \leq [1 + \alpha_j]f_j(t) \quad (5)$$

and is contractually obligated to support the resulting sequence of purchases. The manufacturer in turn passes a replenishment schedule, denoted $\{r(t)\}^M$, to its own supplier. This will obey constraints analogous to Equation (4) above, except with flexibility parameters $(\tilde{\alpha}, \tilde{\omega})$. Thus the parts supplier knows that revisions to $\{f(t)\}^P$ will stay within the $(\tilde{\alpha}, \tilde{\omega})$ bounds, and in turn passes upstream the replenishment schedule $\{r(t)\}^P$ (staying within the $(\hat{\alpha}, \hat{\omega})$ bounds), and so on. This exercise is repeated each period, with all estimates updated in rolling-horizon fashion.

QF contracts are intended to provide a benefit to each party. The supplier formally guarantees the buyer a specific safety cushion in excess of estimated requirements. In return, the buyer agrees to limit its order reductions, essentially a form of minimum purchase agreement. In this way the buyer accepts some of the downside demand risk which, were forecasts completely divorced of commitment, would be left to the supplier. Mutual agreement on the significance of forecasts improves the planning capabilities of both parties. Any favoritism expressed by this arrangement can be mitigated in setting the flexibility limits, as we will demonstrate.

The emergence of QF contracts as a response to certain supply chain inefficiencies is described in Lee et al. (1997). Sun Microsystems uses QF contracts in its purchase of monitors, keyboards, and various other workstation components (cf. Farlow et al. 1995). Nippon Otis, a manufacturer of elevator equipment, implicitly maintains such contracts with Tsuchiya, its supplier of parts and switches (cf. Lovejoy 1998). Solectron, a leading contract manufacturer for many electronics firms, has recently installed such agreements with both its customers and its raw materials suppliers (Ng 1997), implying that benefits may accrue to either end of such a contract. QF-type contracts have also been used by Toyota Motor Corporation (Lovejoy 1998), IBM (Connors et al. 1995), Hewlett Packard, and Compaq (Faust 1996). A similar structure, called a "Take-or-Pay" provision, is often embedded in long-term supply contracts for natural resources (cf. Masten and Crocker 1985, Mondschein 1993, National Energy Board 1993). In addition to being used to govern relations between separate companies, QF structures have also appeared at the interface between the manufacturing and marketing/sales functions (taking the role of supplier and buyer, respectively) within single firms (cf. Magee and Boodman 1967).

While QF contracts are being implemented in industrial practice, the academic literature has thus far had little guidance to offer a firm interested in structuring its supply relationships in this way. This paper seeks to address this need, by pursuing the following objectives: (a) to provide a formal framework for the analysis of such contracts, with explicit consideration of the non-stationarity in demand that drives the desire for flexibility; (b) to propose behavioral models, i.e., forecasting and ordering policies, for buyers who are subject to such constraints in their procurement planning, and for suppliers who promise such flexibility to their customers; and (c) to link these behaviors to local and systemwide performance (e.g., inventory levels and order variability), and therefore guide the negotiation of contracts. In the following discussion, our intent is not necessarily to advocate the QF contract, but to provide conclusions about the implications of its usage.

Section 2 positions this paper in the literature. Sections 3 and 4 introduce the modeling primitives. We will analyze complex systems such as the one in Figure 2 by decomposing the supply chain into modules of simpler structure. All interior nodes, meaning those

(one period prior to purchase) is contained entirely in the former (two periods prior). More generally, requiring Equation (4) at every revision generates a sequence of nested intervals that ultimately converge to the actual purchase. This will become clear when, in §3, we formalize this "cumulative" perspective on the flexibility terms of the contract, taking an alternative view of the per-period incremental flexibilities in Equation (4). Both representations have been observed in industry. The incremental form would be preferred by a buyer, since this constrains the successive updating of its replenishment schedules. The cumulative form would be used by a supplier, since this renders future capacity needs more transparent. But as these forms are mathematically equivalent, our results apply equally well to each.

which have QF contracts on both their input and output sides, can be represented by one node type. Here we will derive a reasonable inventory policy that reconciles the constraints and the commitments implied by the input and output flexibility profiles. Another node type represents the node at the market interface, which has a QF contract on its input side only, but has statistical knowledge about demand on its output side. Here we will suggest an ordering policy that takes into account the market demand dynamics, the relative costs of holding and shortage, and the input-side flexibility parameters. The decision problems of each node type are formidable due to the large number of decision variables and the statistical complexity of customer ordering, so we will utilize heuristic policies. This enables us to explore in §5 the performance properties of supply chains controlled with QF contracts. We investigate the implications of flexibility characteristics for both inventory and service, as well as how order variability propagates along the supply chain. Once these relationships are established, the issue of contract design, i.e., the choice of flexibility parameters, may be pursued. In particular, §6 examines the value of flexibility in the supply chain. We conclude in §7 with discussion of these results and implementation issues. For clarity of exposition, all proofs are deferred to Appendix 1.

2. Literature Review

It is not generally the case that a supply chain composed of independent agents acting in their own best interests will achieve systemwide efficiency, often due to some incongruence between the incentives faced locally and the global optimization problem. In our single-product setting in which the only uncertainty is in the market demand and the only decision is product quantity, this is because overstock and understock risks are visited differently upon the individual parties.

One response is to reconsider the nature of the supply contracts along the chain. (See Tsay et al. (1999) for a recent review.) The general goal is to install rules for materials accountability and/or pricing that will guide autonomous entities towards the globally desirable outcome (cf. Whang 1995, Lariviere 1999). This type of approach recurs in a broad range of settings, for example the economic literature on "vertical restraints" (cf. Mathewson and Winter 1984, Tirole 1988, Katz 1989), the marketing literature of "channel coordination" (e.g., Jeuland and Shugan 1983, Moorthy 1987), and agency theory (cf. Bergen et al. 1992, Van Ackere 1993). Recent examples in the multi-echelon inventory literature include Lee and Whang (1997), Chen (1997), and Iyer and Bergen (1997). When recourse in light of information changes is admitted, results are limited to single-period settings. Contractual structures that have been shown to replicate the efficiency of centralized control in that context include buyback/return arrangements (cf. Pasternack 1985, Donohue 1996, Kandel 1996, Ha 1997, Emmons and Gilbert 1998) and the QF contract (cf. Tsay 1996). In all the above works, information about market demand is common to all parties.

Some flexible supply contracts with risk-sharing intent have been studied in more realistic settings. Bassok and Anupindi (1995) consider forecasting and purchasing behavior when the buyer initially forecasts month-by-month demand over an entire year and then may revise each month's purchase once within specified percentage bounds. Bassok and Anupindi (1997a) analyze a contract which specifies that cumulative purchases over a multi-period horizon exceed a previously (and exogenously) specified quantity, a form of minimum-purchase agreement. Bassok and Anupindi (1997b) study a rolling-horizon flexibility contract similar to our QF structure, focusing on the retailer's ordering behavior when facing an independent and stationary market demand process. Eppen and Iyer (1997) analyze "backup agreements" in which the buyer is allowed a certain backup quantity in excess of its initial forecast at no premium, but pays a penalty for any of these units not purchased. These models do not attempt to demonstrate efficiency of the contract, instead focusing on the operational implications of the specified prices and constraints for the buyer. No consideration is made for how the supplier might best support its obligations, as the upstream decision problem is rendered difficult by the statistical complexity of the demand that is transmitted through. Moreover, the information structure is kept simplified, with the

forecast for a given period's demand updated at most once, if at all.

What little is known about ongoing relationships with information updating is limited to a single node setting with very stylized demand models. For example, Azoury (1985), Miller (1986) and Lovejoy (1990, 1992) consider demand whose structure is known except for a single uncertain parameter that is updated each period in a very specific way (e.g., Bayesian updating, or exponentially smoothed mean). Base stock policies with moving targets turn out to be optimal or near-optimal. While these are quite powerful results, they apply only when delivery is immediate. When lead times are non-zero, a properly made current-period decision would need to account for the behavior of demand over several subsequent periods. Even with these relatively straightforward demand models, the statistics required for the policy calculations become computationally formidable. This is the case even absent supply side flexibility.

Industrially, rolling horizon planning is the most common approach to non-stationary problems with positive lead times, a prominent application being Material Requirements Planning (MRP). As in our setting, MRP seeks a supply schedule that attends to a period-by-period schedule of materials needs. Baker (1993) provides a recent review of lot-sizing studies, for both single and multiple level models. Numerical simulation is the predominant means of evaluating algorithm performance, largely due to the complexity of the setting.

Our primary interest is in the way these studies model demand and how demand information is incorporated into the planning process. In general, the installed policies rarely explicitly account for the temporal dynamics of the underlying demand. The accuracy of the forecasts may be specified as a forecast error that gets incorporated into safety stock factors for each period (cf. Miller 1979, Guererro et al. 1986). However, there is no consideration for how each forecast might change from one period to the next. Typically, either deterministic end demand is assumed (in which case forecast updating is not an issue) or the forecast is frozen over the planning horizon. Either way, the response is reactive. Finding that the "stochastic, sequential, and multi-dimensional nature" of this class of problem defies an optimization-based approach, Heath and Jackson (1994) suggests that this approximates "reasonable" decision-making. We share this view in our pursuit of insights for industrial application.

One limitation of the MRP framework and other conventional models is the notion of a fixed, or what we call "rigid", lead time. In many real systems, the lead times that are loaded into the materials planning model are exaggerated to hedge against uncertainties in the supply process (e.g., queuing or raw materials shortages) (cf. Karmarkar 1989). The QF contract formalizes the reality that a single lead time alone is an inadequate representation of many supply relationships, as evinced by the ability of buyers to negotiate quantity changes even within quoted lead times.

This paper seeks insights for a setting including all of the above features: resources which require advance commitments, non-stationary demand about which information evolves over time, and the possibility of revising the commitments within bounds in reaction to information changes. Because this work evolved from collaboration with an industrial partner competing in a volatile industry, we have avoided as much as possible any dependence on specific statistical assumptions about market demand. In this context, optimal policies are unknown, so we seek behavioral models that mimic rational but potentially suboptimal policymakers. We also consider the perspectives of both parties to each contract. In addition to specifying the buyer's behavior, we recommend how a supplier might economically deliver the promised flexibility, and characterize how the costs of both parties vary with the contract parameters.

3. Analysis of an Interior Node

We first specify the structure and behavior of a *flex node*, which we use to represent an agent which has QF contracts with both its supplier and customer (e.g., the manufacturer or the parts supplier in Figure 2). In §4 we will introduce the *semi-flex node* to handle the case when the customer-side interface is unstructured. We will model multi-stage supply chains by linking these modular units.

At each period t, the node receives $\{f(t)\} = [f_0(t), f_1(t),$

$f_2(t), \ldots]$ as defined in Equation (3), the *release schedule* delineating the downstream node's needs. The node will in turn provide its upstream supplier with a *replenishment schedule* $\{r(t)\} = [r_0(t), r_1(t), r_2(t), \ldots]$ as defined in Equation (2). Note that one node's release schedule is simultaneously the downstream node's replenishment schedule. $I(t)$ is the node's period t ending stock, calculated as $I(t) = I(t-1) + r_0(t) - f_0(t)$. All quantities are measured in end-item equivalents.

The input and output QF parameters are denoted as $(\alpha^{\text{in}}, \omega^{\text{in}})$ and $(\alpha^{\text{out}}, \omega^{\text{out}})$ respectively, superscripted to signify the node's point of reference. Restating Equations (4) and (5) with this notation gives the following ground rules for schedule revisions, termed *Incremental Revision* (IR) constraints:

$$[1 - \omega_j^{out}]\, f_j(t) \le f_{j-1}(t+1) \le [1 + \alpha_j^{out}]\, f_j(t), \quad \text{for all } t, \text{ each } j \ge 1 \tag{6}$$

$$[1 - \omega_j^{in}]\, r_j(t) \le r_{j-1}(t+1) \le [1 + \alpha_j^{in}]\, r_j(t), \quad \text{for all } t, \text{ each } j \ge 1. \tag{7}$$

Naturally, we assume $\alpha_j^{\text{in}}, \alpha_j^{\text{out}} \ge 0$ and $0 \le \omega_j^{\text{in}}, \omega_j^{\text{out}} \le 1$. Since these IR constraints are assumed to hold in all future iterations, the current period's $f_j(t)$ suggests bounds on $f_0(t+j)$, the actual customer purchase in period $(t+j)$. Specifically, Equation (6) implies

$$[1 - \Omega_j^{out}]\, f_j(t) \le f_0(t+j) \le [1 + A_j^{out}]\, f_j(t), \quad \text{for all } t, \text{ each } j \ge 1, \text{ where} \tag{8}$$

$$1 - \Omega_j^{out} \doteq \prod_{q=1}^{j} (1 - \omega_q^{out}) \text{ and}$$

$$1 + A_j^{out} \doteq \prod_{q=1}^{j} (1 + \alpha_q^{out}). \tag{9}$$

Similarly, on the replenishment side, Equation (7) implies

$$[1 - \Omega_j^{in}]\, r_j(t) \le r_0(t+j) \le [1 + A_j^{in}]\, r_j(t), \quad \text{for all } t, \text{ each } j \ge 1, \text{ where} \tag{10}$$

$$1 - \Omega_j^{in} \doteq \prod_{q=1}^{j} (1 - \omega_q^{in}) \text{ and}$$

$$1 + A_j^{in} \doteq \prod_{q=1}^{j} (1 + \alpha_q^{in}). \tag{11}$$

Equations (8) and (10) are termed *Cumulative Flexibility* (CF) constraints. Clearly A_j^{in}, Ω_j^{in}, A_j^{out} and Ω_j^{out} are non-negative and increasing in j, indicating that greater cumulative flexibility is available for periods further out, which is helpful since longer-term projections are generally less informative. As noted in §1, the IR and CF systems of constraints are mathematically equivalent, so that QF contracts may be stated either way. Each perspective has certain advantages, and throughout this paper we will use whatever form is more convenient for the given context.

Replenishment Planning at a Flex Node

The flex node decision problem is to construct the $\{r(t)\}$ to be passed upstream, given the $\{f(t)\}$ faced and the local inventory level. The only policies we deem "admissible" are those that uphold the release-side contract without violating the replenishment-side contract. That is, an admissible policy is one for which, given any arbitrary sequence of $\{f(t)\}$ whose updates obey Equation (6), (a) updates to $\{r(t)\}$ obey (7), and (b) coverage is provided (i.e., $I(t-1) + r_0(t) \ge f_0(t)$ for all t).

The stochastic optimization problem to be solved at period t, called program (F), is:

$$\min_{\{r(t)\},(r_0(t+1),\ldots,r_0(t+H))} \sum_{j=0}^{H} E[G(I(t+j)) \mid \{f(t)\}]$$

subject to (12)

$$I(t+j) = I(t+j-1) + r_0(t+j) - f_0(t+j) \quad \text{for } j = 0, \ldots, H \tag{13}$$

$$I(t+j) \ge 0 \quad \text{for } j = 0, \ldots, H \tag{14}$$

$$(1 - \omega_{j+1}^{in}) r_{j+1}(t-1) \le r_j(t) \le (1 + \alpha_{j+1}^{in}) r_{j+1}(t-1) \quad \text{for } j = 0, \ldots H-1 \tag{15}$$

$$(1 - \Omega_j^{in}) r_j(t) \le r_0(t+j) \le (1 + A_j^{in}) r_j(t) \quad \text{for } j = 0, \ldots, H. \tag{16}$$

$G()$ is some convex cost function (minimized at zero) that is charged against future ending stock levels, so the objective is to minimize expected total cost over H periods for some fixed H. This problem is stochastic because, as suggested by balance Equation (13), $G(I(t$

$+ j)$) depends on the random variables ($f_0(t + 1), \ldots,$ $f_0(t + j)$) conditional on $\{f(t)\}$. The decision variables are $\{r(t)\}$ (the current replenishment schedule, which is all that must be formally stated to the supplier) and, for internal planning purposes, ($r_0(t + 1), \ldots, r_0(t + H)$) (the sequence of intended future purchases, which still enjoys some opportunity for revision).[2] Equation (14) enforces the coverage commitment, Equation (15) states what $\{r(t)\}$ is allowed given $\{r(t - 1)\}$ and the input side IR constraint[3] and Equation (16) then computes the CF bounds on the node's future purchases based on the $\{r(t)\}$ chosen.

Exact solution to (F) is difficult for two primary reasons. First, dimensionality of the decision space is very large, with each decision variable subject to constraints. In particular, Equation (16) acts like a capacity constraint, which precludes closed-form solution in a stochastic setting (cf. Federgruen and Zipkin 1986, Tayur 1992). Here, the added wrinkle is that future capacity limits can not only vary by period, but are actually decision variables that can be dynamically adjusted. Second, and more problematically, the statistical properties of the random variables ($f_0(t + 1)$, $f_0(t + 2)$. . .) are in general very complex, since not only are they ultimately derived from a non-stationary and multivariate market demand/forecast process, they are filtered through the inventory policies of one or more intermediaries (see Figure 2) and all intervening QF constraints. Hence, while the expectation in the objective function may be well-defined in theory, in practice it is intractable, rendering the search for an optimal policy problematic. However, we can identify an open-loop feedback control (OLFC) policy (cf. Bertsekas 1976) that has some satisfying mathematical and intuitive properties. In an OLFC policy, at each period a sequence of actions is computed looking forward and *assuming perfect information*, and the first action is invoked. The information is then updated the following period and another forward-looking sequence of actions is computed, and so forth. In this way, a complex stochastic dynamic program is approximated by a series of deterministic models. Such policies are commonplace in problems with complex or incompletely specified process dynamics. The conventional wisdom is that OLFC is a fairly satisfactory mode of control for many problems. This, in fact, is the approach taken by industry practitioners in their adoption of the MRP paradigm.

To construct an OLFC policy for the control of a flex node, we suppress explicit consideration of future updates to $\{f(t)\}$. Instead, the contractual coverage obligation suggests fixed targets to which the flex node can position. In particular, this node must fill any orders provided that the customer's revisions do not exceed the defined bounds.[4] The resulting deterministic problem, which we denote program (F-OLFC) is:

$$min_{\{r(t)\},(r_0(t+1),\ldots,r_0(t+h))} \sum_{j=0}^{h} G(I(t + j)) \quad \text{subject to}$$

$$I(t + j) = I(t + j - 1) + r_0(t + j) - (1 + A_j^{out}) f_j(t) \quad \text{for } j = 0, \ldots, h \tag{17}$$

$$I(t + j) \geq 0 \quad \text{for } j = 0, \ldots, h \tag{18}$$

$$(1 - \omega_{j+1}^{in}) r_{j+1}(t - 1) \leq r_j(t) \leq (1 + \alpha_{j+1}^{in}) r_{j+1}(t - 1) \quad \text{for } j = 0, \ldots h - 1 \tag{19}$$

$$(1 - \Omega_j^{in}) r_j(t) \leq r_0(t + j) \leq (1 + A_j^{in}) r_j(t) \quad \text{for } j = 0, \ldots, h. \tag{20}$$

$f_0(t + j)$ has been replaced with $(1 + A_j^{\text{out}})\, f_j(t)$ for reasons discussed above. This program also considers a potentially shorter time window, of length $h \leq H$, as a practical consideration. Naturally, this assumes that all flexibility parameters are well-defined for an h-period outlook.

Proposition 1. *The following* $\{r(t)\}$ *is optimal for program* (F-OLFC), *and is admissible:*

$$r_j(t) \doteq max[T_j(t), (1 - \omega_{j+1}^{in}) r_{j+1}(t - 1)] \quad \textit{for } j = 0, \ldots, h, \textit{ where} \tag{21}$$

[2] $\{r(t + 1)\}$, $\{r(t + 2)\}$, etc. need not be specified at this point since any influence they may have are reflected implicitly through Equation (16). Values consistent with any feasible solution can be inferred if desired.

[3] $\{r(t - 1)\}$ is data resulting from the period $(t - 1)$ planning iteration.

[4] This is not the same as guaranteeing to meet all customer demand, since the allowable order is groomed in advance by the flexibility constraints, i.e., it is a truncated version of what the customer might desire otherwise.

$$\bullet\ T_j(t) \doteq \frac{(1 + A_j^{out})\, f_j(t) - l_j(t)}{1 + A_j^{in}} \tag{22}$$

$$\bullet\ l_j(t) \doteq \begin{cases} I(t-1) & \text{for } j = 0 \\ [l_{j-1}(t) + (1 - \Omega_{j-1}^{in}) r_{j-1}(t) \\ \quad - (1 + A_{j-1}^{out}) f_{j-1}(t)]^+ & \text{for } j \geq 1. \end{cases} \tag{23}$$

This is named the *Minimum Commitment* (MC) policy as the present decisions minimize commitment to future costs subject to supporting service obligations. $(r_0(t+1), \ldots, r_0(t+h))$ is not stated explicitly since only $\{r(t)\}$ needs to be provided to the supplier (see Appendix 1 for the complete optimal solution). $l_j(t)$ is the period t projection of inventory assured to be available at period $(t + j)$, anticipating the future actions of the OLFC-optimal decision rule. From here on, we assume that flex nodes use the MC policy. The next section investigates the relationships among flexibility, inventory, and information subject to this behavioral assumption.

The Effect of Flexibility Disparities Across a Flex Node

This section makes rigorous the notion that inventory results from a disparity between input and output flexibility. The intuition is as follows. The goal is for supply to track customer orders as closely as possible. Because of forecast updating, those orders are moving targets and the output flexibility defines the range of potential movement. Meanwhile, the input flexibility represents the node's tracking ability. A node with difficulty in matching upside movement compensates by increasing its general positioning. Inventory accrues when the node is unable to pare down its replenishments as quickly as the customer is allowed to reduce its own requirements.

Proposition 2 demonstrates that a flex node which possesses more flexibility (in CF form) in its supply process than it offers its customer can meet all obligations with zero inventory.

Proposition 2. *If (a) updates to $\{f(t)\}$ obey IR constraints, (b) the MC policy is used, (c) $I(0) = 0$, and (d) $(A^{\text{in}}, \Omega^{\text{in}}) \geq (A^{\text{out}}, \Omega^{\text{out}})$, then $I(t) = 0$ for all t. In the special case that $(A^{\text{in}}, \Omega^{\text{in}}) = (A^{\text{out}}, \Omega^{\text{out}})$, then $r_j(t) = f_j(t)$ for all $j \geq 0$, $t \geq 1$.*

Note that $(\alpha^{\text{in}}, \omega^{\text{in}}) \geq (\alpha^{\text{out}}, \omega^{\text{out}})$ is sufficient, but not necessary, to guarantee that $(A^{\text{in}}, \Omega^{\text{in}}) \geq (A^{\text{out}}, \Omega^{\text{out}})$. The result holds under the latter, less restrictive condition.

This proposition provides insight into one aspect of flexibility contracting. Once the input profile matches the output profile, additional supply side flexibility is wasted and represents an irrational configuration. (Formally, this would be the case if, in addition to condition (d), $A_j^{\text{out}} > A_j^{\text{in}}$ or $\Omega_j^{\text{out}} > \Omega_j^{\text{in}}$ for at least one j.) Such a node "absorbs" flexibility with no benefit to the system, and would be able to provide better service (more flexibility) at no cost to itself (no increase in inventory) by passing its excess flexibility downstream until $(A^{\text{in}}, \Omega^{\text{in}}) = (A^{\text{out}}, \Omega^{\text{out}})$. This will result in a perfect non-distortive conduit of information and materials. Orders are filled exactly, no inventory accumulates, and every schedule received is transmitted straight upstream unaltered (a pure lot-for-lot policy). In all other scenarios, the node serves as an "amplifier" of flexibility, offering more to the customer than it itself receives. Such nodes must carry inventory to meet their contracted goals. The specific inventory requirement will be driven not only by the flexibility profiles, but also the nature of the $\{f(t)\}$ process facing the node.

Analytical results predicting inventory from the installed flexibilities are currently limited. While this question will be addressed for the general setting via numerical simulation in §5, to obtain insight into how inventory builds we consider here the simplest conceivable sequence of $\{f(t)\}$: deterministic and stable release schedules, i.e., $f_j(t) = \hat{f}_j$ for all $j \geq 0$, where the $\hat{f}_j$ are constants which satisfy Equation (6) ($[1 - \omega_j^{\text{out}}]\, \hat{f}_j \leq \hat{f}_{j-1} \leq [1 + \alpha_j^{\text{out}}]\, \hat{f}_j$ for $j \geq 1$). These "stable forecasts" are perfect in that the actual release is exactly $\hat{f}_0$ every time period. Naturally, if this were known in advance, the output flexibility could be eliminated since the customer has no real need for revision capability. However, to investigate the inventory impact of non-zero flexibilities we consider how the MC policy will perform if applied to this predictable process. Inventory will still arise due to the need to cover the possibility of increases.

An equilibrium for a flex node facing stable forecasts consists of an inventory level and replenishment schedule that, once in place as the state variables, persist for all subsequent periods. Proposition 3 provides explicit characterization of the equilibrium behavior.

PROPOSITION 3. *An equilibrium for a flex node facing stable forecasts* $\{\hat{f}\}$ *is* $\{\hat{r}, \hat{I}\}$ *where:*

$$\hat{r}_j = \begin{cases} \dfrac{\hat{f}_0}{1 - \Omega_j^{in}} & \text{for } 0 \le j \le j^* \\ \dfrac{1}{1 - \Omega_j^{in}} \max_{k \ge j} \{z_k\} & \text{for } j^* < j \le h \end{cases} \tag{24}$$

$$\text{and } \hat{I} = \sum_{k=1}^{j^*} [(1 + A_k^{out})\hat{f}_k - \hat{f}_0] - \hat{f}_0 \left[\frac{A_{j^*}^{in} + \Omega_{j^*}^{in}}{1 - \Omega_{j^*}^{in}}\right] \quad \text{where} \tag{25}$$

$$z_j \doteq \left[\frac{(1 + A_j^{out})\hat{f}_j}{1 + A_j^{in}}\right][1 - \Omega_j^{in}] \quad \text{and}$$

$$j^* \doteq \begin{cases} \max \{j: z_j > \hat{f}_0\} \text{ if } \exists j \text{ s.t. } z_j > \hat{f}_0 \\ 0 \text{ otherwise.} \end{cases}$$

The above expressions may be interpreted in the following way. As it is increasing in the output flexibility and decreasing in the input flexibility, z_j reports the relative inadequacy of the input side flexibility over a j-period-away outlook. Based on the z_js, j^* defines the *flexibility shortfall horizon*, the shortest horizon length within which input flexibility constraints bind. Beyond j^*, the z_ks are "small," which may be interpreted as a surplus of input flexibility. Indeed, for these indices, Equation (24) indicates that maximal replenishment flexibility is not exercised. j^* plays a key role in the computation shown in Equation (25), which accumulates period-by-period the amount by which the coverage target exceeds the actual demand over the flexibility shortfall horizon (the last term is a boundary effect adjustment). Inventory results from a non-zero j^*, i.e., the existence of a window within which flexibility is lacking, an insight that extends beyond the "stable forecasts" setting. Comparative statics for the inventory level are cataloged in Proposition 4.

PROPOSITION 4. *Under the conditions of Proposition 3, the following properties apply: (a) Release Schedule: (i)* $\Delta\hat{I}/\Delta\hat{f}_0 \le 0$, *(ii)* $\Delta\hat{I}/\Delta\hat{f}_j \ge 0$ *for* $j \ge 1$ *(the inequality is strict for* $j \le j^*$*); (b) Upside Output Flexibility:* $\Delta\hat{I}/\Delta A_j^{\text{out}} \ge 0$ *for* $j \ge 1$ *(the inequality is strict for* $j \le j^*$*); (c) Downside Output Flexibility:* $\Delta\hat{I}/\Delta\Omega_j^{\text{out}} = 0$ *for all* j*; (d) Upside Input Flexibility:* $\Delta\hat{I}/\Delta A_j^{\text{in}} < 0$ *for* $j = j^*$, $\Delta\hat{I}/\Delta A_j^{\text{in}} = 0$ *otherwise; (e) Downside Input Flexibility:* $\Delta\hat{I}/\Delta\Omega_j^{\text{in}} < 0$ *for* $j = j^*$, $\Delta\hat{I}/\Delta\Omega_j^{\text{in}} = 0$ *otherwise.*

Proposition 4 may be interpreted as follows. First, the inventory level is determined by the size of the actual release relative to the upside coverage targets. In (a.i), increasing $\hat{f}_0$ suggests that the demand outcome materializes higher relative to forecast, which decreases inventory. Increasing the forward-looking components of the release schedule as in (a.ii) necessitates inflation of corresponding replenishments, hence potentially more inventory. Comparing (b) to (a.ii) suggests that $\hat{f}_j$ and A_j^{out} have similar effects, which follows since only the product $(1 + A_j^{\text{out}})\hat{f}_j$ plays into the MC logic. As Ω_j^{out} appears nowhere in Proposition 3, $\Delta\hat{I}/\Delta\Omega_j^{\text{out}} = 0$, which may seem counterintuitive. However, (c) assumes that $\{\hat{f}\}$ remains constant. In reality, a rational downstream customer should increase its $\{\hat{r}\}$ (which becomes this flex node's $\{\hat{f}\}$) in response to an increase in its downside input flexibility (this flex node's Ω^{out}). Hence the net effect would actually be more consistent with that described in (a), a network phenomenon not captured in this single-node analysis. Items (d) and (e) show that improvements in input flexibility reduce inventory, but only on the boundary of the flexibility shortfall horizon. Adding within the horizon does not help, since the constraint that defines the boundary continues to bind. Beyond the boundary additional flexibility only contributes to an existing surplus. Of course, with more realistic release schedule dynamics, j^* will move about, so that increasing any component of the input flexibility would likely be beneficial. This and all other insights reported above have been corroborated by numerous simulation experiments.

4. The Market Interface

A QF contract delineates conditions under which all orders will be filled. However, at the market interface this may be an inappropriate representation of the supply relationship. For example, consider a retailer that serves the external market, which is not a single entity with which a contract of this sort may be written. There is no rationale for limiting a customer's entitlement to product, nor is there a customer-provided forecast to

which to tie a minimum purchase requirement. We represent this situation with a "semi-flex node".

Like a flex node, the semi-flex node has replenishment governed by a QF contract. However, there is no such structure on the release side. $\{\mu(t)\} = [\mu_0(t), \mu_1(t), \mu_2(t), \ldots]$ represents information at period t regarding the period-by-period demand, as defined in Equation (1). The construction of $\{\mu(t)\}$ is exogenous to the node but will certainly impact performance. As with the flex node, the decision is $\{r(t)\}$, with updates governed by the IR constraints in Equation (7). Ending inventory is updated by $I(t) = I(t-1) + r_0(t) - \mu_0(t)$, which assumes complete backordering.

The optimization problem faced by a semi-flex node is analogous to program (F) faced by a flex node, except that the expectation in the objective function Equation (12) would be conditional on $\{\mu(t)\}$ rather than $\{f(t)\}$, and $\mu_0(t+j)$ should appear in Equation (13) in place of $f_0(t+j)$. The same issues that complicate the solution of (F) and motivate an OLFC approach (dimensionality and statistical complexity) also apply here. Hence, following the logic applied at the flex node, we formulate program (S-OLFC) as the open-loop version of the semi-flex node's decision problem:

$$\min_{\{r(t)\},(r_0(t+1),\ldots,r_0(t+h))} \sum_{j=0}^{h} E[G(I(t+j)) \mid \{\mu(t)\}]$$

subject to

$$I(t+j) = I(t+j-1) + r_0(t+j) - \mu_0(t+j) \quad \text{for } j = 0, \ldots, h \tag{26}$$

$$(1 - \omega_{j+1}^{in}) r_{j+1}(t-1) \le r_j(t) \le (1 + \alpha_{j+1}^{in}) r_{j+1}(t-1) \quad \text{for } j = 0, \ldots, h-1 \tag{27}$$

$$(1 - \Omega_j^{in}) r_j(t) \le r_0(t+j) \le (1 + A_j^{in}) r_j(t) \quad \text{for } j = 0, \ldots, h. \tag{28}$$

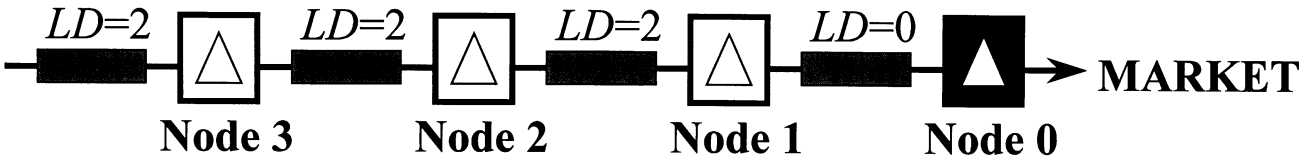
Whereas for the flex node the release-side contractual obligation induced a *deterministic* schedule of future releases on which to focus, here there is no such commitment, reflected in the lack of an analog to Equation (18). Hence, in contrast to (F-OLFC), this open-loop objective function still involves an expectation, which will be based on the distribution of $(\mu_0(t+1), \ldots, \mu_0(t+h))$ conditional on $\{\mu(t)\}$. The open-loop approach is to suppress consideration of how $\{\mu(t)\}$ might be updated over time.

Even with IID market demand and a $G()$ of simple structure, (S-OLFC) is difficult to solve analytically due to the dimensionality and the constraint structure. Instead, we have considered a number of computationally attractive, heuristic approaches based on relaxations of (S-OLFC), and performed a series of numerical simulation tests, assuming a specific market demand process. In particular, since flexibility is most meaningful when tracking a non-stationary process, for all studies in this paper we have used an *Exponentially Weighted Moving Average* (EWMA) process (cf. Box et al 1994). In an EWMA process, period t demand is $\mu_0(t) = \bar{\mu}_1(t-1) + \xi_t$. $\xi_t \sim N(0, \sigma^2)$ is an IID normal forecasting noise with known variance, and $\bar{\mu}_1(t-1)$ is the mean of period t's demand, which follows exponential smoothing dynamics: $\bar{\mu}_1(t) = (1-\delta) \cdot \bar{\mu}_1(t-1) + \delta \cdot \mu_0(t)$. $0 \le \delta \le 1$, with $\delta = 0$ corresponding to IID demand and larger values of δ indicating more volatile demand environments. The demand and forecast process then has two parameters of volatility, δ and σ, and tests were conducted for numerous parameter combinations. Based on the discussion and simulation analysis detailed in Appendix 2, we propose the following heuristic.

The "Sequential Fractile" (SF) policy constructs $\{r(t)\}$ as follows. Define $S_0^*(t) = \mu_0(t)$ and $S_j^*(t) \doteq \text{argmin}_{S_j} E[G(S_j - D_j(t)) \mid \{\mu(t)\}]$, where $D_j(t) \doteq \sum_{q=0}^{j} \mu_0(t+j)$ is the cumulative demand for periods t through $(t+j)$. Letting $y \perp [a, b]$ denote the point in the interval $[a, b]$ closest to y, for $j = 0, \ldots, h$, select:

$$r_j(t) = \frac{r_0(t+j)}{(2 + A_j^{in} - \Omega_j^{in})/2} \perp [(1 - \omega_{j+1}^{in}) r_{j+1}(t-1), (1 + \alpha_{j+1}^{in}) r_{j+1}(t-1)], \text{ where} \tag{29}$$

Figure 3 Supply Chain for System Performance Analysis



$$r_0(t + j) = \left\{S_j^*(t) - I(t - 1) - \sum_{q=0}^{j-1} r_0(t + q)\right\}$$
$$\perp [(1 - \Omega_{j+1}^{in})r_{j+1}(t - 1),$$
$$(1 + A_{j+1}^{in})r_{j+1}(t - 1)].$$

It is straightforward to verify that in a conventional scenario of a fixed lead-time with no flexibility, this reduces to the classical policy of maintaining stock on-hand plus on-order at a critical fractile of cumulative demand over the lead-time. In fact, the SF policy may be viewed as a generalization of multi-period newsvendor logic, known to be optimal with IID demand, to rolling horizon planning in the presence of flexibility. Replenishment policies based on IID logic but applied to real (almost certainly not IID) demand processes have been demonstrated both in research and practice to be very effective, if not optimal (cf. Lovejoy 1990, 1992). We make no claim that the SF policy is optimal in more general settings, only that it includes logic approximating the behavior of a reasonable practitioner and has intuitive appeal. Bassok and Anupindi (1997b) propose alternative OLFC semi-flex node policies under slightly different assumptions, which allow for the development of certain performance bounds. The computationally intensive nature of their policies underscores the need for simplifying heuristics.

5. Performance Properties of QF Supply Chains

We are now prepared to explore the performance properties of multi-level supply chains controlled with QF contracts, which can be modeled by linking together the individual node building blocks presented in §2 and §3. Below we characterize the following metrics: (i) system-wide inventory patterns, (ii) variability of orders placed at each node, and (iii) service provided at the market interface. In particular, the comparative statics of each of these with respect to the market demand volatility and system flexibility characteristics will be provided.

Modeling Supply Chains

Inventory points whose replenishments and releases are both controlled by QF contracts are represented by flex nodes (cf. §2). Only the single node furthest downstream in the chain may deviate from this structure, and semi-flex structure (cf. §3) accommodates its distinctive features.

The link between two nodes is described by the *flexibility profile* of the QF contract and, if desired, a *logistical delay* (LD). The LD allows the representation of delay that is truly unavoidable (e.g., for ocean transit). As in MRP explosion calculus, a buyer node's replenishment schedule becomes its supplier's release forecast, differing by the intervening *LD* time offset: $f_{j-LD}^{supplier}(t) \rightarrow r_j^{buyer}(t)$ for $j \geq LD$. A non-zero *LD* also leads the parties to perceive the QF contract differently. Along with the time offset, i.e. $(\alpha_{j-LD}^{out}, \omega_{j-LD}^{out})^{supplier} \leftrightarrow (\alpha_j^{in}, \omega_j^{in})^{buyer}$, the immutability of orders within the incoming logistical pipeline is represented by $(\alpha_j^{in})^{buyer} = (\omega_j^{in})^{buyer} = 0$ for $j \leq LD$. Hence, a logistical delay may be regarded as an extreme form of inflexibility.

Supply Chain Performance

For the following experiments we consider the serial chain depicted in Figure 3. Nodes 1–3 are flex nodes and node 0 is a semi-flex node. Logistical delays are as labeled.

Figure 4 presents the assumed system flexibility characteristics, stated in CF form since the computational algorithms were easier to implement this way. Conversion back to IR form is easy, via Equations (9) and (11). Parameter values were chosen to provide

Figure 4 Base-Case System Flexibilities

	j	1	2	3	4	5	6	7	8	9	10
Node 1	A_j^{out} and Ω_j^{out}	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
	A_j^{in} and Ω_j^{in}	0.00	0.00	0.04	0.08	0.12	0.16	0.20	0.24	0.28	0.32
Node 2	A_j^{out} and Ω_j^{out}	0.04	0.08	0.12	0.16	0.20	0.24	0.28	0.32		
	A_j^{in} and Ω_j^{in}	0.00	0.00	0.03	0.06	0.10	0.13	0.16	0.19		
Node 3	A_j^{out} and Ω_j^{out}	0.03	0.06	0.10	0.13	0.16	0.19				
	A_j^{in} and Ω_j^{in}	0.00	0.00	0.03	0.05	0.08	0.10				

Figure 5 Summary of Experiments and Observations

System Parameter Under Consideration	Observations and Conclusions: *Inventory*	*Variability of Orders*	*Node 0 Cost & Service Level*
1. Demand forecast error. σ is increased incrementally.	increases at every node (Fig. 6)	over all σ considered, upstream variability $<$ market demand variability (Fig. 10)	both cost and fill rate worsen with σ (Fig. 14)
2. Parameter governing movement of mean demand. δ is increased incrementally.	increases at every node (Fig. 7)	for low δ, upstream variability $<$ market demand variability; as δ increases, bullwhip effect eventually occurs (Fig. 11)	both cost and fill rate worsen with δ (Fig. 15)
3. Flexibility between flex nodes. Components of $(A^{out}, \Omega^{out})^{Node2}$ are increased incrementally. $\{\delta, \sigma\} = \{0.3, 20\}$	decreases at Node 1, increases at Node 2; impact on Node 3 is minor (Fig. 8)	upstream variability is fairly robust to small perturbations of internal flexibility parameters (Fig. 12)	NOT APPLICABLE
4. Flexibility between flex node and semi-flex node. Components of $(A^{out}, \Delta^{out})^{Node1}$ are increased incrementally. $\{\delta, \sigma\} = \{0.3, 20\}$	decreases at Node 0, increases at Nodes 1 and 2; impact on Node 3 is minor (Fig. 9)	order variability is apparently fairly robust to small perturbations of internal flexibility parameters (Fig. 13)	more supply-side flexibility improves both cost and fill rate (Fig. 16)

flexibility amplification (cf. Proposition 2) at each flex node, with upside-downside symmetry in each profile. This network configuration will be referred to as the *Base-Case*. We again use the EWMA demand and forecast process detailed in Appendix 2, with $\bar{\mu}_1(0) = 100$ and $(c_o, c_u) = (30, 150)$.

In a series of simulation experiments, we consider the relationship between key parameters and performance outcomes. The parameters studied are: (1) σ, the demand forecast error, (2) δ, the parameter governing movement of the mean demand, (3) the flexibility profile between two flex nodes (Nodes 1 and 2), and (4) the flexibility profile between a flex node and a semi-flex node (Nodes 1 and 0, respectively). The outcomes reported for each node are: (1) average inventory, and (2) variability of orders (i.e., $StdDev(r_0())$). The investigation of variability is motivated by concern for the "bullwhip" effect, an empirically common phenomenon in which the variability of replenishment orders placed by a node exceeds the variability of customer orders encountered. That is, order variability exceeds market demand variability, and increases on moving upstream. Lee et al. (1997) reports that the QF contract has appeared in industry as a counter-measure to the bullwhip effect.

For stated combinations of the system parameters we report the performance metrics over 100 separate 500-period simulation runs. The four experiments and observations are summarized in Figure 5, and illustrated in Figures 6–16.

Note that increasing the flexibility between flex nodes (Experiment 3 in Figure 5) has no bearing on Node 0 performance. This is because Node 0 continues to receive the same flexibility from Node 1, regardless

of what happens further upstream. Of course, we would expect that in a real supply chain an increase in upstream flexibility should potentially benefit even downstream parties further removed. This would occur if, for instance, Node 1 were to be willing to pass to Node 0 some of the inventory savings enabled by the improved flexibility provided by Node 2. This could be in some combination of increased flexibility and lower unit cost. Such behaviors are not considered within the scope of these experiments.

Figures 6 and 7 validate our intuitions regarding demand variability and inventory. Figure 8 is consistent with the intuitions developed in Proposition 4. Node 1 is receiving improved service (higher input flexibility), therefore can meet its commitments with less inventory. Node 2 is in turn promising a higher level of service, and carries more inventory as a result. From this we note that all else equal, increasing the parameters of the QF contract reduces the customer's costs at the expense of the supplier. This conflict of preferences provides the tension in the contract negotiation process. Even though Node 3's flexibility status is unaltered, its inventory situation does change. The effects are carried upstream via changes in the dynamics of the information vector. Each flexibility profile transforms the information flow, so changes in any profile will have ramifications for all nodes upstream no matter how far removed. As with Figure 8, Figure 9 shows that increasing the flexibility between two nodes (this time a flex node and a semi-flex node) shifts inventory upstream. Slight upward pressure is also expressed at Node 2, which apparently gets damped out before reaching Node 3. At this point it is still unclear where inventory, and by implication flexibility, should best be positioned from a system-optimizing perspective. This design question requires additional structure describing the relative economic implications of holding inventory at the various locations, which we do not pursue in this paper. A methodology for addressing this issue is provided in Tsay (1995).

The next several figures investigate the prevalence of the bullwhip effect in QF environments. In Figure 10, which has IID market demand, no bullwhip occurs.

Figure 6 Inventory vs. σ, with $\delta = 0$

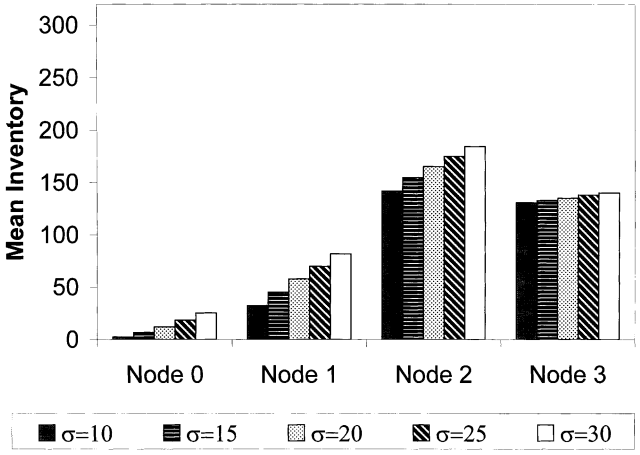


Figure 7 Inventory vs. δ, with $\sigma = 20$

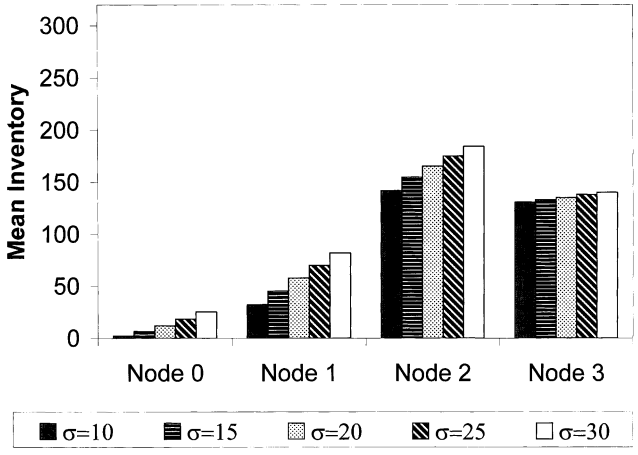
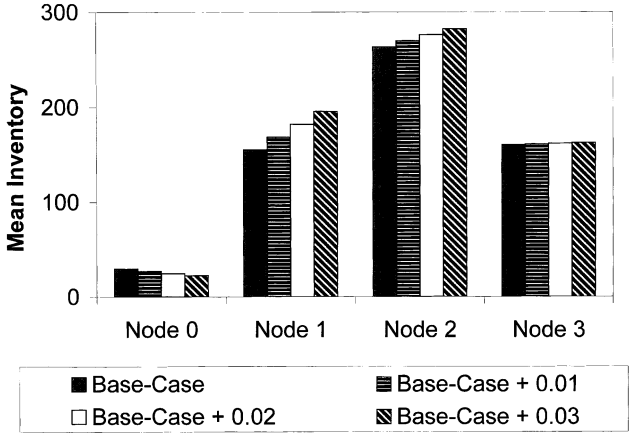


Figure 8 Inventory vs. $(A^{out}, \Omega^{out})^{Node\ 2}$



This was not unexpected since the phenomenon is usually associated with non-stationary demand. However, *dampening* of variability is achieved. When demand is non-stationary (Figure 11), increasing volatility in the market demand and forecasts eventually overwhelms the variability-diffusing capability of the installed flexibility. However, a true bullwhip, which would correspond to an upward-sloping curve, is not always present. Figures 10 and 11 confirm that at each node $StdDev(r_0())$ increases with either demand variability parameter. Figures 12 and 13 suggest that the patterns of variability are fairly robust to small perturbations of flexibility parameters.

We conclude that the presence of flexibility can dampen the transmission of order variability up the chain. This is because an entire replenishment schedule can move in response to changes in the demand environment. For example, suppose demand forecasts are revised upwards in a given period, which would lead a node to generally increase the elements of its replenishment schedule. If the demand forecasts are revised back down in the next period, the node has the opportunity to undo some of the previous increases in the replenishment schedule. The ability to dynamically adjust the estimates is what enables a node to recover from some of the overreacting that becomes a bullwhip

Figure 9 Inventory vs. $(A^{out}, \Omega^{out})^{Node\ 1}$

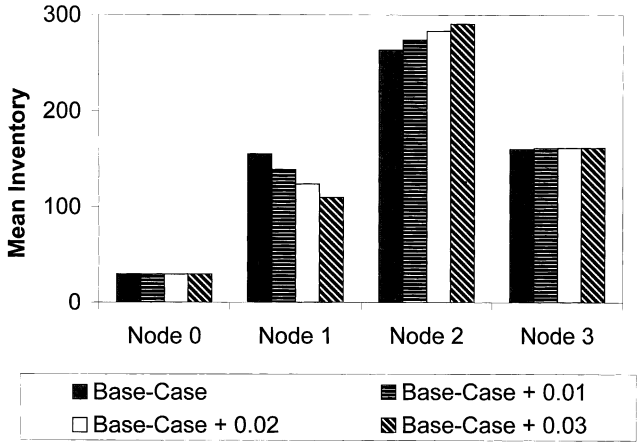


Figure 10 System Variability vs. σ, with $\delta = 0$

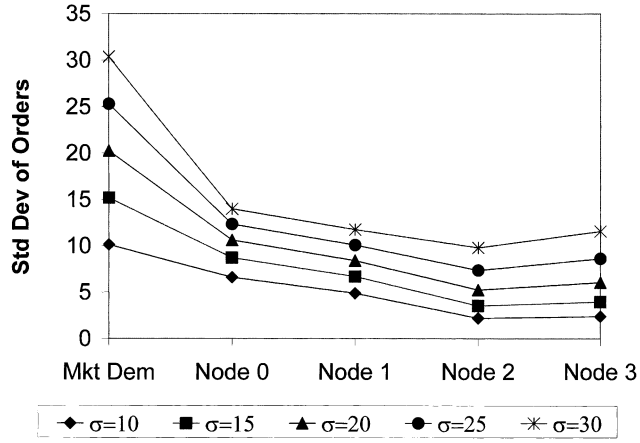


Figure 11 System Variability vs. δ, with $\sigma = 20$

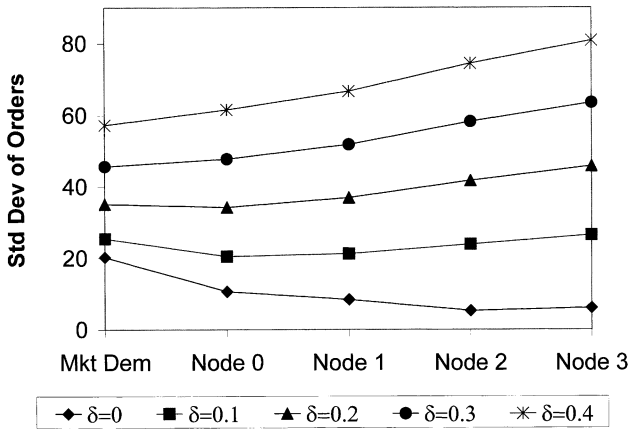
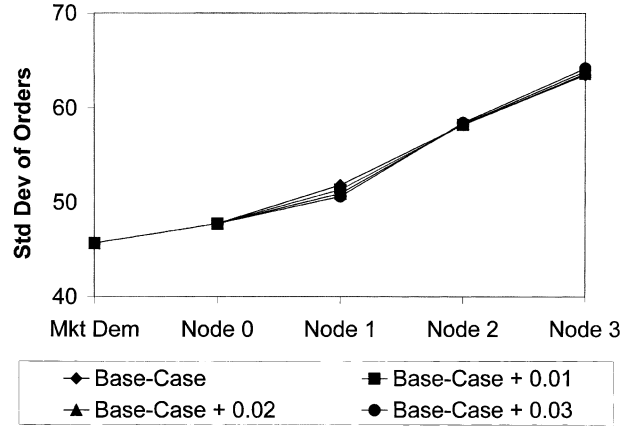


Figure 12 System Variability vs. $(A^{out}, \Omega^{out})^{Node\ 2}$, with $\{\delta, \sigma\} = \{0.3, 20\}$



effect in rigid lead-time settings. As market demand becomes more volatile, the dampening capabilities of the installed flexibilities are eventually overwhelmed, and a bullwhip-type of effect may then be expressed.

As the semi-flex node (Node 0) has distinct structure due to its interface with the external market, additional performance metrics are appropriate. Figures 14 through 16 report this node's average holding and backorder cost per period and service performance (defined as a fill rate) for the relevant experiments. As we would expect, increasing market demand uncertainty and forecast volatility (Figures 14 and 15) cause both the cost and fill rate to worsen, and increased input flexibility (Figure 16) enables an improvement in both.

Natural performance benchmarks are apparent only for the semi-flex node. These include a single-location model with immediate replenishment (extreme flexibility) and one with a fixed lead time of $H > 0$ (zero flexibility), which are well understood in IID demand settings (this approach is taken in Bassok and Anupindi 1997b). However, what remains lacking is some basis for evaluating the absolute magnitudes of the performance outcomes observed at individual flex nodes and across the system. Are there ways to control

Figure 13 System Variability vs. $(A^{out}, \Omega^{out})^{Node\ 1}$, with $\{\delta, \sigma\} = \{0.3, 20\}$

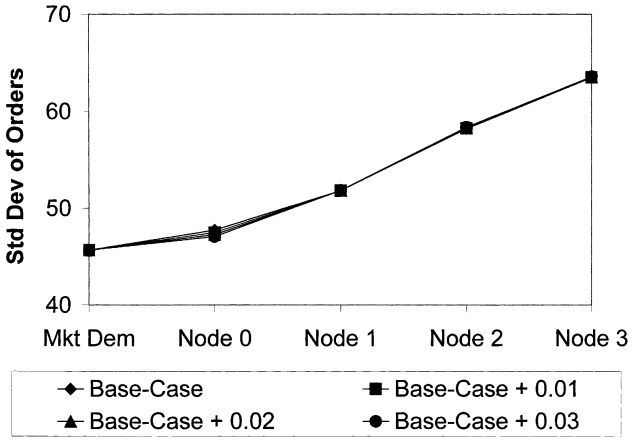


Figure 14 Node 0 Performance vs. σ, $\delta = 0$

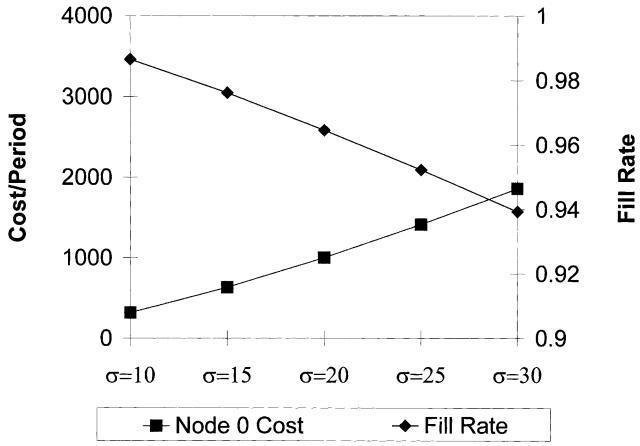


Figure 15 Node 0 Performance vs. δ, $\sigma = 20$

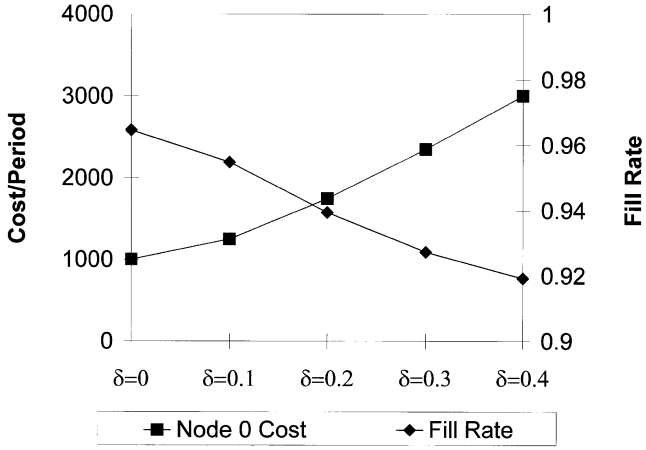
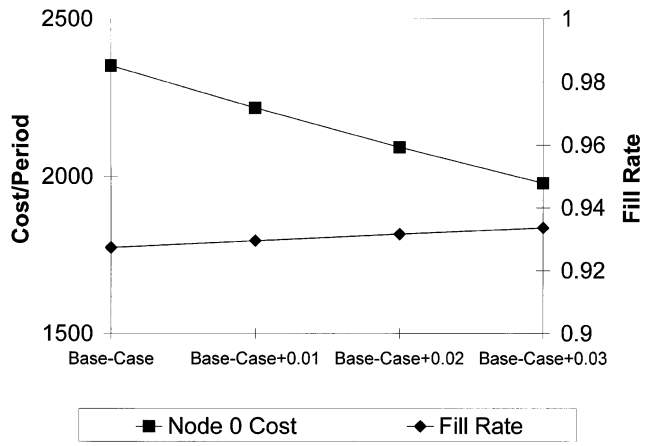


Figure 16 Node 0 Performance vs. $(A^{in}, \Omega^{in})^{Node\ 0}$, with $\{\delta, \sigma\} = \{0.3, 20\}$



the same supply chain which will result in lower inventory levels across the board? Would these methods increase or decrease the order variability? Models of behavior and performance under alternative control schemes are necessary. To the best of our knowledge, these remain open research areas.

6. Contract Design

Thus far we have provided primitives for modeling supply chains controlled by QF contracts and characterized system performance for fixed flexibility parameters. We now consider these as decision variables, since this will be a manager's ultimate interest.[5] Our goal is to provide the "willingness-to-pay" for increments of flexibility, which a materials manager can then compare against the menu of flexibility vs. unit procurement cost combinations offered by a vendor or pool of vendors, as well as other cost considerations not included in this analysis.

To illustrate our methodology we use the simple tandem chain depicted in Figure 17, in which a single flex node (Node 1) feeds into a semi-flex node (Node 0) located at the market interface. Given a contract between Node 0 and Node 1 of (A, Ω), we wish to place a value on Node 1's supply-side flexibility, denoted as $(\tilde{A}, \tilde{\Omega})$. Both contracts have $h = 4$. While we use a multi-level system for greater realism in the dynamics of the materials and information flows, the results and intuitions that follow are not materially different from those obtained for a single node model.

[5]In general, the planning horizon H should also be open to negotiation, and the method we present could easily handle this simply by increasing the dimensionality of the experiment design (i.e., repeating the process for alternative values of H).

The general methodology is straightforward, in that we incrementally increase and record the corresponding reductions in Node 1's inventory cost given a holding cost per period of 15, using the method of §4 to compute average inventory levels in each case. Rather than varying $(\tilde{A}, \tilde{\Omega})$ along its eight degrees of freedom independently, here we limit consideration to a specific parametric form: $\tilde{A} = \tilde{\Omega} = \{0.04s, 0.08s, 0.12s, 0.16s\}$ with $s = 0, \ldots, 5$. Using $\bar{\mu}_1(0) = 100$ and $\sigma = 20$, this procedure was repeated for δ values of {0.3, 0.5, 0.7}. The cost outcomes are reported in Figure 18 as Node 1's average inventory cost per unit of demand, which is appropriate for comparison against unit procurement cost.

The left figure reports how inventory costs vary with the external contract, while on the right is the same data in terms of savings relative to the zero-flexibility case ($s = 0$). This describes the buyer's "willingness to pay" (WTP) for positive increments of flexibility relative to a rigid supply lead time. The cost curves indicate that for any external contract the costs are increasing with the market's δ. Each cost curve is decreasing in s, as would be expected. As s becomes arbitrarily large the cost approaches zero since demand can be tracked perfectly with infinite flexibility. The WTP curves suggest, for example, that in a market with $\delta = 0.7$ the materials manager of Node 1 should to be willing to pay the external vendor an additional \$7.60/unit to go from a no-flexibility contract ($s = 0$) to an $s = 5$ supply contract. The curves shift upward with δ, which we expect since flexibility, the ability to track a moving target, should increase in value with the extent of movement to be tracked. More generally, flexibility cannot be valued without an environmental context. For example, the WTP curve will be uniformly zero in a world of completely deterministic demand as long

Figure 17 Tandem Supply Chain for Contract Evaluation

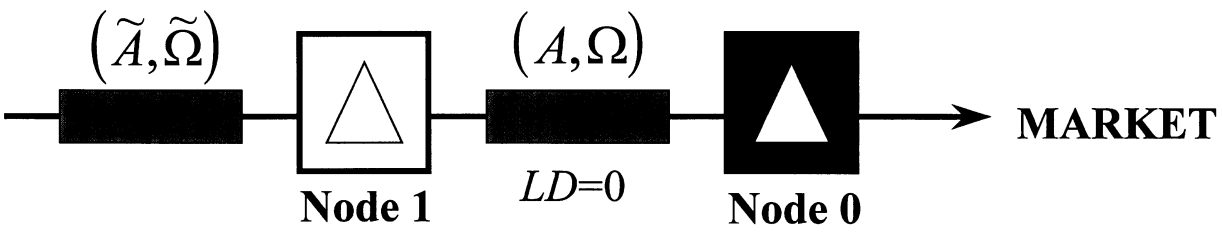
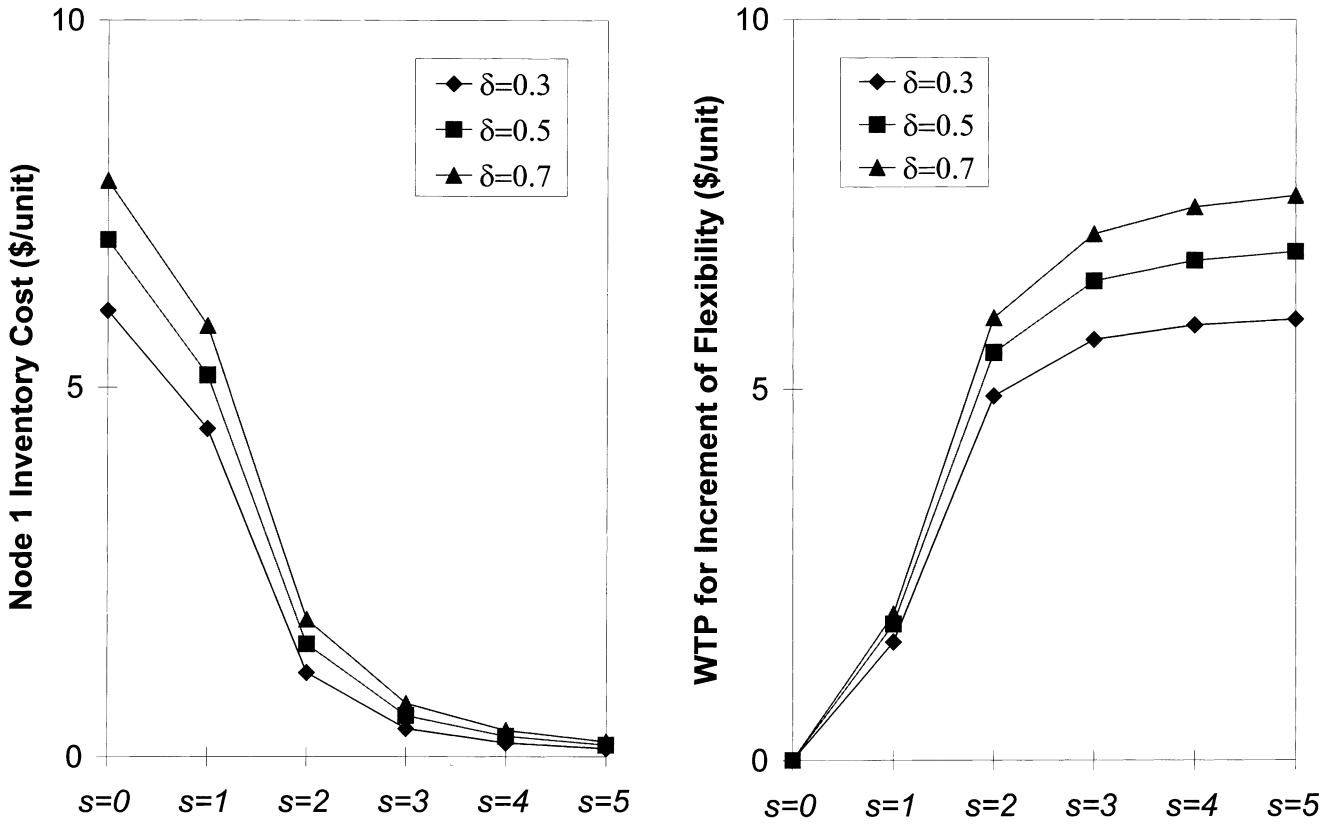


Figure 18 **Node 1 Inventory Cost, Willingness-to-Pay (WTP) (per unit) vs. Supply Flexibility**



as the internal contracts are specified properly. In each demand environment there appears to be a point of diminishing returns beyond which additional flexibility becomes practically worthless, suggesting that there is already sufficient flexibility on hand to suitably respond to the degree of schedule volatility encountered. A buyer always prefers more flexibility, but should be happy to settle for less if the price is right.

7. Concluding Remarks

This paper proposes a framework for performance analysis and design of QF supply chains. We have provided local policies that, in addition to suggesting a rational way to make use of flexible supply, dictate what actions must be taken to support flexibility promised to a customer. While these are not necessarily optimal in the traditional sense, we feel they provide a reasonable compromise in light of their computational properties and the complexity of the general problem.

We have developed the notion of inventory as a consequence of disparities in flexibility. In particular, inventory is the cost incurred in overcoming the inflexibility of a supplier so as to meet a customer's desire for flexible response, which we call flexibility amplification. All else equal, increasing a node's input flexibility reduces its costs. And all else equal, promising more output flexibility comes at the expense of greater inventory costs. We therefore recommend that inventory management should be viewed as the management of process flexibilities.

The modular design of our local nodal models enables multi-echelon analysis, which has been lacking in the literature of flexible supply contracts. Our experiences have revealed that the distribution of the inventory burden across QF supply chains is determined

by the system flexibility characteristics and the volatility in the market demand and forecast process. We have found in addition that QF contracts can dampen the transmission of order variability throughout the chain, thus potentially retarding the well-known "bullwhip effect".

We provide a methodology for computing a materials manager's "willingness-to-pay" for flexibility from an external vendor, which has certain properties. These include the notions that flexibility increases in value as the market environment becomes more volatile, and that flexibility observes a principle of diminishing returns. The buyer always prefers more flexibility, but should be careful to make the appropriate cost-benefit assessment in negotiating the contract.

As firms have experimented with QF contracts, certain implementation issues have come to light. The QF contract represents a radical change in procurement practice for some firms, and change rarely comes without organizational resistance.

Materials buyers may present one source of opposition. Some are accustomed to manipulating orders without perceived consequence, and are reluctant to surrender this position. For others it is the formality of the flexibility limits, rather than the particular latitudes specified, that inspires discontent. Some of these individuals thrive on the thrill and challenge of the dynamic bargaining process, and have confidence in their ability to extract greater concessions in an ad-hoc system than any supplier would actually commit to formally. A large part of this problem is in the difficulty of understanding just how much flexibility is actually needed and how much is available in the relationship. More fundamentally, it can be problematic for a materials organization to recalibrate its intuitions and business practices around specifying flexibilities instead of inventories. The intent of this paper has been to inform these issues.

Depending on what behavior is being replaced, it is unclear whether the move to a QF arrangement will drive procurement prices down or up. Even if these increase, this may still be the best solution in terms of total costs. Yet this can be obstructed by a conflict of interest within the buyer organization. The QF contract is precisely about trading off procurement price for inventory cost, yet in many firms different groups are held accountable for each of these. In Sun Microsystems, for example, the Supplier Management organization is responsible for the unit price, while the Materials organization owns the inventory (cf. Farlow et al. 1995). Will the group concerned with procurement price pay for the supply flexibility that will help the factory operate with less inventory?

A similar conflict can occur within the supplier organization. The supplier benefits from the more honest forecasts that the buyer may provide due to the QF contract, but in exchange may need to lower its selling price and carry additional inventory to meet its promise of coverage. Resistance may result if inventory and price (which now affects revenue) are concerns of different groups.

These, and other cultural and organizational considerations, will join efficiency and valuation issues in determining the popularity of QF contracts over time.[6]

Appendix 1. Proofs of Propositions

Proof of Proposition 1.

We solve (F-OLFC) in several steps, outlined as follows. First, we momentarily relax the upper bounds in Constraints (19) and (20) to avoid potential infeasibility. The relaxed solution is not unique in $\{r(t)\}$, so we pick the option that has the lowest values componentwise. Finally, we show that if updates to $\{f(t)\}$ satisfy the required IR constraints, our solution to the relaxed program automatically satisfies the upper bounds of Equations (19) and (20), and hence is admissible as well as being optimal for (F-OLFC). We now proceed in this fashion.

(F-OLFC) is potentially infeasible since the upper bounds in Equations (19) and (20), which act like capacity constraints, may preclude coverage. The problem is that in converting to a deterministic problem, the information indicating that updates to $\{f(t)\}$ are also bounded is lost. So for the moment we relax these upper bounds, in which case Equations (19) and (20) can be combined into $(1 - \Omega_{j+1}^{in})r_{j+1}(t - 1) \leq r_0(t + j)$, and the optimal $(r_0(t + 1), \ldots, r_0(t + h))$ can be stated as:

[6]The authors would like to thank a number of individuals. Timothy Eckert and Richard Goldstein of Sun Microsystems engaged us in many meaningful conversations in the model design stage. Professors J. Michael Harrison, Warren Hausman, Martin Lariviere, Hau Lee, James Patell, Evan Porteus, Seungjin Whang and Robert Wilson have provided many insightful comments. Seminar participants at Duke University, Santa Clara University, Stanford University, the University of Michigan, and Washington University (St. Louis) have greatly assisted in the refining of our ideas. Last, but not least, we are grateful to the referees and editors for thoughtful and timely review. Any errors remain the responsibility of the authors.

$$r_0^*(t + j) \doteq max\{(1 + A_j^{out}) f_j(t) - \bar{I}_j(t),$$
$$(1 - \Omega_{j+1}^{in}) r_{j+1}(t - 1)\} \text{ for } j = 0, \ldots, h \quad (30)$$

$$\text{where } \bar{I}_0(t) \doteq I(t - 1) \text{ and } \bar{I}_j(t) \doteq \bar{I}_{j-1}(t) + r_0^*(t + j - 1) - (1 + A_{j-1}^{out}) f_{j-1}(t). \quad (31)$$

The formal proof is a straightforward application of Kuhn-Tucker conditions (cf. Rockafellar 1972). See Tsay (1995) for details. In fact, this solution is readily apparent from the problem's economic structure. (F-OLFC) without the upside constraints is essentially an MRP-style lot-sizing problem with minimum lot sizes. With no fixed cost per lot and a holding cost for any material taken earlier than absolutely necessary, a lot-for-lot policy (modified for minimum lot size requirements) will be appropriate. The sequential algorithm stated in Equations (30) and (31) does precisely this, with the construct $\bar{I}_j(t)$ extrapolating the beginning inventory for period $(t + j)$.

While above we have computed the desired *future* replenishments, denoted by $(r_0^*(t + 1), \ldots, r_0^*(t + h))$, the *present* decision is $\{r(t)\}$, which is not uniquely determined by (F-OLFC). Because an $r_j(t)$ (in conjunction with the input flexibility parameters) simply stakes out a region within which $r_0^*(t + j)$ may lie, there will be many $\{r(t)\}$ that can enable the above $(r_0^*(t + 1), \ldots, r_0^*(t + h))$. Since $\{r(t)\}$ defines the lower IR bounds in subsequent periods, a minimal choice of each $r_j(t)$ reduces the risk of unnecessary future inventory. (20) requires $r_0^*(t + j) \le (1 + A_j^{in}) r_j(t)$ (one of the two constraints we relaxed earlier), so choosing an $r_j(t) \ge r_0^*(t + j)/(1 + A_j^{in})$ is necessary. To guarantee this without violating (19), we select:

$$r_j(t) \doteq max[r_0^*(t + j)/(1 + A_j^{in}),$$
$$(1 - \omega_{j+1}^{in}) r_{j+1}(t - 1)] \text{ for } j = 0, \ldots, h \quad (32)$$

The policy that results from applying this rule *every period* may be stated in a more compact and analytically convenient form that gives $\{r(t)\}$ as a direct function of $\{f(t)\}$, bypassing the intermediate calculation of $(r_0^*(t + 1), \ldots, r_0^*(t + h))$ in (30) and (31). Detailed proof of this equivalence is omitted, however the general idea is as follows. Direct substitution of (30) and (31) into (32) is followed by a straightforward but tedious inductive argument that $\bar{I}_j(t)$ (as defined in (31)) and $l_j(t)$ (as defined in (23)) are equivalent for all j when (32) is applied at every t.

To show admissibility, we first prove Lemma 1, which states a property of $l_j(t)$.

Lemma 1.
In rolling from period $(t - 1)$ to period t, if: (a) $I(t - 1) \ge 0$; (b) $\{f(t)\}$ obeys the upside of the output IR constraints; and (c) the $\{r(t)\}$ generated by the MC policy obeys the downside of the input IR constraints, then $l_j(t) \ge l_{j+1}(t - 1)$ for all $j \ge 0$.

Proof of Lemma 1. This property follows from induction on j. Details are omitted due to space limitations. Instead we offer the following intuition. From the period $(t - 1)$ perspective, $l_{j+1}(t - 1)$ is the most conservative (i.e., lowest) estimate for the period $(t + j)$ inventory. That is, it assumes maximal demand and minimal replenishment in all intervening periods. One period's demand and schedule revision outcome is resolved with each horizon roll, and cannot result in inventory any lower than in the extreme scenario.

Admissibility requires that if all updates to $\{f(t)\}$ obey their IR constraints, then for all t, $I(t) \ge 0$ and replenishment side IR constraints are observed. Proof is by induction on t. At period $(t - 1)$, (21) implies $r_{j+1}(t - 1) \ge T_{j+1}(t - 1) \doteq [(1 + A_{j+1}^{out}) f_{j+1}(t - 1) - l_{j+1}(t - 1)]/(1 + A_{j+1}^{in})$ for all $j \ge 0$, which may be rewritten as $(1 + \alpha_{j+1}^{in}) r_{j+1}(t - 1) \ge [(1 + A_j^{out})(1 + \alpha_{j+1}^{out}) f_{j+1}(t - 1) - l_{j+1}(t - 1)]/(1 + A_j^{in})$ (see (9) and (11)). Since $f_j(t) \le (1 + \alpha_{j+1}^{out}) f_{j+1}(t - 1)$ (IR constraint) and $l_j(t) \ge l_{j+1}(t - 1)$ (Lemma 1), this suggests $(1 + \alpha_{j+1}^{in}) r_{j+1}(t - 1) \ge [(1 + A_j^{out}) f_j(t) - l_j(t)]/(1 + A_j^{in}) \doteq T_j(t)$. Thus, $r_j(t) \doteq max[T_j(t), (1 - \omega^{in}{}_{j+1}) r_{j+1}(t - 1)] \le (1 + \alpha_{j+1}^{in}) r_{j+1}(t - 1)$, so the upper bound in (19) is obeyed. Furthermore, $r_j(t) \ge T_j(t)$ for all $j \ge 0$ by construction. At $j = 0$, this is $r_0(t) \ge T_0(t) \doteq f_0(t) - I(t - 1)$, or equivalently, $0 \le I(t - 1) + r_0(t) - f_0(t) \doteq I(t)$. Thus, admissibility conditions are satisfied at every t. ■

Proof of Proposition 2. The MC policy can be stated as follows:

$$r_j(t) \doteq \frac{1}{1 - \Omega_j^{in}} max_{k \ge j}[(1 - \Omega_k^{in}) T_k(t - (k - j))]$$
$$\text{for } j \ge 0, \text{ with } T_k() \text{ from (22)} \quad (33)$$

The equivalence of this more analytically convenient form can be verified by induction on j.

We next establish that inventory is non-increasing with time. Using (33) at $j = 0$:

$$r_0(t) \doteq \frac{1}{1 - \Omega_0^{in}} max_{k \ge 0} \left[(1 - \Omega_k^{in}) \frac{f_k(t - k)(1 + A_k^{out}) - l_k(t - k)}{1 + A_k^{in}}\right]$$
$$\le max_{k \ge 0} \left[f_k(t - k)(1 - \Omega_k^{out}) \left(\frac{1 + A_k^{out}}{1 + A_k^{in}}\right)\left(\frac{1 - \Omega_k^{in}}{1 - \Omega_k^{out}}\right)\right] \le f_0(t)$$

The former inequality holds because $l_k()$ is non-negative and $\Omega_0^{in} = 0$. The latter is due to the output CF constraint and $[(1 + A_k^{out})/(1 + A_k^{in})(1 - \Omega_k^{in})/(1 - \Omega_k^{out})] \le 1$, which follows from condition (d). Thus $I(t) \doteq I(t - 1) + r_0(t) - f_0(t) \le I(t - 1)$. Furthermore, $I(t)$ remains non-negative by the admissibility of the MC policy. So if the inventory is initialized at zero, it will remain there.

The results for the specific case of $(A^{in}, \Omega^{in}) = (A^{out}, \Omega^{out})$ follow from induction on j. We have shown that $I(t) = I(t - 1) = 0$ for all $t \ge 1$. As $I(t) \doteq I(t - 1) + r_0(t) - f_0(t)$ for all $t \ge 1$, this implies $r_0(t) = f_0(t)$. Also, $l_0(t) \doteq I(t - 1) = 0$ for all $t \ge 1$.

Next, suppose that $l_{j-1}(t) = 0$ and $r_{j-1}(t) = f_{j-1}(t)$ for some $j \ge 1$. Then

$$l_j(t) \doteq [l_{j-1}(t) + (1 - \Omega_{j-1}^{in}) r_{j-1}(t) - (1 + A_{j-1}^{out}) f_{j-1}(t)]^+$$
$$= [-(\Omega_{j-1}^{in} + A_{j-1}^{out}) r_{j-1}(t)]^+ = 0$$

We also know that $l_j(t) \geq l_{j+q}(t-q) \geq 0$ for all $q \geq 0$, where the first inequality is due to Lemma 1 and the second reflects the nonnegativity of these entities. Consequently, $l_{j+q}(t-q) = 0$ for all $q \geq 0$. Or, with the change of variable $k = j + q$, $l_k(t-(k-j)) = 0$ for all $k \geq j$. Then, beginning with (33), we have

$$r_j(t) = \frac{1}{1-\Omega_j^{in}} \max_{k \geq j} \left[(1-\Omega_k^{in}) \frac{(1+A_k^{out}) f_k(t-(k-j)) - l_k(t-(k-j))}{1+A_k^{in}} \right]$$

$$= \frac{1}{1-\Omega_j^{out}} \max_{k \geq j} [(1-\Omega_k^{out}) f_k(t-(k-j))]$$

$$= \frac{(1-\Omega_j^{out}) f_j(t)}{1-\Omega_j^{out}} = f_j(t)$$

The second equality is due to (33) and the assumption that $A_k^{in} = A_k^{out}$ and $\Omega_k^{in} = \Omega_k^{out}$ for all k. By the lower output IR constraint, $f_k(t-(k-j)) \geq (1-\omega_{k+1}^{out}) f_{k+1}(t-(k+1-j))$ for all k, or equivalently, $(1-\Omega_k^{out}) f_k(t-(k-j)) \geq (1-\Omega_{k+1}^{out}) f_{k+1}(t-(k+1-j))$. This delivers the third equality as the maximization must then occur at $k = j$. ■

Proof of Proposition 3. The proof, as detailed in Tsay (1995), entails a single, purely mechanical iteration through the MC policy, and is omitted due to space limitations. ■

Proof of Proposition 4. The explicit functional forms of the differences are computed in a tedious but straightforward manner from the results of Proposition 3. ■

Appendix 2. Analysis of Semi-Flex Node Policy

Our approach to obtaining a reasonable and computationally efficient policy for the semi-flex node will be as follows. The solution to (S-OLFC) with (27) and (28) relaxed is relatively straightforward to obtain. We will then consider several alternative heuristic approaches for reconciling this with (27) and (28), and select one for use in network performance analysis based on numerical simulation studies.

Noting that $I(t+j) = I(t-1) + \Sigma_{q=0}^{j} r_0(t+q) - \Sigma_{q=0}^{j} \mu_0(t+q)$ and defining $S_j \doteq (I(t-1) + \Sigma_{q=0}^{j} r_0(t+q))$ and $D_j(t) \doteq \Sigma_{q=0}^{j} \mu_0(t+j)$, the objective in (S-OLFC) can be restated as $min_{\{r(t)\},(S_0,\ldots,S_h)} \Sigma_{j=0}^{h} E[G(S_j - D_j(t)) \mid \{\mu(t)\}]$. If (27) and (28) are relaxed, then clearly $S_0^*(t) = \mu_0(t)$ and $S_j^*(t) \doteq argmin_{S_j} E[G(S_j - D_j(t)) \mid \{\mu(t)\}]$ for $j \geq 1$ will be optimal since the summation in the objective can be decomposed. The corresponding optimal $r_0^*(t+j)$ would then be obtained as $r_0^*(t) = S_0^*(t) - I(t-1)$ and $r_0^*(t+j) = S_j^*(t) - S_{j-1}^*(t)$ for $j \geq 1$. However, in general the attainment of this solution will be obstructed by some of the constraints. We therefore seek a feasible point that is "close" to this ideal in some sense. Our candidate heuristics each have two steps: (Step 1) projecting $(S_0^*(t), \ldots, S_h^*(t))$ into a feasible $(r_0(t+1), \ldots, r_0(t+h))$, and (Step 2) constructing $\{r(t)\}$ to declare to the supplier based on this $(r_0(t+1), \ldots, r_0(t+h))$. Below are two proposed alternatives for each step.

Step 1: ***(Option a) Component-wise projection.*** By the above argument, the ideal would be to achieve $r_0(t) = S_0^*(t) - I(t-1)$ and $r_0(t+j) = S_j^*(t) - S_{j-1}^*(t)$ for $j \geq 1$. However, (27) and (28) together require that $(1-\Omega_{j+1}^{in}) r_{j+1}(t-1) \leq r_0(t+j) \leq (1+A_{j+1}^{in}) r_{j+1}(t-1)$ for all j. So one approach is to get as close as possible term-wise, subject to this constraint, i.e.,

$$r_0(t+j) = \begin{cases} (S_0^*(t) - I(t-1)) \perp [(1-\Omega_1^{in}) r_1(t-1), \\ (1+A_1^{in}) r_1(t-1)] \text{ for } j = 0 \\ (S_j^*(t) - S_{j-1}^*(t)) \perp [(1-\Omega_{j+1}^{in}) r_{j+1}(t-1), \\ (1+A_{j+1}^{in}) r_{j+1}(t-1)] \text{ for } j \geq 1 \end{cases}$$

(Option b) Lexicographic projection. Here the projection is performed sequentially, with the index j target taking into account what has been installed for all preceding terms. So, for all j,

$$r_0(t+j) = \left(S_j^*(t) - \left(I(t-1) + \sum_{q=0}^{j-1} r_0(t+q) \right) \right) \perp [(1-\Omega_{j+1}^{in}) r_{j+1}(t-1), (1+A_{j+1}^{in}) r_{j+1}(t-1)]$$

The rationale for this approach is that the consequences of decision variables for near-term replenishments exceed those for periods further off. Also, the latitude for change is less broad for periods closer in. So it makes sense to first position $r_0(t)$ as close to its ideal value as possible, then compensate for discrepancies in that match when $r_0(t+1)$ is selected, and so on.

Step 2: ***(Option a) Minimum commitment.*** This is the same approach as at the flex node: $r_j(t) \doteq max[r_0(t+j)/(1+A_j^{in}), (1-\omega_{j+1}^{in}) r_{j+1}(t-1)]$ for $j = 0, \ldots, h$. The $(r_0(t+1), \ldots, r_0(t+h))$ chosen at Step 1 takes into account the relative impacts of overage and underage. Here we install the (component-wise) minimum allowable $\{r(t)\}$ that renders those targets attainable.

(Option b): Centering. The selection of $r_j(t)$ induces $[(1-\Omega_j^{in}) r_j(t), (1+A_j^{in}) r_j(t)]$ as the feasible range for $r_0(t+j)$. This option positions that interval so that the target $r_0(t+j)$ sits as close to the midpoint $(r_j(t)[(1-\Omega_j^{in}) + (1+A_j^{in})]/2)$ as is allowed by (27): $r_j(t) \doteq r_0(t+j)/[(2+A_j^{in}-\Omega_j^{in})/2] \perp [(1-\omega_{j+1}^{in}) r_{j+1}(t-1), (1+\alpha_{j+1}^{in}) r_{j+1}(t-1)]$. Whereas minimum commitment logic was used at the flex node because maximum potential customer requests are already incorporated into the targets, at a semi-flex node the updates to $\{\mu(t)\}$ are unconstrained. There is uncertainty as to the direction and extent that the desired $r_0(t+j)$ will move going forward in time, so this method tries to keep the latest target at the middle of the window to leave room to track it in either direction.

The above alternatives suggest the following four distinct heuristics, labeled SF1–SF4:

		Step 2: $(r_0(t), \ldots, r_0(t+h)) \rightarrow \{r(t)\}$	
		Min. commitment	Centering
Step 1: $(S_0^*(t), \ldots, S_h^*(t)) \rightarrow (r_0(t), \ldots, r_0(t+h))$	Component-wise	SF1	SF2
	Lexicographic	SF3	SF4

We compare these methods via numerical simulation, using the EWMA process defined in §3. For this process, an unbiased and minimum mean-squared-error estimate of period $(t + k)$ demand is provided by setting $\mu_k(t) = E[\mu_0(t + k) \mid \bar{\mu}_1(t)] = \bar{\mu}_1(t)$ for $k \geq 1$ (the last equality is true since $\mu_0(t + k) = \bar{\mu}_1(t) + \delta \Sigma_{m=1}^{k-1} \xi_{t+m} + \xi_{t+k}$, cf. Box et al 1994). Cumulative demand is $D_j(t) = \mu_0(t) + j \cdot \bar{\mu}_1(t) + \Sigma_{n=0}^{j-1} (\delta n + 1)\xi_{t+j-n}$, a normal variate with moments $E[D_j(t)] = \mu_0(t) + j \cdot \bar{\mu}_1(t)$ and $Var[D_j(t)] = j\sigma^2[\delta^2(j - 1)(2j - 1)/6 + \delta(j - 1) + 1]$. (Calculation of the latter uses identities $\Sigma_{n=1}^{k} n^2 = k(k + 1)(2k + 1)/6$ and $\Sigma_{n=1}^{k} n = k(k + 1)/2$.)

We assume $G(x) = c_o[x]^+ + c_u[x]^-$, where c_o and c_u are respectively the linear holding and backorder costs, in which case the $S_j^*(t)$ are easily obtained. Specifically, $S_0^*(t) = \mu_0(t)$ and, by newsvendor logic (cf. Heyman and Sobel 1984), $S_j^*(t) = F_{D_j(t)}^{-1}(c_u/(c_o + c_u))$ where $F_{D_j(t)}()$ is the distribution of $D_j(t)$. For the EWMA process, the above analysis suggests that $S_j^*(t) = \mu_0(t) + j \cdot \bar{\mu}_1(t) + (\kappa\sqrt{j} \cdot \sigma)\sqrt{\delta^2(j - 1)(2j - 1)/6 + \delta(j - 1) + 1}$ where $\kappa = \Phi^{-1}(c_u/(c_o + c_u))$ and $\Phi()$ is the standard normal distribution function.

We compare the heuristics over scenarios distinguished by values used for δ, σ, and (A^{in}, Ω^{in}): $\delta \in \{0.3, 0.7\}$, $\sigma \in \{10, 20\}$, and $(A^{in}, \Omega^{in}) \in \{SY, UD, DD\}$ as described below. Profile SY has $A^{in} = \Omega^{in} = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$, symmetrical in upside and downside flexibility. UD is upside dominant, with $A^{in} = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$ and $\Omega^{in} = \{0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45\}$. DD is downside dominant, with $A^{in} = \{0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45\}$ and $\Omega^{in} = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$. Cost parameters $(c_o, c_u) = (30, 150)$ are used. The performance of each heuristic is evaluated by the average cost over 100 sample paths, each path representing 500 periods. $\bar{\mu}_1(0) = 100$ in all cases.

The outcomes of the 12 scenarios support the following conclusions, with numerical details omitted due to space limitations (see Tsay 1995). *SF3* and *SF4* are each uniformly superior to both *SF1* and *SF2* by far, with results that are statistically significant with p-values no greater than 1×10^{-17} in all cases (and typically even lower). So Lexicographic projection dominates Component-wise projection for Step 1 regardless of the option taken at Step 2, presumably for its handling of the interrelationships between periods. There is no dominant approach at Step 2, with relative performance varying with the flexibility structure. We thus select *SF3* as the semi-flex node operating policy, acknowledging the existence of alternatives that are equally easy to implement and give superior performance in some settings.

References

Azoury, K. S. 1985. Bayes solution to dynamic inventory models under unknown demand distribution. *Management Sci.* **31** 1150–1160.

Baker, K. R. 1993. Requirements Planning. S. C. Graves, A. H. G. Rinnooy Kan, P. H. Zipkin, eds. *Handbooks in Operations Research and Management Science, Vol. 4 (Logistics of Production and Inventory).* Elsevier Science Publishing Company B.V., Amsterdam, The Netherlands.

Bassok, Y., R. Anupindi. 1995. Analysis of supply contracts with forecasts and flexibility. Working Paper, Northwestern University.

——, ——. 1997a. Analysis of supply contracts with total minimum commitment. *IIE Trans.* **29** 373–381.

——, ——. 1997b. Analysis of supply contracts with commitments and flexibility. Working Paper, Northwestern University.

Bergen, M., S. Dutta, O. C. Walker. 1992. Agency relationships in marketing: A review of the implications and applications of agency and related theories. *J. Marketing* **56** 3 1–24.

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

Box, G. E. P., G. M. Jenkins, G. C. Reinsel. 1994. *Time Series Analysis: Forecasting and Control.* Prentice Hall, Englewood Cliffs, NJ.

Chen, F. 1997. Decentralized supply chains subject to information delays. Working paper, Graduate School of Business, Columbia University.

Connors, D., C. An, S. Buckley, G. Feigin, A. Levas, N. Nayak, R. Petrakian, R. Srinivasan. 1995. Dynamic modeling for re-engineering supply chains. Research report, IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY.

Donohue, K. L. 1996. Supply contracts for fashion goods: Optimizing channel profits. Working paper, Department of OPIM, The Wharton School, University of Pennsylvania.

Emmons, H., S. M. Gilbert. 1998. Note: The role of returns policies in pricing and inventory decisions for catalogue goods. *Management Sci.* **44** 2 276–283.

Eppen, G. D., A. V. Iyer. 1997. Backup agreements in fashion buying: The value of upstream flexibility. *Management Sci.* **43** 1469–1484.

Farlow, D., G. Schmidt, A. A. Tsay. 1995. Supplier management at Sun Microsystems. Case Study, Graduate School of Business, Stanford University, Stanford, CA.

Faust, M. 1996. Personal communication from a product manager at one of Compaq's suppliers of memory chips.

Federgruen, A., P. Zipkin. 1986. An inventory model with limited production capacity and uncertain demands—I: The average-cost criterion/II: The discounted-cost criterion. *Math. Oper. Res.* **11** 193–215.

Guererro, H. H., K. R. Baker, M. H. Southard. 1986. The dynamics of hedging the master schedule. *Internat. J. Production Res.* **24** 1475–1483.

Ha, A. Y. 1997. Supply contract for a short-life-cycle product with demand uncertainty and asymmetric cost information. Working paper, Yale School of Management.

Heath, D. C., P. L. Jackson. 1994. Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Trans.* **26** 17–30.

Heyman, D., M. Sobel. 1984. *Stochastic Models in Oper. Res., Volume II (Stochastic Optimization)* McGraw Hill, New York.

Iyer, A., M. E. Bergen. 1997. Quick response in manufacturer-retailer channels. *Management Sci.* **43** 4 559–570.

Jeuland, A. P., S. M. Shugan. 1983. Managing channel profits. *Marketing Sci.* **2** 239–272.

Kandel, E. 1996. The right to return. *J. Law and Economics* **39** 329–356.

Karmarkar, U. S. 1989. Getting control of just-in-time. *Harvard Business Review* September–October 122–131.

Katz, M. L. 1989. Vertical contractual relations. R. Schmalensee, R. D. Willig, eds. *Handbook of Industrial Organization: Volume I.* Elsevier Science Publishers B.V., New York.

Lariviere, M. A. 1999. Supply chain contracting and coordination with stochastic demand. S. Tayur, R. Ganeshan, M. Magazine, eds. *Quantitative Methods for Supply Chain Management.* Kluwer Academic Publishers, Norwell, MA.

Lee, H. L., P. Padmanabhan, S. Whang. 1997. The bullwhip effect in supply chains. *Sloan Management Rev.* **38** 3 93–102.

——, S. Whang. 1997. Decentralized multi-echelon inventory control systems: Incentives and information. Working Paper, Stanford University, Stanford, CA.

Lovejoy, W. S. 1990. Myopic policies for some inventory models with uncertain demand distributions. *Management Sci.* **36** 724–738.

——. 1992. Stopped myopic policies in some inventory models with generalized demand processes. *Management Sci.* **38** 688–707.

——. 1998. *Integrated Operations,* Southwestern College Publishing, Cincinnati, Ohio, Forthcoming.

Magee, J. F., D. M. Boodman. 1967. *Production Planning and Inventory Control.* McGraw-Hill Book Company, New York.

Masten, S. E., K. J. Crocker. 1985. Efficient adaptation in long-term contracts: Take-or-pay provisions for natural gas. *American Economic Rev.* **75** 1083–1093.

Mathewson, G. F., R. A. Winter. 1984. An economic theory of vertical restraints. *Rand J. Economics* **15** 1 27–38.

Miller, B. L. 1986. Scarf's state reduction method, flexibility, and a dependent demand inventory model. *Oper. Res.* **36** 83–90.

Miller, J. G. 1979. Hedging the master schedule. L. P. Ritzman et al., eds. *Disaggregation Problems in Manufacturing and Service Organizations.* Martinus Nifhoff, Boston, MA.

Mondschein, M. 1993. Negotiating product supply agreements. *National Petroleum News.* **85** 45.

Moorthy, K. S. 1987. Managing channel profits: Comment. *Marketing Sci.* **6** 4 375–379.

Nahmias, S. 1997. *Production and Operations Analysis.* Irwin, Homewood, IL.

National Energy Board. 1993. Natural gas market assessment: Long-term Canadian natural gas contracts. *Gas Energy Review* **21** 8–11.

Ng, S. 1997. Supply chain management at Solectron. Presentation. *Industrial Symposium on Supply Chain Management.* Stanford University, June.

Pasternack, B. A. 1985. Optimal pricing and returns policies for perishable commodities. *Marketing Sci.* **4** 166–176.

Rockafellar, R. T. 1970. *Convex Analysis.* Princeton University Press, Princeton, NJ.

Tayur, S. 1992. Computing the optimal policy for capacitated inventory models. *Comm. Statist. Stoch. Models* **9** 585–598.

Tirole, J. 1988. *The Theory of Industrial Organization.* The MIT Press, Cambridge, MA.

Tsay, A. A. 1995. *Supply Chain Control with Quantity Flexibility.* Ph.D. Dissertation, Graduate School of Business, Stanford University, Stanford, CA.

——. 1996. The quantity flexibility contract and supplier-customer incentives. Working Paper, Leavey School of Business, Santa Clara University.

——, S. Nahmias, N. Agrawal. 1999. Modeling supply chain contracts: A review. S. Tayur, R. Ganeshan, M. Magazine, eds. *Quantitative Methods for Supply Chain Management.* Kluwer Academic Publishers, Norwell, MA.

Van Ackere, A. 1993. The principal/agent paradigm: Its relevance to various functional fields. *Eur. J. Oper. Res.* **70** 83–103.

Whang, S. 1995. Coordination in operations: A taxonomy. *J. Oper. Management,* **12** 413–422.

Accepted by Paul Zipkin; received January 26, 1998. This paper has been with the authors 45 days for 2 revisions. The average review cycle time was 32.3 days.