Select Papers on Supply Chain Management

Institute for Operations Research and the Management Sciences

INFORMS Select Papers on Supply Chain Management

Table of Contents

Article 1:
Extended Enterprise Supply-Chain Management at IBM Personal Systems Group and Other Divisions
Interfaces, Vol. 30, No. 1, January-February 2000 (pp. 7-25).

Article 2:
Xilinx Improves Its Semiconductor Supply Chain Using Product and Process Postponement
Interfaces, Vol. 30, No. 4, July-August 2000 (pp. 65-80).

Article 3:
Stock Positioning and Performance Estimation in Serial Production-Transportation Systems
Manufacturing & Service Operations Management, Vol. 1, No. 1, 1999 (pp. 77-88).

Article 4:
Quantity Flexibility Contracts and Supply Chain Performance
Manufacturing & Service Operations Management, Vol.1, No. 2, 1999 (pp. 89-111).

Article 5:
Optimizing Strategic Safety Stock Placement in Supply Chains
Manufacturing & Service Operations Management, Vol. 2, No. 1, Winter 2000 (pp. 68-83).

Article 6:
A Dynamic Model for Requirements Planning with Application to Supply Chain Optimization
Operations Research, Vol. 46, Supp. No.3, May-June 1998 (pp. S35-S49).

Article 7:
Development of a Rapid-Response Supply Chain at Caterpillar
Operations Research, Vol. 48, No. 2, March-April 2000 (pp. 189-204).

Extended-Enterprise Supply-Chain Management at IBM Personal Systems Group and Other Divisions

Grace Lin, Markus Ettl
Steve Buckley, Sugato Bagchi
IBM T. J. Watson Research Center
Yorktown Heights, New York 10598

David D. Yao
Columbia University
New York, New York 10027

Bret L. Naccarato
IBM Printing Systems Company
Endicott, New York 13760

Rob Allan, Kerry Kim
Lisa Koenig
IBM Personal Systems Group
Research Triangle Park, North Carolina 27709

In 1994, IBM began to reengineer its global supply chain. It wanted to achieve quick responsiveness to customers with minimal inventory. To support this effort, we developed an extended-enterprise supply-chain analysis tool, the Asset Management Tool (AMT). AMT integrates graphical process modeling, analytical performance optimization, simulation, activity-based costing, and enterprise database connectivity into a system that allows quantitative analysis of extended supply chains. IBM has used AMT to study such issues as inventory budgets, turnover objectives, customer-service targets, and new-product introductions. We have implemented it at a number of IBM business units and their channel partners. AMT benefits include over $750 million in material costs and price-protection expenses saved in 1998.

As the world's largest company providing computer hardware, software, and services, IBM makes a wide variety of products, including semiconductors, processors, hard disks, personal computers, printers, workstations, and mainframes. Its manufacturing sites are linked with tens of thousands of suppliers and distribution channels all over the world. A single product line may involve thousands of part numbers with multilevel bills of materials, highly varied lead times and costs, and dozens to hundreds of manufacturing and distribution sites

0092-2102/00/3001/0007/$05.00
1526–551X electronic ISSN
This paper was refereed.

INVENTORY/PRODUCTION—APPLICATIONS
INDUSTRIES—COMPUTERS
MANUFACTURING—SUPPLY CHAIN MANAGEMENT

INTERFACES **30:** 1 January–February 2000 (pp. 7–25)

linked by different transportation modes. Facing the challenges of increasing competition, rapid technology advance, and continued price deflation, the company launched an internal reengineering effort in 1993 to streamline business processes in order to improve the flow of material and information. The reengineering effort focused on improving customer satisfaction and market competitiveness by increasing the speed, reliability, and efficiency with which IBM delivers products to the marketplace.

In 1994, IBM launched an asset-management reengineering initiative as part of the overall reengineering effort. The objectives were to define the supply-chain structure, to set strategic inventory and customer-service targets, to optimize inventory allocation and placement, and to reduce inventory while meeting customer-service targets across the enterprise. The company formed a cross-functional team with representatives from manufacturing, research, finance, marketing, services, and technology. The team identified five areas that needed modeling support for decision making: (1) design of methods for reducing inventory within each business unit; (2) development of alternatives for achieving inventory objectives for senior-management consideration; (3) development and implementation of a consistent process for managing inventory and customer-service targets, including tool deployment, within each business unit; (4) complete evaluation of such assets as service parts, production materials, and finished goods in the global supply network; and (5) evaluation of cross-brand product and unit synergy to improve the management of inventory and risk.

We developed the Asset Management Tool (AMT), a strategic decision-support tool, specifically to address these issues. The integration of AMT with the other asset-management reengineering initiatives has resulted in the successful implementation of extended-enterprise supply-chain management within IBM.

The Asset Management Tool

An extended-enterprise supply chain is a network of interconnected facilities through which an enterprise procures, produces, distributes, and delivers products and services to its customers. As procurement, distribution, and sales have become increasingly global, the supply chains of large companies have become deeply intertwined and interdependent. Today's extended-enterprise supply chains are in fact networks of many supply chains representing the interests of many companies, from supplier's suppliers to customer's customers. Because of this interdependency, a company with an extended supply chain performs well only when it collaborates and cooperates actively with its suppliers and resellers.

A company with an extended supply chain performs well only when it collaborates with its suppliers and resellers.

In high-technology industries, management of the extended-enterprise supply chain becomes very important. At its best, it keeps operating costs low and profits high. But a poorly managed supply chain can reverse that relationship, eroding profits, compromising innovation, and ham-

pering business growth. Early in our efforts, we realized that there were two fundamental keys to overhauling IBM's supply chain. First, we had to reduce and manage uncertainty to promote more accurate forecasts. Second, we had to improve supply-chain flexibility to facilitate quick adaptation to changes in the marketplace. From the outset, we focused on the intrinsic interdependency of an extended-enterprise supply chain. We knew our system would perform as desired only if it reflected the policies and processes used by our suppliers and channels, integrating their value chains with our own. This perspective helped to shape our vision: an integrated modeling and analysis tool for extended-enterprise supply chains. It would be a tool with new methodologies to handle the uncertainties inherent in demand, lead time, supplier reliability, and other factors. It would be scalable, so that it could handle the vast amounts of data describing product structure, supply-chain processes, and component stock information that typify the industry. Finally, the new tool would be equally effective at modeling basic types of supply-chain policies and their interactions, because different companies may use different policies.

We designed AMT to address all of these issues. It is a modeling and analysis system for strategic and tactical supply-chain planning that emerged from various earlier internal IBM reengineering studies [Bagchi et al. 1998; Buckley 1996; Buckley and Smith 1997; Feigin et al. 1996]. It supports advanced modeling, simulation, and optimization capabilities for quantitative analysis of multiechelon inventory systems, along with such features as enterprise database connectivity and internet-based communication. AMT is built on six functional modules: a data-modeling module, a graphical user interface, an experiment manager, an optimization engine, a simulation engine, and a report generator.

The data-modeling module provides a relational data interface, including product structures, lead times, costs, demand forecast and the associated variability information. It has built-in explosion of bills of materials and data-reduction capabilities, and automatic checks for data integrity. It provides access to IBM's global and local operational databases through data bridges.

The graphical user interface (GUI) combines supply-chain modeling with dialog-based entry of supply-chain data. It allows users to build supply networks by dragging and dropping model components, such as manufacturing nodes, distribution centers, and transportation nodes, onto the work space.

The experiment manager facilitates the organization and management of data sets associated with supply-chain experiments. It allows users to view and interactively modify parameters and policies. In addition, it provides automated access to output data generated during experiments and supports a variety of file-management operations.

The optimization engine performs AMT's main function, quantifying the trade-off between customer-service targets and the inventory in the supply network. This module can be accessed from the GUI pull-down menu or called by the simulation engine.

The simulation engine simulates the

performance of the supply chain under various parameters, policies, and network configurations, including the number and location of suppliers, manufacturers, and distribution centers; inventory and manufacturing policies, such as base-stock control, days of supply, build-to-stock, build-to-order, and continuous or periodic replenishment policies. The simulation engine contains an animation module that helps users to visualize the operation of the supply chain or vary parameters and policies while monitoring the simulation output reports.

The report generator offers a comprehensive view of the performance of the supply chain under study, including average cycle times, customer-service levels, shipments, fill rates, and inventory. It also generates financial results, including revenues, inventory capital, raw-material costs, transportation costs, and activity-based costs, such as material handling and manufacturing.

The Optimization Engine

The central function of the optimization engine is to analyze the trade-off between customer-service and inventory investment in an extended-enterprise supply chain. The objective is to determine the safety stock for each product at each location in the supply chain to minimize the investment in total inventory. We view the supply chain as a multiechelon network in which we model each stocking location as a queuing system. In addition to the usual queuing modeling, we incorporated into the model an inventory-control policy: the base-stock control, with the base-stock levels being decision variables. To numerically evaluate such a network, we developed an approach based on decomposition. The key idea is to analyze each stocking location in the network individually and to capture the interactions among different stocking locations through their so-called actual lead times.

We modeled each stocking location as a queue with batch Poisson arrivals and infinite servers with service times following general distributions, denoted as $M^X/G/\infty$ in queueing notation. To do this, we had to specify the arrival and the service processes. We obtained the arrival process at each location by applying the standard MRP demand explosion technique to the production structure. The batch Poisson arrival process has three main parameters: the arrival rate, and the mean and the variance of the batch size. It thus accommodates many forms of demand data; for instance, demand in a certain period can be characterized by its minimum, maximum, and most likely value. The service time is the actual lead time at each stocking location. The actual lead time at a stocking location can be derived from its nominal lead time (for example, the manufacturing or transportation time) along with the fill rate of its suppliers. In particular, when a supplier has a stock-out, we have to add the resulting delay to the actual lead time. This delay is the time the supplier takes to produce the next unit to supply the order. In our model, we derive the additional delay from Markov-chain analysis.

AMT embodies a creative coupling of optimization, performance evaluation, and simulation.

With the arrival and service processes in place, we can analyze the queue and derive performance measures, such as inventory, back-orders, fill rates, and customer-service levels. The key quantity in the analysis of a stocking location, i, is the number of jobs in the $M^X/G/\infty$ queue, denoted N_i, which can be derived from standard queueing results [Liu, Kashyap, and Templeton 1990]. To alleviate the computational burden in large-scale applications, we approximated N_i by a normal distribution. This way, we need to derive only the mean and the variance of N_i, both of which depend on the actual lead time, which is the service time in the queuing model. Figure 1 shows a snapshot of the dynamics at a stocking location.

The objective of the optimization model is to minimize the total expected inventory capital in the supply network. This total is a summation over all stocking locations, each of which carries two types of inventory: finished goods (on-hand) inventory, and work-in-process (on-order) inventory. The constraints of the optimization model are the required customer-service targets. They are represented as the probability, say 95 or 99 percent, that customer orders are filled by a given due date. Our formulation allows users to specify customer-service targets separately for each demand stream. We first derive the fill rates for each end product to meet the required customer-service target. These fill rates relate to the actual lead times of all upstream stocking locations, via the bills-of-materials structure of the network, and to the actual lead times. The model thus captures the interdependence of different stocking locations, in particular the effect of base-stock levels and fill rates on customer-service. Related models in supply-chain and distribution networks include those of Lee and Billington [1993], Arntzen et al. [1995], Camm et al. [1997],

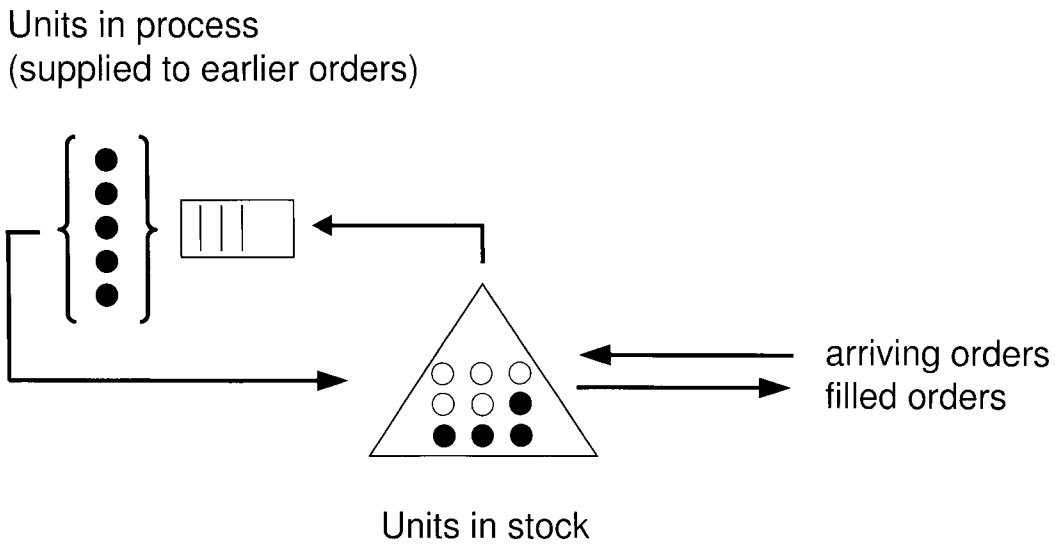


Figure 1: In this snapshot of the system dynamics at a stocking location, the base-stock level is nine, and when there are four units in stock, the other five units have been supplied to earlier orders, which translates into the five jobs in process.

Kruger [1997], Graves, Kletter, and Hetzel [1998], and Andersson, Axsaeter, and Marklund [1998].

To allow fast execution of the optimization, we derived analytical gradient estimates in closed form and implemented a gradient search algorithm to generate optimal solutions. Technical details of this work are presented by Ettl et al. [1998] and in the Appendix. In addition to the gradient search, we developed a heuristic optimization procedure based on product clustering. To validate the solution approach, we compared it against exhaustive searches for test problems of moderate size. For large-scale, industry-size applications, the model has been extensively tested at several IBM business units.

The Simulation Engine

The simulation engine allows users to simulate various supply-chain policies and in particular to verify and fine-tune the performance of the solutions generated by the optimization engine. We built the simulation engine upon SimProcess [Swegles 1997], a general-purpose business-process simulator that was developed jointly by IBM Research and CACI Products Company. The simulation engine preserves the capabilities of SimProcess while adding a supply-chain modeling functionality. Specifically, it provides modeling functions for the following supply-chain processes:

—The customer process represents outside customers that issue orders to the supply chain, based on the modeled customer demand. It can also model information about the desired customer-service target and priority for the customer.

—The manufacturing process models assembly processes, buffer policies, and replenishment policies. It can also be used to model suppliers.

—The distribution process models distribution centers and can also be used to model retail stores.

—The transportation process models transportation time, vehicle loading, and transportation costs.

—The forecasting process represents product forecasts, including promotional and stochastic demand, for future periods.

—The inventory-planning process models periodic setting of inventory target levels. Underlying this process is the AMT optimization engine that computes recommended inventory levels at the various stocking locations in a supply chain based on desired customer-service target.

The simulation engine allows the user to vary a set of input parameters while monitoring output reports to obtain the best set of output values. All input and output parameters reside in the AMT modeling database. Users provide input parameters for the simulation in the form of random variables with stochastic distributions; these include manufacturing lead times, transportation times, material-handling delay times, demand forecasts, product quantity required in a bill of material, and supplier reliability. The stochastic distribution functions supported include beta, Erlang, exponential, gamma, normal, lognormal, Poisson, triangular, uniform, Weibull, and user-defined distributions.

We designed the simulation engine to enable scenario-based analyses in which supply-chain parameters, such as the number and location of suppliers, manufacturers, and distribution centers, inven-

tory levels, and manufacturing, replenishment, and transportation policies (build-to-plan, build-to-order, assemble-to-order, continuous replenishment, periodic replenishment, full truckload, less-than-truckload, and so forth) are varied across simulation runs. For each simulation run, the user can specify a planning horizon, the number of replicating scenarios (sample runs), and a warm-up period during which statistics are not retained. The length of the planning horizon depends on the particular application in question and the availability of historical demand forecasts. We typically choose a horizon that is between six and 12 months.

The simulation-run outcome is in the form of measurement reports that can be generated for turnaround times, customer-service, fill rates, stock-out rates, shipments, revenue, safety stock, and work-in-process. To analyze financial impacts, users can employ the following items, all of which are monitored during the simulation: cost of raw material; revenue from goods sold; activity-based costs, such as material handling and manufacturing; inventory-holding costs; transportation costs; penalties for incorrectly filled or late orders delivered to customers; credits for incorrectly filled or late deliveries from suppliers; cost of goods returned by customers; and credits for goods returned to suppliers.

System Integration and Technical Innovations

We integrated the six functional modules of AMT in a system architecture that is flexible enough to accommodate users' varying computational needs. The architecture is based on a client-server programming model in which one can conduct experiments using the resources of a computer network (Figure 2). The AMT client side provides a set of functions for viewing the graphical user interface and dialog-based data entry. The AMT server side, which typically resides on a powerful workstation or midrange computer, provides the full modeling and analysis functionality. For users with access to low-powered computers, such as laptops, we developed an architecture in which the AMT client side is implemented as a platform-independent Java application or applet; web-enabled clients allow users to access AMT through a web browser.

To manage supply-chain operations, AMT requires data about the different stages and processes that products go through. This data is accessible through a relational modeling database that is connected to the server through a relational interface. The database stores the information associated with the various modeling scenarios, including the supply-chain structure, product structure, manufacturing data, and demand forecasts. The product structures are derived from a top-down bills-of-materials explosion that is processed for each end product. We extracted all product data from corporate databases and from local site data sources.

To facilitate data extraction, we developed a number of database connectivity modules that provide automated database access, extract production data, and feed them into the modeling database. All connectivity modules have built-in bills-of-materials explosion functionality. To detect inconsistencies in data recording caused by missing or incomplete informa-

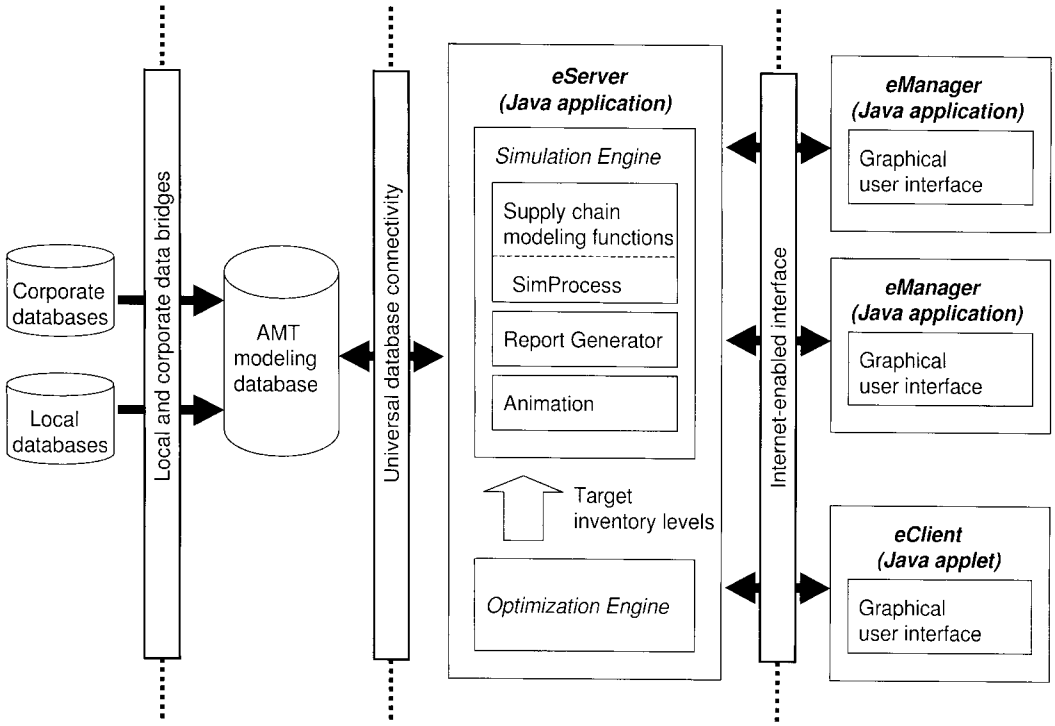


Figure 2: AMT is implemented using a client-server architecture in which the modeling functionality is separated from the graphical user interface. The modeling engines reside on a server computer (eServer). The graphical user interfaces are piped to client computers that are implemented as either Java applications (eManager) or Java applets (eClient). The AMT modeling database can be accessed through a relational database interface. It contains such supply chain data as bills of materials, demand forecasts, lead times, costs, inventory policies, and customer-service requirements. Local and corporate data bridges provide automated access to enterprise data sources.

tion pertaining to the bills of materials, we added database consistency checks that generate missing data reports and reduce the data set to a consistent level that can be downloaded to the modeling database. The data-collection process allows the user to supply missing data in relational tables that can be merged with the output of the explosion. To keep the complexity of the bills-of-materials explosion manageable, we implemented data-reduction routines through which one can eliminate noncritical components automatically, based on the item's value class or annual requirements cost.

AMT's graphical user interface allows modelers to build supply networks for a variety of supply chains by dragging and dropping generic supply-chain components on the workspace (Figure 3). Sophisticated algorithms are encapsulated in the components. For instance, clicking the "PSG manufacturing" node will bring up screens for the user to specify parameters and policies, such as delay time, manufacturing lead times, bills of materials, and such manufacturing policies as build to order or build to plan. AMT also supports

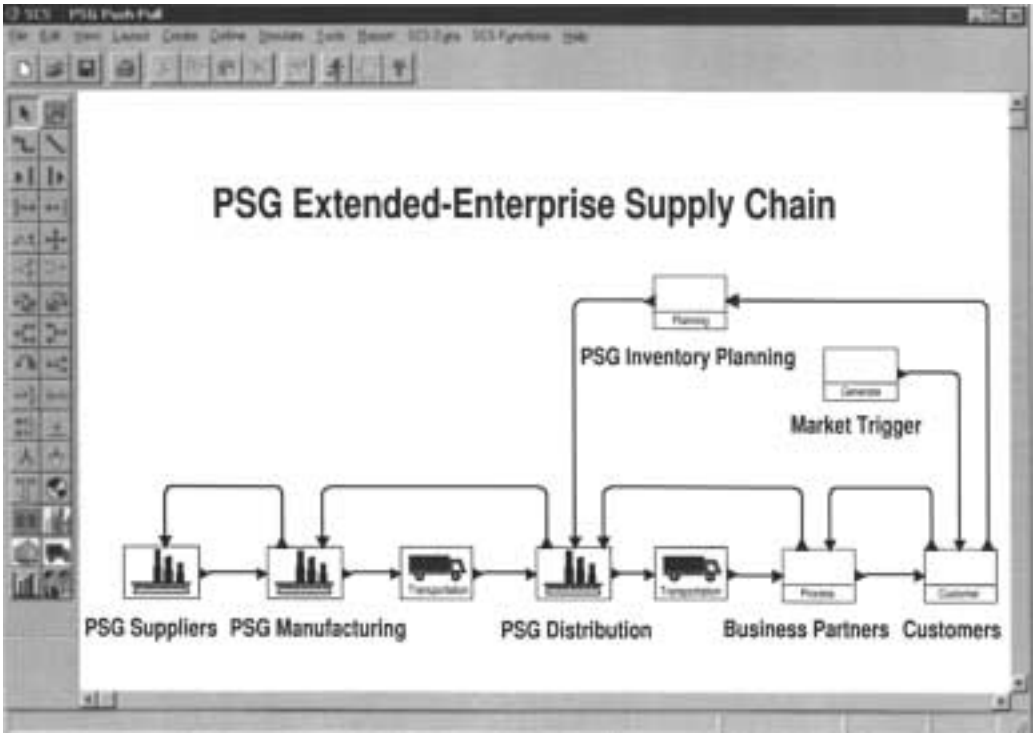


Figure 3: AMT provides a graphical user interface that allows one to interactively construct supply chain scenarios. In this example of an extended-enterprise supply chain, business partners (PSG Business Partners) send orders to a distribution center (PSG Distribution). The distribution center processes the orders and sends products to a transportation node that ships the products to the business partners. The distribution center needs to replenish its stock from time to time, so that it sends replenishment orders to the manufacturing site (PSG Manufacturing) that assembles finished products. The manufacturing site in turn replenishes its parts supply by sending orders to its suppliers (PSG Suppliers). An inventory-planning node (PSG Inventory Planning) representing the AMT optimization engine computes optimal inventory levels for the distribution center based on forecasts of customer demand.

hierarchical process modeling. The user can drill down to include other layers of the supply chain, adding scalability to the modeling approach. The customer node captures demand, forecast, and customer-service requirements. We built in animation to help users visualize the supply-chain activities of orders, goods, and trucks moving between nodes. As the simulation is running, users can view reports, such as service or inventory reports, to see the current status of the simulation. In addition to these real-time reports, AMT also offers the financial and performance reports that we discussed earlier.

An important feature of AMT is the complementary functionality of the optimization and simulation engines. With the optimization engine, the user can perform fast yet very deep what-if analyses, which are beyond the capability of any standard simulation tool. With the simulation en-

gine, the user can invoke the inventory module to perform periodic recalculations of optimal inventory levels while simulating dynamic supply-chain processes and policies. The user can run simulations on optimized solutions, observing how different supply-chain policies at different locations affect the supply chain's performance. Simulation results can also be used to adjust parameters of the simulation or optimization runs. An automated interface between the simulation engine and the optimization engine allows users to invoke optimization periodically during a simulation run, for example to recalculate target inventory levels according to the latest forecast of demand. Users can also use the optimization engine to periodically generate build plans in a mixed push-pull manufacturing environment, taking into account service targets and system uncertainty.

In summary, AMT embodies a creative coupling of optimization, performance evaluation, and simulation, integrated with data connectivity and an Internet-enabled modeling framework. This makes it a powerful and versatile tool for capturing the stochastic and dynamic environment in large-scale industrial supply chains. We model extended-enterprise supply chains as networks of inventory queues, using a decomposition scheme and queuing analysis to capture the performance of each stocking location. We developed multiechelon, constrained inventory-optimization algorithms that use conjugate gradient and heuristic searches for efficient large-scale applications. We developed a supply-chain simulation library consisting of an extensive set of supply-chain processes and policies for modeling various supply-chain environments with little programming effort. It offers performance measures, financial reports, and activity-based costing down to the level of individual stock-keeping units. It also gives the user a way to validate and fine-tune supply-chain parameters based on analytical results.

Extended-Enterprise Supply Chain Management at IBM Personal Systems Group

The IBM Personal Systems Group (PSG) is responsible for the development, manufacture, sale, and service of personal computers (for example, commercial desktops, consumer desktops, mobile products, workstations, PC servers, network PCs, and related peripherals). PSG employs over 18,500 workers worldwide. Sales and marketing groups are located in major metropolitan areas, with manufacturing plants located in the United States, Latin America, Europe, and Asia. In 1998, PSG sold approximately 7.7 million computers under such brand names as IBM PC, Aptiva, ThinkPad, IntelliStation, Netfinity, and Network Station.

Increased competition from such PC manufacturers as Dell and Gateway, which use a direct, build-to-order business model, prompted PSG in 1997 to reevaluate its business practices and its relationships with its supply-chain partners. The goal was to design and implement a hybrid business model, one that incorporated the best features of the direct model (build to order, custom configuration, and inventory minimization) and the best features of the indirect model (final configuration, high customer service, and support), sell-

ing products through multiple channels.

PSG formed a cross-functional team in April 1997 with the task of quantifying the relationship between customer service and inventory throughout the extended supply chain under the existing business model and under various proposed channel-assembly alternatives. We used production data from a subset of PSG's commercial desktop products to develop a baseline supply-chain model in AMT. The model was triggered by end-user demand, reseller ordering behavior, IBM manufacturing and inventory policies, supplier performance, and lead-time variability. We collected actual end-user sales data for 22 reseller locations over five months. Resellers' ordering behavior was influenced by many factors, such as gaming strategies, marketing incentives, confidence in supplier reliability, and stocking for large customer purchases. Modeling each individual activity would have been too complex. Our model captured the aggregate ordering for each PSG reseller by substituting alternative ordering policies, representing current levels of sales activity in the channel. For example, if a particular reseller held an average of 60 days of inventory, the model established a target base-stock level representing 60 days of channel inventory for this reseller. To see what would happen if resellers changed their ordering policies, we changed the levels of channel inventory in the AMT model and ran different what-if scenarios. For each ordering policy, we assumed that a reseller would stock a product at a given level of days of supply.

During the normal course of business, PSG forecasts its manufacturing volumes over a rolling 13-week horizon. The current week's forecast becomes the build plan, which then pushes products built at PSG's manufacturing sites to the distribution warehouse where they are held until the products are eventually ordered, or pulled, by a reseller. This type of replenishment policy captured the logic of PSG's hybrid push-pull manufacturing and ordering system in which PSG built products to a forecast and held them as finished goods in the warehouse until it received orders from its resellers. This system is not a true pull system because product availability influences reseller ordering. Likewise, the system is not a true push system because the backlog of resellers' orders influences the schedules at PSG manufacturing sites. To effectively capture variability caused by component shortages, capacity constraints, and requirements for minimum lot sizes, we analyzed the range of the 13-week forecasts.

PSG's channel look-back expenses dropped by more than $100 million.

PSG set a service target for customer deliveries of three days, 95 percent of the time, which translated directly into the customer-service constraint required by the AMT optimization engine. Combining the simulation engine with the optimization engine, the model recalculated the base-stock levels every week, according to the latest available forecast of demand so that customer orders could be filled within three days 95 percent of the time. This replenishment policy formed the basis for PSG's supplier orders for components and

subassemblies and for its subsequent manufacturing activity. In Phase 1 of the project, we used a reduced data set to construct a simplified prototype model of PSG's supply chain to test assumptions, to investigate alternative modeling algorithms, and to better understand possible limitations of the AMT application.

In Phase 2 of the project, we developed more detailed modeling scenarios to vary channel inventory and to incorporate a channel-assembly policy at the resellers. PSG delivers to its resellers two types of products, (1) standard machine-type models (MTMs), which are fully configured and tested computers, and (2) so-called open-bay machines, which are nonfunctional, basic computers without such preconfigured components as memory, hard files, and CD ROMs. These open bays allow resellers to assemble machines according to specific customer requirements. We found that some resellers converted open bays into standard MTMs as needed and then sold them to their customers. We refer to this as an example of flexibility because resellers can use their current open-bay inventory to fill orders for standard MTMs, instead of stocking open bays exclusively to fill orders for nonstandard MTMs. Other resellers stockpiled open-bay inventory, and if they needed standard MTMs to fill an order, they would reorder from PSG instead of configuring an open bay already in stock (an example of inflexibility). Both methods affect inventory and customer service. Because reseller flexibility could not be defined accurately, we designed different sets of simulation experiments with the intent to bound, or frame, the true impact of channel assembly within the two extreme cases of 100 percent reseller flexibility and 100 percent reseller inflexibility.

We validated the accuracy of the AMT models by comparing the outputs of the simulation runs to historical PSG data. We adjusted our modeling assumptions and parameters as necessary and ran multiple simulations using different parameters and policies. The key results of the study can be summarized as follows:

—Implementing channel assembly based on PSG's existing product structure, low volume environment, and present supply-chain policy reduces inventory very little (inflexible reseller channel behavior).

—Allowing resellers to configure any MTM from their stock of components could improve customer service by two percent and simultaneously reduce inventories by 12 percent (flexible reseller channel behavior).

—Consolidating the demand at 22 configuration sites into three large hubs could improve customer service by six percent and reduce inventories by five percent.

—Based on the existing push-pull supply-chain policy, PSG can reduce channel inventory by 50 percent without affecting its customer-service level. The overall supply-chain inventory levels were far in excess of the optimum needed to maintain PSG's service target.

This and subsequent projects brought together four functional groups—marketing and sales, manufacturing, distribution, and development—to seek a company-wide consensus on PSG's strategic direction and subsequent actions. Our studies contributed directly to PSG's advanced fulfillment initiative (AFI), an effort to in-

crease flexibility in the reseller channel by improving parts commonality in PSG's product structure [Narisetti 1998]. Also, PSG management endorsed the reduction of the number of configuration sites, as a result of changing channel price-protection terms and conditions. The specific terms and conditions were tied to the output of the AMT model, and they were implemented in November 1997 after a series of related enhancements to the logistics process.

PSG has based many of its decisions on how to prioritize project deployment and manage channel inventory on the results of subsequent AMT analyses. While the analysis that drove PSG's initial business transformation was conducted in 1997, the 1998 business benefits were substantial.

The more accurate a reseller's forecasts, the higher the level of service.

PSG reduced its overall inventory by over 50 percent from year-end 1997 to year-end 1998. As a direct consequence of this inventory reduction, PSG's channel look-back expenses dropped by more than $100 million from 1997 levels. Look-back expenses account for payments to distributors and business partners that compensate for price actions on the inventory they are holding. In addition, by selling products four to six weeks closer to when the components are procured, PSG saved an additional five to seven percent on product cost. This equates to more than $650 million of annual savings.

In the months following the original assessment, we conducted further supply-chain studies, including analyses that (1) incorporated the supply chains of business partners; (2) modeled additional geographies; (3) assessed the impact on inventory and customer service of delaying final assembly to the reseller's distribution facilities; and (4) estimated the impact on inventory of reducing manufacturing cycle times. These studies have helped PSG's business partners make more informed decisions on supply-chain policy. In particular, they have led IBM and its major business partners to establish a colocation policy. In colocation, a business partner locates its distribution space inside of IBM, eliminating the need for costly handling and transportation among different sites. Finally, because we found that forecast accuracy greatly affected inventory and customer service, PSG used the AMT to determine the level of service it would promise to its business partners, based on their ability to provide accurate forecasts. The more accurate a reseller's forecasts, the higher the level of service PSG would provide to that reseller. This policy is unprecedented in the industry and has been favorably received by PSG's business partners. Overall, PSG believes that the AMT has been an invaluable asset in developing and implementing world-class supply-chain-management policies.

Other AMT Applications Across IBM

AMT has also been applied and deployed in other IBM manufacturing divisions, including the printing systems division (PSC), the midrange computer division (AS/400), the office workstation division (RS/6000), the storage systems division (SSD), the mainframe computer division (S/390), and PSG's European mar-

ket. A number of PSG's business partners have used AMT, including Pinacor, GE Capital, and Best Buy. IBM's Industry Solution Unit uses the tool externally for consulting engagements. Following are brief descriptions of three recent AMT engagements:

The IBM Printing Systems Company (PSC) is a leading supplier of printer solutions for business enterprises. The product line includes printers for office printing to high-volume production printing. The company employs approximately 4,550 people, with total gross revenue for 1998 of $1.95 billion. In 1996, PSC conducted an intensive testing process on the AMT over a five-month period. In its assessment report, the testing team concluded that AMT produces accurate results, provides productivity improvements over existing supply-chain-management and inventory tools, and improves PSC's precision in validating and creating inventory budgets and turnover objectives. PSC then used AMT to study the effects of forecast accuracy, product structure, the introduction of a new distribution center, and different business scenarios on the performance of the supply network for different product families. In one of the cases alone, it reported inventory savings of $1.6 million, which represented 30 percent of the total inventory holding cost.

Financial savings amount to more than $750 million at PSG in 1998.

IBM's AS/400 division manufactures midrange business computers and servers, providing more than 150 models and upgrades with up to 1,000 features. Assembling these systems requires several thousand unique part numbers, approximately 1,000 of them used at the highest level of assembly just prior to building a complete system. Providing customers with the flexibility to customize the equipment they order by selecting features creates manufacturing complexity and efficiency challenges. The division used AMT to analyze and quantify the impact on inventory and on-time delivery of feature reduction, feature substitution, parts commonality, and delayed customization. The analysis showed that eliminating low-volume parts would improve inventory turnover by 15 percent and that substituting and postponing their final assembly would improve inventory turnover by approximately 30 percent. The AS/400 division has reduced its feature count by approximately 30 percent since 1998 with steady growth in total revenue.

In 1995, IBM established a quick-response service program to provide rapid delivery for customers buying selected mid-range computer memory, storage, and features. In September 1998, IBM instituted the quick-response program as a front end to provide real e-commerce for our large business partners. IBM used AMT to analyze the trade-off between service and inventory in choosing an optimum performance point. It later used it to assess the impact of the quick-response program on allocating inventory between manufacturing and distribution centers. The results helped IBM to maximize business efficiency and contributed to doubling the growth of quick-response revenue in 1998.

Conclusions

The AMT effort uses advanced OR techniques and combines technical innovations with practical and strategic implementations to achieve significant business impacts. IBM has used AMT to address a wide range of business issues, including inventory management, supply-chain configuration, product structure, and replenishment policies. AMT has been implemented in a number of IBM business units and their business partners. Financial savings through the AMT implementations amount to more than $750 million at PSG in 1998 alone. Furthermore, AMT has helped IBM's business partners to meet their customers' requirements with much lower inventory and has led to a co-location policy with many business partners. It has become the foundation for a number of supply-chain-reengineering initiatives. Several IBM business partners view the AMT analyses as key milestones in their collaboration with IBM in optimizing the extended-enterprise supply chain.

Acknowledgments

We gratefully acknowledge the contributions and support of the following people: Chae An, Ray Bessette, Rick Bloyd, Harold Blake, Richard Breitwieser, J. P. Briant, Bob Chen, Feng Cheng, Arthur Ciccolo, Daniel Connors, Ian Crawford, Anthony Cyplik, John Eagen, Brian Eck, Gerry Feigin, Angela Gisonni, John Konopka, Tom Leiser, Tony Levas, Nikorn Limcharoen, Joe Magliula, Barbara Martin, Larry McLaughlin, Bob Moffat, Gerry Murnin, Nitin Nayak, Jim Nugent, Lynn Odean, Krystal Reynolds, Richard Shore, Mukundan Srinivasan, Jayashankar Swaminathan, David Thomas, Bill Tulskie, Burnie Walling, Wen-Li Wang, and James Yeh.

APPENDIX

Optimization of Multi-Echelon Supply Networks with Base-Stock Control

Here we provide a brief overview of the key points of the mathematical model in the optimization engine. Ettl, Feigin, Lin and Yao [1998] give the full details, including topics that we do not touch upon here, such as the treatment of nonstationary demands, the related rolling-horizon implementation, the derivation of the gradients, and many preprocessing and post-processing steps.

We specify the configuration of the supply network using the bills-of-materials structure of the products. Each site in the network is either a plant or a distribution center. Associated with each site and each product processed at the site is a multi-level bill of material. Each site has storage areas, which we refer to as stores, to hold both components that appear on the bills of materials and finished products, which correspond to input stores and output stores. The subscripts i and j index the stores, and S denotes the set of all stores in the network. We assume a distributed inventory-control mechanism whereby each store follows a base-stock control policy for managing inventory. The policy works as follows: When the inventory position (that is, on-hand plus on-order minus backorder) at store i falls below some specified base-stock level, R_i, a site places a replenishment order. In our model, R_i is a decision variable.

For each store i, there is a nominal lead time, L_i, with a given distribution. The nominal lead time corresponds to the production time or transshipment time at the site where the store resides, assuming there is no delay (due to stock-out) in any upstream output stores. The actual lead time, $\tilde{L}_i$, in contrast, takes into account possible additional delays due to stock-

out. Whereas L_i's are given data, $\tilde{L}_i$'s are derived performance measures.

To analyze the performance of each store i, we use an inventory-queue model, for example, Buzacott and Shanthikumar's [1993]. Specifically, we combine the base-stock control policy with an $M^X/G/\infty$ queue model, where arrivals follow a Poisson process with rate λ_i, and each arrival brings in a batch of X_i units, or orders. The batch Poisson arrival process is a good trade-off between generality and tractability. In particular, it offers at least three parameters to model the demand data: the arrival rate and the first two moments of the batch size (whereas a simple Poisson arrival process has only one parameter).

To derive the performance measures at each store i, we need to first generate the input process to the $M^X/G/\infty$ queue. To do this, we take the demand stream (forecast or real) associated with each class, translate it into the demand process at each store by going through the bills-of-materials structure level by level, and shift the time index by the lead times at each level. This process is quite similar to the explosion and offsetting steps in standard MRP analysis. A second piece of data needed for the $M^X/G/\infty$ queue is the service time, which we model as the actual lead time.

Let N_i be the total number of jobs in the queue $M^X/G/\infty$ in equilibrium. Following standard queueing results [Liu, Kashyap, and Templeton 1990], we can derive the mean and the variance of N_i, denoted as μ_i and σ_i^2. We then approximate N_i with a normal distribution:

$$N_i = \mu_i + \sigma_i Z, \tag{1}$$

where Z denotes the standard normal variate. Accordingly, we write the base-stock level as follows:

$$R_i = \mu_i + k_i\sigma_i, \tag{2}$$

where k_i is the so-called *safety factor*. As R_i and k_i relate to each other via the above relation, either can serve as the decision variable. Let I_i be the level of on-hand inventory, and B_i the number of back-orders at store i. These relate to N_i and R_i as follows:

$$I_i = [R_i - N_i]^+ \quad \text{and} \quad B_i = [N_i - R_i]^+, \tag{3}$$

where $[x]^+ = \max(x,0)$. We can then derive the expectations:

$$E[I_i] = \sigma_i H(k_i), \text{ and } E[B_i] = \sigma_i G(k_i) \tag{4}$$

where

$$H(k_i) = \int_{ki}^{\infty} (k_i - z)\phi(z)dz \text{ and}$$

$$G(k_i) = \int_{ki}^{\infty} (z - k_i)\phi(z)dz, \tag{5}$$

and $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$ is the density function of Z. Furthermore, writing $\Phi(x) = \int_0^x \phi(z)dz$, the distribution function of Z, and $\bar{\Phi}(x) = 1 - \Phi(x)$, we can derive the stock-out probability p_i and the fill rate f_i at store i as follows:

$$p_i = \bar{\Phi}(k_i), \text{ and}$$

$$f_i = 1 - \sigma_i\phi(k_i)/\mu_i - \bar{\Phi}(k_i). \tag{6}$$

All of the above performance measures involve the actual lead time at store i, which can be expressed as follows:

$$\tilde{L}_i = L_i + \max_{j \in S > i}(\tau_j), \tag{7}$$

where $S_{>i}$ denotes the set of stores that supply the components needed to build the units in store i, and τ_j denotes the additional delay at store $j \in S_{>i}$. As τ_j is quite intractable in general, with queueing analysis, we have derived the following approximation:

$$\tau_j = \tilde{L}_j r_j \text{ where } r_j := \frac{E[B_j]}{p_j(R_j + 1)}. \tag{8}$$

Intuitively, $E[B_j]/p_j$ is the average number of back-orders at location j conditioned

upon a stock-out there, and each of these back-orders requires an average time of $\tilde{L}_j/(R_j + 1)$ to fill, that is, during the stock-out, on the average, there are $(R_j + 1)$ outstanding orders in process.

Customer demands are supplied from a set of end stores, S_0, stores at the boundary of the network. Consider a particular customer class, and suppose its demand is supplied by one of the end stores, $i \in S_0$. Let W_i denote the waiting time to receive an order. The required customer-service target is

$$P[W_i \leq \beta_i] \geq \alpha_i, \tag{9}$$

where β_i and α_i are given data. When the demand is supplied from on-hand inventory, the delay is simply the transportation time T_i, time to deliver the finished products to customers, which is given; otherwise, there is an additional delay of τ_j. Hence,

$$P[W_i \leq \beta_i] = f_i P[T_i \leq \beta_i] + (1 - f_i)P[T_i + \tau_i \leq \beta_i].$$

For the above to be at least α_i, we need to set f_i, the fill rate, to the following level:

$$f_i = \frac{\alpha_i - P[T_i + \tau_i \leq \beta_i]}{P[T_i \leq \beta_i] - P[T_i + \tau_i \leq \beta_i]}. \tag{10}$$

The quantity τ_i involved in the right-hand side of the above equation can be expressed as $\tau_i = \tilde{L}_i r_i$, following (8). Since r_i involves B_i and R_i, both of which are functions of k_i, and so is f_i, we need to solve a fixed-point problem defined by the equation in (8) to get f_i (or k_i). In the iterations involved in the optimization procedure, however, this fixed-point problem can be avoided by simply using the r_i value obtained from the previous iteration. Once we derive f_i and k_i, the base-stock level (2) and the stock-out probability (6) then follow.

The objective of our optimization model is to minimize the total expected inventory capital throughout the supply network while satisfying customer-service requirements. Each store has two types of inventory: on-hand inventory and work-in-process (WIP) inventory. (The WIP includes the orders in transition, that is, orders being transported from one store to another.) From the above discussion, the expected on-hand inventory at store i is $E[I_i] = \sigma_i H(k_i)$, and the expected WIP is $E[N_i] = \mu_i$. Therefore, the objective function takes the following form:

$$C(\mathbf{k}) = \sum_{i \in S} [c_i' \mu_i + c_i \sigma_i H(k_i)], \tag{11}$$

where c'_i and c_i denote the inventory capital per unit of the on-hand and WIP inventory, respectively, with c_i assumed given, and c'_i derived from the c_i's along with the BOM. We want to minimize $C(\mathbf{k})$, subject to meeting the fill-rate requirements in (10), for all the end stores: $i \in S_0 \subset S$. This is a constrained nonlinear optimization problem. We derive the partial derivatives $\partial/\partial k_j\, C(k)$, all in explicit analytical forms based on the relations derived above (and others). We use these in a conjugate-gradient search routine, for example that of Press et al. [1994]. As the surface of the objective function is quite rugged, to avoid local optima, we also implemented several heuristic search procedures. For instance, evaluate a set of randomly generated initial points and pick the best one (in terms of the objective value) to start the gradient search.

References

Andersson, J.; Axsaeter, S.; and Marklund, J. 1998, "Decentralized multi-echelon inventory control," *Production and Operations Management*, Vol. 7, No. 4, pp. 370–386.

Arntzen, B. C.; Brown, G. G.; Harrison, T. P.; and Trafton, L. L. 1995, "Global supply chain management at Digital Equipment Corporation," *Interfaces*, Vol. 25, No. 1, pp. 69–93.

Bagchi, S.; Buckley, S.; Ettl, M.; and Lin, G. 1998, "Experience using the supply chain simulator," *Proceedings of the Winter Simulation Conference*, Washington, DC, December, pp. 1387–1394.

Buckley, S. 1996, "Supply chain modeling," *Proceedings of the Autofact Conference*, Detroit, Michigan, pp. 749–756.

Buckley, S. and Smith, J. 1997, "Supply Chain Simulation," *Georgia Tech Logistics Short Course*, Atlanta, Georgia, pp. 1–17.

Buzacott, J. A. and Shanthikumar, J. G. 1993, *Stochastic Models of Manufacturing Systems*, Prentice-Hall, Englewood Cliffs, New Jersey.

Camm, J. D.; Chorman, T. E.; Dill, F. A.; Evans, J. R.; Sweeney, D. J.; and Wegryn, G. W. 1997, "Blending OR/MS, judgment, and GIS: Restructuring P&G's supply chain," *Interfaces*, Vol. 27, No. 1, pp. 128–142.

Ettl, M.; Feigin, G.; Lin, G.; and Yao, D. D. 1998, "A supply network model with base-stock control and service requirements," *Operations Research*, forthcoming.

Feigin, G.; An, C.; Connors, D.; and Crawford, I. 1996, "Shape up, ship out," *ORMS Today*, Vol. 23, No. 2, pp. 1–5.

Graves, S.; Kletter, D. B.; and Hetzel, W. B. 1998, "A dynamic model for requirements planning with application to supply chain optimization," *Operations Research*, Vol. 46, No. 3, pp. S35–S49.

Kruger, G. A. 1997. "The supply chain approach to planning and procurement management," *Hewlett-Packard Journal*, Vol. 48, No. 1, pp. 1–9.

Lee, H. L. and Billington, C. 1993, "Material management in decentralized supply chains," *Operations Research*, Vol. 41, No. 5, pp. 835–847.

Liu, L.; Kashyap, B. R. K.; and Templeton, J. G. C. 1990, "On the $GI^X/G/\infty$ system," *Journal of Applied Probability*, Vol. 27, No. 3, pp. 671–683.

Narisetti, R. 1998, "How IBM turned around its ailing PC Division," *Wall Street Journal*, October 3, p. B1.

Press, W. H.; Teukolsky, S. A.; Vettering, W. T.; and Flannery, B. P. 1994, *Numerical Recipes in C*, second edition, Cambridge University Press, New York.

Swegles, S. 1997, "Business process modeling with SIMPROCESS," *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 606–610.

Bob Moffat, general manager for manufacturing, procurement, and fulfillment at IBM Personal Systems Group, said during the presentation of the paper at the Edelman competition: "We reduced our channel inventory from over three months to approximately one month. As a direct consequence of this inventory reduction, our division has reduced 1998 price protection expenses by over $100M from the previous year. Price protection expenses are what we reimburse business partners whenever we take a price action on products they are holding. We had reduced our end-to-end inventory from four and a half months to less than two months by the end of 1998. By closing the gap between component procurement and product sale by four to six weeks, there is a savings on product cost of at least five percent. This equates to more than $650 million of annual savings. AMT has improved our relationships with business partners, making them more efficient, more productive, and ultimately more powerful in the marketplace. I believe this will lead to a fundamental change in our business culture, a unification of basic value among suppliers, manufacturers, and resellers."

Jean-Pierre Briant, IBM vice president for integrated supply chain, further explained: "The AMT tool has found application in almost every supply chain within IBM. It helps us understand our extended supply chain—from our suppliers' suppliers to our customers' customers. We have deployed the AMT tool to assist external companies in managing their supply chains, with very effective results."

Jim Manton, president and COO of Pinacor, said: "The results that the [AMT] team delivered on the supply chain analy-

sis helped Pinacor identify opportunities for optimizing the product flow between our companies. . . . I am pleased to see that both IBM and Pinacor are focusing on the recommendations to make the necessary improvements. . .".

Mac McNeill, senior vice president of global operations for GE Capital IT Solutions, who sponsored a four-month project using the AMT to model GE Capital's personal-computer supply chain commented: "The modeling allowed us to develop a base case using actual end-user customer sales and then to quickly model and optimize many alternatives based on various levels of GE forecast accuracy, IBM fill rates, transit times, in-bound and out-bound delays, and commonality of parts. The optimization results will allow us to develop action plans to balance improved levels of serviceability with lower levels of inventory."

Xilinx Improves Its Semiconductor Supply Chain Using Product and Process Postponement

Alexander O. Brown — *Owen Graduate School of Management, Vanderbilt University, 401 21st Avenue South, Nashville, Tennessee 37203*

Hau L. Lee — *Graduate School of Business and the Department of Management Science and Engineering, Stanford University, Stanford, California 94305*

Raja Petrakian — *Xilinx, Inc., 2100 Logic Drive, San Jose, California 95124*

The semiconductor firm Xilinx uses two different postponement strategies: product postponement and process postponement. In product postponement, the products are designed so that the product's specific functionality is not set until after the customer receives it. Xilinx designed its products to be programmable, allowing customers to fully configure the function of the integrated circuit using software. In process postponement, a generic part is created in the initial stages of the manufacturing process. In the later stages, this generic part is customized to create the finished product. Xilinx manufactures a small number of generic parts and holds them in inventory. The use of these generic parts allows Xilinx to hold less inventory in those finished products that it builds to stock. And for some finished products, Xilinx can perform the customization steps quickly enough to allow it to build to order.

High technology industries, such as semiconductors and computers, are characterized by short product life cycles and proliferating product variety. Faced with such challenges, companies in these industries have found that delaying the point of product differentiation can be an effective technique to cut supply-chain costs and improve customer service. This postponement technique is a powerful way to enable cost-effective mass customization [Feitzinger and Lee 1997]. To use

0092-2102/00/3004/0065/$05.00
1526–551X electronic ISSN
This paper was refereed.

INVENTORY—PRODUCTION—MULTI-ITEM, ECHELON, STAGE
INDUSTRIES—COMPUTER—ELECTRONICS

INTERFACES **30:** 4 July–August 2000 (pp. 65–80)

postponement effectively, companies must carefully design their products and processes. Through careful design of the product and the process, many electronics and computer companies have been able to delay the point of product differentiation, either by standardizing some components or processes or by moving the customization steps to downstream sites, such as distribution centers or retail channels. Lee [1993, 1996]; Lee, Billington, and Carter [1993]; Lee, Feitzinger, and Billington [1997]; and Lee and Sasser [1995] give examples.

Postponement concepts have also been applied in other industries, such as the automobile industry [Whitney 1995] where product modularity enables delayed customization of auto parts. Indeed, Ulrich [1995] showed that a high degree of product modularity coupled with component-process flexibility could render postponement possible and effective. Lee, Padmanabhan, and Whang [1997] also said that both product and process modularity support postponement. Modular designs for products or modular processes (a manufacturing process that can be broken down into subprocesses that can be performed concurrently or in different sequential order) are techniques that enable postponement.

The semiconductor industry has been plagued by a proliferation of product variety because of the overlapping product life cycles—companies introduce new or enhanced versions of products before existing products reach the ends of their life cycles. In the programmable-logic segment of the industry, new customers will use the enhanced versions in their products, but some existing customers may delay adopting the new versions despite their improved performance and price. Periods of appreciable demand for a version of a product may range from six months to two years, with products sometimes having an extended period of very low end-of-life demand. Thus, semiconductor companies must offer many products simultaneously. The product-variety problem is compounded by unpredictable demands and long manufacturing lead times.

Semiconductor firms face unpredictable demand, in large part, because of their upstream position in the supply chain. An integrated circuit (IC) made by a semiconductor firm is a component of other subassemblies or final products. Thus, it must pass through other companies, such as contract manufacturers, distributors, and resellers, before the final product reaches the end consumer. Lee, Padmanabhan, and Whang [1997] describe the "bullwhip effect" in which demand fluctuations increase as you travel upstream in the supply chain. Since semiconductor firms are located far upstream in the supply chain, they often face such large fluctuations.

Manufacturing cycle times in the semiconductor industry are still very long despite advances in the process technology. The manufacturing process, consisting of wafer fabrication, packaging, and testing, takes about three months. With such long manufacturing lead times, the semiconductor companies must hold large inventories of finished goods or their customers—computer assemblers, telecommunication manufacturers, or other electronics manufacturers—must hold large

inventories to hedge against demand uncertainties.

Product variety, long production lead times, and demand unpredictability negatively affect the manufacturing efficiency and performance of both semiconductor companies and their customers. These characteristics also affect the customer's product-development processes. For example, one part of a telecommunications-equipment manufacturer's product-development process might be the custom design of application-specific integrated circuits (ASICs). The design process often includes creating a number of prototypes before settling on a final working design. Because of the long production times, there is often a significant delay between designing and receiving prototypes. Since time to market is a key factor in the success of high-tech products, this delay may be very costly for the manufacturer. To compress the cycle, such manufacturers may request many prototypes towards the beginning of the design process, resulting in additional design and development costs.

Semiconductor companies offer many products simultaneously.

Product variety, long lead times, and demand unpredictability are all unavoidable and problematic characteristics of the semiconductor industry. However, some companies are finding new ways to cope with them. Xilinx, Inc., uses innovative design principles of postponement to avoid excessive inventory while providing great service to its customers. It uses both product and process postponement extensively.

In product postponement, the firm designs the product so that it can delay its customization, often by using standardized components. Xilinx relies on a more extreme form of product postponement. Instead of the firm performing the final configuration during manufacture or even distribution, it designs the ICs so that its customers perform the final configuration using software. Consequently, Xilinx greatly shortens the product-development cycles of its customers, as the customers do not have to specify the full features and functionalities of the ICs before production.

Using proprietary design technologies, Xilinx creates many types of ICs, differentiated by such general features as speed, number of logic gates, package type, pin count, and grade. Although the customers perform the final configuration of the logic, they must order products with the appropriate general features. For example, a customer with a large and complex design requiring high speed must select a physical device type with a large number of logic gates and a high speed. Later the customer can configure the logic of the device using software, creating an enormous number of possible designs. Product postponement is very suitable for programmable devices because a near-infinite number of designs can be created from a few thousand physical-product permutations.

In process postponement, the firm designs the manufacturing and distribution processes so that it can delay product differentiation, often by moving the push-pull boundary or decoupling point toward

the final customer. A push-pull boundary is the point in the manufacturing-and-distribution process at which production control changes from push to pull. Early in the process, prior to the push-pull boundary, the firm builds to forecast. Later in the process, after the push-pull boundary, it builds to order. Often, process designs allow manufacturers to change their push-pull boundaries. A highly celebrated example of process postponement is the case of Benetton, which used to make sweaters by first dyeing the yarns and then knitting them into finished garments of different colors. Its push-pull boundary used to be at finished sweaters—all production was built to forecast. Benetton resequenced its production process so that it first knits undyed garments, and then dyes them (and thereby customizes them to the different color versions) on demand. Hence, its new push-pull boundary is between knitting and dyeing [Dapiran 1992].

To improve its manufacturing process, Xilinx focused on creating a new push-pull boundary, working with its suppliers. Rather than going through all the steps to create an IC in its finished form from a raw silicon wafer, Xilinx divides the process in two stages. In the first step, its wafer-fabrication supply partners manufacture unfinished products, called dies, and hold inventory of this material. This inventory point is the push-pull boundary. Based on actual orders from the customers, another set of supply partners pull dies from inventory and customize them into finished ICs.

The Xilinx Supply Chain

Digital semiconductor devices can be broadly grouped into three categories: memory, microprocessors, and logic. While the general-purpose microprocessors can execute almost any logical or mathematical operation, logic devices provide specific functionality at lower cost and greater speed. However, the traditional method of defining the functions of a logic device is to configure it during the fabrication process. Recently, with the introduction of programmable logic devices, it has become possible to customize a generic but more expensive logic device using software after the logic device has been completely manufactured and packaged.

Xilinx was one of the first to use a virtual business model.

Founded in 1984, Xilinx developed the field-programmable gate array (FPGA), a programmable logic device, and it has become one of the two largest suppliers of programmable logic solutions in the world. The company's revenues in 1997 were $611 million and the gross margin was around 62 percent. Xilinx was one of the first semiconductor companies to use a virtual business model: it subcontracts out logistics, sales, distribution, and most manufacturing to long-term partners. Xilinx's only manufacturing facilities are its California and Ireland facilities that just perform some final testing. It meets about 74 percent of its total demand through distributors, whose expertise has evolved beyond traditional warehousing and inventory management to include engineering functions, such as helping customers design Xilinx parts into their systems. Xilinx

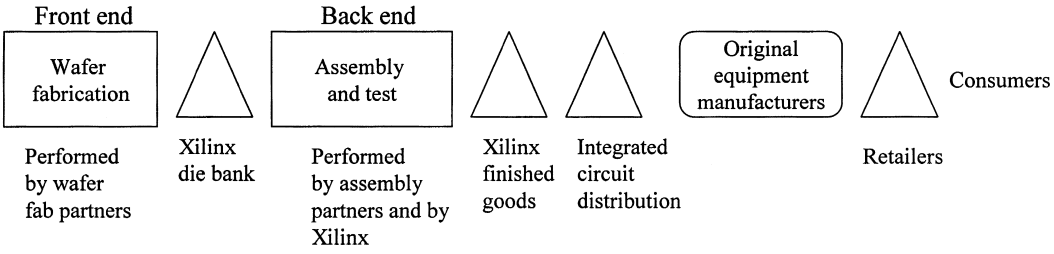


Figure 1: In the Xilinx supply chain, supply partners perform the wafer fabrication and assembly, while Xilinx manages production levels and the inventory levels in die bank and finished goods. After production, a distributor buys the integrated circuits and supplies them to original equipment manufacturers that incorporate the integrated circuit into their products. Consumers purchase the products through retailers. Triangles represent inventory stocking locations, and squares represent manufacturing processes.

keeps certain core functions in-house, such as technology research, circuit design, marketing, manufacturing engineering, customer service, demand management, and supply-chain management. This virtual business model provides Xilinx with a high degree of flexibility at low cost. Its partners benefit because Xilinx uses standard manufacturing and business processes and aggressively drives process improvements through technical innovation and re-engineering. Although the virtual model has strategic risks (the core competencies becoming commodity-like) and operational risks (unexpected lack of available capacity at suppliers), it has proven highly successful in the industry [Lineback 1997].

Today, most of Xilinx's competitors have access to the same fabrication process technology through their own wafer-fabrication partners. The technology and manufacturing gap between members of the industry is closing. Consequently, Xilinx sees management of the demand-and-supply chain as providing it with a competitive advantage in the market. In 1996, Xilinx executive management initiated a major initiative to overhaul the company's practices and processes for managing supply and demand.

In the Xilinx supply chain, the flow of materials begins with the fabrication process (front end), where raw silicon wafers are started and manufactured using hundreds of complex steps that typically take two months (Figure 1). Anywhere from 20 to 500 integrated circuits come from each fabricated wafer. In the last process steps of the front end, the wafers are sorted and tested for basic electrical characteristics. Although precise information is not available until the final test after assembly, this step provides some useful indications of the proportion of good integrated circuits on the wafer and the speed mix that they are likely to yield. After sorting and preliminary testing, wafers are stored in inventory—the die-bank. Planning wafer starts to ensure proper die-bank inventory is a major challenge, requiring such information as demand forecasts, projected yields, and work in process to determine the volume and mix of wafers needed to meet the demand and inventory targets.

The next link in the supply chain is the back end, a term that refers to both the assembly and test processes. In the back-end, wafers are first cut into dies, or individual "raw" integrated circuits. There are approximately 100 different types of dies. To be usable, the integrated circuits must be placed in a package, a plastic casing with electric lead pins, that allows them to be later mounted in a circuit board. There are usually about 10 to 20 package types from which a customer can select for a given die. The dies are wire bonded to form a permanent electrical contact with the package. The packaged dies are then tested electrically to determine if they meet stringent design and quality requirements and to determine their speed. There are usually about five to 10 different possible speed grades. The packaged devices that pass the quality tests are then stored in finished-goods inventory. With lead times of three weeks for assembly and test, planning of back-end starts is difficult, requiring information on both the backlog of orders and demand forecasts. One complexity involves the issue of device speed. Although Xilinx understands the expected fraction of dies that will yield to each speed level, the actual fraction for any given die is different. Thus, planning using the expected fraction of dies at each speed level will often result in a mismatch of supply and demand. To meet the demand, Xilinx will start more material in the back end and pick wafers intelligently using measurements collected in the fabrication-and-sort step.

Most Xilinx customers are serviced through distributors who maintain inventories of Xilinx finished-goods parts. The advantages distributors provide to Xilinx are that they have cost-effective means for handling large numbers of small to medium-size customer orders and they offer such value-added services as inventory consolidation, inventory management, and procurement-program support. The cost of Xilinx is that they add an extra link to the supply chain, causing a potential distortion in demand information. The lack of end-demand visibility can be partially offset when distributors provide Xilinx with systematic data regarding point of sale (POS), bookings, backlog, and inventory. Most Xilinx customers are original equipment manufacturers (OEMs) that put one or more Xilinx parts on a circuit board and then assemble a large system using the board and other components. The OEMs then sell these systems to other customers using various marketing and distribution channels. The Xilinx supply chain is further complicated by the practice of many OEMs of subcontracting the board assemblies to specialized vendors.

On-time delivery is emphasized at Xilinx. As a result, Xilinx has often resolved the trade-off between inventory and on-time delivery by adding inventory. One of the key goals of the supply-chain-management initiative is to achieve the same levels of customer service with lower inventory costs throughout the supply chain.

Product Postponement: The Programmable Logic Devices

Before recent developments in programmable logic, logic devices were primarily ASICs in which the logic was built in during wafer fabrication. Typically, the OEM customer would design an ASIC as part of

a larger design of the system board on which the ASIC would be mounted. The OEM customer submitted a design for the ASIC to a semiconductor manufacturer, who fabricated a prototype of the device according to design specifications. The characteristics of ASICs were fully determined during fabrication, and hence the OEM customer receiving an ASIC could use it only for the intended design. Yet, because of changes in the system specifications or design flaws, design iterations were very common in such product-development projects in the high-technology industries (Figure 2). Any change in the design of an ASIC required both modifying the semiconductor-fabrication process and manufacturing additional prototype ASICs using the modified process. A change in the fabrication process could cost hundreds of thousands of dollars and manufacturing prototype ASICs could take over three months. As a result, design iterations in systems using ASICs were very time consuming [Trimberger 1994].

With programmable logic devices, the OEM customer receives a "generic" device. These devices are not completely generic—each type has features that cannot be customized. Thus, once a customer chooses a generic die type, the customer can customize within a certain range of parameters. The features that create these hard design limits include die packaging, speed grade, maximum number of logic gates, voltage, power, maximum die input and output, and software programming methodology:

—The customer chooses from a set of possible package types and lead-pin counts. Different packages have different thermal and protective properties and have different maximum electrical input and output characteristics.

—The customer chooses from a set of speed grades, each of which produces a different clock rate. Higher speeds may be required for some applications.

—The customer chooses from a set of possible device sizes, specified by the number of logic gates. The number of logic gates determines the size and complexity of the logic design that can be implemented.

—The customer selects from a variety of voltages used to power the device (usually 2.5 V, 3.3 V, or 5V).

—Each generic device type has different power constraints.

—Each generic device has different maxi-

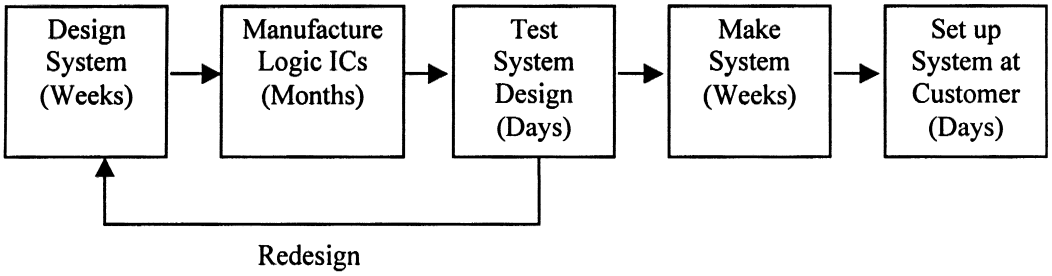


Figure 2: When building a system using an ASIC, the manufacturer incorporates the logic when the integrated circuit is manufactured. Thus, the designer must wait for a new integrated circuit to be manufactured to make design changes.

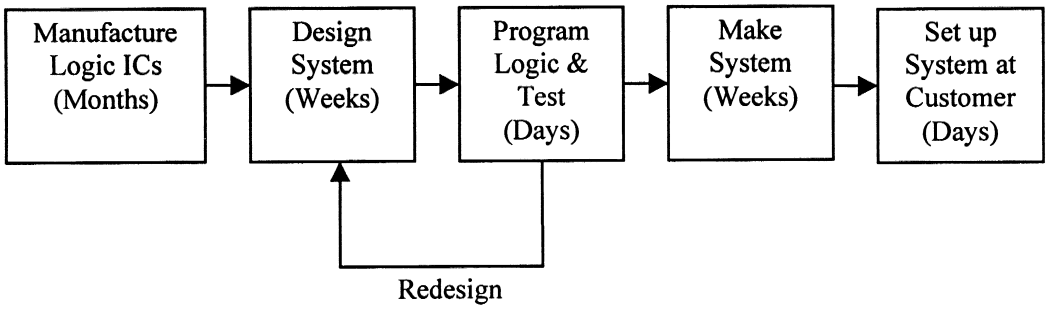


Figure 3: When building a system with a programmable logic device, the customer incorporates the logic using software after the integrated circuit is manufactured. Thus, design changes can be made quickly using software. In contrast to Figure 2, the steps "manufacture logic IC" and "design system" are reversed.

mum input and output electrical characteristics, for example, the maximum level of current that the device can put out.
—The customer may select a device that uses a familiar programming methodology.

Although the customer must decide on some characteristics in advance, the essential characteristic of the final device, the logic function of the device, is not defined in physical processing. Instead, the OEM customer programs, in minutes or hours, the programmable logic device using software running on a personal computer. The user downloads the information into the generic die and thus completes a fully customized logic device. With such a programmable logic device, the process for designing an end system is now dramatically different (Figure 3). Each design iteration takes less time as does the overall design and development process.

Besides shortening the design-process time, product postponement can improve the operational efficiency of the supply chain by reducing the procurement lead times. ASIC suppliers often operate under a build-to-order system, not maintaining finished-goods inventory (but they may have some in-process inventory). As a result, the procurement lead times for OEM customers are sometimes two to three months long. Since accurate forecasting of demand at the specific ASIC device level over such a long horizon is difficult, OEM customers using ASICs often keep large inventories of the ASICs. Programmable logic suppliers can afford to keep inventory in finished-goods form or in the die bank because programmable logic devices are more generic with more predictable demand. Thus, lead times for procuring programmable logic devices are in days or weeks so OEM customers who use them need less inventory.

In-system programming (ISP) allows even greater product postponement. With this capability, customers can easily program or reprogram the logic even after the device is installed in the system (Figure 4). For example, electronic systems such as multi-use set-top boxes, wireless-telephone cellular base stations, communications satellites, and network-management systems, can now be fixed, modified, or upgraded after they have been installed.

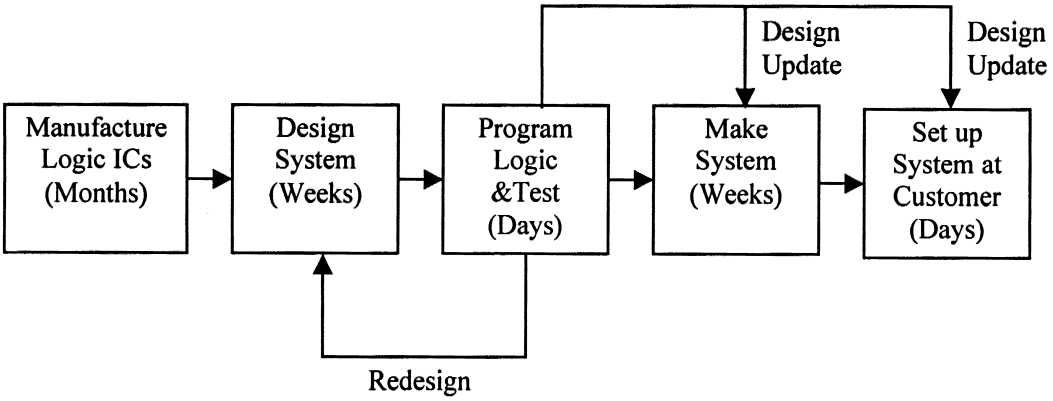


Figure 4: In a system built with a programmable logic device with in-system programming capability, the logic can be incorporated after the system is set up with the customer.

Process Postponement: The Die Bank as the Push-Pull Boundary

The use of product postponement allows Xilinx's customers to create a near-infinite number of different products (different logic designs) from a few thousand types of physical products (Xilinx finished goods). However, since demand for each finished good is usually very uncertain and manufacturing takes around three months, achieving excellent service with reasonable overall inventory levels was a challenge with this many different finished goods. Since many of the finished goods use the same type of die, Xilinx recognized an opportunity to implement process postponement to simultaneously reduce inventory and increase service responsiveness.

Its revised process using postponement works as follows. Instead of using the projected demands for individual finished goods to determine production at the front end, Xilinx aggregates the demands for finished goods into die demands and uses the projected die demands to determine the front-end production starts. After completing the front-end stage, it decides how to customize the dies into different finished goods in the back-end stage. It thus postpones product differentiation, moving it from the beginning to the end of the front-end stage. It still bases customization in the back-end stage on demand forecast (push), with inventory being held in finished-goods form. Thus, the push-pull boundary remains at the end of the process. Since the point of product differentiation moves forward but the push-pull boundary is still at the end of the process, we refer to this approach as partial postponement. Eppen and Schrage [1981] initially proposed this approach in a multilevel distribution setting; it is equally applicable to this manufacturing setting.

Although partial postponement provides benefits, moving the push-pull boundary to an earlier point in the process can increase them. In full die-bank push-pull postponement, the generic dies are held in inventory (the die bank) immediately after the front-end stage, and this die bank becomes the new push-pull boundary. No inventory is held in finished-

goods form; instead, the dies are customized according to customer orders.

We compared die-bank push-pull postponement and the no-postponement approach by analyzing the inventory and service trade-off for each approach using data from a family of finished goods derived from the one die type. We assumed independent and normally distributed demands and a weekly periodic review base-stock policy. For the no-postponement approach, we modeled the system as independent inventory nodes, each representing a finished goods part. We calculated the minimum inventory required to meet a service constraint (maximum expected back orders) for each node and summed the inventory across nodes. For a given level of safety stock, we estimated the expected back orders for each node using the demand uncertainty and the planning lead time [Nahmias 1993]. For the die-bank push-pull postponement approach, we modeled the system as a single inventory node at the die bank. We estimated expected back orders at this node using the demand uncertainty of the aggregated die demand. We showed that the die-bank push-pull strategy offers significant improvements (Figure 5).

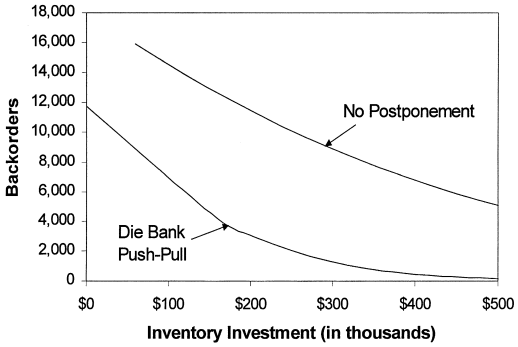


Figure 5: The graph shows the expected number of back orders as a function of the total inventory for two approaches: the no-postponement approach and the die-bank push-pull approach. For the same level of inventory investment, the expected number of back orders is much lower under the die-bank push-pull approach.

Although this die-bank push-pull postponement strategy offers performance improvements, it is not acceptable for customers that require fast deliveries. Thus, if the back-end lead time is two weeks and the customer needs delivery in one, Xilinx could not meet the customer's requirement. Xilinx wanted to move from a partial postponement approach to the die-bank push-pull approach and still satisfy such customer requirements. Thus, it has adopted a hybrid approach. Xilinx has been reducing back-end lead times, and the times for the majority of products are now shorter than customers usually require. It builds these products for the die bank according to customer orders (the die-bank push-pull strategy). It builds finished goods with longer back-end lead times and shorter delivery time to forecast (the partial-postponement strategy).

To determine the distribution of inventory between finished goods and die bank, we used the same number of finished goods as in the previous analysis. We assumed each finished-goods part had one of two back-end lead times: a time equal to the customer-response time and one longer than the customer-response time (set at the average for the parts with lead times greater than the customer-response time). We increased the percent of parts with the short lead time from 0 to 100 percent to generate the results. To avoid concerns about the order in which we selected

finished goods for back-end lead time reduction, we assumed equal demands for all finished goods. So that we could use Eppen and Schrage's [1981] model to analyze the partial postponement approach, we assumed all parts had the same coefficient of variation.

For parts with the short back-end lead time, we used the die-bank push-pull approach and determined the minimum die-bank inventory to maintain the desired level of service. For the parts with the longer back-end lead time, we used the partial-postponement approach. For these parts, we determined the inventory levels required for the given service level using Eppen and Schrage's results [1981]. Their results are for just such a partial-postponement structure (under a different name), and they allow us to calculate the effective demand uncertainty as a function of the individual finished-goods uncertainty levels and the front-end and back-end lead times. Using these results, we calculate the total safety stock in finished goods for a maximum level of expected backorders.

When few parts have short lead times, we must manage most parts using the partial-postponement approach, keeping most inventory in finished goods. As the number of products with short back-end lead times increases, we can build more parts from the die bank to meet customer orders, decreasing inventory in finished goods and increasing that at the die bank. The decrease in finished-goods inventory is much more rapid than the increase in die-bank inventory. Thus, moving towards the pure die-bank push-pull approach reduces inventory and dramatically reduces

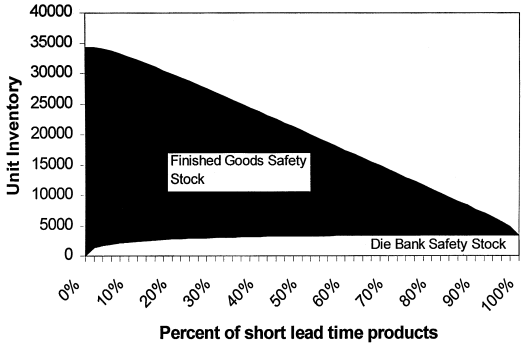


Figure 6: The figure illustrates the inventory distribution between die bank (white) and finished goods (black) when adopting a hybrid strategy. The horizontal axis is the proportion of finished goods that have back-end lead times within the customer-response-time window. As this proportion grows, more of the products can be built to order. Thus, the total inventory decreases significantly and the mix of inventory becomes more heavily weighted to die bank. Results are shown for a constant service level (as measured by expected back orders).

cost since the cost of finished goods is 40 percent more that the die cost.

Table 1 summarizes the four process-postponement approaches. The primary driver of the benefits of process postponement is the risk pooling or statistical pooling that occurs when aggregating demands for many finished goods into demand for fewer dies. The aggregate demand is less uncertain, and thus the firm can hold less inventory to provide the same level of service. The risk-pooling effect is large when the number of finished goods for each die type is large and the correlation between finished-goods demands is small. A large correlation between two finished goods means that if demand is larger than expected for one finished good, it will likely also be larger than expected for the second finished good. Fortunately, at Xilinx, there are a

Strategy	Postponement of product decision	Inventory at die bank	Inventory at finished goods
No postponement			■
Partial postponement	■		■
Die bank push-pull	■	■	
Hybrid	■	■	■

Table 1: For each of four approaches to managing Xilinx's process, the table indicates whether or not postponement is used and where inventory is held—in the intermediate and generic form at die bank or in the final form at finished goods. Xilinx adopted the hybrid model, allowing it to reduce inventories and maintain a high level of customer service.

large number of finished goods for each die (50 to 150) and the average correlation between the finished goods was found to be only 0.018.

Using postponement and holding most inventory at die bank has a number of additional benefits. Inventory held at the die bank is less costly than that at finished goods. About 30 to 50 percent of each product's total value is added in the back-end stage. Inventory held at the die bank also has a lower risk of obsolescence. Many finished goods have just a few customers. If demand drops unexpectedly, Xilinx may be left with inventory of these goods that it cannot sell to anyone else. Die inventory, however, has not yet been customized, and its flexibility greatly reduces the risk of obsolescence. Obsolescence costs in the industry are often about five percent of gross inventory per year, nearly all for finished goods. Postponement makes inventory management easier. In practice, inventory cannot be managed solely by a model-based system. Its decisions must be adjusted for issues beyond the model's scope. With process postponement, management can focus on managing the inventory of the 100 dies rather than trying to make decisions for 10,000 finished goods.

Implementing Process Postponement

Implementing process postponement often requires redesigning current products while trying to keep the changes transparent to the customer. Fortunately this can be done fairly easily in high-technology manufacturing because of the short life of products. To redesign a product to enable process postponement, a manufacturer can simply wait the short time until the next product-generation release when many customers will convert their designs to take advantage of speed and price benefits.

Xilinx designs products to allow for the use of process postponement, keeping the degree of customization low through the front-end stage. For a few general product categories, the die options (for example, many options for logic cell count) are numerous but packaging options are few. Thus, process postponement provides minimal advantage, and little can be done from a design perspective because some features (such as logic cell count) can be created only during the front-end stage.

Xilinx has pursued three process-related initiatives to make process postponement more effective—inventory modeling, supply-mix prediction, and back-end cycle-time reduction. It uses inventory

modeling to determine the appropriate push-pull boundaries for finished goods and to determine inventory levels at various stocking locations. For parts in finished-goods stock, it is optimal to keep inventory in the die bank for quick replenishment instead of using pure partial postponement. Xilinx uses inventory models to improve the hybrid strategy and to determine the optimal level of inventory to hold in the die bank to replenish finished goods and to fill orders for build-to-order parts. It currently uses a multi-echelon model developed jointly with IBM [Ettl et al. forthcoming; Brown et al. 1999].

In the supply-mix-prediction initiative, Xilinx uses statistical models to predict the speed mix of the die-bank inventory. Customer orders specify the desired speed. To customize dies from the die bank to meet customer orders or to replenish finished-goods stock, Xilinx must know how many dies are in each speed yield in the die-bank inventory. Xilinx can easily predict the average fraction of die per wafer that will be of each speed. However, due to slight perturbation in the wafer-fabrication process, the actual fraction for each individual wafer will be different. The objective of the supply-mix initiative is to predict this fraction. Although the true speed of a device is not known until it completes the assembly and test stages, Xilinx can get initial data using a test on die-bank inventory called wafer sort. Using this data, Xilinx applies regression and other statistical methods to estimate speed yield distributions quite accurately [Ehteshami and Petrakian 1998]. This knowledge enables a planner to choose wafers from the die-bank inventory that closely match the order requirements, thus reducing the wasted dies and improving response times.

> Xilinx reduced its inventory levels without harming overall customer service.

The third initiative to improve process postponement is a continuing process to reduce back-end lead times. Xilinx has worked with its manufacturing partners to reduce the wafer-fabrication time from three to one-and-a-half months. For Xilinx to make the die bank the push-pull boundary, the back-end lead time must be short. With a shorter back-end lead time, Xilinx can satisfy a larger proportion of customer orders using the die bank as the push-pull boundary instead of finished goods. Much of the back-end lead time is administrative time. Thus, Xilinx has been able to streamline the process and reduce the lead time through information technology and closer supplier (for assembly and testing) involvement. Internal planning and order fulfillment systems have been made more responsive and electronic data interchange or Extranet web-based tools have been used to expedite the exchange and processing of information between Xilinx and its worldwide vendors.

Conclusion

Xilinx has created tremendous values through product and process postponement. In the case of product postponement, it has found the value of ISP and IRL to be tremendous. For example, Hewlett-Packard Company used a Xilinx field-programmable gate array, a powerful variety of programmable logic devices,

when it designed the LaserJet Companion, reducing its design cycle by an estimated six to 12 months [Rao 1997]. For the electronics industry, Reinertsen [1983] estimated that a six-month delay in the development time of a product reduces the profits generated over the product's life cycle by a third.

Firms are only beginning to realize the potential of product postponement. Rao [1997] describes how IBM designed asynchronous transfer mode (ATM) networking switches when the industry had not yet fully developed standards and protocols. Using programmable logic devices with ISP capabilities, it was able to deliver systems to its customers that could easily be upgraded to the latest standards with no hardware changes. With more recent technological advances, firms can even provide these upgrades through the Internet for systems that are online. Villasenor and Mangione-Smith [1997] describe how FPGAs are changing the field of computing, possibly resulting in major technological breakthroughs. They envision computing devices that adapt their hardware almost continuously in response to changing input. They also predict that configurable computing is likely to play a growing role in the development of high-performance computing systems, resulting in faster and more versatile machines than are possible with either microprocessors or ASICs. With such technology, firms can postpone the definition of a product without limit, an ultimate form of product postponement.

Process postponement has also significantly improved financial performance at Xilinx. Although Xilinx has not kept performance metrics since it first introduced process postponement, its refinement of the process-postponement hybrid from the third quarter of 1996 to the third quarter of 1997 helped it to reduce corporate inventory from 113 dollar days to 87 dollar days (dollar days is the net inventory divided by the cost of goods sold for the quarter times 90 days per quarter). This translates directly into cost savings and improvements in the company's return on assets. At the same time, customer service, measured by the percentage of times that customer orders are filled on time, has remained the same. This is particularly impressive because during that period, Xilinx released an unusually large number of new products. Despite the proliferation of product variety and the increase in service back orders associated with technical problems with the new products, Xilinx reduced its inventory levels without harming overall customer service. During this time period, the inventory levels at the key competitors increased to well over 140 dollar days.

Gaining acceptance of the models took time and effort.

Currently, Xilinx is working closely with its partners to further reduce lead times at both the front-end and back-end stages. Clearly, reducing front-end lead times will result in even less safety stock needed in the die bank; while reducing the back-end lead times will enable Xilinx to satisfy more customer orders by using the die bank as the push-pull boundary.

Implementing postponement at Xilinx requires tremendous organizational sup-

port. The change from stocking primarily in finished goods to stocking primarily in die bank initially created some nervousness among the sales and logistics personnel who dealt with customers' orders. Although the company realized that it needed to use scientific inventory models to manage inventory levels effectively, gaining acceptance of the actual models took time and effort. We ran extensive computer simulations to demonstrate the effectiveness of the model and conducted intensive training and education programs with various functions within the company to create confidence in the model and acceptance of this new approach. The results showed that all these efforts were worthwhile, and postponement is now a key part of Xilinx's overall supply-chain strategy.

Acknowledgments

We thank Chris Wire, a key figure in driving the demand-and-supply-chain initiative at Xilinx, for his general input. We also thank John McCarthy and Dean Strausl for their support and vision in the projects and Donald St. Pierre for providing the engineering details of in-system programming for logic devices.

References

Brown, A. O.; Ettl, M.; Lin, G. Y.; and Petrakian, R. 1999, "Implementing a multi-echelon inventory system at a semiconductor company: Modeling and results," IBM Watson Labs technical report, Yorktown, New York.

Dapiran, P. 1992, "Benetton—Global logistics in action," *Asian Pacific International Journal of Business Logistics*, Vol. 5, No. 3, pp. 7–11.

Ehteshami, B. and Petrakian, R. 1998, "Speed yield prediction," Working paper, Xilinx, Inc., San Jose, California.

Eppen, G. D. and Schrage, L. 1981, "Centralized ordering policies in a multi-warehouse system with lead times and random demand," in *Multi-Level Production/Inventory Systems: Theory and Practice*, ed. L. B. Schwarz, North-Holland, Amsterdam and New York.

Ettl, M.; Feigin, G. E.; Lin, G. Y.; and Yao, D. D. forthcoming, "A supply network model with base-stock control and service requirements," *Operations Research*.

Feitzinger, E. and Lee, H. L. 1997, "Mass customization at Hewlett-Packard: The power of postponement," *Harvard Business Review*, Vol. 75, No. 1, pp. 116–121.

Lee, H. L. 1993, "Design for supply chain management: Concepts and examples," in *Perspectives in Operations Management: Essays in Honor of Elwood S. Buffa*, ed. R. Sarin, Kluwer Academic Publishers, Boston, Massachusetts, pp. 45–65.

Lee, H. L. 1996, "Effective inventory and service management through product and process redesign," *Operations Research*, Vol. 44, No. 1, pp. 151–159.

Lee, H. L.; Billington, C.; and Carter, B. 1993, "Hewlett-Packard gains control of inventory and service through design for localization," *Interfaces*, Vol. 23, No. 4, pp. 1–11.

Lee, H. L.; Feitzinger, E.; and Billington, C. 1997, "Getting ahead of your competition through design for mass customization," *Target*, Vol. 13, No. 2, pp. 8–17.

Lee, H. L.; Padmanabhan, V.; and Whang, S. 1997, "The bullwhip effect in supply chains," *Sloan Management Review*, Vol. 38, No. 3, pp. 93–102.

Lee, H. L. and Sasser, M. 1995, "Product universality and design for supply chain management," *Production Planning and Control: Special Issue on Supply Chain Management*, Vol. 6, No. 3, pp. 270–277.

Lineback, R. J. 1997, "The foundry/fabless model could become dominant," *Semiconductor Business News*, Vol. 1, No. 5, p. 1.

Nahmias, S. 1993, *Production and Operations Analysis*, second edition, Richard D. Irwin, Inc., Homewood, Illinois.

Rao, S. S. 1997, "Chips that change their spots," *Forbes*, Vol. 160, No. 1, pp. 294–296.

Reinertsen, D. G. 1983, "Whodunit? The search for new-product killers," *Electronic Business*, Vol. 9, No. 7, pp. 34–39.

Trimberger, S. M. 1994, *Field-Programmable Gate Array Technology*, Kluwer Academic Publishers, Boston, Massachusetts.
Ulrich, K. 1995, "The role of product architecture in the manufacturing firm," *Research Policy*, Vol. 24, No. 3, pp. 419–440.
Villasenor, J. and Mangione-Smith, W. H. 1997, "Configurable computing," *Scientific American*, Vol. 276, No. 6, pp. 66–71.
Whitney, D. E. 1995, "Nippondenso Co. Ltd.: A case study of strategic product design," Working paper, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Randy Ong, Vice-President, Operations, Xilinx Inc., 2180 Logic Drive, San Jose, California 95124–3400, writes: "This is to certify that the supply-chain efforts at Xilinx as described by the authors . . . have indeed been carried out. We have observed tremendous payoffs via such efforts, improving the efficiencies and effectiveness of our supply-chain and order-fulfillment processes. As a fabless semiconductor company, Xilinx has to rely on tight integration with our supply partners, distributors, and customers to remain competitive. Demand and supply-chain management is a cornerstone of our manufacuring strategy, and we are pleased to see such efforts creating great values for the company. I am also pleased to report that we are continuing our efforts to build supply-chain excellency so that Xilinx can become the leading edge supply-chain company in the semiconductor industry."

Stock Positioning and Performance Estimation in Serial Production-Transportation Systems

Guillermo Gallego • Paul Zipkin
Department of Industrial Engineering & Operations Research, Columbia University, New York, New York 10027
The Fuqua School of Business, Duke University, Durham, North Carolina 27708

This paper considers serial production-transportation systems. In recent years, researchers have developed a fairly simple functional equation that characterizes optimal system behavior, under the assumption of constant leadtimes. We show that the equation covers a variety of stochastic-leadtime systems as well. Still, many basic managerial issues remain obscure: When should stock be held at upstream stages? Which system attributes drive overall performance, and how? To address these questions, we develop and analyze several heuristic methods, inspired by observation of common practice and numerical experiments. One of these heuristics yields a bound on the optimal average cost. We also study a set of numerical examples, to gain insight into the nature of the optimal solution and to evaluate the heuristics. (*Inventory/Production*; *Multistage*; *Solutions and Heuristics*)

1. Introduction

Consider a serial production-transportation system.

• There are several *stages*, or stocking points, arranged in series. The first stage receives supplies from an external source. Demand occurs only at the last stage. Demands that cannot be filled immediately are backlogged.

• There is *one product*, or more precisely, one per stage.

• To move units to a stage from its predecessor, the goods must pass through a *supply system*, representing production or transportation activities. The cost for a shipment to each stage is linear in the shipment quantity.

• There is an inventory-holding cost at each stage and a backorder-penalty cost at the last stage. The horizon is infinite, all data are stationary, and the objective is to minimize total average cost. Information and control are centralized.

We focus on a basic system, where time is continuous, demand is a Poisson process, and each stage's supply system generates a constant leadtime. However, virtually all the results remain valid for a discrete-time system with i.i.d. demands, for compound-Poisson demand in continuous time, and for more complex supply systems with stochastic leadtimes. Also, since an assembly system can be reduced to an equivalent series system (Rosling 1989), the results apply there too.

Clark and Scarf (1960) initiated the analysis of this system, assuming discrete time with a finite horizon and nonstationary data. They showed that the optimal policy has a simple, structured form (an echelon base-stock policy) and developed a tractable scheme to compute it. Federgruen and Zipkin (1984) adapted the results to the stationary, infinite-horizon setting and pointed out that the algorithm becomes simpler there. Rosling (1989) and Langenhoff and Zijm (1990) provided streamlined statements of the results. Chen and Zheng (1994) further streamlined the results and extended them to continuous time. The outcome is a fairly simple functional equation, Equation (5) in §2,

1523-4614/99/0101/0077$05.00

that characterizes the optimal policy. See Federgruen (1993) for a review of this literature.

There is another, very different stream of research on multistage systems, one that emphasizes policy evaluation. It assumes a particular policy type, usually a base-stock policy, and estimates key performance measures, especially average inventories and backorders. Those measures are used to construct an optimization model, whose solution yields the best such policy. The supply systems can be fairly complex, indeed some generate stochastic leadtimes. In most cases the performance estimates are approximations. The system structure too can be more complex; in addition to series systems, the approach applies to distribution and assembly systems. This literature begins with the METRIC model of Sherbrooke (1968). Recent contributions include Graves (1985), Sherbrooke (1986), and Svoronos and Zipkin (1991). Reviews can be found in Nahmias (1981) and Axsäter (1993). We explain in §3 that, despite these differences, the solution to Equation (5) also yields the best base-stock policy for such a system, up to the approximation.

Still, many basic managerial issues concerning such systems remain obscure: When should stock be held at upstream stages? Which system attributes drive overall performance, and how? To address these questions, we study several heuristic methods (§4), inspired by observation of common practice and numerical experiments, including one that yields a bound on the optimal average cost. Sensitivity analysis of this result reveals interesting features of system behavior. We also study a set of numerical examples (§5), both to gain insight into the nature of the optimal solution and to evaluate the heuristics.

Section 6 presents our conclusions. A key finding is that system performance is fairly insensitive to stock positioning, provided the overall system inventory is near optimal. In particular, certain heuristic policies which concentrate stock at a few locations perform quite well.

We also discuss a broader system *design* problem, as in Gross et al. (1981). Here, the stages are *potential* storage locations, but none have yet been built. The design problem is to select a subset among them and then to determine a control policy for the resulting network. There is a cost to open each facility, and such costs appear in either the objective function or a constraint. There may be several products sharing the same facilities. This is a hard problem, but several of our heuristics apply to it as well.

2. Base-Stock Policy Evaluation and Optimization

This section reviews the basic facts concerning policy evaluation and optimization.

2.1. Stages

For now, assume Poisson demand and constant leadtimes. Denote

J = number of stages
j = stage index, $j = 1, \ldots, J$
λ = demand rate
L_j = supply leadtime to stage j
L = total system leadtime = $\Sigma_j L_j$.

The numbering of stages follows the flow of goods; stage 1 is the first, and stage J is the last, where demand occurs. The external source, which supplies stage 1, has ample stock; it responds immediately to orders.

2.2. Base-Stock Policies

In a single-stage system, a *base-stock policy* aims to keep the inventory position constant. The target inventory position is a policy variable, the *base-stock level,* denoted s. When the inventory position falls below s, the policy orders enough to raise the inventory position to s; otherwise, it does not order. Thus, once the inventory position hits s, orders precisely equal demands.

In a multi-stage system, there are two classes of base-stock policy, local and echelon. Although they seem quite different, the two classes are equivalent (Axsäter and Rosling 1993).

A *local base-stock policy* is a decentralized control scheme, where each stage monitors its own local inventory position and places orders with its predecessor. Each stage j follows a standard, single-stage base-stock policy with parameter

s_j' = local base-stock level for stage j,

a nonnegative integer. The overall policy is characterized by the vector $\mathbf{s}' = (s_j')_{j=1}^J$.

An *echelon base-stock policy* is a centralized control

scheme. It monitors each stage's echelon inventory (the stage's own stock and everything downstream), and determines external orders and inter-stage shipments according to a base-stock policy. The policy parameters are

s_j = echelon base-stock level for stage j,

also a nonnegative integer. Let $\mathbf{s} = (s_j)_{j=1}^J$. As shown by Chen and Zheng, given stationary parameters, such a policy is optimal in either a periodic-review or a continuous-review setting.

Given a local base-stock policy $\mathbf{s}'$, an equivalent echelon base-stock policy has parameters $s_j = \Sigma_{i \geq j} s_i'$. Conversely, starting with an echelon base-stock policy $\mathbf{s}$, one can construct an equivalent local policy, setting $s_j^- = \min_{i \leq j} \{s_i\}$ and $s_j' = s_j^- - s_{j+1}^-$, where $s_{J+1}^- = 0$. (Also, the echelon base-stock policy $\mathbf{s}^- = (s_j^-)_{j=1}^J$ is equivalent to $\mathbf{s}$.)

2.3. Cost

Denote

$E[\cdot]$ = expectation
$[x]^+$ = $\max\{0, x\}$
$D(t)$ = cumulative demand in the interval $(0, t]$.
$V[\cdot]$ = variance
$[x]^-$ = $\max\{0, -x\}$

The following are state random variables in equilibrium:

I_j' = local on-hand inventory at stage j
B_j' = local backorders at stage j
B = customer backorders = B_J'
IT_j = inventory in transit to stage j (units in j's supply system)
I_j = echelon inventory at stage j = $I_j' + \Sigma_{i>j} (IT_i + I_i')$
IN_j = echelon net inventory at stage j = $I_j - B$.

Also, let

D_j =
leadtime demand for stage j, a generic random variable having the distribution of $D(L_j)$. The D_j are independent.

The cost factors are

b = backorder penalty-cost rate
h_j' = local inventory holding-cost rate at stage j
h_j = echelon inventory holding-cost rate at stage j = $h_j' - h_{j-1}'$,

where $h_0' = 0$.

The usual accounting scheme for in-transit inventories charges h_j' on IT_{j+1} as well as I_j'. We exclude such costs, in order to facilitate comparison among policies and systems. Thus, the total average cost, expressed in local terms, is

$$C(\mathbf{s}') = E[\Sigma_{j=1}^J h_j' I_j' + bB]. \tag{1}$$

The equivalent expression in echelon terms is

$$C(\mathbf{s}) = E[\Sigma_{j=1}^J h_j IN_j + (b + h_J')B] - E[\Sigma_{j=1}^J h_j' D_{j+1}]. \tag{2}$$

(Here, $D_{J+1} = 0$. The second term is necessary, because the first includes the usual in-transit holding cost, and $E[IT_j] = E[D_j]$.)

2.4. Local Policy Evaluation

For any policy $\mathbf{s}'$, the equilibrium local backorder variables satisfy the following recursion:

$$\begin{aligned} B_0' &= 0 \\ B_j' &= [B_{j-1}' + D_j - s_j']^+. \end{aligned} \tag{3}$$

And,

$$I_j' = s_j' - (B_{j-1}' + D_j) + B_j'. \tag{4}$$

(See, e.g., Graves 1985.) From these, we can compute $E[B]$ and $E[I_j']$ and thus the average cost [Equation (1)].

2.5. Echelon Policy Optimization

We now present a method to determine an optimal echelon base-stock policy, denoted $\mathbf{s}^*$. This is the Clark-Scarf algorithm, essentially as stated by Chen and Zheng:

Set $\underline{C}_{J+1}(x) = (b + h_J')[x]^-$. For $j = J, J-1, \ldots, 1$, given $\underline{C}_{j+1}$, compute

$$\begin{aligned} \hat{C}_j(x) &= h_j x + \underline{C}_{j+1}(x) \\ C_j(y) &= E[\hat{C}_j(y - D_j)] \\ s_j^* &= \text{argmin } \{C_j(y)\} \\ \underline{C}_j(x) &= C_j(\min\{s_j^*, x\}). \end{aligned} \tag{5}$$

At termination, set $C^* = C_1(s_1^*) - E[\Sigma_{j=1}^J h_j' D_{j+1}]$. This is the optimal cost.

A similar calculation can be used to evaluate any policy **s**. Just omit the optimization step, and use s_j in place of s_j^* in the last step. One can show that this method is equivalent to Equations (2) through (4). Conversely, one can show directly that Equation (5) optimizes over policies evaluated by Equations (2) through (4). This point underlies the extensions of §3. (To our knowledge, these observations are new here.)

Recursion (5) deserves to be called *the fundamental equation of supply-chain theory*. It captures the basic dynamics and economics of serial systems. It omits much, but any more comprehensive theory must build on it. We know little about its solution, however. The remainder of the paper begins to investigate it.

2.6. Decreasing Holding Costs

Examination of Equation (5) reveals that, for $j < J$, if $h_{j+1} \leq 0$, then $s_{j+1}^* = \infty$, which implies $s_j'^* = 0$. In this case, we can eliminate stage j, replacing L_{j+1} by $L_{j+1} + L_j$ and h_{j+1} by $h_{j+1} + h_j$. (Rosling observes this.) Continue to eliminate stages in this way, until all the remaining $h_j > 0$. Thus, *a stage holds stock **only** when it is cheaper to hold it there than anywhere downstream*. This makes sense intuitively; downstream inventory provides more direct, effective protection against customer backorders than upstream inventory. The only possible advantage of upstream inventory is lower inventory-holding cost.

3. Other Demand and Supply Processes

The same methods can be used to evaluate and optimize, exactly or approximately, under a variety of other model assumptions.

3.1. Compound-Poisson Demand

Suppose that demand is a compound-Poisson process, and each increment of demand can be filled separately. *All* the results above remain valid. Here, each D_j has a compound-Poisson distribution, but that is the only difference.

3.2. Exogenous, Sequential Supply Systems

Consider a system like that of Svoronos and Zipkin (1991), specialized to a series structure: Each stage's supply system is stochastic. Stage j's system generates a *virtual leadtime* $L_j(t)$; a shipment to j initiated at time t arrives at $t + L_j(t)$. The system processes orders *sequentially*, so shipments arrive in the same sequence as the corresponding orders; that is, $t + L_j(t)$ is nondecreasing in t. Each supply system is *exogenous*, i.e., its internal state and $L_j(t)$ are stochastic processes, but they are unaffected by shipments. Each system is *ergodic*, i.e., $L_j(t)$ approaches a steady-state random variable L_j, regardless of initial conditions. Finally, these systems, and hence the $L_j(t)$ and L_j, are independent over j.

Svoronos and Zipkin show that Equations (3) through (4) evaluates a base-stock policy. Here, D_j has the distribution of $D(L_j)$, the demand over the (stochastic) virtual leadtime L_j, so $E[D_j] = \lambda E[L_j]$ and $V[D_j] = \lambda E[L_j] + \lambda^2 V[L_j]$. These D_j are again independent. Consequently, as explained in §2.5, Equation (5) finds the best base-stock policy.

3.3. Independent Leadtimes

Return to the Poisson-demand case. Suppose that each stage's leadtimes are i.i.d. random variables; in effect, each supply system consists of multiple identical processors in parallel. Let L_j be the generic leadtime random variable for stage j. In this context, Equations (3) through (4) remain valid with IT_j in place of D_j.

It is difficult to characterize the IT_j, in general. There is one case where it is easy, namely, when $\mathbf{s} = \mathbf{s}' = \mathbf{0}$. There, the system is equivalent to a tandem network of queues with Poisson input, where each node j has an infinite number of servers with service times L_j. So, IT_j has the Poisson distribution with mean $\lambda E[L_j]$, and the IT_j are independent. (See, e.g., Kelly 1979.). For general $\mathbf{s}' \geq \mathbf{0}$ we can use this same distribution to *approximate* the IT_j. This is, in fact, the key approximation underlying the METRIC procedure (see Sherbrooke 1968, 1986 and Graves 1985), specialized to series systems. It is quite accurate.

With this approximation, using D_j to stand for the approximate IT_j, Equations (3) through (4) evaluate a local policy. Therefore, Equation (5) computes the best base-stock policy, up to the approximation.

3.4. Limited-Capacity Supply Systems

Now, suppose each supply system consists of a single processor and its queue. The processing times at stage

j are i.i.d., distributed exponentially with rate μ_j. Assume $\lambda < \mu \equiv \min_j \{\mu_j\}$. Recursion (3), with IT_j in place of D_j, applies here too. Again, it is difficult to characterize the IT_j in general, but easy in the case $\mathbf{s}' = \mathbf{0}$. Here, IT_j has the geometric distribution with parameter $\rho_j = \lambda/\mu_j$, and the IT_j are independent (Kelly). This works well as an approximation for the general case, as shown by Buzacott et al. (1992), Lee and Zipkin (1992), and Zipkin (1995). So, Equation (5) again finds the (approximately) best base-stock policy.

4. Bounds and Heuristics

4.1. The Restriction-Decomposition Approximation

This section presents a fairly simple way to determine a useful heuristic policy and an upper bound on the optimal cost. The approach involves *restriction* of the policy space and *decomposition* of the resulting model into single-stage submodels. Accordingly, we call it the restriction-decomposition or RD approximation. This approach, or something like it, is widely used in practice. It is striking that this simple idea actually bounds the original system.

Let $\mathbf{J}_+$ be any subset of stages that includes J. We construct an approximation for any choice of $\mathbf{J}_+$ and then select the best $\mathbf{J}_+$. Index these stages in order by $j(m)$, $m = 1, \ldots, M$. So, $j(M) = J$. Also, denote $j(0) = 0$. Let

$$D_{(i,j]} = D_{i-1} + D_{i+2} + \ldots + D_j, \quad 0 \le i < j \le J,$$
$$D^m = D_{(j(m-1),j(m)]}, \quad m = 1, \ldots, M.$$

First, *restrict* $s'_j = 0$, $j \notin \mathbf{J}_+$, so that only stages in $\mathbf{J}_+$ are allowed to hold stock. Using Equation (3), one can readily show that

$$B'_{j(m)} = [B'_{j(m-1)} + D^m - s'_{j(m)}]^+, \; m = 1, \ldots, M.$$

The next steps effectively *decompose* the system at stages $\mathbf{J}_+$. It is easy to show that

$$B'_{j(m)} \le B'_{j(m-1)} + [D^m - s'_{j(m)}]^+, \quad m = 1, \ldots, M$$
$$B = B'_{j(M)} \le \sum_{m=1}^{M} [D^m - s'_{j(m)}]^+$$
$$I'_{j(m)} \le [s'_{j(m)} - D^m]^+, \quad m = 1, , \ldots, M.$$

Consequently,

$$C(s') \le \sum_{m=1}^{M} E[h'_{j(m)} [s'_{j(m)} - D^m]^+ + b[D^m - s'_{j(m)}]^+].$$

Equivalently, let

$$\hat{C}'_j(x) = h'_j[x]^+ + b[x]^-$$
$$C_{(i,j]}(y) = E[\hat{C}'_j(y - D_{(i,j]})], \quad i < j.$$

Then,

$$C(\mathbf{s}') \le \sum_{m=1}^{M} C_{(j(m-1),j(m)]}(s'_{j(m)}).$$

Each term in this sum is the cost of a single-stage system. It charges the full penalty cost b to local backorders at each stage $j(m)$, while ignoring the effects of those backorders on downstream stages. In this sense it splits the system into separate subsystems.

Now, let $s_{(i,j]}$ minimize $C_{(i,j]}(y)$, and denote the minimal cost by $C^*_{(i,j]}$. Then,

$$C^* \le \sum_{m=1}^{M} C^*_{(j(m-1),j(m)]}.$$

This relation holds for any $\mathbf{J}_+$. To find the best such bound over all possible $\mathbf{J}_+$, consider the following network: The nodes are $\{0,1, \ldots, J\}$, the arcs are (i,j), $i < j$, and the arc lengths are $C^*_{(i,j]}$. The best bound, then, is the length of the shortest path from 0 to J. This problem has precisely the same structure as the dynamic economic lot-size problem of Wagner and Whitin (1958), and can be solved using the same algorithm.

From the best $\mathbf{J}_+$ one can construct a plausible heuristic policy: Set $s'_j = 0$, $j \notin \mathbf{J}_+$, and for $j = j(m) \in \mathbf{J}_+$, set $s'_j = s_{(j(m-1),j(m)]}$. The actual cost of this policy is no more than the upper bound. (Alternatively, use Equation (5) to find the optimal policy for the system restricted to $\mathbf{J}_+$. We have not tested this more refined approach.)

The RD approximation extends directly to the design problem: If there is a fixed cost k_j to build stage j, just add k_j to each $C^*_{(i,j]}$. Also, if several products share the network, compute the $C^*_{(i,j]}$ for each product, and then sum them over the products. The algorithm above then provides a heuristic solution and an upper bound.

We remark that the complexity of the RD heuristic appears to be $O(J^2)$, compared to $O(J)$ for the optimizing algorithm Equation (5). Indeed, we have observed that, for very large J, the heuristic can take longer than Equation (5). For smaller, plausibly-sized systems, however, the heuristic is usually much faster. And, it is a tractable method for the design problem.

Here is a further useful approximation: Scarf (1958) and Gallego and Moon (1993) show that

$$C^*_{(i,j]} \le (bh'_j)^{1/2}\sigma_{(i,j]} \equiv C^+_{(i,j]},$$

where $\sigma_{(i,j]}$ is the standard deviation of $D_{(i,j]}$. Using the $C^+_{(i,j]}$ in the calculations above yields a *distribution-free* bound, one that depends only on two moments of leadtimes and demands, not their actual distributions. Call this the *maximal RD approximation*. The same analysis yields a heuristic solution of the form $s^+_{(i,j]} = E[D_{(i,j]}] + z'_j\sigma_{(i,j]}$, where z'_j is a safety factor depending on b and h'_j, whose cost is no more than the upper bound. (This approach is much faster than the original RD heuristic, since $C^+_{(i,j]}$ is easier to compute than $C^*_{(i,j]}$.)

This simple form facilitates sensitivity analysis: Observe that, in the Poisson-demand, constant-leadtime case, each $C^+_{(i,j]}$ depends on λ through a factor $\sqrt{\lambda}$. Thus, the shortest path is independent of λ, and the cost bound is proportional to $\sqrt{\lambda}$. That is, *the heuristic's choice of stocking points is independent of the demand volume,* and *the true optimal cost is bounded above by a function proportional to* $\sqrt{\lambda}$. A similar analysis of $s^+_{(i,j]}$ suggests that the overall safety stock is proportional to $\sqrt{\lambda}$. The same is true for stochastic, independent leadtimes (§3.3). For exogenous, sequential leadtimes (§3.2), however, $\sigma_{(i,j]} = (\lambda E[L_{(i,j]}] + \lambda^2 V[L_{(i,j]}])^{1/2}$, so the optimal cost is bounded by a *linear* function of λ, as is the safety stock.

Likewise, the shortest path is independent of b, and the cost bound is proportional to $\sqrt{b}$.

The leadtime L_k affects $\sigma_{(i,j]}$ for all $(i,j]$ with $i < k \le j$. It has the biggest impact on the $\sigma_{(i,j]}$ for short intervals $(i,j]$ around k. Thus, for small k, L_k has a major impact only on terms $C^+_{(i,j]}$ with small j and hence low h'_j. Conversely, L_k for large k affects terms with large h'_j. This suggests that *downstream leadtimes have a greater impact on system performance than upstream ones.*

The familiar normal approximation yields an approximation to $C^*_{(i,j]}$ of the same form as $C^+_{(i,j]}$, namely, a factor depending on the cost parameters, times $\sigma_{(i,j]}$. It also yields a solution of the same form as $s^+_{(i,j]}$. Call this the *normal RD approximation*. So, the observations above about λ and the L_k remain valid. The cost factors, however, grow more slowly in b than $\sqrt{b}$.

Some additional bounds for two-stage systems can be found in Gallego and Zipkin (1994).

4.2. The Zero-Safety-Stock Heuristic

This approach (the ZS heuristic, for short) sets $s'_j = E[D_j]$, $j < J$, and then optimizes over s'_J. More precisely, to cover the case of non-integral $E[D_j]$, the heuristic sets $s'_j = \lceil \Sigma_{i \le j} E[D_i] \rceil - s'_{j-1}$, $j < J$. Then, using Equation (3), it computes the distribution of B'_{J-1}. Finally, it chooses s'_J to minimize the stage-J holding and penalty costs, a single-stage problem. (This method was inspired by some preliminary numerical results, in which the optimal s'_j was near $E[D_j]$, $j < J$.) Evidently, this is an $O(J)$ calculation, and it is very fast in practice.

4.3. The Two-Stage Heuristic

This approach (the TS heuristic) restricts inventory to two stages, the last one J and some single $j < J$. Given $j < J$, it finds the optimal policy for the resulting two-stage system. It then selects the best such policy over $j < J$. (This method too was based on empirical observations, namely, that restricting the number of locations sometimes has little cost impact.)

This technique requires solving $J - 1$ two-stage problems, nearly as much work as the full optimization algorithm Equation (5). The purpose of the heuristic is not speed. Rather, it is a tool to investigate stock-positioning issues: Where is stock most useful? And, how costly is the restriction to two stages? This approach also extends easily to the design problem; in that context it is a plausible heuristic for systems with large fixed facility costs.

5. Numerical Results

This section presents some numerical examples, to provide insight into the behavior of the optimal policy and the performance of the heuristics.

5.1. Specification

5.1.1. System Structure and Parameters. We assume Poisson demand and constant leadtimes. Without loss of generality, we fix the time scale so that the total leadtime is $L = 1$, and the monetary unit so that the last stage's holding cost is $h'_J = 1$. The stages are spaced symmetrically, so each stage j's leadtime is $L_j = 1/J$. We consider four numbers of stages, $J = 1, 4, 16, 64$; two demand rates, $\lambda = 16, 64$; and two penalty costs, $b = 9, 39$ (corresponding to fill rates of 90%, 97.5%).

5.1.2. Holding Cost Forms. We consider several forms of holding costs h_j', depicted in Figure 1. The simplest form has *constant* holding costs, where all $h_j' = 1$. Here, there is no cost added from source to customer. This is a rather unrealistic scenario, but it is a useful starting point to help understand other forms. The *linear* holding-cost form has $h_j' = j/J$, or $h_j = 1/J$. Here, cost is incurred at a constant rate as the product moves from source to customer. This is quite realistic. *Affine* holding costs, where $h_j' = \alpha + (1 - \alpha)j/J$ for some $\alpha \in (0,1)$, are even more realistic. Here, the material at the source has some positive cost, and the system then adds cost at a constant rate. This form is a combination of the constant and linear forms. In Figure 1 and the calculations below, $\alpha = 0.75$.

The last two forms represent deviations from linearity. The *kink* form is piecewise linear with two pieces. The system incurs cost at a constant rate for a while, but at some point shifts to a different rate, which remains constant from then on. Here, the kink occurs halfway through the process, at stage $J/2$. So, for some $\alpha \in (-1,1)$, $h_j = (1 - \alpha)/J$, $j \leq J/2$, and $h_j = (1 + \alpha)/J$, $j > J/2$. Again, we set $\alpha = 0.75$. Finally, in the *jump* form, cost is incurred at a constant rate, except for one stage with a large cost. Here, the jump occurs just after stage $J/2$. So, $h_j = \alpha + (1 - \alpha)/J$, $j = J/2 + 1$, and $h_j = (1 - \alpha)/J$ otherwise, for some $\alpha \in (0,1)$. We can view this as linear cost before $J/2$ and affine cost after. Here again, $\alpha = 0.75$.

5.2. Optimal Policy

5.2.1. Constant Holding Costs. The optimal policy in this case is simple: For $j < J$, $s_j'^* = 0$; only the last stage carries inventory. Stage J, in effect, becomes a single-stage system with leadtime L. The optimal policy is the same for all J. This is also the optimal policy for $J = 1$ under any other holding-cost form.

5.2.2. Linear Holding Costs. Figure 2 shows the optimal policy $\mathbf{s}^*$ for $J = 64$ and two values each of λ and b. Several observations are worth noting: The curves are *smooth* and *nearly linear;* the optimal policy does *not* lump inventory in a few stages, but rather spreads it quite evenly. The departures from linearity are interesting too: The curves are *concave.* Thus, the policy focuses safety stock at stages nearest the customer.

5.2.3. Affine Holding Costs. Figure 3 shows the optimal policy. For $j > 1$, the curves follow the same pattern as in Figure 2. (Indeed, the curves for $b = 9$ here are identical to those for linear costs and $b = 39$, because these two cases have identical ratios $h_j/(b + h_j')$, $j > 1$.) However, the curves break down sharply at $j = 1$ (because h_1 is large). Therefore, the equivalent policy $\mathbf{s}^-$ is flat for small j, and so the policy holds *no* inventory at early stages. This solution is intermediate between those for constant and linear costs. As α increases and the costs move upwards, stocks shift toward the customer. The total system stock decreases

Figure 1 Holding Cost Forms

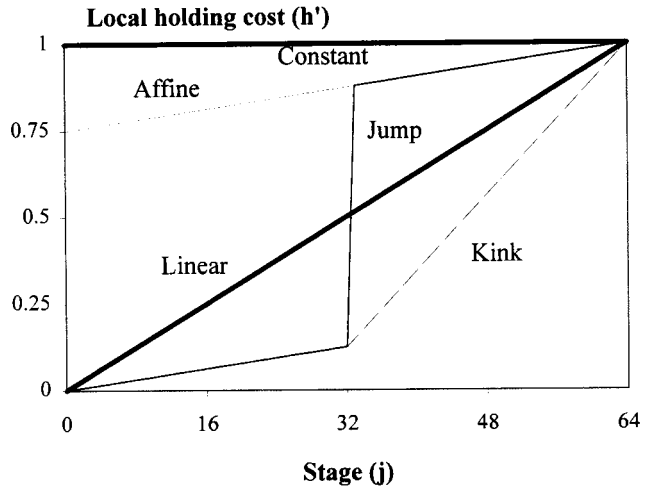
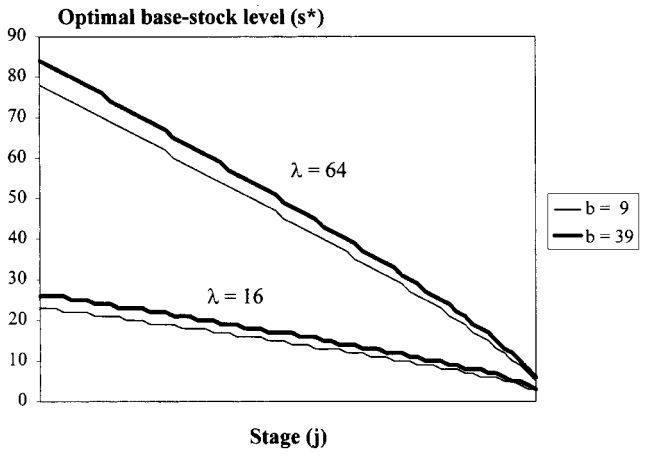


Figure 2 Optimal Policy: Linear Holding Costs



slightly. But, perhaps surprisingly, stocks near the customer actually increase.

5.2.4. Kink Holding Costs. Figure 4 displays **s***. Downstream from the kink (before Algorithm (5) encounters it), the curves exhibit the same pattern as in the linear case. Upstream from the kink, the policy again follows the linear pattern, almost as if the kink were the last stage. The net result is substantial stock at and just before the kink, where holding costs are low relative to later stages.

5.2.5. Jump Holding Costs. Figure 5 displays **s***. From the jump on, the policy behaves much as in the affine case: smooth, concave decrease beyond the jump, but a sharp break downwards at the jump. Upstream from the jump, the policy again follows the pattern of the linear case. Thus, there is substantial stock just before the jump and none just after it.

5.3. Sensitivity Analysis

5.3.1. Number of Stages. Figure 6 compares the **s*** for different *J*s, each with linear holding costs, $\lambda = 64$, and $b = 39$. The curves follow the same patterns as before, as closely as the restricted numbers of stages allow. Indeed, the actual echelon stock at a stocking point is nearly identical to the $J = 64$ case. Closer inspection shows that the total system stock is slightly

Figure 3 Optimal Policy: Affine Holding Costs

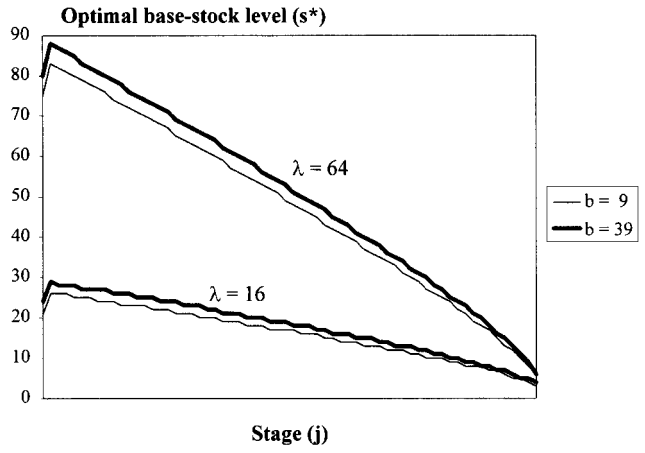


Figure 4 Optimal Policy: Kink Holding Costs

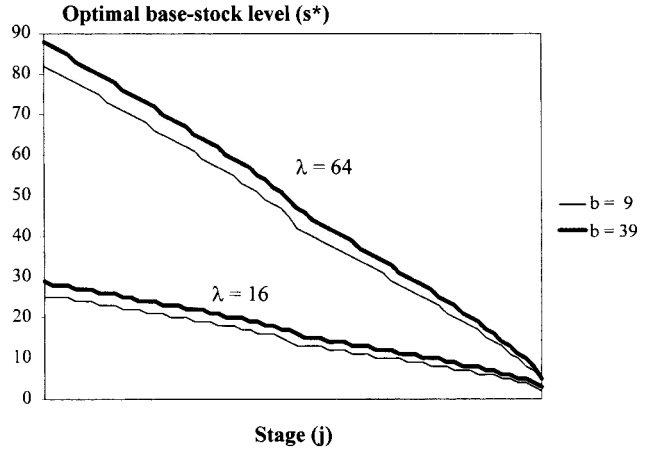


Figure 5 Optimal Policy: Jump Holding Costs

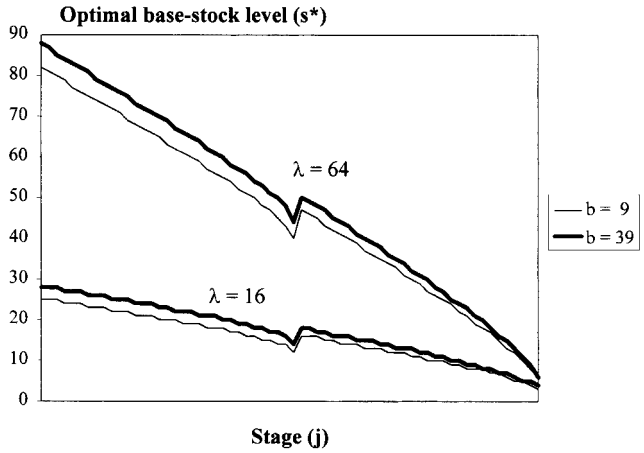
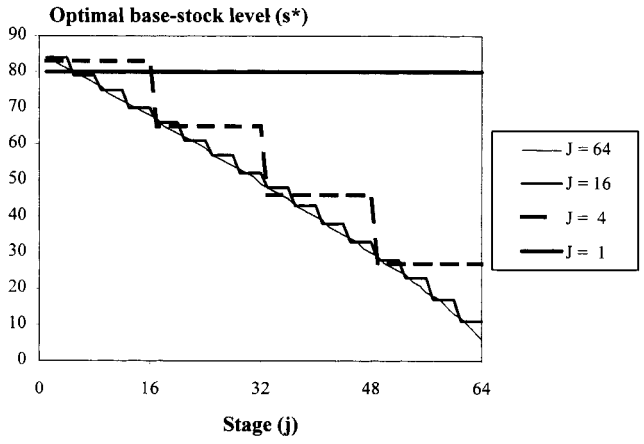


Figure 6 Optimal Policy: Effects of *J*



higher for larger J. Likewise, the optimal cost decreases in J, but quite slowly, as shown in Figure 7.

Similar results hold for affine holding costs. Indeed, the optimal cost is even less sensitive to J. For kink holding costs (Figure 8), the optimal cost is significantly lower at $J = 4$ than at $J = 1$, due to the availability of the low-cost stocking point at the kink. Larger Js yield relatively minor improvements. The jump form displays a similar pattern. Thus, for these two forms, it is important to position stock at the kink (or jump). Otherwise, the cost is quite insensitive to J.

These results suggest that *the system cost is relatively insensitive to stock positioning*, provided the overall stock level is about right, and obvious low-cost stocking points are exploited. We shall see further evidence for this below.

Figure 7 **Optimal Cost: Linear Holding Costs**

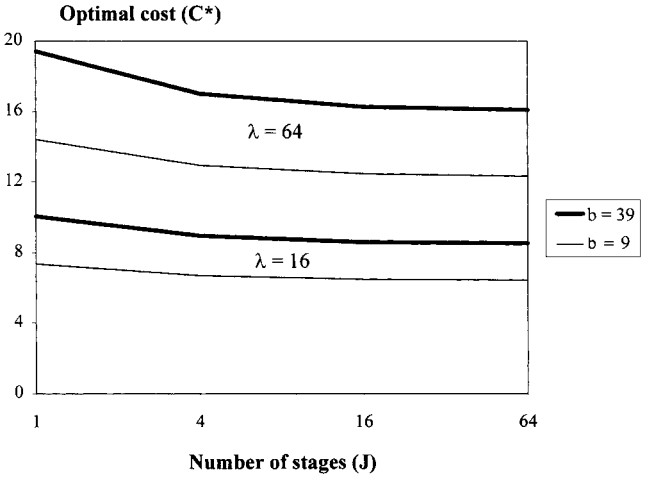
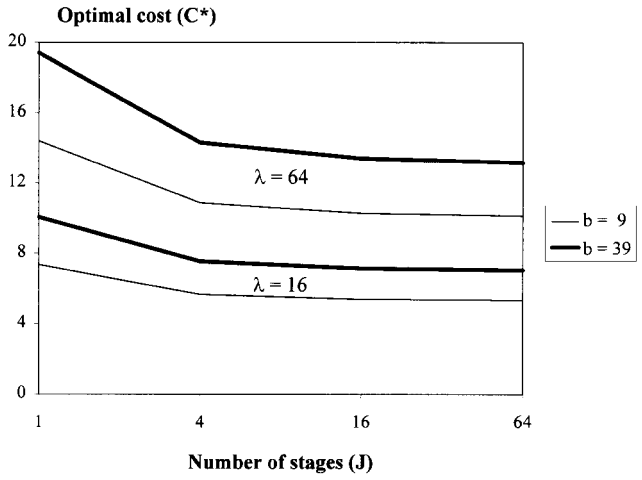


Figure 8 **Optimal Cost: Kink Holding Costs**



5.3.2. Demand Rate. In Figures 7 and 8 the optimal cost for $\lambda = 64$ is about twice that for $\lambda = 16$ in every case. This is consistent with the notion that the optimal cost is nearly proportional to $\sqrt{\lambda}$, as suggested in §4.1. We have also plotted, but omit here, the cumulative safety stocks $\Sigma_{i \le j}\, s_j'^* - \lambda j/J$. The curves for $\lambda = 64$ are about twice those for $\lambda = 16$. So, the safety stocks too are nearly proportional to $\sqrt{\lambda}$.

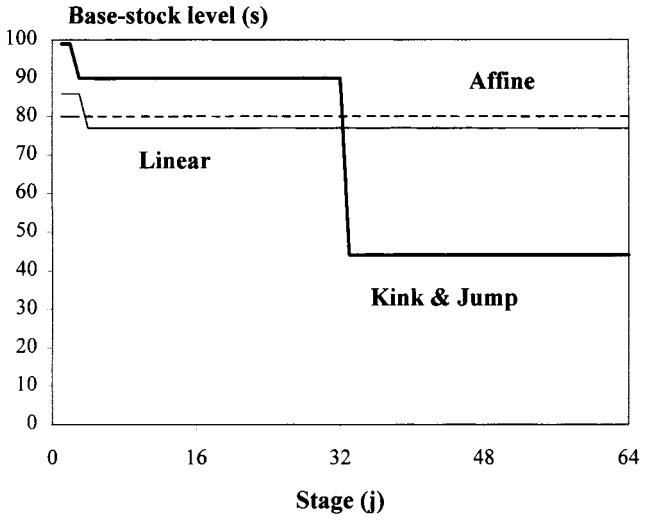
5.3.3. Backorder Cost. The figures above indicate that the base-stock levels and optimal cost are increasing in b. The policy, however, is not very sensitive to b. The cost, though rather more sensitive, grows considerably slower than $\sqrt{b}$, as suggested by the normal RD approximation.

5.3.4. Leadtimes. Figures 7 and 8 provide some evidence for the notion that downstream leadtimes are more important than upstream ones. Starting with linear holding costs, contract the downstream leadtimes and expand the upstream ones, keeping L and the h_j' fixed. The result looks much like the kink form with $\alpha \in (0,1)$. And, the kink form has lower optimal cost for $J > 1$.

5.4. Performance of Bounds and Heuristics

5.4.1. The RD Approximation. Figure 9 shows the policies chosen by the RD heuristic in one case ($J = 64$, $\lambda = 64$, $b = 39$) for all four holding-cost forms. (The same policy is chosen for the kink and jump forms.) These policies are quite different from the corresponding optimal ones; they concentrate stock in just a few stages. For the linear form, the policy places a small inventory near the source (9 units at stage 3) and a large one (77) at the last stage. For the affine form, the policy is even more extreme, placing all its stock (80) at the end. For the kink and jump forms, the policy places substantial inventory (46) at stage 32, just before the cost increase, a little near the source (9 at stage 2), and the rest (44) at the end. Also, the total system stocks are slightly larger than optimal. The results for other J, λ, and b are similar.

Figure 9 RD Heuristic's Policies



Even so, the RD heuristic and the cost bound perform fairly well. Table 1 shows the percentage errors for all three heuristics. For example, for the linear form, the RD policy's cost exceeds the optimal by 10%–20%. (The errors tend to increase slowly in J, λ, and b.) These errors are far smaller than the cost differences between systems. The cost bound is usually just a bit more than the actual heuristic policy's cost.

Thus, the RD approximation provides crude but robust estimates of system performance. It is certainly accurate enough for rough-cut design studies. This fact, coupled with the gross differences between the RD and optimal policies, is further evidence of the insensitivity of performance to stock positioning.

5.4.2. The ZS Heuristic. The ZS heuristic, by definition, sets s_j' to the average leadtime demand up to the last stage. It sets s_j' larger than the optimal policy does, to compensate for the lower stocks at earlier stages. It generates the same policy for all four cost forms. It works very well for linear holding costs, rather less well for affine costs, and not so well for the kink and jump forms.

5.4.3. The TS Heuristic. For $J = 64$, for linear costs, the TS heuristic places stock just past the middle of the system, in addition to stage J. Specifically, for $\lambda = 64$ and $b = 39$, it chooses $j = 36$. For affine costs, the heuristic places stock further downstream, at $j = 48$. (The locations are just slightly different for the other λ and b.) For the kink and jump forms, it selects $j = 32$, just before the cost increase, in all cases. The results are similar for smaller J. As Table 1 indicates, this method performs quite well; it is the best among the three heuristics. This is yet more evidence of the insensitivity of performance to stock positioning.

Table 1 Heuristics' Percentage Costs over Optimal

Form	RD	ZS	TS
Linear	10–20%	2–8%	4–11%
Affine	1–3%	3–14%	0–2%
Kink	9–22%	11–25%	5–17%
Jump	5–7%	11–15%	1–3%

6. Conclusions

We have seen that the optimal policy depends on the growth of holding costs between source and customer. For constant costs, the policy puts all stock at the last stage. For linear costs, the policy distributes stock quite evenly, though favoring downstream sites. In other cases the policy can be understood as a systematic combination and variation of these patterns.

On the other hand, although it is important to optimize the system-wide inventory and to exploit especially low holding costs, system performance is otherwise fairly insensitive to stock positioning. One can deviate substantially from the optimal policy for a rather small cost penalty, as in the restriction to smaller J and the heuristics. In particular, the RD and TS heuristics work fairly well; they capture the gross behavior of the optimal policy, though differing substantially in detail. Consequently, they are reasonable heuristics for the design problem.

The sensitivity of the system to its parameters is similar in many ways to the familiar single-stage system. For instance, with constant leadtimes, the optimal cost and safety stocks increase as the square root of the demand rate. Multistage systems have certain additional characteristics, however. For example, downstream leadtimes have greater impacts on performance than upstream ones.

We have presented these results to several groups of

managers in different industries. Their reactions are worth reporting. They showed considerable interest in the *forms* of the figures as diagnostic devices. For example, they wanted to plot their own holding costs in the style of Figure 1, to see where cost accrues quickly and where slowly. (This type of diagram is called a *time-cost profile* by Fooks (1993) and Schraner (1994). Observe that the in-transit holding cost is essentially the area under each curve.) Likewise, a plot of actual inventories in the manner of Figures 2 through 5 is a convenient way to see just where stock is concentrated.

Many managers at first resisted the notion that stock should be concentrated close to customers. After all, the downstream sites are the most expensive ones. But, following discussion of the sites' different degrees of stockout protection, as in §2.6, most agreed that the optimal policy was at least plausible. Several noted that their own firms' stock-positioning policies were quite different, and planned to investigate the alternative suggested by the model. Similarly, many had embraced the idea of reducing total leadtime, and were dubious that downstream leadtimes could be more important than upstream ones. Once the logic was explained, however, they accepted it.

Finally, none of the managers found it hard to believe that the heuristics perform well. Indeed, they preferred solutions that concentrate stock in only a few locations, and they appreciated the simplicity of the heuristics. Their experience suggested that *all* real systems incur some fixed costs, as in the design problem.

Several questions remain: Are there better heuristics? Do the results extend to more complex systems, such as distribution systems and systems with fixed order costs? These are subjects of ongoing research.[1]

References

Arrow, K., S. Karlin, H. Scarf, eds. 1958. *Studies in the Mathematical Theory of Inventory and Production*. Stanford University, Stanford, CA.

Axsäter, S. 1993. Continuous review policies for multi-level inventory systems with stochastic demand. S. Graves, A. Rinnooy Kan, P. Zipkin, eds. *Logistics of Production and Inventory*. Elsevier (North-Holland), Amsterdam, The Netherlands. Chapter 4.

——, K. Rosling. 1993. Installation vs. echelon stock policies for multilevel inventory control. *Management Sci.* **39** 1274–1280.

[1]We are grateful to Jing-Sheng Song for helpful comments on earlier versions of this paper.

Buzacott, J., S. Price, J. Shanthikumar. 1992. Service level in multistage MRP and base stock controlled production systems. G. Fandel, T. Gulledge, A. Jones, eds. *New Directions for Operations Research in Manufacturing*. Springer, Berlin, Germany.

Chen, F., Y. Zheng. 1994. Lower bounds for multi-echelon stochastic inventory systems. *Management Sci.* **40** 1426–1443.

Clark, A., H. Scarf. 1960. Optimal policies for a multi-echelon inventory problem. *Management Sci.* **6** 475–490.

Federgruen, A. 1993. Centralized planning models for multi-echelon inventory systems under uncertainty. S. Graves, A. Rinnooy Kan, P. Zipkin, eds. *Logistics of Production and Inventory*. Elsevier (North-Holland), Amsterdam, The Netherlands. Chapter 3.

——, P. Zipkin. 1984. Computational issues in an infinite-horizon, multiechelon inventory model. *Oper. Res.* **32** 818–836.

Fooks, J. 1993. *Profiles for Performance*. Addison-Wesley, New York.

Gallego, G., I. Moon. 1993. The distribution-free newsboy problem: Review and extensions. *J. Oper. Res. Soc.* **44** 825–834.

——, P. Zipkin. 1994. Qualitative analysis of multi-stage production-transportation systems: Stock positioning and performance estimation. Working paper, Columbia University, New York.

Graves, S. 1985. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Sci.* **31** 1247–1256.

——, A. Rinnooy Kan, P. Zipkin, eds. 1993. *Logistics of Production and Inventory*. Handbooks in Operations Research and Management Science, Volume 4, Elsevier (North-Holland), Amsterdam, The Netherlands.

Gross, D., R. Soland, C. Pinkus. 1981. Designing a multi-product, multi-echelon inventory system. L. Schwarz ed. *Multi-Level Production/Inventory Control Systems: Theory and Practice*. North-Holland, Amsterdam, The Netherlands. Chapter 1.

Kelly, F. 1979. *Reversibility and Stochastic Networks*. Wiley, New York.

Langenhoff, L., W. Zijm. 1990. An analytical theory of multi-echelon production/distribution systems. *Statist. Neerlandica* **44** 3, 149–174.

Lee, Y., P. Zipkin. 1992. Tandem queues with planned inventories. *Oper. Res.* **40** 936–947.

Nahmias, S. 1981. Managing reparable item inventory systems: A review. L. Schwarz ed. *Multi-Level Production/Inventory Control Systems: Theory and Practice*. North-Holland, Amsterdam, The Netherlands. Chapter 13.

Rosling, K. 1989. Optimal inventory policies for assembly systems under random demands. *Oper. Res.* **37** 565–579.

Scarf, H. 1958. A min-max solution of an inventory problem. K. Arrow, S. Karlin, H. Scarf, eds. *Studies in the Mathematical Theory of Inventory and Production*. Stanford University, Stanford, CA. Chapter 12.

Schraner, E. 1994. Optimal production operations sequencing. Working paper, Stanford University, Stanford, CA.

Schwarz, L., ed. 1981. *Multi-Level Production/Inventory Control Systems: Theory and Practice*. North-Holland, Amsterdam, The Netherlands.

Sherbrooke, C. 1968. METRIC: A multi-echelon technique for recoverable item control. *Oper. Res.* **16** 122–141.

——. 1986. VARI-METRIC: Improved approximations for multi-indenture, multi-echelon availability models. *Oper. Res.* **34** 311–319.

Svoronos, A., P. Zipkin. 1991. Evaluation of one-for-one replenishment policies for multiechelon inventory systems. *Management Sci.* **37** 68–83.

Wagner, H., T. Whitin. 1958. Dynamic version of the economic lot size model. *Management Sci.* **5** 89–96.

Zipkin, P. 1995. Processing networks with planned inventories: Tandem queues with feedback. *European J. Oper. Res.* **80** 344–349.

Accepted by Stephen Graves; received November 18, 1996. This paper has been with the authors 8 $\frac{1}{2}$ months for 4 revisions.

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*)

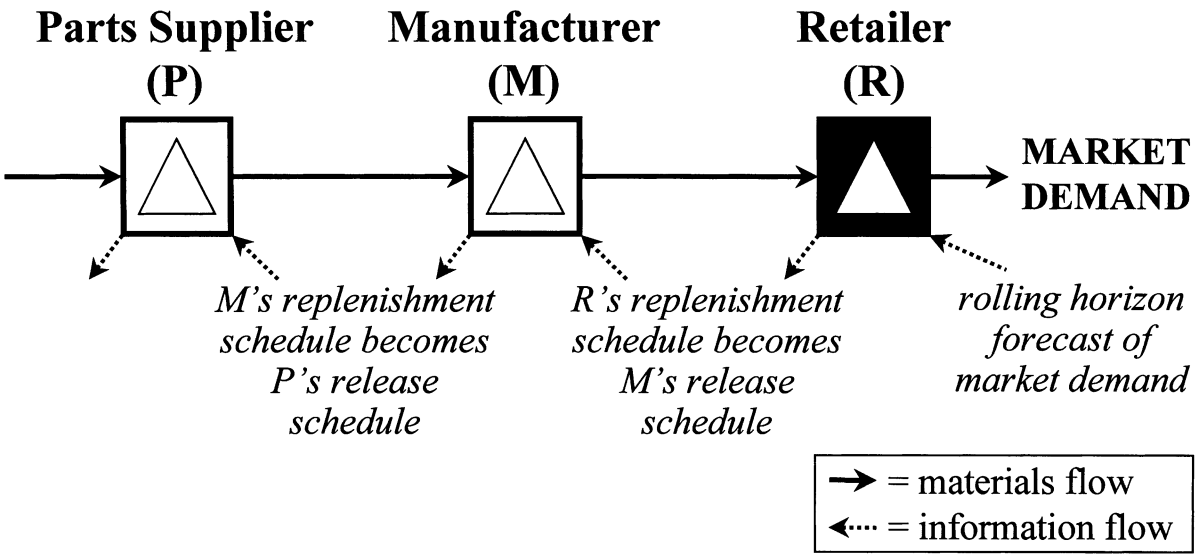
1. Introduction

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).

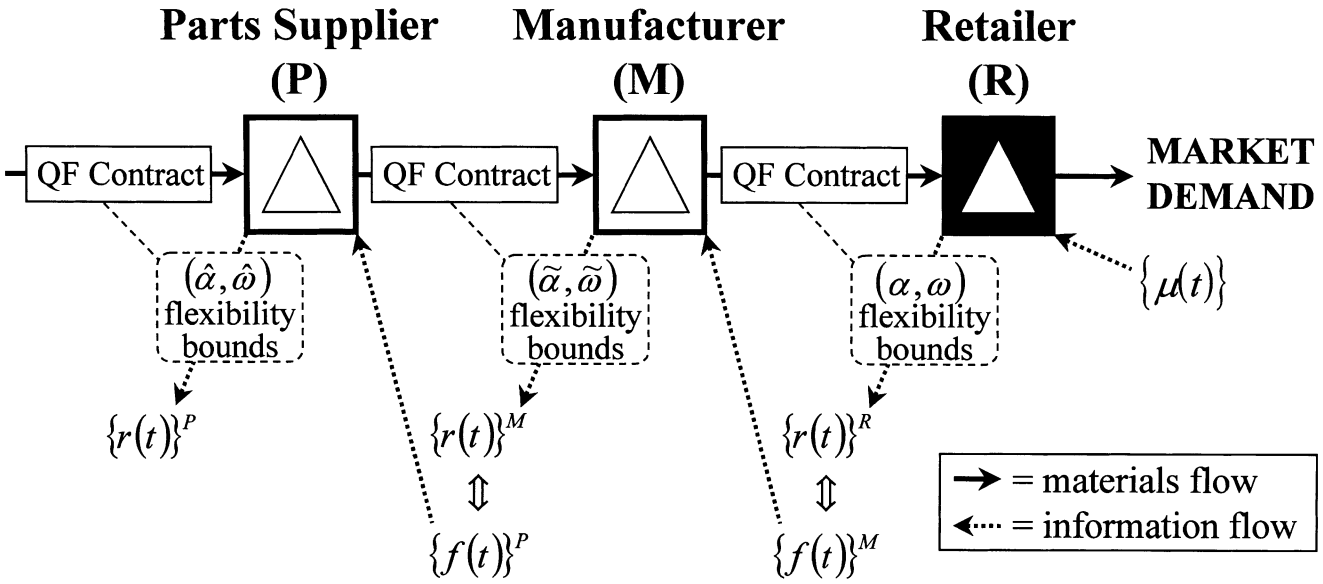
Various remedies to this well-known inefficiency have been attempted, a number of which are noted in §2. One approach that has become popular in many industries is the Quantity Flexibility (QF) contract, which attaches a degree of commitment to the forecasts by installing constraints on the buyer's ability 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

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

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

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

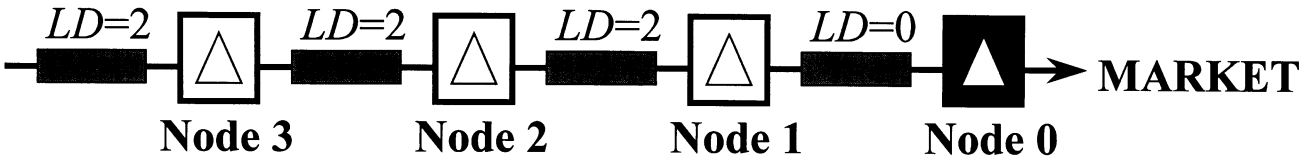
Whereas for the flex node the release-side contractual obligation induced a *deterministic* schedule of future releases on which to focus, here there is no such commitment, reflected in the lack of an analog to Equation (18). Hence, in contrast to (F-OLFC), this open-loop objective function still involves 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 $\delta = 0$

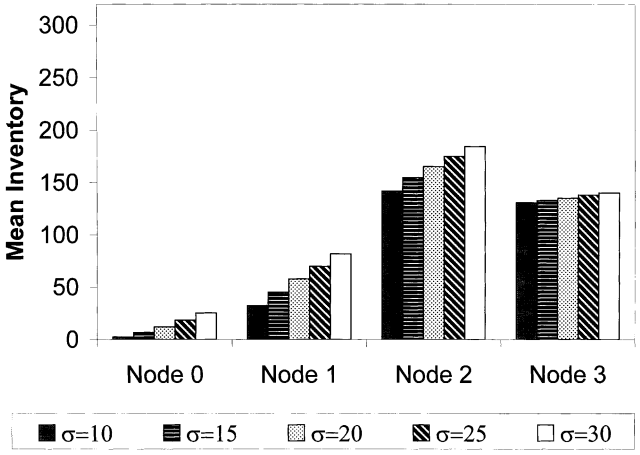


Figure 7 Inventory vs. δ, with $\sigma = 20$

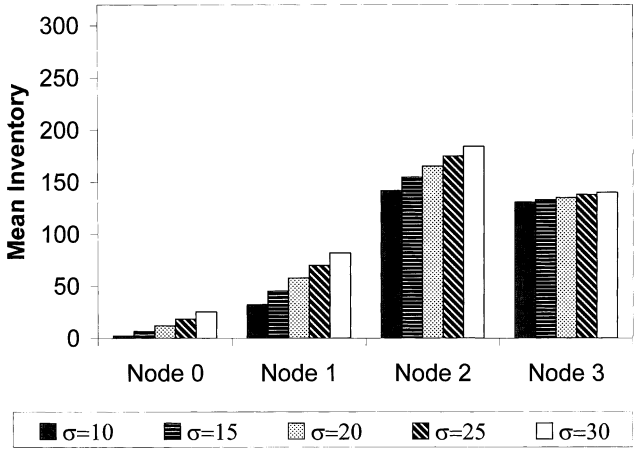
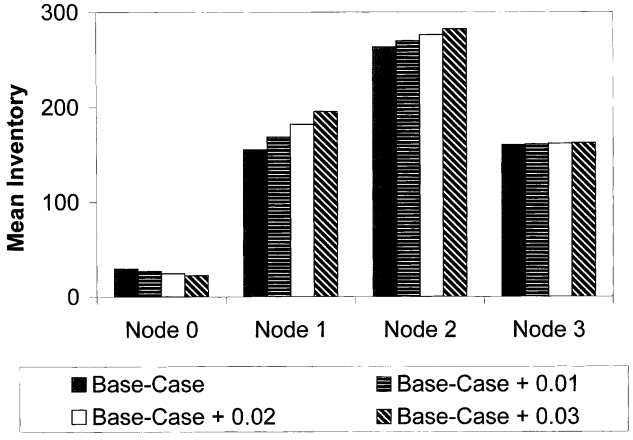


Figure 8 Inventory vs. $(A^{out}, \Omega^{out})^{Node\ 2}$



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. $(A^{out}, \Omega^{out})^{Node\ 1}$

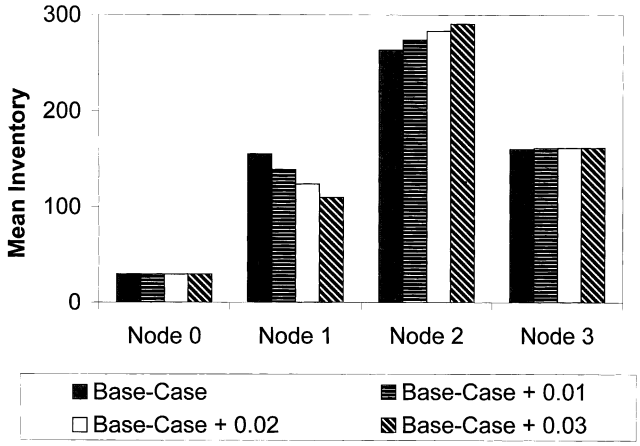


Figure 10 System Variability vs. σ, with $\delta = 0$

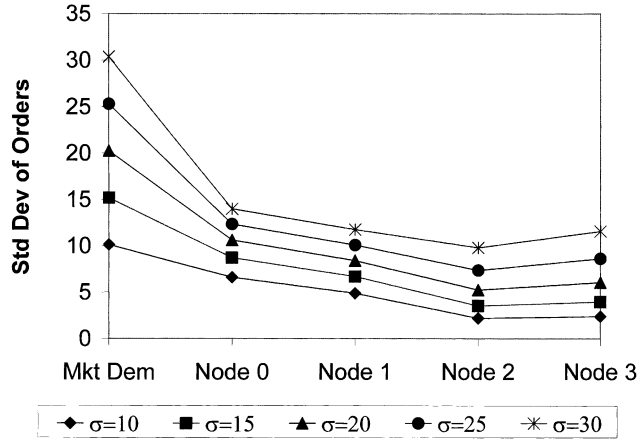


Figure 11 System Variability vs. δ, with $\sigma = 20$

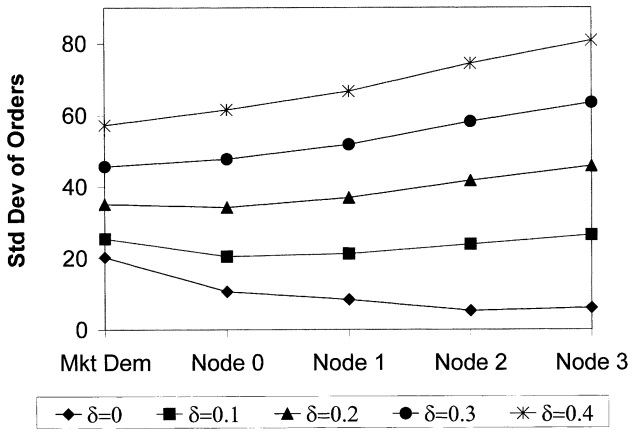
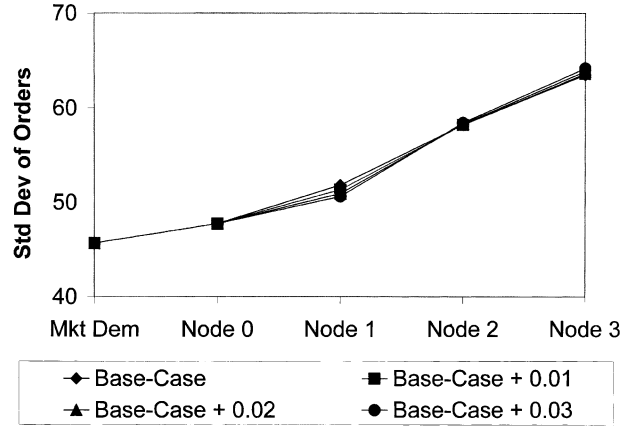


Figure 12 System Variability vs. $(A^{out}, \Omega^{out})^{Node\ 2}$, with $\{\delta, \sigma\} = \{0.3, 20\}$



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 $\{\delta, \sigma\} = \{0.3, 20\}$

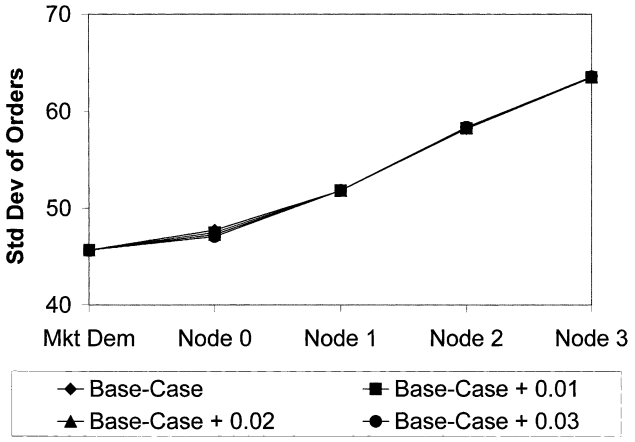


Figure 14 Node 0 Performance vs. σ, $\delta = 0$

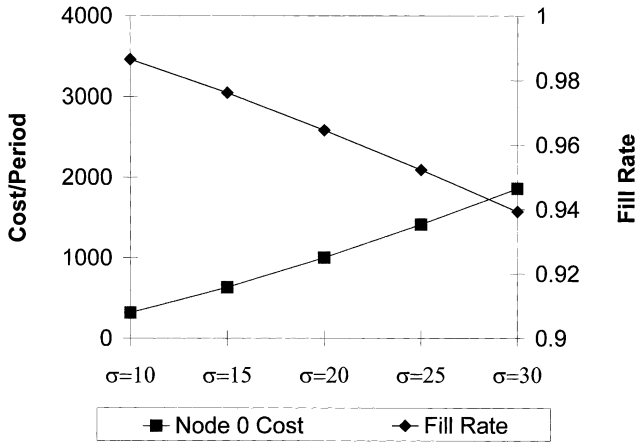


Figure 15 Node 0 Performance vs. δ, $\sigma = 20$

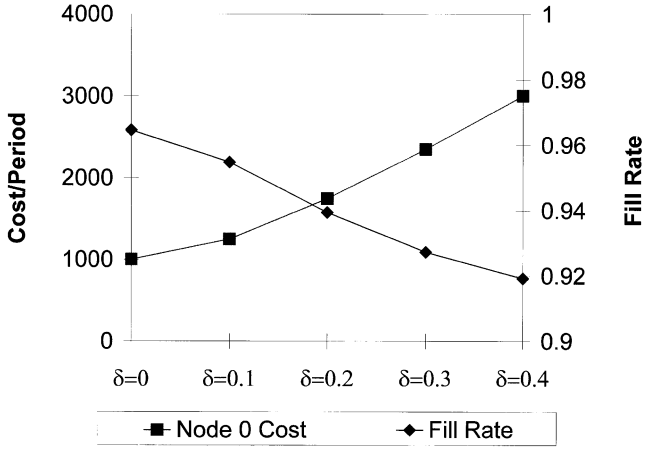
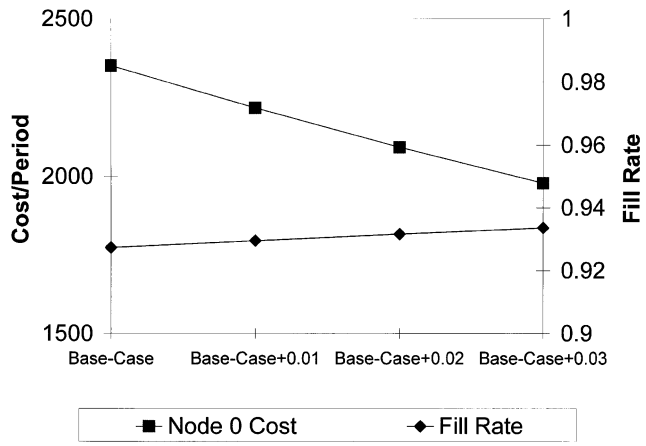


Figure 16 Node 0 Performance vs. $(A^{in}, \Omega^{in})^{Node\ 0}$, with $\{\delta, \sigma\} = \{0.3, 20\}$



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

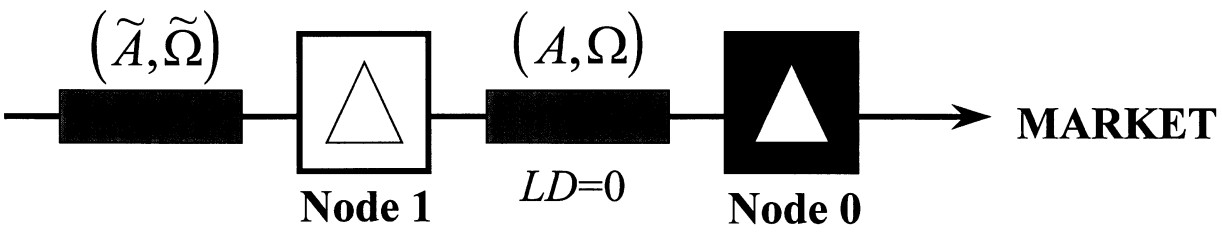
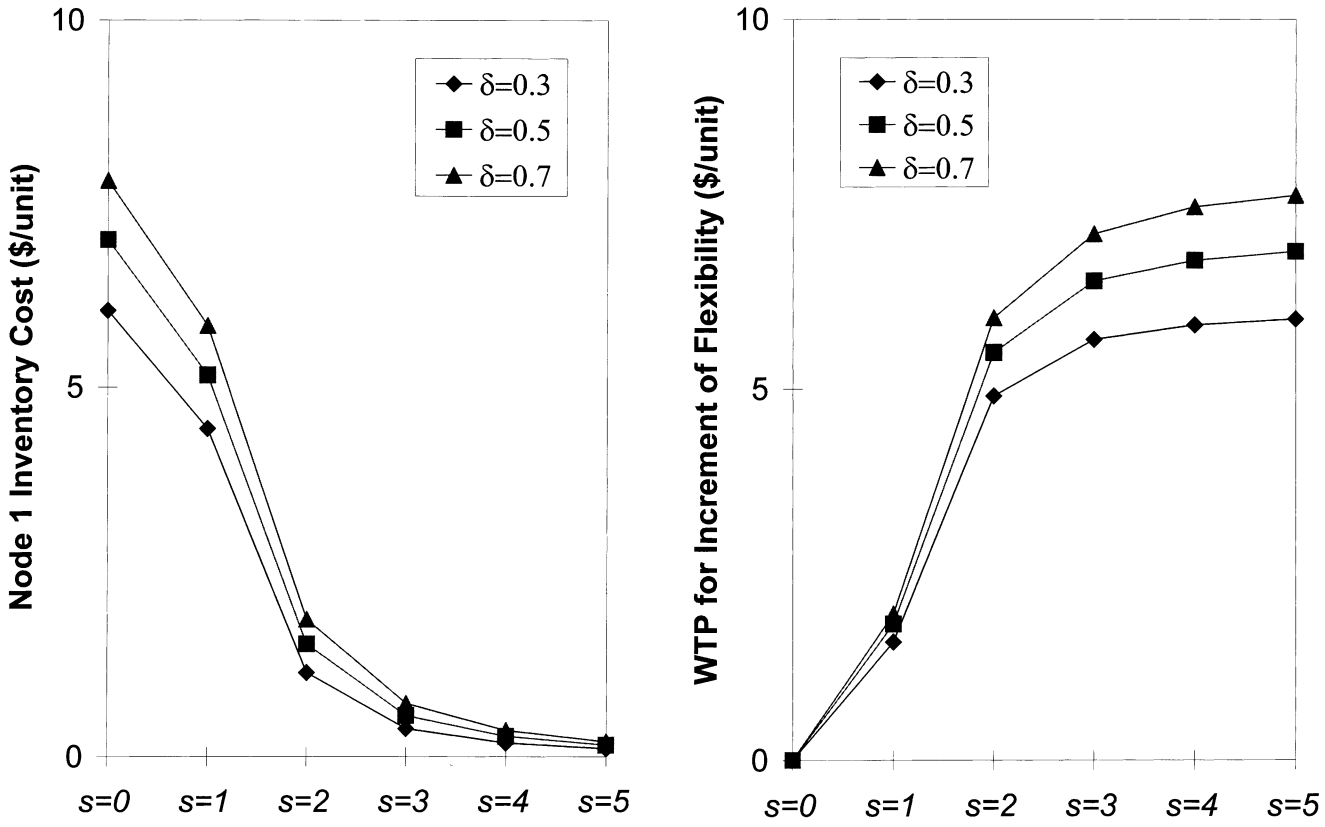


Figure 18 **Node 1 Inventory Cost, Willingness-to-Pay (WTP) (per unit) vs. Supply Flexibility**



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.

Optimizing Strategic Safety Stock Placement in Supply Chains

Stephen C. Graves • Sean P. Willems
Leaders for Manufacturing Program and A. P. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, sgraves@mit.edu
College of Business Administration, University of Cincinnati, Cincinnati, Ohio 45221-0130

Manufacturing managers face increasing pressure to reduce inventories across the supply chain. However, in complex supply chains, it is not always obvious where to hold safety stock to minimize inventory costs and provide a high level of service to the final customer. In this paper we develop a framework for modeling strategic safety stock in a supply chain that is subject to demand or forecast uncertainty. Key assumptions are that we can model the supply chain as a network, that each stage in the supply chain operates with a periodic-review base-stock policy, that demand is bounded, and that there is a guaranteed service time between every stage and its customers. We develop an optimization algorithm for the placement of strategic safety stock for supply chains that can be modeled as spanning trees. Our assumptions allow us to capture the stochastic nature of the problem and formulate it as a deterministic optimization. As a partial validation of the model, we describe its successful application by product flow teams at Eastman Kodak. We discuss how these flow teams have used the model to reduce finished goods inventory, target cycle time reduction efforts, and determine component inventories. We conclude with a list of needs to enhance the utility of the model.
(*Base-Stock Policy*; *Dynamic Programming Application*; *Multi-echelon Inventory System*; *Multi-Stage Supply-Chain Application*; *Safety Stock Optimization*)

1. Introduction

Manufacturing firms are subject to pressure to do everything faster, cheaper, and better. Firms are expected to continue to improve customer service by increasing on-time deliveries and reducing delivery lead-times. At the same time, they must provide this service more cheaply and utilize fewer assets. Increasingly, firms need to do this for a global marketplace.

This pressure to improve forces companies to look at their operations from a supply-chain perspective and to seek improvements from better coordination and communication across the supply chain. A supply-chain perspective is essential to avoid some of the local suboptimization that occurs when each step in a process operates independently with its own metrics and rewards. Using a supply chain as a focusing mechanism challenges an organization to examine cross-functional solutions to address some of the barriers that inhibit improvements.

The primary intent of this research is to develop a tactical tool to help cross-functional teams in their efforts to model and improve a supply chain. We provide a framework for modeling a supply chain and develop an approach, within the framework, to optimize the inventory in a supply chain. More specifically, we provide an optimization algorithm for finding the optimal placement of safety stock in a supply chain, modeled as a spanning tree and subject to uncertain demand. Key assumptions for the optimization are that each stage of the supply chain operates with a

1523-4614/00/0201/0068$05.00
1526-5498 electronic ISSN

periodic-review, base-stock policy, that each stage quotes a guaranteed service time to its customers, and that demand is bounded.

We refer to this effort as the placement of "strategic" safety stock. As will be seen, the optimization model leads to the determination of where to place decoupling inventories that protect one part of the supply chain from another. In particular, a decoupling safety stock is an inventory large enough to permit the downstream portion of the supply chain to operate independently from the upstream, provided that the upstream portion replenishes the external demand. In this sense, the determination of where to place these decoupling points in a supply chain is a major design decision and is "strategic" in nature. Furthermore, this terminology is consistent with that used in industry.

In order to have an opportunity to test the research and validate its utility for industry, we have built a commercial-quality software application to implement the model described in this paper. The software can be downloaded from our website, http://web.mit.edu/lfmrg3/www/.

In the remainder of this section we briefly discuss related literature. In §2, we present our framework for modeling a supply chain by stating and discussing the key assumptions. We introduce the model for a single stage in §3; this serves as the building block for the multi-stage model described in §4. In §5 we develop the optimization algorithm for safety stock placement in a supply chain modeled as a spanning tree. We present an overview of our application experience with the model in §6, and conclude in §7 with thoughts on how to improve the tool.

Related Literature.

There is an extensive literature on inventory models for multi-stage or multi-echelon systems with uncertain demand; much of this literature is applicable to supply chains as now defined. We refer the reader to the survey articles by Axsäter (1993), Federgruen (1993), Inderfurth (1994), van Houtum et al. (1996), and Diks et al. (1996). Within this vast literature, we mention two sets of papers that are most related to our work.

First, we note the work by Simpson (1958) who determined optimal safety stocks for a supply chain modeled as a serial network. Our work is based on similar assumptions about the demand process and about the internal control policies for the supply chain. Our work is also closely related to that of Inderfurth (1991, 1993), Inderfurth and Minner (1998), and Minner (1997), who also build off Simpson's framework for optimizing safety stocks in a supply chain. We extend the work of Simpson and of Inderfurth and Minner by treating a more general network, namely spanning trees. We also provide a different, and we believe richer, interpretation of the framework and its applicability to practice. We provide new results in the appendix on the form of the optimal policies when we relax a constraint on the internal control policy for the supply chain.

Second, our work is closely related in intent to Lee and Billington (1993), Glasserman and Tayur (1995), and Ettl et al. (2000). Each of these papers examines the determination of the optimal base-stock levels in a supply chain, and tries to do so in a way that is applicable to practice. Glasserman and Tayur (1995) show how to use simulation and infinitesimal perturbation analysis to find the optimal base-stock levels for capacitated multi-stage systems. Both Lee and Billington (1993) and Ettl et al. (2000) develop performance evaluation models of a multi-stage inventory system, where the key challenge is how to approximate the replenishment lead-times within the supply chain. They then formulate and solve a nonlinear optimization problem that minimizes the supply chain's inventory costs subject to user-specified requirements on the customer service level. Our work is similar in that we also assume base-stock policies and focus on minimizing the inventory requirements in a supply chain. The resulting models and algorithms are much different, though, due to different assumptions about the demand process and different constraints on service levels within the supply chain.

2. Assumptions

Multi-Stage Network.

We model a supply chain as a network where nodes are stages in the supply chain and arcs denote that an upstream stage supplies a downstream stage. A stage represents a major processing function in the supply chain. A stage might represent the procurement of a raw material, or the production of a component, or the

manufacture of a subassembly, or the assembly and test of a finished good, or the transportation of a finished product from a central distribution center to a regional warehouse. Each stage is a potential location for holding a safety-stock inventory of the item processed at the stage.

We associate with each arc a scalar ϕ_{ij} to indicate how many units of the upstream component i are required per downstream unit j. If a stage is connected to several upstream stages, then its production activity is an assembly requiring inputs from each of the upstream stages. A stage that is connected to multiple downstream stages is either a distribution node or a production activity that produces a common component for multiple internal customers.

Production Lead-Times.
For each stage, we assume a known deterministic production lead-time; call it T_i. When a stage reorders, the production lead-time, is the time from when all of the inputs are available until production is completed and available to serve demand. The production lead-time includes the waiting and processing time at the stage, plus any transportation time to put the item into inventory. For instance, suppose stage k requires inputs from stage i and j; then for a production request made at time t, stage k completes the production at time $t + T_k$, provided that there are adequate supplies of i and j at time t.

We assume that the production lead-time is not impacted by the size of the order; hence, in effect, we assume that there are no capacity constraints that limit production at a stage.

Periodic-Review Base-Stock Replenishment Policy.
We assume that all stages operate with a periodic-review base-stock replenishment policy with a common review period. For each period, each stage observes demand either from an external customer or from its downstream stages, and places orders on its suppliers to replenish the observed demand. There is no time delay in ordering; hence, in each period the ordering policy passes the external customer demand back up the supply chain so that all stages see the customer demand.

Demand Process.
Without loss of generality, we assume that external demand occurs only at nodes that have no successors, which we term demand nodes or stages. For each demand node j, we assume that the end-item demand comes from a stationary process for which the average demand per period is μ_j.

An internal stage has only internal customers or successors; its demand at time t is the sum of the orders placed by its immediate successors. Since each stage orders according to a base-stock policy, the demand at internal stage i is:

$$d_i(t) = \sum_{(i,j)\in A} \phi_{ij}\, d_j(\mathrm{t})$$

where $d_j(t)$ denotes the realized demand at stage j in period t and A is the arc set for the network representation of the supply chain. For every arc $(i, j) \in A$, stage j orders an amount $\phi_{ij}\, d_j(t)$ from upstream stage i in time period t. The average demand rate for stage i is:

$$\mu_i = \sum_{(i,j)\in A} \phi_{ij}\, \mu_j.$$

We assume that demand at each stage j is *bounded* by the function $D_j(\tau)$, for $\tau = 1, 2, 3, \ldots M_j$, where M_j is the maximum replenishment time for the stage.[1] That is, for any period t and for $\tau = 1, 2, 3, \ldots M_j$, we have

$$D_j(\tau) \geq d_j(t - \tau + 1) + d_j(t - \tau + 2) + \cdots + d_j(t).$$

We define $D_j(0) = 0$ and assume that $D_j(\tau)$ is increasing and concave on $\tau = 1, 2, 3, \ldots M_j$.

Discussion of Assumption of Bounded Demand.
The assumption of bounded demand is contrary to most of the literature on stochastic-demand inventory models, and as such, is controversial. We need to discuss this assumption in the context of the intent of the research, namely to provide tactical guidance for where to position safety stock in a supply chain.

We presume that it is possible to establish a meaningful upper bound on demand over varying horizons for each end item. By meaningful, we mean in the context of setting safety stocks: the safety stock is set to cover all demand realizations that fall within the upper bounds. If demand exceeds the upper bounds, then the safety stock, by design, is not adequate. In such extraordinary cases, a manager resorts to other tactics to

[1]The maximum replenishment time for node j is defined as $M_j = T_j + \max\{M_i \mid i{:}(i,j) \in A\}$.

handle the excess demand. A manager might use expediting, subcontracting, premium freight transportation, and/or overtime to accommodate the windfall of demand. In specifying the demand bounds, a manager indicates explicitly a preference for how demand variation should be handled—what range is covered by safety stock and what range is handled by other actions or responses.

As an example, consider a typical assumption where demand for end item j is normally distributed each period and i.i.d., with mean μ and standard deviation σ. Then, for the purposes of positioning safety stock, a manager might specify the demand bounds at the demand node by:

$$D_j(\tau) = \tau\mu + k\sigma\sqrt{\tau} \quad (1)$$

where k reflects the percentage of time that the safety stock covers the demand variation. The choice of k indicates how frequently the manager is willing to resort to other tactics to cover demand variability.

In some contexts there may be natural bounds on the end-item demand. For instance, suppose the end item is a component or subassembly for a manufacturing process whose production is limited by capacity constraints or by a frozen master schedule. An example is a supply chain that supplies components to an automobile assembly line or an OEM subassembly to a system integrator. In these cases, bounded demand for the component corresponds to the maximum consumption by the manufacturing process over various time horizons.

For each internal stage we assume that we can also establish meaningful demand bounds. If stage i has a single successor, say stage j, then $D_i(\tau) = \phi_{ij} D_j(\tau)$ for all relevant τ. For stages with more than one successor, we require some judgment for deciding how to combine the demand bounds for the downstream stages to obtain a relevant demand bound for the upstream stage for the purposes of positioning the safety stock. One possibility is just to sum the downstream demand bounds; however, this approach assumes that there is no risk pooling from combining the demand of multiple end items. An alternative approach is to assume that there will be some relative reduction in variability as we combine demand streams, i.e., some risk pooling. For instance, we might infer the demand bounds for internal stages by means of an expression like

$$D_i(\tau) = \tau\mu_i + p\sqrt{\sum_{(i,j)\in A} \{\phi_{ij}(D_j(\tau)' - \tau\mu_j)\}^p} \quad (2)$$

where $p \geq 1$ is a given constant. Larger values of p correspond to more risk pooling. Setting $p = 1$ models the case of no risk pooling. If we were to model the end-item demand bounds by Equation (1), then setting $p = 2$ equates to combining standard deviations of independent demand streams.

We do not attempt to model what happens when actual demand exceeds the bounds. When this happens, we assume that the supply chain responds with an equally extraordinary measure, as noted above. We regard this as beyond the scope of the model, given the stated intention to provide tactical decision support. See Kimball (1988), Simpson (1958), and Graves (1988) for further discussion of this assumption.

Finally we note that there are *no assumptions made about the demand distribution.*

Guaranteed Service Times.
We assume that each demand node j promises a *guaranteed service time* S_j by which the stage j will satisfy customer demand. That is, the customer demand at time t, $d_j(t)$, must be filled by time $t + S_j$. Furthermore, we assume that stage j provides *100% service* for the specified service time: stage j delivers exactly $d_j(t)$ to the customer at time $t + S_j$.

Similarly, an internal stage i quotes and guarantees a service time S_{ij} for each downstream stage j, $(i, j) \in A$. Given a base-stock policy, stage j places an order equal to $\phi_{ij} d_j(t)$ on stage i at time t; then stage i delivers exactly this amount to stage j at time $t + S_{ij}$.

For the initial development, we assume that stage i quotes the same service time to all of its downstream customers; that is, $S_{ij} = S_i$ for each downstream stage j, $(i, j) \in A$. Graves and Willems (1998) describe how to extend the model to permit customer-specific service times. In brief, if there is more than one downstream customer, we can insert zero-cost, zero-production lead-time dummy nodes between a stage and its customers to enable the stage to quote different service times to each of its customers. The stage quotes the same service time to the dummy nodes and each dummy node is free to quote any valid service time to its customer stage.

The service times for both the end items and the internal stages are *decision variables* for the optimization model, as will be seen in §4. However, as a model input, we may impose bounds on the service times for each stage. In particular, we assume that for each end item we are given a maximum service time, presumably set by the marketplace.

Discussion of Assumption of Guaranteed Service Times.

We assume that there are no violations of the guaranteed service times; each stage provides perfect or 100% service to its customers. As such, we do not explicitly model a tradeoff between possible shortage costs and the costs for holding inventory. Rather, we pose the problem as being how to place safety stocks across the supply chain to provide 100% service for the assumed bounded demand with the least inventory holding cost.

In defense of this assumption, it is often very difficult in practice to assess shortage costs for an external customer. Similarly, when we have asked managers for their desired service level, more often than not the response is that there should be no stock-outs for external customers. We have found that managers seem more comfortable with the notion of 100% service for some range of demand; they accept the fact that if demand exceeds this range, they will have shortages unless they can somehow expand the response capability of their supply chain. The assumptions for the model presented herein are consistent with this perspective.

For an internal customer, guaranteed service times need not be optimal in terms of least inventory costs. Indeed we show in the Appendix how to relax this assumption for a serial network, and report the cost impact of this assumption for a set of 36 test problems: the safety stock holding cost is 26% higher on average, while the total inventory cost is 4% higher on average. However, guaranteed service times are very practical in contexts where there is a need to coordinate replenishments. For instance, any assembly or subassembly stage requires the concurrent availability of multiple components, not all of which might be explicitly included in the model. When we assume guaranteed service times, we make the challenge of coordinating the availability of these components much easier.

3. Single-Stage Model

In this section we present the single-stage model (see Kimball 1988 or Simpson 1958) that serves as the building block for modeling a multi-stage supply chain.

Inventory Model

We assume the inventory system starts at time 0 with initial inventory $I_j(0)$. Given our assumptions, we can express the finished inventory at stage j at the end of period t as

$$I_j(t) = B_j - d_j(t - SI_j - T_j,\ t - S_j) \qquad (3)$$

where $B_j = I_j(0) \geq 0$ denotes the base stock, $d_j(a, b)$ denotes the demand at stage j over the time interval $(a, b]$, and SI_j is the inbound service time for stage j. Since we assume a discrete-time demand process, we understand $d_j(a, b)$ to be

$$d_j(a, b) = d_j(a + 1) + d_j(a + 2) + \cdots + d_j(b)$$

for $a < b$, where $d_j(t) = 0$ for $t \leq 0$. When $a \geq b$, we define $d_j(a, b) = 0$.

The inbound service time SI_j is the time for stage j to get supplies from its immediate suppliers and to commence production. In period t, stage j places an order equal to $\phi_{ij}\ d_j(t)$ on each upstream stage i for which $\phi_{ij} > 0$. Stage j cannot start production to replenish $d_j(t)$ until all inputs have been received; thus we have $SI_j \geq \max\{S_i \mid i{:}(i, j) \in A\}$.

We permit $SI_j > \max\{S_i \mid i{:}(i, j) \in A\}$ to allow for the possibility that the replenishment time for the inventory at stage j is less than its service time S_j; that is, the case when

$$\max\{S_i \mid i{:}(i, j) \in A\} + T_j < S_j.$$

In this case we would delay the orders to the suppliers by $S_j - \max\{S_i \mid i{:}(i, j) \in A\} - T_j$ periods, so that the supplies arrive exactly when needed. To account for this case, we set the inbound service time so that the effective replenishment time for the inventory at stage j, namely $SI_j + T_j$, equals the service time S_j, i.e., $SI_j + T_j = S_j$. Thus, we define the inbound service time as

$$SI_j = \max\{S_j - T_j,\ \max\{S_i \mid i{:}(i, j) \in A\}\}.$$

If the inbound service time is such that $SI_j > S_i$ for some $(i, j) \in A$, then by convention stage j delays orders from stage i by $SI_j - S_i$ periods to avoid unnecessary inventory.

Now, to explain Equation (3), we observe that in period t stage j completes the replenishment of the demand observed in period $t - SI_j - T_j$. By the end of period t, the cumulative replenishment to the inventory at stage j equals $d_j(0, t - SI_j - T_j)$. In period t, stage j fills the demand observed in time period $t - S_j$ from its inventory. By the end of period t the cumulative shipment from the inventory at stage j equals $d_j(0, t - S_j)$. The difference between the cumulative replenishment and the cumulative shipment is the inventory shortfall, $d_j(t - SI_j - T_j, t - S_j)$. The on-hand inventory at stage j is the initial inventory or base stock minus the inventory shortfall, as given by Equation (3).

Determination of Base Stock.
In order for stage j to provide 100% service to its customers, we require that $I_j(t) \geq 0$; from (3) we see that this requirement equates to

$$B_j \geq d_j(t - SI_j - T_j, t - S_j).$$

Since demand is bounded, we can satisfy the above requirement with the least inventory by setting the base stock as follows:

$$B_j = D_j(\tau) \quad \text{where } \tau = SI_j + T_j - S_j. \tag{4}$$

Any smaller value does not assure that $I_j(t) \geq 0$, and thus cannot guarantee 100% service.

In words, the base stock equals the maximum possible demand over the *net* replenishment time for the stage. The *net* replenishment time for stage j is the replenishment time ($SI_j + T_j$) minus its service time (S_j). At any time t, stage j has filled its customers' demand through time $t - S_j$, but has only been replenished for demand through time $t - SI_j - T_j$. The base stock must cover this time interval of exposure, namely the net replenishment time.

Safety Stock Model.
We use Equations (3) and (4) to find the expected inventory level $E[I_j]$:

$$\begin{aligned} E[I_j] &= B_j - E[d_j(t - SI_j - T_j, t - S_j)] \\ &= D_j(SI_j + T_j - S_j) - (SI_j + T_j - S_j)\mu_j. \end{aligned} \tag{5}$$

The expected inventory represents the safety stock held at stage j, and depends on the net replenishment time and the demand bound. As an example, suppose the demand bound is given by Equation (1); then the safety stock is $E[I_j] = k\sigma\sqrt{SI_j + T_j - S_j}$.

Pipeline Inventory.
In addition to the safety stock, we may want to account for the in-process or pipeline stock at the stage. Following the development for Equation (3), we observe that the work-in-process inventory at time t is given by

$$W_j(t) = d_j(t - SI_j - T_j, t - SI_j).$$

That is, the work-in-process corresponds to T_j periods of demand. The expected work-in-process depends only on the lead-time at stage j and is not a function of the service times:

$$E[W_j] = T_j\mu_j.$$

Hence, in posing an optimization problem in the next section, we ignore the pipeline inventory and only model the safety stock. This is not to say that the work-in-process is not a significant part of the inventory in a supply chain. But for the purposes of this work, we assume that the lead-time of a stage, as well as the demand rate, are input parameters and thus the pipeline stock is predetermined. Nevertheless, in any application, we account for both the safety stock and the pipeline stock as both will contribute to the total supply chain inventory.

4. Multi-Stage Model

To model the multi-stage system, we use Equation (5) for every stage, but where the inbound service time is a function of the outbound service times for the upstream stages; to wit, the model for stage j is

$$E[I_j] = D_j(SI_j + T_j - S_j) - (SI_j + T_j - S_j)\mu_j, \tag{6}$$

$$SI_j + T_j - S_j \geq 0, \tag{7}$$

$$SI_j - S_i \geq 0 \quad \text{for all } (i, j) \in A. \tag{8}$$

Equation (6) expresses the expected safety stock as a function of the net replenishment time. Equation (7) assures that the net replenishment time is nonnegative. Equation (8) constrains the inbound service time to equal or exceed the service times for the upstream stages.

We assume that the production lead-times, the means and bounds of the demand processes, and the

maximum service times for the demand nodes are known input parameters. This suggests the following optimization problem **P** for finding the optimal service times:

$$\mathbf{P} \min \sum_{j=1}^{N} h_j\{D_j(SI_j + T_j - S_j) - (SI_j + T_j - S_j)\mu_j\}$$

$$\text{s. t.} \quad S_j - SI_j \le T_j \quad \text{for } j = 1, 2, \ldots, N,$$

$$SI_j - S_i \ge 0 \quad \text{for all } (i, j) \in A,$$

$$S_j \le s_j \quad \text{for all demand nodes } j,$$

$$S_j, SI_j \ge 0 \text{ and integer} \quad \text{for } j = 1, 2, \ldots, N,$$

where h_j denotes the per-unit holding cost for inventory at stage j, and s_j is the maximum service time for demand node j. The objective of problem **P** is to minimize the holding cost for the safety stock in the supply chain. The constraints assure that the net replenishment times are nonnegative, the inbound service time equals the maximum supplier service time, and the end-item stages satisfy their service guarantee. The decision variables are the service times.

Problem **P** is a nonlinear optimization problem. The objective function is a concave function, provided that the demand bound $D_j(\,)$ is a concave function for each stage j. The feasible region is convex but not necessarily bounded; however, one can show that the optimal service times need not exceed the sum of the production lead-times, provided that $D_j(\,)$ is a nondecreasing function for each stage j. Thus, problem **P** is the minimization of a concave function over a closed, bounded convex set. An optimum for such problems is at an extreme point of the feasible region, e.g., Luenberger 1973.

Simpson (1958) considered a serial-line supply chain, where he assumed that the guaranteed service time for the external customer is zero. Simpson showed that there is an optimal extreme point solution for **P** for which $S_i = 0$ or $S_i = S_{i+1} + T_i$, where stage $i + 1$ supplies stage i. Thus, there is an "all or nothing" optimal solution; a stage either has no safety stock ($S_i = S_{i+1} + T_i$) or has sufficient safety stock ($S_i = 0$) to de-couple it from its downstream stage. Gallego and Zipkin (1999) provide supporting evidence that "all or nothing" policies can be near optimal in serial systems under more traditional assumptions where demand is not bounded.

Graves (1988) observed that the serial-line problem can be solved as a shortest path problem. In a series of papers. Inderfurth (1991, 1993), Inderfurth and Minner (1998), and Minner (1997) show how to solve problem **P** by dynamic programming when the supply chain is an assembly network or a distribution network. Graves and Willems (1996) developed similar results for assembly and distribution networks. In the next section we present a dynamic programming algorithm for the more general case of a spanning tree.

5. Algorithm for Spanning Tree

We describe in this section how to solve **P** by dynamic programming when the underlying network for the supply chain is a spanning tree, like in the Figure 1.

We solve **P** by decomposing the problem into N stages where N is the number of nodes in the spanning tree and there is one stage for each node. For a spanning tree, there is not a readily-apparent ordering of the nodes by which the algorithm would proceed. Indeed, we label the nodes in a spanning tree (and thus sequence the algorithm) so that there will be a single state variable for the dynamic programming recursion. However, the state variable for the dynamic program will be either the inbound service time at a stage or its outbound service time, where the determination depends on the topology of the network.

We first present the algorithm for labeling the nodes. Next we present the functional equations for the dynamic programming recursions, and then state the algorithm.

Labeling the Nodes.

The algorithm for labeling or re-numbering the nodes is as follows:

1. Start with all nodes in the unlabeled set, U.
2. Set $k := 1$.
3. Find a node $i \in U$ such that node i is adjacent to at most one other node in U. That is, the degree of node i is 0 or 1 in the subgraph with node set U and arc set A defined on U.
4. Remove node i from set U and insert into the labeled set L; label node i with index k.

5. Stop if U is empty: otherwise set $k := k + 1$ and repeat steps 3– 4.

For a spanning tree, there is always an unlabeled node in step 3 that is adjacent to at most one other unlabeled node. As a consequence, the algorithm will eventually label all of the nodes in N iterations. Indeed, each node labeled in the first $N - 1$ steps is adjacent to exactly one other node in set U. Thus, the nodes with labels 1, 2, . . . , $N - 1$ have one adjacent node with a higher label, denoted by $p(k)$ for $k = 1, 2, \ldots, N - 1$. Node N has no adjacent nodes with larger labels.

As an illustration, we renumber the nodes in Figure 1 to produce Figure 2. Note that the labeling is not unique as there may be multiple choices for node i in step 3.

For each node k we define N_k to be the subset of nodes {1, 2, . . . , k} that are connected to k on the subgraph with node set {1, 2, . . . , k}. We will use N_k to explain the dynamic programming recursion. We can determine N_k by the following equation:

$$N_k = \{k\} + \bigcup_{i<k,(i,k)\in A} N_i + \bigcup_{j<k,(k,j)\in A} N_j.$$

For instance, in Figure 2, N_k is {3} for $k = 3$, {1, 2, 3, 9} for $k = 9$, {1, 2, 3, 4, 5, 9, 11} for $k = 11$, and {6, 7, 8, 10, 12} for $k = 12$. We can compute N_k as part of the labeling algorithm.

Figure 1 **Spanning Tree**

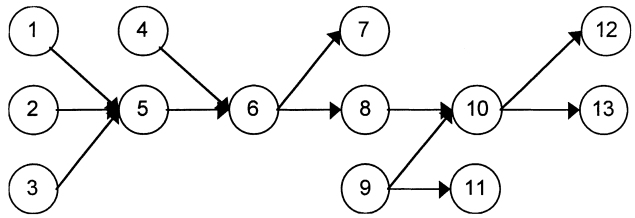
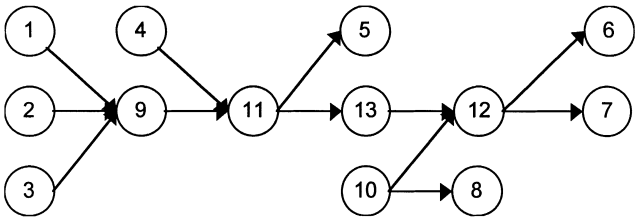


Figure 2 **Renumbered Spanning Tree**



Functional Equations.

The dynamic program evaluates a functional equation for each node in the spanning tree, where we have renumbered the nodes as described above. There are two forms for the functional equation. First, the function $f_k(S)$ is the minimum holding cost for safety stock in a subnetwork with node set N_k, where we assume that the outbound service time for stage k is S. Second, the function $g_k(SI)$ is the minimum holding cost for safety stock in a subnetwork with node set N_k, where we assume that the inbound service time for stage k is SI.

At node k (or stage k) for $1 \le k \le N - 1$, the algorithm determines either $f_k(S)$ or $g_k(SI)$, depending upon the location of the node with higher label that is adjacent to k. If $p(k)$ is downstream [upstream] of node k, then we evaluate $f_k(S)$ [$g_k(SI)$]. For node N, we can evaluate either functional equation.

To develop the functional equations, we first define the minimum inventory holding cost for the subnetwork with node set N_k as a function of both the inbound service time and the outbound service time at node k:

$$c_k(S, SI) = h_k\{D_k(SI + T_k - S) - (SI + T_k - S)\mu_k\} + \sum_{\substack{(i,k)\in A \\ i<k}} f_i(SI) + \sum_{\substack{(k,j)\in A \\ j<k}} g_j(S).$$

The first term is the holding cost for the safety stock at node k as a function of S and SI.

The second term corresponds to the nodes in N_k that are upstream of k. For each node i that supplies node k, we include the minimum inventory holding costs for the subnetwork with node set N_i, as a function of SI. The inbound service time to node k, SI, is an upper bound for the outbound service time for node i. We can show that $f_i(\ )$, the inventory holding costs for the subnetwork with node set N_i, is nonincreasing in the service time at node i. Hence, we equate the outbound service time at i to the inbound service time at k without loss of generality.

The third term corresponds to the nodes in N_k that are downstream of k. For each node j, $j \in N_k$ and $(k, j) \in A$, we include the minimum inventory holding costs for the subnetwork with node set N_j, as a function of S. The outbound service time for node k, S, is a lower bound for the inbound service time for node j. We can

show that $g_j(\ )$, the inventory holding costs for the subnetwork with node set N_j, is nondecreasing in the inbound service time to node j; and thus we equate the inbound service time at j to the outbound service time at k without loss of generality.

We solve the following optimization by enumeration to find the functional equation $f_k(S)$:

$$f_k(S) = \min_{SI}\{c_k(S, SI)\}$$

$$\text{s. t. } \max(0, S - T_k) \le SI \le M_k - T_k, \text{ and } SI \text{ integer,}$$

where M_k is the maximum replenishment time for node k. The lower bound on SI comes from **P**, while the definition of M_k gives the upper bound.

The functional equation for $g_k(SI)$ is very similar in structure:

$$g_k(SI) = \min_{S}\{c_k(S, SI)\}$$

$$\text{s. t. } 0 \le S \le SI + T_k, \text{ and } S \text{ integer.}$$

If node k is a demand node, then we also constrain S by its maximum service time, i.e., $S \le s_k$. The minimization can be done by enumeration.

Dynamic Program.
The dynamic programming algorithm is now as follows:

1. For $k := 1$ to $N - 1$
   2. If $p(k)$ is downstream of k, evaluate $f_k(S)$ for $S = 0, 1, \ldots, M_k$.
   3. If $p(k)$ is upstream of k, evaluate $g_k(SI)$ for $SI = 0, 1, \ldots, M_k - T_k$.
4. For $k := N$ evaluate $g_k(SI)$ for $SI = 0, 1, \ldots, M_k - T_k$.
5. Minimize $g_N(SI)$ for $SI = 0, 1, \ldots, M_N - T_N$ to obtain the optimal objective function value.

This procedure finds the optimal objective function value; we can find an optimal set of service times by the standard backtracking procedure for a dynamic program.

To summarize, at each stage of the dynamic program, we find the minimum inventory holding costs for the subnetwork with node set N_k, as a function of a state variable. The state variable depends on the direction of the arc that connects the subnetwork N_k to the rest of the network. When the connecting arc originates in N_k, then the state variable is the outbound service time (step 2); otherwise, the state variable is the inbound service time (step 3). We number the nodes so that we have previously determined the functions required to evaluate either $f_k(S)$ or $g_k(SI)$. At stage N (step 4), we determine the inventory costs for the entire network as a function of the inbound service time to node N. At step 5, we optimize over the inbound service time to find the optimal inventory cost.

The computational complexity of the algorithm is of order NM^2 where M is the maximum service time, which is bounded by the sum of the production leadtimes $\sum_{j=1}^{N} T_j$. We have implemented the algorithm for a PC in the C++ programming language. The run times for real problems with 25 to 30 nodes are effectively instantaneous on a Pentium PC with a 100 megahertz Intel processor.

6. Application

This section presents an application of the model at the Eastman Kodak Company. Starting in 1995, Kodak has applied the model to more than eleven finished products across two of its assembly sites within its equipment division. We first present the model's application to the internal supply chain for a high-end digital camera,[2] and then summarize Kodak's financial results, as of 1997 year-end.

Product Background.
The key subassemblies for the digital camera are a traditional 35 mm camera, an imager, and a circuit-board assembly. The 35mm camera is procured from an outside vendor. The imager (a charge-coupled device) and the circuit-board assembly are produced internally. The 35mm camera supplies the lens, shutter, and focus functions for the digital camera. The imager captures and digitizes the picture, and the circuit-board assembly processes and stores the image. To produce the digital camera, the back of the 35mm camera is removed and replaced with a housing containing the imager

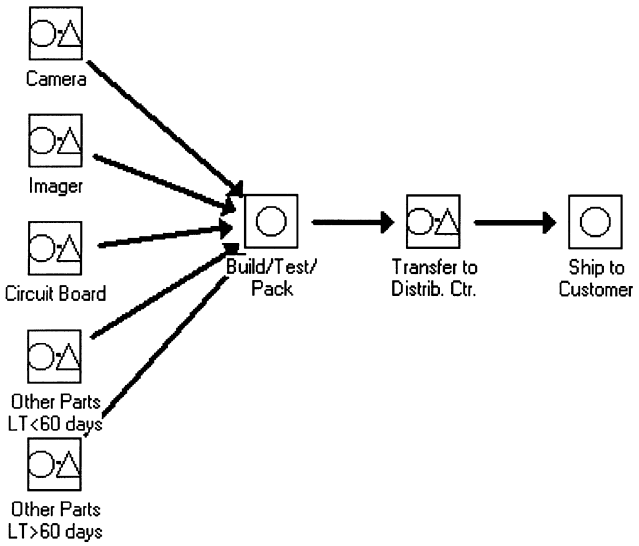
[2]The data presented in this section has been altered to protect proprietary information. However, the resulting qualitative relationships and insights drawn from this example are the same as they would be from using the actual data.

and circuit board. The camera is then tested to make sure that there are no defects in the imager. Once the camera passes the quality tests, the product is shipped to the distribution center. From the distribution center, the camera is shipped to the final customers, which for our purposes are high-end photography shops and computer superstores.

In Figure 3, we provide a high-level depiction of this supply chain. In addition to the three key subassemblies, we include the remaining parts in order to accurately represent the product's cost structure; since there are nearly 100 additional parts in a camera, modeling them in any level of detail would greatly expand the size of the model. Hence, we group these parts into two aggregate stages of the supply chain, where one stage represents all of the parts with long procurement lead-times (greater than 60 days) and the other stage represents the short lead-time parts (less than 60 days).

We also aggregate certain operations. As seen in Figure 3, we combine the build operation for a camera with the test operation and the packing operation. The imager stage and circuit board stage are also aggregates as each represents the flow through a separate department. The circuit-board stage entails circuit board assembly and test. The imager stage consists of a semiconductor operation to produce wafers, followed by packaging and testing of the semiconductors, followed by an assembly operation.

Figure 3 **Implemented Safety Stock Policy for Digital Camera. All Stages Have a Circle That Denotes the Processing Activity at the Stage. A Triangle Denotes That the Stage Holds a Safety Stock of Its Finished Goods**



Implementation Approach.

The product's supply chain crosses several functional boundaries within Kodak. Functional areas like circuit-board assembly and imager assembly are separate departments and act as suppliers to an assembly group that performs final assembly and test. Distribution is a separate organization and owns the product once it leaves the final assembly area. To improve coordination across the departments, the equipment division at Kodak has set up product flow teams with the general charge to optimize their supply chains.

The product flow team for the digital camera relied on the model to identify opportunities for better coordination and improved asset utilization. The team implemented the model in phases. The implementation strategy was to start simple and get experience with the model; once there was some evidence of the utility of the model, the team extended the application in increments to capture more and more of the supply chain.

Phase One.

The initial goal was to optimize the safety stock levels for the stages that were under the direct control of the final assembly area. The decision to start with the final assembly area was based on the product's high material cost and its relatively simple supply chain structure, as described above. The (disguised) costs and production lead-times are:

Table 1 **Phase One Digital Camera Information**

Stage Name	Production Lead-Time	Cost Added
Camera	60	750
Imager	60	950
Circuit Board	40	650
Other Parts LT $<$ 60 days	60	150
Other Parts LT $>$ 60 days	150	200
Build/Test/Pack	6	250
Transfer to DC	2	50
Ship to Customer	3	0

The demand bound was estimated by Equation (1) where $\mu = 11$, $\sigma = 7$ and k = 1.645. From looking at historical demand and future demand estimates, Kodak felt that this function realistically captured the range of demand for which they wanted to use safety stock.

This demand characterization excluded large one-time orders from the government and some large corporations. These orders are typically for 200 – 300 units with delivery scheduled less than a month from when the order is placed. However, since there is advance warning about these orders and they are independent of the other demand for the product, we developed a separate anticipatory stock policy to deal with large, infrequent orders.

Marketing determined that the maximum service time to the final customer is five days.

Finally, the assembly group imposed the constraint that a safety stock of imagers must be held on-site at final assembly. Therefore, we set the service time for the imager stage to be zero; the effect of this constraint increased the total safety stock cost by 8.7%.

In the optimal solution, the subassembly stages, the aggregate parts stages, and the build/test/package stages hold safety stocks and quote zero service times. The ship-to-distribution and ship-to-customer stages each quote their maximum feasible service times, two and five days, respectively. The annual holding cost for the safety stock is \$78,000. Thus, the optimal solution holds an inventory of components, subassemblies, and completed cameras at the manufacturing site, but holds no inventory in the distribution center. In effect, the distribution center would act only as an order processing and transshipment center. This is feasible since it is possible to get the product from the assembly area through the distribution center and to the final customer within the maximum service time of five days.

The product flow team decided to explore some near-optimal solutions because they felt that there were some additional organizational constraints not captured in the model; in particular, distribution would want to hold safety stock on-site. To ameliorate the situation, the team suggested that both manufacturing and distribution hold safety stock and quote zero service times. However, the model showed that the cost for the safety stock would increase to \$89,000. The team also investigated a policy in which the distribution center would hold safety stock but the manufacturing site would not. The safety stock cost for this policy was \$81,000, which was deemed to be acceptable as it was quite close to the unconstrained optimum and satisfied distribution's desire to hold inventory. This policy, as shown in Figure 3, was implemented at the end of phase one of the application.

Phase Two.

After the initial phase, the product flow team expanded the model to incorporate the internal supply chain for the imager. The resulting supply chain is shown in Figure 4.

Prior to this study, safety stocks of (in-process) imagers had been held at each stage of the supply chain. By application of the model, the team decided to remove safety stocks from two stages in the supply chain for the imagers, as shown in the figure. This required some increase in the downstream safety stocks of finished imagers, but overall the supply chain's safety stock of imagers (measured in terms of finished imagers) was more than halved.

Now that the model has been successfully piloted with an internal supplier, the product flow team is in the process of extending the model to incorporate other key internal and external suppliers.

Financial Results.

Table 2 contains the financial summary for two assembly sites that use the model. Site A has applied the

Figure 4 Digital Camera Supply Chain

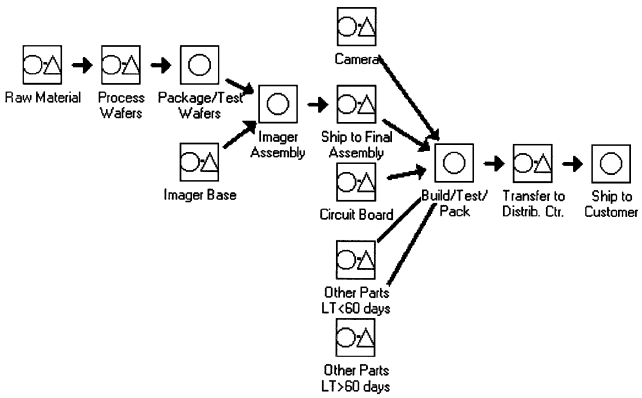


Table 2 Financial Summary for Assembly Sites *A* and *B*

Assembly Site *A*	*Y/E* 95	*Y/E* 96	*Y/E* 97
Worldwide FGI	\$6.7*m*	\$3.3*m*	\$3.6*m*
Raw Material & WIP	\$5.7*m*	\$5.6*m*	\$2.9*m*
Delivery Performance	80%	94%	97%
Manufacturing Operation	MTS	RTO	RTO
Assembly Site *B*			
Worldwide FGI	\$4.0*m*	\$4.0*m*	\$3.2*m*
Raw Material & WIP	\$4.5*m*	\$1.6*m*	\$2.5*m*
Delivery Performance	Unavailable	78%	94%
Manufacturing Operation	MTS	RTO	RTO

model to each of its eight products and Site B has applied the model to each of its three product families. The sales volume has remained relatively constant over the three years.

At the start of 1996, the sites moved from a make-to-schedule (MTS) to a replenish-to-order (RTO) system. The modeling effort began at the end of 1995 and was used to help guide the transition to replenish-to-order. The increase in worldwide finished goods inventory for 1997 is due to a marketing promotion that was underway in Europe. By our estimate, this promotion increased the finished goods inventory by as much as \$0.5 million. In the first year of the project, the emphasis was on reducing the areas directly under the control of final assembly. Over the following year, the effort was on reducing the raw material costs and WIP in the manufacturing supply chain. The total value of the inventory for these products has been reduced by over one third over the two years.

Kodak's product flow teams have also used the model for a variety of purposes other than setting safety stocks. Some products have tens of components with long procurement lead-times. The model has helped to prioritize the suppliers with whom to work to reduce these lead-times. The teams have used the model to determine the cost effectiveness of lead-time reduction efforts in manufacturing. One can compare the investment required to reduce a lead-time versus the cost savings from the reductions in pipeline and safety stock cost. Finally, manufacturing and marketing personnel have used the model to help quantify the cost of quoting a specific maximum service time to the final customer. With the model, the supply-chain team can accurately estimate the costs of a one-day, one-week, or two-week guaranteed service time to the customer, and weigh the costs of the policy against the marketing benefits of the policy.

Another benefit of the model is that it provides a common, objective framework with which a cross-functional supply-chain team can work. In particular, we note that it provides a standard terminology and set of assumptions for these teams to use as they work together to improve or optimize a supply chain. As such the model has been a very effective communication vehicle or platform.

7. Conclusion

In this paper we introduce and develop a model for positioning safety stock in a supply chain. We model the supply chain as a network, where the nodes of the network are the stages of a supply chain. We assume that each stage uses a base-stock policy to control its inventory. We also assume that each stage quotes a service time to its customers, both internal and external, and that each stage provides 100% service for these quoted service times. Finally we assume that external customer demand is bounded.

We show how to evaluate the inventory requirements at each stage as a function of the service times. For supply chains that can be modeled as spanning trees, we develop an optimization algorithm for finding the service times that minimize the holding cost for the safety stock in the supply chain.

As a form of validation, we describe an application of the model at Kodak to an internal supply chain for a digital camera. This application helped Kodak to reposition its inventories in this supply chain to reduce its inventory and increase its service performance. In particular, Kodak realized the benefit from creating a few strategic locations to hold safety stocks rather than spreading the safety stock across the entire supply chain. We have also applied the model to a number of other related products at Kodak and at two other companies (Black 1998, Coughlin 1998, Felch 1997, Wala 1999).

As with any research, we end with a number of unresolved issues and new questions. We discuss these

in the relative order of importance, based on our experience in applying the research to date.

Stochastic Lead-Times. We assume that associated with each stage is a known, deterministic lead-time. In practice, this is often not true; for instance, component procurement times are often highly uncertain. It will be of value to capture this in the model. We know how to extend the model in an approximate way for stages that procure raw materials or components from an outside vendor. In effect, for such a stage we just need to approximate its inventory requirements as a function of the outbound service time quoted by the stage and the stochastic procurement time. But it is less clear how to extend the model, either exactly or approximately, to permit stochastic lead-times at stages whose function is not procurement.

Non-Stationary Demand. We assume that the end-item demand processes are stationary. Yet virtually all of the products with which we have worked have short lifetimes over which demand is never really stationary. In practice, one runs the model under various (stationary) scenarios to see how sensitive the safety stock is to the demand characterization (Coughlin 1998). Fortunately, we have found empirically that where the model locates safety stock in the supply chain is fairly insensitive to the demand. The size of the safety stock, though, does depend directly on the demand characterization. We currently are conducting research to understand these observations better, to extend the model to treat non-stationary demand.

Different Review Periods. We assume that each stage operates with a base-stock policy with a common review period. In many supply chains different stages will operate with different reorder frequencies. That is, whereas one stage may place replenishment orders on a daily basis, another stage may do this weekly. In other cases, a stage may operate with a continuous-review policy so that the time between orders varies. We can extend the model to evaluate nested periodic-review base-stock polices in which whenever one stage reorders, all stages downstream also reorder. That is, the review period for an upstream stage is an integer multiple of the review period of its immediate customers. However, we have not yet built the software to implement this extension, as it is a major programming task and it may only be a partial fix to the issue.

Capacity Constraints. In the model we ignore capacity constraints. For certain stages in a supply chain, the consideration of a capacity limit may be necessary in order to get a credible model for determining safety stock requirements. At this time, we do not have a good idea of how to add this to the model.

General Networks. In this paper, we have developed and implemented an optimization algorithm for supply chains that can be modeled as spanning trees. We describe in Graves and Willems (1998) how to extend this algorithm to general networks. However, we have not done a systematic study of this extension beyond some exploratory work. More research is needed to test and refine these ideas as well as uncover better approaches.[3]

Appendix

In this appendix we examine the assumption that each internal stage quotes a guaranteed service time to its customers. To get some insight, we consider a serial system for which we can determine the optimal policy when we relax the assumption of guaranteed service times for internal customers. We then compare the inventory holding costs for the optimal policies with and without this assumption for a small set of test problems.

Consider a serial supply chain with N stages where stage 1 is the demand node and stage i supplies stage $i - 1$ for $i = 2, \ldots, N$. The same assumptions hold as in the original model, except that we do not require guaranteed service times to internal customers. There are no restrictions on the service level that stage i provides to its customer, stage $i - 1$ for $i = 2, \ldots, N$; rather, these internal service levels depend on the base stocks, which are chosen to minimize the inventory holding costs for the entire supply chain. We do assume that stage 1 provides a 100% service level to the external customer, and, without loss of generality, we assume that the service time quoted to the external customer is zero.

For ease of presentation, we assume that $\phi_{i,i-1} = 1$ for $i = 2, \ldots, N$. We let $d(t)$ denote the end-item demand in period t; $d(a, b]$ denote the end-item demand over the time interval $(a, b]$; and $D(\tau)$ denote the maximum possible end-item demand over a time interval of τ periods.

[3]This research has been supported in part by the Eastman Kodak Company; by the MIT Leaders for Manufacturing Program, a partnership between MIT and major U.S. manufacturing firms; and by the MIT Integrated Supply Chain Management consortium. The authors acknowledge and thank Dr. John Ruark who contributed significantly to this research effort; John played a lead role in developing the software application for implementing the results of this research. We also wish to thank the editors and referees for their very helpful and constructive feedback on earlier versions of the paper.

For each stage i, we define $Q_i(t)$ to be the shortfall or backlog at time t, namely the amount that has been ordered by the stage's customer but not yet delivered. We assume at $t = 0$, $I_i(t) = B_i \geq 0$ and $Q_i(t) = 0$ for all stages.

We can show for $i = 1, 2, \ldots, N$ that the on-hand inventory and backlog at time t are:

$$I_i(t) = [B_i - d(t - T_i, t) - Q_{i+1}(t - T_i)]^+,$$

$$Q_i(t) = [d(t - T_i, t) + Q_{i+1}(t - T_i) - B_i]^+, \qquad \text{(A1)}$$

where $[x]^+ = \max(0, x)$, and $Q_{N+1}(t) = 0$ by definition. Equation (A1) requires that each stage has a deterministic lead-time and that each stage follows a base-stock policy in which, for each period, each stage observes end-item demand and places a replenishment order for this amount. The essence of the argument is to observe that the net inventory on hand at a stage equals the stage's base stock minus the inventory on order. For stage i, the inventory on order at time t is the backlog as of time $t - T_i$, plus all of the demand over the interval $(t - T_i, t]$.

From Equation (A1) we can show by induction that for $i = 1,2, \ldots, N$,

$$\begin{aligned} Q_i(t) = \max[0,\ & d(t - T_i, t) - B_i,\ d(t - T_i - T_{i+1}, t) \\ & - B_i - B_{i+1}, \ldots, d(t - T_i - T_{i+1} - \cdots - T_N, t) \\ & - B_i - B_{i+1} - \cdots - B_N]. \end{aligned} \qquad \text{(A2)}$$

In order for the supply chain to provide 100% service to the external customer, we must never have a backlog at stage 1; thus, we must select base stocks so that $Q_1(t) = 0$ for all t. From Equation (A2) we see that $Q_1(t) = 0$ is assured if the base stocks satisfy the following constraints:

$$B_1 + B_2 + \cdots + B_i \geq D(T_1 + T_2 + \cdots + T_i)$$

$$\text{for } i = 1, 2, \ldots, N. \qquad \text{(A3)}$$

Thus, if the base stocks satisfy Equation (A3), there will never be a shortfall at stage 1 and end-item demand will be satisfied with 100% service. As we assume that the demand bounds can be realized, then the constraint set (A3) provides not just sufficient but also necessary conditions for assuring 100% service for end-item demand.

In order to select the base stocks to minimize the inventory holding costs for the supply chain, we must develop an expression for the inventory holding costs; we note from Equation (A1) that the net inventory on hand at stage i is given by:

$$I_i(t) - Q_i(t) = B_i - d(t - T_i, t) - Q_{i+1}(t - T_i). \qquad \text{(A4)}$$

From Equation (A4) we can write the inventory holding costs for the supply chain as:

$$\sum_{i=1}^{N} h_i E[I_i(t)] = \sum_{i=1}^{N} h_i\{B_i - \mu T_i + E[Q_i(t)] - E[Q_{i+1}(t - T_i)]\} \qquad \text{(A5)}$$

where μ is the expected demand rate, and $E[\ ]$ denotes expectation.

We now pose an optimization problem to select the base stocks; namely, we minimize Equation (A5) subject to Equation (A3) and nonnegativity constraints on the base stocks. After dropping constant terms in Equation (A5) and noting that $Q_1(t) = 0$ for any feasible solution, we write the optimization as

$$\min \sum_{i=1}^{N} h_i B_i - \sum_{i=2}^{N} e_{i-1} E[Q_i]$$

P* s.t.

$$B_1 + B_2 + \cdots + B_i \geq D(T_1 + T_2 + \cdots + T_i)$$

$$\text{for } i = 1, 2, \ldots, N,$$

$$B_i \geq 0 \quad \text{for } i = 1, 2, \ldots, N,$$

where $e_i = h_i - h_{i+1}$ is the echelon holding cost. We note from Equation (A2) that $E[Q_i]$ is a nonlinear function of $B_i, \ldots, B_N$ for $i = 1, 2, \ldots, N$.

Our main result is that there is an optimal solution to **P*** in which all the constraints in Equation (A3) are binding. More formally we state the following:

Result. If the echelon holding costs are nonnegative and if $D(\ )$ is a nondecreasing function, then an optimal solution to **P*** is given by

$$B_1 = D(T_1),$$

$$B_i = D(T_1 + \cdots + T_i) - D(T_1 + \cdots + T_{i-1})$$

$$\text{for } i = 2, \ldots, N. \qquad \text{(A6)}$$

Proof. We note that the solution given by Equation (A6) is nonnegative and satisfies the constraints in Equation (A3) as equalities; thus it is a feasible solution to **P***. To prove that this is also an optimal solution, we will argue that there must be an optimal solution in which the constraints in Equation (A3) are binding.

Suppose we have a solution $B_1^*, \ldots, B_N^*$ such that Equation (A3) holds as a strict inequality for one or more constraints. Suppose the kth constraint is the first constraint that is not binding and that $k < N$; we will treat the case when $k = N$ later. Thus, we assume

$$B_1^* + B_2^* + \cdots + B_i^* = D(T_1 + T_2 + \cdots + T_i)$$

$$\text{for } i = 1, 2, \ldots, k - 1 \text{ and}$$

$$B_1^* + B_2^* + \cdots + B_k^* > D(T_1 + T_2 + \cdots + T_k).$$

We now define a new solution $B_1^{**}, \ldots, B_N^{**}$ in which the first k constraints are satisfied as equalities, and show that its objective value is no worse than that for $B_1^*, \ldots, B_N^*$:

$$B_i^{**} = B_i^* \quad \text{for } i = 1, \ldots, N \text{ and } i \neq k, k + 1,$$

$$B_k^{**} = B_k^* - \Delta$$

$$B_{k+1}^{**} = B_{k+1}^* + \Delta,$$

where

$$\Delta = B_k^* - D\left(\sum_{i=1}^{k} T_i\right) + D\left(\sum_{i=1}^{k-1} T_i\right).$$

We first observe that $\Delta > 0$ due to the supposition that the solution $B_1^*, \ldots, B_N^*$ satisfies the kth constraint in Equation (A3) as a strict inequality. Thus, we have $B_{k+1}^{**} \geq 0$. We also see that $B_k^{**} \geq 0$ since $D(\ )$ is nondecreasing. Hence the new solution $B_1^{**}, \ldots, B_N^{**}$ is nonnegative. By construction, the new solution satisfies the kth constraint as an equality, and there are no changes in the remaining constraints. Thus, the new solution $B_1^{**}, \ldots, B_N^{**}$ is a feasible solution.

To express the objective function for the new solution, we decompose it into two parts. The first part of the objective function is

$$\sum_{i=1}^{N} h_i B_i^{**} = (-h_k + h_{k+1})\Delta + \sum_{i=1}^{N} h_i B_i^{*} = -e_k\Delta + \sum_{i=1}^{N} h_i B_i^{*}. \quad \text{(A7)}$$

For the second part of the objective function, let $E[Q_i]^*$ and $E[Q_i]^{**}$ denote the expected backlog at stage i for the first and second solution. Then we find from Equation (A2) that

$$E[Q_i]^{**} = E[Q_i]^* \quad \text{for } i > k + 1,$$

$$E[Q_i]^* \leq E[Q_i]^{**} \leq E[Q_i]^* + \Delta \quad \text{for } i < k + 1, \text{ and}$$

$$E[Q_{k+1}]^* \geq E[Q_{k+1}]^{**} \geq E[Q_{k+1}]^* - \Delta.$$

Thus, for nonnegative echelon holding costs, we can bound the second part of the objective function as follows:

$$-\sum_{i=2}^{N} e_{i-1} E[Q_i]^{**} \leq -\sum_{i=2}^{N} e_{i-1} E[Q_i]^* + e_k\Delta. \quad \text{(A8)}$$

By combining Equations (A7) and (A8), we see that the objective function for the second solution is no greater than the objective for the first. Thus, we have found a new solution in which the first k constraints in Equation (A3) are binding and whose objective value is no worse than that for the first solution. This argument can be continued in this fashion to construct a solution in which the first $N - 1$ constraints in Equation (A3) are binding and whose objective value is no worse than that for the solution $B_1^*, \ldots, B_N^*$. The argument for the case when $k = N$ is similar in structure but easier; we just have to reduce the base stock for stage N until the Nth constraint is binding, which can be done with no penalty to the objective function.

Hence, there is a feasible solution that satisfies all the constraints in Equation (A3) as equalities and that has an objective value no higher than that for the solution $B_1^*, \ldots, B_N^*$. Furthermore, this new solution must be given by Equation (A6), as it is easy to see that it is the unique binding solution to Equation (A3). Finally we conclude that Equation (A6) must be an optimal solution, as its objective value equals or is less than that for any interior solution $B_1^*, \ldots, B_N^*$. This completes the proof.

We note that the optimal base-stock policy does not depend at all on the holding costs. All we need to know is that the holding costs do not decrease as we move down the supply chain, closer to the customer. We also note that this result generalizes to assembly systems by means of the transformation given by Rosling (1989); namely, we can transform an assembly system into an equivalent serial system, and the result applies.

We use this result to compare the performance of the base stock policies with and without the assumption of guaranteed service times for internal customers. The test problems were all for a 3-stage serial system; the problems differed according to their demand process, their production lead-times, and their holding costs.

For the demand process, we start with a Poisson demand distribution with mean λ and with a specified percentile α to truncate the demand. For each time window of length τ, we set the demand bound $D(\tau)$ as the smallest integer such that the cumulative probability for the Poisson random variable with mean $\lambda\tau$ exceeds α. We then normalize the demand distribution over the truncated range. We consider four possible demand processes: $\lambda = 10$, $\alpha = 0.90$; $\lambda = 10$, $\alpha = 0.98$; $\lambda = 50$, $\alpha = 0.90$; $\lambda = 50$, $\alpha = 0.98$.

We permit three settings for the production lead-times and three settings for the holding costs, as follows:

$$(T_1, T_2, T_3) = (4, 4, 4); \quad (1, 3, 8); \quad (8, 3, 1).$$

$$(h_1, h_2, h_3) = (1, 0.5, 0.2); \quad (1, 0.66, 0.33); \quad (1, 0.8, 0.5).$$

By evaluating all combinations we have a total of 36 test problems. For each test problem we determine the optimal policy for the model with guaranteed internal service times and the optimal policy (given by the result above) for the model without this requirement. For each instance, we evaluate the base stocks, the safety-stock holding cost and the total inventory holding cost. The safety stock holding cost is given by the objective function of **P** for the model with guaranteed internal service times and by Equation (A5) for the model without this requirement. The total inventory holding cost is the sum of the safety-stock holding cost plus the pipeline-stock holding cost. The expected pipeline stock at stage i is μT_i; we assume that the holding cost for the pipeline stock at stage i is $(h_i + h_{i+1})/2$.

For the 36 test problems we find that the safety-stock holding cost for the model with guaranteed internal service times is on average 26% higher than that for the model without this requirement; the range is between 7% and 43%. The size of the gap is insensitive to the choice of demand process. However, the gap becomes larger as the production lead-time at stage 1 increases and as the echelon holding cost at stage 1 increases.

The impact on the total inventory holding cost is less dramatic. The difference in holding costs is 4% on average, with a range from less than 1% to 14%. The gap increases as the holding cost of the pipeline stock decreases, namely as the production lead-time at stage 1 decreases and as the demand rate decreases.

From the limited computational study we see that there can be a significant increase in safety stock due to the assumption of guaranteed internal service times. Relative to the total inventory, this increase does not look as large. Nevertheless, there is a cost in terms of higher inventories from the requirement of guaranteed internal service times. This cost needs to be considered in light of the practical benefits, as discussed in the body of the paper, from imposing this requirement. Based on our observations from industrial projects, this requirement, and the resulting increase in safety stock, has not been an issue as the assumption of guaranteed internal service times is already ingrained in practice.

References

Axsäter S. 1993. Continuous review policies for multi-level inventory systems with stochastic demand. S. C. Graves. A. H. Rinnooy

Kan, P. H. Zipkin, eds. *Handbooks in Oper. Res. and Management Sci. Vol 4., Logistics of Production and Inventory*. North-Holland Publishing Company, Amsterdam, The Netherlands. Chapter 4.

Black, B. E. June 1998. Utilizing the principles and implications of the base stock model to improve supply chain performance. S.M. Thesis, Leaders for Manufacturing Program, MIT, Cambridge, MA.

Coughlin, R. L. June 1998. Optimization and measurement of a world-wide supply chain. S.M. Thesis, Leaders for Manufacturing Program, MIT, Cambridge, MA.

Diks, E. B., A. G. de Kok, A. G. Lagodimos. 1996. Multi-echelon systems: A service measure perspective: *European J. Oper. Res.*, **95** 241–263.

Ettl, M., G. E. Feigin, G. Y. Lin, D. D. Yao. 2000. A supply network model with base-stock control and service requirements. *Oper. Res.* **48**, March-April.

Federgruen, A. 1993. Centralized planning models for multi-echelon inventory systems under uncertainty. S. C. Graves, A. H. Rinnooy Kan, P. H. Zipkin, eds. *Handbooks in Operations Research and Management Science, Vol 4., Logistics of Production and Inventory*, North-Holland Publishing Company, Amsterdam, The Netherlands, Chapter 3.

Felch, J. A. June 1997. Supply chain modeling for inventory analysis. S. M. Thesis, Leaders for Manufacturing Program, MIT, Cambridge, MA.

Gallego, G., P. Zipkin. 1999. Stock positioning and performance estimation in serial production-transportation systems. *Manufacturing & Service Oper. Management*. **1** 77–88.

Glasserman, P., S. Tayur. 1995. Sensitivity analysis for base-stock levels in multiechelon production-inventory systems *Management Sci.* **41** 263–281.

Graves, S. C. 1988. Safety stocks in manufacturing systems. *J. of Manufacturing and Oper. Management* **1** 67–101.

——, D. B. Kletter, W. B. Hetzel. 1998. A dynamic model for requirements planning with application to supply chain optimization. *Oper. Res.* **46** S35–S49.

——, S. P. Willems. 1996. Strategic safety stock placement in supply chains. *Proceedings of the 1996 MSOM Conference*, Hanover, NH.

——, ——. August 1998. Optimizing strategic safety stock placement in supply chains. Working Paper available from http://web.mit.edu/sgraves/www/papers/.

Inderfurth, K. 1991. Safety stock optimization in multi-stage inventory systems. *Interna. J. of Production Econom.* **24** 103–113.

—— 1993. Valuation of leadtime reduction in multi-stage production systems. G. Fandel, T. Gulledge, A. Jones, eds. *Oper. Res. in Production Planning and Inventory Control*, Springer, Berlin, Germany, 1993, 413–427.

—— 1994. Safety stocks in multistage, divergent inventory systems: A survey. *International J. of Production Economics* **35** 321–329.

——, S. Minner. 1998. Safety stocks in multi-stage inventory systems under different service measures. *European J. Oper. Res.* **106** 57–73.

Kimball, G. E. 1988. General principles of inventory control. *J. of Manufacturing and Oper. Management*. **1** 119–130.

Lee, H. L., C. Billington. 1993. Material management in decentralized supply chains. *Oper. Res.*, **41** 835–847.

Luenberger, D. G. 1973. *Introduction to Linear and Nonlinear Programming*. Addison Wesley, Reading, MA, 1973.

Minner, S. 1997. Dynamic programming algorithms for multi-stage safety stock optimization. *OR Spektrum* **19** 261–271.

Rosling, K. 1989. Optimal inventory policies for assembly systems under random demands. *Oper. Res.*, **37** 565–579.

Simpson, K. F. 1958. In-process inventories. *Oper. Res.*, **6** 863–873.

Van Houtum, G. J., K. Inderfurth, W. H. J. Zijm. 1996. Materials coordination in stochastic multi-echelon systems. *European J. of Oper. Res.*, **95** 1–23.

Wala, T. June 1999. Build-to-order/direct ship model for professional digital cameras. S. M. Thesis, Leaders for Manufacturing Program, MIT, Cambridge MA.

The consulting Senior Editor for this manuscript was Paul Zipkin. This manuscript was received on June 1, 1998, and was with the authors 127 days for 3 revisions. The average review cycle time was 53 days.

A DYNAMIC MODEL FOR REQUIREMENTS PLANNING WITH APPLICATION TO SUPPLY CHAIN OPTIMIZATION

STEPHEN C. GRAVES

Massachusetts Institute of Technology, Cambridge, Massachusetts

DAVID B. KLETTER

Booz, Allen & Hamilton Inc., New York, New York

WILLIAM B. HETZEL

Merck & Co., Inc., Whitehouse Station, New Jersey

(Received March 1994; revisions received July 1995, January 1996; accepted June 1996.)

This paper develops a new model for studying requirements planning in multistage production-inventory systems. We first characterize how most industrial planning systems work, and we then develop a mathematical model to capture some of the key dynamics in the planning process. Our approach is to use a model for a single production stage as a building block for modeling a network of stages. We show how to analyze the single-stage model to determine the production smoothness and stability for a production stage and the inventory requirements. We also show how to optimize the tradeoff between production capacity and inventory for a single stage. We then can model the multistage supply chain using the single stage as a building block. We illustrate the multistage model with an industrial application, and we conclude with some thoughts on a research agenda.

Most discrete parts manufacturing firms plan their production with MRP (materials requirements planning) systems, or at least, with logic based on the underlying assumptions of MRP. A typical planning system starts with a multiperiod forecast of demand for each finished good or end item. The planning system then develops a production plan (or master schedule) for each end item to meet the demand forecast. These production plans for the end items, after offsetting for lead times, then act as the requirement forecasts for the components needed to produce the end items. The requirements forecast for each component gets translated into production plans for the component, similar to how the production plan for the end items was created. The planning system continues in this way, developing requirement forecasts and production plans for each level of the bill of materials.

Implicit in this planning process are assumptions about the production and demand process. The production plan is developed assuming that the forecast is accurate and will not change. Within the production process, requirements are generated assuming that there are deterministic production lead times and deterministic yields. Needless to say, these assumptions of a benevolent world do not match reality. Inevitably, the forecast changes, and uncertainties in the production process arise that result in deviations from the plan. To respond to these changes, most planning systems will completely revise their plan after some time period, say a week or a month. Again, the planning process starts with the (new) forecast and repeats the steps necessary to regenerate a plan for each level in the products' bills of materials.

The intent of this paper is to present a model that captures the basic flavor of this planning process, and does so in such a way that it can be used to look at various tradeoffs within the production and planning systems. In particular, we model the forecasts for the planning system as a stochastic process. In this way, we try to represent a dynamic input to the planning system, namely, how forecasts change and evolve over time. The forecast process is a key input for the model. Another key is how the forecasts get converted into production plans or master schedules. We model this process as a linear system, with which we can represent the logic for MRP systems and from which we get significant analytical tractability. Finally, the model is structured so that it can describe multistage production-inventory systems.

We are not aware of very much work that is directly related to the dynamic modeling of requirements planning. Baker (1993) provides a nice survey and critique of the literature relevant to the general topic of requirements planning. However, most of the work deals with specific issues like lot sizing or determination of buffer levels. Karmarkar (1993) discusses tactical issues of lot sizing, order release, and lead times in the context of dynamic planning systems. But neither of these papers reports on work that attempts to model a dynamic forecast process. One exception is Graves et al. (1986), in which we modeled a two-stage production-inventory system with a dynamic forecast

Subject classifications: Inventory/production: multi-item, multi-stage supply chain with uncertain demand and dynamic forecast revisions; application to film production; dynamic requirements planning and supply chain optimization.
Area of review: MANUFACTURING OPERATIONS.

0030-364X/98/4603-S035 $05.00

process. In contrast with the present paper, Graves et al. (1986) focused on issues of how to disaggregate an aggregate plan in the two-stage context. Although this paper does not consider the disaggregation issue, it does provide a more powerful model that is applicable to general multistage systems.

Another exception is Heath and Jackson (1994), who considered the same dynamic forecast process as this paper as part of a simulation model that was used to analyze safety stock levels in a multiproduct production/distribution system.

The model for converting the forecast into a production plan is related to earlier work by the first coauthor, in that it uses linear systems for a production-inventory context. (See Graves 1986, 1988a, 1988b, 1988c, and Fine and Graves 1989.)

Lastly, we note Lee and Billington (1993), who develop a model for supply chain optimization and describe its application at Hewlett Packard. Our work complements their work but differs as we try to model the process of requirements planning.

In the next section we develop the model for a single production-inventory stage. As part of the development, we present our model for the forecast process, and we develop the analyses to generate three performance measures for the stage: production smoothness, production stability, and inventory requirements. In the second section we examine an optimization for the tradeoff between production capacity and inventory for a single stage. Although the development is somewhat involved, the final results are surprisingly simple and, we believe, of interest. This section can be omitted by the reader without loss of continuity. In the third section, we show how the model for the single stage can serve as a building block in modeling a general acyclic network of multiple stages. We report on an application of the model to a supply-chain study in the fourth section. The application demonstrates the value of a system-wide perspective for optimizing the supply chain. In the final section we briefly summarize the paper, and then lay out a research agenda for further work.

1. SINGLE-STAGE MODEL

In this section we present the model for a single production stage that produces one (aggregate) product and serves demand from a finished good inventory. The single-stage model serves as a building block for creating models of multistage, multiitem systems. We first describe the forecast process and state our assumptions about how the forecast evolves over time. We then give a model for determining the schedule for production outputs from the production stage, and we show how to manipulate this model to obtain three measures of interest: (1) the production variance as a measure of production smoothing, (2) the inventory variance as a measure of safety stock, and (3) the stability of the production schedule as a measure for the forecast process passed on to any upstream stages.

1.1. Forecast Process

We assume that there is a forecast horizon H such that in each time period t we have forecasts for the requirements for the next H periods. Let $f_t(t + i)$ be the forecast made at time t for the requirements in period $t + i$, $i = 1, 2, \ldots H$. We denote the demand observed in period t by $f_t(t)$, the forecast made in period t for requirements in period t. Beyond the forecast horizon, there is no specific information about requirements. In effect, for $i > H$ we assume that $f_t(t + i) = \mu$, where μ equals the long-run average demand rate.

We propose a stochastic model of this forecast process and show that the forecasts are unbiased, the forecasts improve as they are revised, and the forecast error over the forecast horizon matches the inherent variability in the demand process.

We assume that, each period, we generate a new set of forecasts $f_t(t + i)$ that incorporates new information about future demand. We define the updates of the forecasts from period to period by the *forecast revision,* $\Delta f_t(t + i)$:

$$\Delta f_t(t + i) = f_t(t + i) - f_{t-1}(t + i) \quad \text{for } i = 0, 1, \ldots H, \tag{1}$$

where $f_{t-1}(t + H) = \mu$ by assumption.

Let $\underline{\Delta f}_t$ be the vector for the revisions to the forecast process, where $\Delta f_t(t + i)$ is the $i + 1$st element, $i = 0, 1, 2, \ldots H$. We assume that $\underline{\Delta f}_t$ is an i.i.d. random vector with $E[\underline{\Delta f}_t] = 0$ and $\text{Var}[\underline{\Delta f}_t] = \Sigma$, the covariance matrix. Thus, for a fixed index i, $\Delta f_t(t + i)$ is an i.i.d. random variable over time t with zero mean, and the forecast process is a martingale. We note that if we can observe the forecast process, then we can assess whether or not the forecasts are unbiased (i.e., $E[\underline{\Delta f}_t] = 0$) with independent revisions and we can estimate the covariance matrix Σ.

This model of the forecast process is the same as that of Graves et al. (1986) and Heath and Jackson (1994). We have validated this model as part of field studies at AT&T and at Kodak. And this forecast model is descriptive of the forecast process at nearly all of the discrete-part manufacturing contexts we have encountered.

The *i-period forecast error* is the difference between the actual demand in period t and the forecast of this demand made i periods earlier:

$$f_t(t) - f_{t-i}(t) = \Delta f_t(t) + \Delta f_{t-1}(t) + \cdots + \Delta f_{t-i+1}(t).$$

We can now demonstrate the following properties for this model of the forecast process:

1. The i-period forecast, $f_{t-i}(t)$, is an unbiased estimate of demand in period t.
2. The variance of the i-period forecast error is no greater than the variance of the $(i + 1)$-period forecast error, for $i = 1, 2, \ldots H$.
3. The trace of the covariance matrix Σ equals the variance of the demand process.

We see that the first property must be true by observing that the expectation of the i-period forecast error is zero, since $E[\Delta f_{t-s}(t)] = 0$, for $s = 0, \ldots i - 1$.

We now prove the second property. Since $\Delta f_{t-s}(t)$ for $s = 0, 1, \ldots i - 1$ are independent random variables, the variance of the i-period forecast error is given by:

$$\begin{aligned}\text{Var}\,[f_t(t) - f_{t-i}(t)] &= \text{Var}\,(\Delta f_t(t)) + \text{Var}\,(\Delta f_{t-1}(t)) \\ &\quad + \cdots + \text{Var}\,(\Delta f_{t-i+1}(t)) \\ &= \sigma_0^2 + \sigma_1^2 + \cdots + \sigma_i^2,\end{aligned}$$

where $\sigma_j^2 = \text{Var}(\Delta f_{t-j}(t))$ is the $j + 1$st element on the diagonal of the covariance matrix Σ, for $j = 0, 1, \ldots H$. Thus, since $\sigma_j^2 \geq 0$ for $j = 0, 1, \ldots H$, each forecast revision improves the forecast, in that it reduces the variance of the forecast error.

For the third property, we observe from the above expression that the variance of the $(H + 1)$-period forecast error equals $\sigma_0^2 + \sigma_1^2 + \cdots + \sigma_H^2$, i.e., the trace of Σ. Since by assumption $f_{t-H-1}(t) = \mu$, we have,

$$\text{Var}\,[f_t(t) - f_{t-H-1}(t)] = \text{Var}\,[f_t(t)],$$

which proves the third property.

Since the demand variance is an exogenous parameter, this imposes a constraint on the forecast process: namely, the variance of the forecast error over the forecast horizon must equal the demand variance.

1.2. Schedule for Production Outputs

Given the forecast vector for period t, we need to convert it into a schedule or plan for production. This is often termed the master schedule. We focus on production outputs from the production stage. Later we will discuss how to translate a plan for production outputs into production starts. Production starts will be of interest, since they serve as the requirements forecast for the next upstream production stage.

Let $F_t(t + i)$ equal the *planned production outputs* for period $t + i$ as of period t, where $F_t(t)$ is the actual production completed in period t. We assume that the production plan extends out only for the next H periods, and that beyond this horizon the plan is just to produce the average demand, that is, $F_t(t + i) = \mu$ for $i > H$.

Each period, after we obtain the new forecast, we update or revise the plan for production outputs. We define $\Delta F_t(t + i)$ as the *plan revision*:

$$\Delta F_t(t + i) = F_t(t + i) - F_{t-1}(t + i).$$

From this definition and the fact that $F_t(t + i) = \mu$ for $i > H$, we see that:

$$F_t(t + i) = \mu + \Delta F_{t+i-H}(t + i) + \cdots + \Delta F_t(t + i) \quad \text{for } i = 0, 1, \ldots H. \quad (2)$$

Thus to model the production plan, we need to model the plan revision $\Delta F_t(t + i)$. To do this, we first define the inventory process. For I_t being the inventory at time t, the inventory balance equation is:

$$I_t = I_{t-1} + F_t(t) - f_t(t). \quad (3)$$

The *planned inventory* at time $t + i$ is the expected level of inventory in a future period given the current forecast and the current production plan as of time t:

$$\begin{aligned}I_t(t + i) &= I_t + F_t(t + 1) + \cdots + F_t(t + i) \\ &\quad - f_t(t + 1) - \cdots - f_t(t + i). \quad (4)\end{aligned}$$

We assume that for each time t, we set the production plan $F_t(t + i)$, $i = 0, 1, \ldots H$, so that the planned inventory at the end of the planning horizon, $I_t(t + H)$, is a given constant. That is, we will set the production plan and maintain it from period to period so that the end-of-horizon inventory neither grows nor decreases, but remains constant. We term the level to which the inventory is targeted as the *safety stock*. In a later section we will discuss how to set this level. For now, all we need to know is that this level remains constant.

From (3) and (4), we obtain by equating $I_{t-1}(t - 1 + H)$ and $I_t(t + H)$ that:

$$\begin{aligned}&\Delta F_t(t) + \Delta F_t(t + 1) + \cdots + \Delta F_t(t + H) \\ &\quad = \Delta f_t(t) + \Delta f_t(t + 1) + \cdots + \Delta f_t(t + H). \quad (5)\end{aligned}$$

That is, to assure that the end-of-horizon inventory remains constant, we require that the cumulative revision to the production plan should equal the cumulative forecast revision in each period.

Each period we revise the production schedule to ensure (5). To do this, we model the schedule update as a linear system:

$$\Delta F_t(t + i) = \sum_{j=0}^{H} w_{ij} \Delta f_t(t + j) \quad \text{for } i = 0, 1, \ldots H, \quad (6)$$

where w_{ij} denotes how the forecast revision affects the schedule. In particular, w_{ij} is the proportion of the forecast revision for period $t + j$ that is added to the schedule of production outputs for period $t + i$.

We expect that $0 \leq w_{ij} \leq 1$. To ensure that (5) is true, we require that for each j:

$$\sum_{i=0}^{H} w_{ij} = 1.$$

We refer to w_{ij} as a *weight* or proportion. We can interpret these weights either as decision variables in a prescriptive model or as parameters in a descriptive model. On the one hand, we can view these weights as control or smoothing parameters and use the model for prescription. To smooth production we set the weights w_{ij} for a fixed j to be as nearly constant as possible (e.g., $w_{ij} = 1/(H + 1)$ for $i = 0, 1, \ldots H$). To minimize inventory, we set the weights so that the production plan tracks the forecast as closely as possible (e.g., for fixed j, $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ otherwise). In this way, specification of the weights permits one to balance the tradeoff between production smoothing and inventory requirements, as will be seen.

On the other hand, we can view the weights as parameters for a descriptive model of an existing planning system.

In particular, we can use (6) to model how most implementations of MRP systems translate forecast revisions into schedule revisions. For instance, in the simplest case at time t the schedule is frozen for periods $t + j$, $j = 0, 1, 2 \ldots k$ for some value of $k < H$, and is totally free to change for periods $t + j$, $j = k + 1, \ldots H$. Then, any revision to the forecast within the frozen zone results in a schedule revision for the first period beyond the frozen zone; i.e., for $0 \leq j \leq k$, $w_{ij} = 0$ for $i \neq k + 1$ and $w_{ij} = 1$ for $i = k + 1$. Any revision to the forecast beyond the frozen zone results in a one-for-one schedule revision in the same period: for $k + 1 \leq j \leq H$, $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ otherwise. Occasionally there is an intermediate zone (between the frozen and free zones) in which changes to the schedule are permitted but are restricted in size, e.g., no more than 10% increase or decrease in the scheduled amount. The model given by (6) cannot exactly capture this policy, but it can approximate its behavior by using fractional weights.

In matrix notation, we can rewrite (6) as:

$$\underline{\Delta F}_t = \mathbf{W}\, \underline{\Delta f}_t, \tag{7}$$

where $\mathbf{W} = \{w_{ij}\}$ is an $(H + 1) \times (H + 1)$ matrix, and $\underline{\Delta F}_t$ and $\underline{\Delta f}_t$ are column vectors with elements $\Delta F_t(t + i)$ and $\Delta f_t(t + i)$, for $i = 0, 1, \ldots H$. From this, we observe that $\underline{\Delta F}_t$ is an independent random vector, has zero mean, and has a covariance matrix $\mathbf{W}\Sigma\mathbf{W}'$. (We will see later that this is an important observation for the extension to multiple stages: we will derive the forecast revision for upstream stages from $\underline{\Delta F}_t$.)

We can express the production plan in matrix notation by:

$$\underline{F}_t = \mathbf{B}\, \underline{F}_{t-1} + \mu \underline{U}_{H+1} + \underline{\Delta F}_t, \tag{8}$$

where $F_t(t + i)$ is the $i + 1$st element of the vector $\underline{F}_t$ for $i = 0, 1, \ldots H$; $\underline{U}_{H+1}$ is a unit vector with $u_i = 0$ for $i = 1, \ldots H$ and $u_{H+1} = 1$; and $\mathbf{B}$ is a matrix with elements $b_{ij} = 1$ for $j = i + 1$, and $b_{ij} = 0$ else. Premultiplying a column vector by $\mathbf{B}$ replaces the ith element in the vector with the $i + 1$st element and replaces the last element with a zero.

From (7) and (8) and repeated substitution, we obtain:

$$\begin{aligned} \underline{F}_t &= \mathbf{B}\underline{F}_{t-1} + \mu \underline{U}_{H+1} + \mathbf{W}\underline{\Delta f}_t \\ &= \mathbf{B}^{H+1}\underline{F}_{t-H-1} + \underline{\mu} + \sum_{i=0}^{H} \mathbf{B}^i \mathbf{W} \underline{\Delta f}_{t-i}, \end{aligned} \tag{9}$$

where $\underline{\mu}$ is the vector with each element equal to μ, and the superscript i in $\mathbf{B}^i$ denotes the ith power of $\mathbf{B}$. We can simplify (9) by noting that premultiplying an $(H + 1) \times 1$ vector by $\mathbf{B}^{H+1}$ gives the null vector:

$$\underline{F}_t = \underline{\mu} + \sum_{i=0}^{H} \mathbf{B}^i \mathbf{W} \underline{\Delta f}_{t-i}. \tag{10}$$

1.3. Measures of Interest

There are three categories of measures for the single-stage model: the *smoothness* of the production outputs, the *safety stock* for the end-item inventory, and the *stability* of the production plan.

The *smoothness* of the production outputs is of interest because more variable (less smooth) production is expected to require more production resources or capacity. Furthermore, we can influence the smoothness of production via our inventory and control policies.

An output of the model is the variability of the inventory process, which will dictate how much *safety stock* is needed to ensure an acceptable service level. If the inventory process is more variable, more safety stock will be needed.

The *stability* of the production output plan is of interest, since the output plan determines the plan for production starts, which determines the requirements forecast for upstream stages. We will, in effect, equate the stability of the production plan to the accuracy of the forecast process for the upstream stages. More stability means a more accurate forecast process upstream. This measure is critical as we try to understand the workings of a multistage system, since the inventory requirements and the variability of the production outputs for a stage will depend heavily on the accuracy of the forecast process.

We first develop the measures for production smoothing and for the stability of the production plan. We will need a more extensive development to obtain the variability of the inventory process in order to set the safety stock.

Production Smoothing. A common measure of production smoothing is the variance of the production output, $\mathrm{Var}[F_t(t)]$. From (10) we see immediately that the random vector $\underline{F}_t$ has mean $\underline{\mu}$ and has a covariance matrix given by:

$$\mathrm{Var}\,(\underline{F}_t) = \sum_{i=0}^{H} \mathbf{B}^i \mathbf{W}\Sigma\mathbf{W}'\mathbf{B}'^i. \tag{11}$$

We can use the covariance matrix to obtain the first measure of the production smoothing, $\mathrm{Var}[F_t(t)]$. Indeed, one can show that:

$$\mathrm{Var}\,[F_t(t)] = \mathrm{tr}(\mathbf{W}\Sigma\mathbf{W}'), \tag{12}$$

where $\mathrm{tr}(\mathbf{A})$ is the trace of matrix $\mathbf{A}$.

A second measure of production smoothing is given by $F_t(t) - F_{t-1}(t-1)$, the change in production outputs from one period to the next. In matrix notation we see from (9) that:

$$\underline{F}_t - \underline{F}_{t-1} = \mu \underline{U}_{H+1} + \mathbf{W}\underline{\Delta f}_t - [\mathbf{I} - \mathbf{B}]\underline{F}_{t-1},$$

where $\mathbf{I}$ is the identity matrix. Since $\underline{\Delta f}_t$ and $\underline{F}_{t-1}$ are independent of each other, we find that the covariance matrix for $\underline{F}_t - \underline{F}_{t-1}$ is given by:

$$\begin{aligned} \mathrm{Var}\,(\underline{F}_t - \underline{F}_{t-1}) &= \mathbf{W}\Sigma\mathbf{W}' \\ &\quad + \sum_{i=0}^{H} [\mathbf{I} - \mathbf{B}]\mathbf{B}^i \mathbf{W}\Sigma\mathbf{W}'\mathbf{B}'^i[\mathbf{I} - \mathbf{B}]' \\ &= (\mathbf{I} - \mathbf{B})\ \mathrm{Var}\,(\underline{F}_t) + \mathrm{Var}\,(\underline{F}_t)(\mathbf{I} - \mathbf{B})'. \end{aligned} \tag{13}$$

From this covariance matrix, we can determine the second measure of production smoothing, namely Var $[F_t(t) - F_{t-1}(t-1)]$.

Production Stability. For the stability of the production plan, we use $\underline{\Delta F_t}$: the random vector for the one-period revision to the production plan, which is the basis for the revision to the forecast of requirements for upstream stages. (The production starts, as described earlier, would generate the actual forecast seen by the upstream stages; but since the starts are usually just the production plan offset by the lead time, we can use the revision to the production plan for defining stability.) From (7) we obtain its expectation and covariance matrix:

$$\begin{aligned} E(\underline{\Delta F_t}) &= \underline{0} \\ \text{Var } (\underline{\Delta F_t}) &= \mathbf{W}\Sigma\mathbf{W}'. \end{aligned} \tag{14}$$

We propose the covariance matrix $\mathbf{W}\Sigma\mathbf{W}'$ as a measure of the stability of the production plan. A more stable production plan will have a smaller covariance matrix, and will yield more accurate forecasts for the upstream stages. When analyzing the upstream stages, the dynamics of the requirements depend upon this covariance matrix. In this sense, for the upstream stages, the covariance matrix in (14) is analogous to Σ for the downstream stage, namely it is the covariance matrix for the relevant requirements forecast process.

One measure of the size of a covariance matrix is its trace. We note that with this interpretation the tr$(\mathbf{W}\Sigma\mathbf{W}')$ signifies not only the stability of the production plan, but also the variance of the requirements forecast for the upstream stages over the planning horizon. Furthermore, we see that, according to the proposed measures (12) and (14), smoothing production is essentially equivalent to stabilizing the production plan and requirements forecast for the upstream stages.

Inventory. We focus on the end-item inventory for the single stage, namely the random variable I_t given in (3). We assume that the requirements for the single stage are to be met from the end-item inventory and that typical service expectations apply, e.g., the inventory should stock out in no more than 2% of the periods, or that the inventory should provide a 97% fill rate. We will find the expectation $E(I_t)$ and variance Var(I_t), from which we can determine the safety stock required to achieve a desired service level, under suitable distributional assumptions. For instance, if the forecast errors are normally distributed, then we will see that I_t has a normal distribution. For a desired service level expressed as the stockout probability, we need to set the safety stock level so that:

$$E(I_t) > k\sigma(I_t), \tag{15}$$

where k is such to ensure the service level, and $\sigma(\,)$ denotes the standard deviation.

Recall that in (4) we defined $I_t(t+i)$ to be the planned inventory level in period $t+i$ as of time t; that is, $I_t(t+i)$ is the expected inventory in period $t+i$, where the expectation is as of period t. For notational convenience, $I_t(t)$ denotes the actual inventory in period t, i.e., $I_t(t)$ is the same as I_t. As stated in the earlier development of (5), we assume that the end-of-horizon inventory $I_t(t+H)$ is targeted to equal some constant, which we call the safety stock and denote by ss.

The inventory flow equation for the planned inventory is:

$$\begin{aligned} I_t(t+i) = I_t(t) + F_t(t+1) + \cdots + F_t(t+i) \\ - f_t(t+1) - \cdots - f_t(t+i). \end{aligned} \tag{16}$$

Define $\Delta I_t(t+i) = I_t(t+i) - I_{t-1}(t+i)$. From (3) and (16), we find that

$$\begin{aligned} \Delta I_t(t+i) = \Delta F_t(t) + \cdots + \Delta F_t(t+i) \\ - \Delta f_t(t) - \cdots - \Delta f_t(t+i), \end{aligned}$$

for $i = 0, 1, \ldots H-1$. By assumption, since we keep the inventory constant at ss beyond the horizon, we have:

$$\Delta I_t(t+H) = I_t(t+H) - I_{t-1}(t+H) = ss - ss = 0.$$

In matrix notation, let $\underline{I_t}$ be an $(H+1) \times 1$ column vector with $I_t(t+i)$ as its $i+1$st element. Then

$$\underline{\Delta I_t} = \mathbf{T}[\underline{\Delta F_t} - \underline{\Delta f_t}],$$

where $\mathbf{T}$ is a matrix with element $t_{ij} = 1$ for $i \geq j$ and $t_{ij} = 0$ else. We can now write the inventory random vector as

$$\underline{I_t} = \mathbf{T}[\underline{\Delta F_t} - \underline{\Delta f_t}] + \mathbf{B}\underline{I_{t-1}} + ss\underline{U}_{H+1}, \tag{17}$$

where we use the fact that $I_{t-1}(t+H) = ss$. We can simplify (17) by repeated substitution, by substitution of (7), and by noting that premultiplication of an $(H+1) \times 1$ vector by $\mathbf{B}^{H+1}$ gives the null vector:

$$\underline{I_t} = \sum_{i=0}^{H} \mathbf{B}^i\mathbf{T}[\mathbf{W} - \mathbf{I}]\underline{\Delta f_{t-i}} + \underline{ss}, \tag{18}$$

where $\underline{ss}$ denotes the column vector with each element equal to ss. From (18) we see that the random vector $\underline{I_t}$ has mean equal to $\underline{ss}$, and has a covariance matrix given by:

$$\text{Var } (\underline{I_t}) = \sum_{i=0}^{H} \mathbf{B}^i\mathbf{T}[\mathbf{W} - \mathbf{I}]\Sigma[\mathbf{W} - \mathbf{I}]'\mathbf{T}'\mathbf{B}'^i. \tag{19}$$

We can use (19) to find Var$[I_t(t)]$, which is necessary to determine how to set the safety stock level ss. From (19), we can show with some effort that

$$\text{Var } [I_t(t)] = \text{tr}(\mathbf{T}[\mathbf{W} - \mathbf{I}]\Sigma[\mathbf{W} - \mathbf{I}]'\mathbf{T}') = \sum_{k=0}^{H}\sum_{i=0}^{k}\sum_{j=0}^{k} q_{ij}, \tag{20}$$

where

$$\mathbf{Q} = \{q_{ij}\} = [\mathbf{W} - \mathbf{I}]\Sigma[\mathbf{W} - \mathbf{I}]'.$$

Now from (15) we set the safety stock by $ss = k\ \sigma[I_t(t)]$, where $\sigma[I_t(t)]$ is obtained from (20) and k is such to provide the desired service level from the inventory.

2. OPTIMAL WEIGHTS FOR SINGLE-STAGE MODEL

For a single stage it is natural to wonder how to choose the weights in (6) that determine how a forecast revision is converted into a revision of the production plan. To gain some insight into this question, we pose and solve an optimization problem for choosing the weights for the simple case of uncorrelated demand. The tradeoff between production smoothing in the stage and the end-item inventory requirements should govern the choice of weights. This tradeoff is the basis for stating the optimization problem:

$$\text{Min } \sigma[F_t(t)],$$

subject to:

$$\sigma[I_t(t)] \leq K,$$

$$\sum_{i=0}^{H} w_{ij} = 1 \quad \forall j. \tag{21}$$

The optimization problem minimizes production smoothing, as given by the standard deviation of the production output, subject to a constraint on the standard deviation of the inventory and the requirement that the weights sum to one. We interpret the objective as minimizing required production capacity. We view the nominal capacity required at the stage as being the expected production requirements, plus some number of standard deviations. (See Graves 1988a for further discussion.) The constraint on the standard deviation of the inventory is effectively a constraint on the amount of safety stock required, where we assume that the safety stock is a multiple of $\sigma[I_t(t)]$. An alternative formulation would be to minimize the standard deviation of the inventory, equivalently minimize the safety stock, subject to a constraint on the standard deviation of the production output.

There are no restrictions in the optimization on the weights, other than the convexity constraint. We have not imposed any nonnegativity constraints, nor any restrictions on the weights due to a fixed production lead time. Rather, we allow the weights to be totally free. In this sense, the optimization will produce a lower bound for the case with fixed lead times.

To develop some insights on the optimal weights, we transform the original optimization problem (21) into an equivalent form by restating it in terms of the variances of the production and inventory variables:

$$\text{Min Var } [F_t(t)], \tag{21a}$$

subject to:

$$\text{Var } [I_t(t)] \leq K^2,$$

$$\sum_{i=0}^{H} w_{ij} = 1 \quad \forall j.$$

To analyze this equivalent problem, we consider the Lagrangian relaxation:

$$L(\lambda) = \text{Min Var } [F_t(t)] + \lambda \text{ Var } [I_t(t)] - \lambda K^2, \tag{21b}$$

subject to:

$$\sum_{i=0}^{H} w_{ij} = 1 \quad \forall j.$$

By solving this problem over a range of positive values for the Lagrange multiplier λ, we can find the tradeoff surface between production smoothing and inventory requirements for a single stage. We will also obtain some intuition for the form of the optimal weighting function.

In the remainder of this section we will focus on solving (21b). To solve (21a), and equivalently (21), we would need to search over λ until the solution to (21b) satisfies the relaxed constraint.

We only consider the case when the covariance matrix for the forecast revision process is diagonal. That is, the forecast revisions are uncorrelated, and $\text{Var}[\underline{\Delta f_t}] = \Sigma = \{\sigma_i^2\}$, where $\sigma_i^2 = \text{Var}[\Delta f_t(t+i)]$ is the $i+1$st element on the diagonal, $i = 0, \ldots H$.

For this case, we can simplify (12) and (20) to be:

$$\text{Var } [F_t(t)] = \sum_{i=0}^{H} \sum_{j=0}^{H} (w_{ij}\sigma_j)^2, \tag{12*}$$

and

$$\text{Var } [I_t(t)] = \sum_{i=0}^{H} \sum_{j=0}^{H} (b_{ij}\sigma_j)^2, \tag{20*}$$

where

$$\begin{aligned} b_{ij} &= w_{1j} + \cdots + w_{ij} && \text{for } i < j, \\ &= w_{1j} + \cdots + w_{ij} - 1 && \text{for } i \geq j. \end{aligned} \tag{22}$$

By substituting (12*) and (20*) into (21b), we observe that the minimization problem separates into $H+1$ subproblems, one for each period j:

$$L(\lambda) = \sum_{j=0}^{H} L_j(\lambda) - \lambda K^2, \tag{21c}$$

where

$$L_j(\lambda) = \text{Min} \sum_{i=0}^{H} (w_{ij}\sigma_j)^2 + \lambda \sum_{i=0}^{H} (b_{ij}\sigma_j)^2, \tag{23}$$

subject to:

$$\sum_{i=0}^{H} w_{ij} = 1.$$

We now characterize the solution to $L_j(\lambda)$ with a series of propositions.

Proposition 1. *The optimal weights in* (23) *are independent of* σ_j^2.

Proof. Each term in the objective function of $L_j(\lambda)$ in (23) is proportional to σ_j^2, which can then be factored out. □

Thus, we can determine the optimal weights in the Lagrangian (21b) without knowing the covariance matrix for the forecast revision. We only need to know that the covariance matrix is diagonal. However, to solve the original

problem, (21) or (21a), does require knowledge of the covariances to ensure satisfaction of the inventory constraint.

The Kuhn-Tucker conditions for (23) consist of the convexity constraint over the weights, plus the following set of equations:

$$w_{ij} + \lambda \sum_{k=i}^{H} (w_{0j} + \cdots + w_{kj} - u_{kj}) = \gamma \text{ for } i = 0, \ldots H, \tag{24}$$

where $u_{kj} = 1$ if $k \geq j$, $u_{kj} = 0$ if $k < j$, and γ is the (scaled) dual variable for the single convexity constraint in (23). Since (23) is a convex program, the Kuhn-Tucker conditions are both sufficient and necessary, and they identify a unique solution.

To find the solution, we equate (24) for $i - 1$ and i to obtain:

$$w_{ij} = w_{i-1,j} + \lambda(w_{0j} + \cdots + w_{i-1,j} - u_{i-1,j}) \quad \text{for } i = 1, \ldots H. \tag{25}$$

We can construct a solution to (24) by selecting a value for w_{0j} and repeatedly applying (25). To satisfy the convexity constraint, we could search over values for w_{0j}. Alternatively, we describe in the next two propositions how to find w_{0j} analytically.

Proposition 2. *For a given value of* λ, *the solution to* (25) *for* w_{ij} *is a linear function of* w_{0j} *given by*:

$$w_{ij} = P_i(\lambda)w_{0j} \qquad \text{for } i = 0, 2 \ldots j, \tag{26a}$$

$$w_{ij} = P_i(\lambda)w_{0j} - R_{i-j}(\lambda) \quad \text{for } i = j + 1, \ldots H, \tag{26b}$$

where $P_i(\lambda)$ *is a polynomial in* λ *of degree* i, *and* $R_{i-j}(\lambda)$ *is a polynomial in* λ *of degree* $i - j$. *In particular, we can show by induction that for* $n = 0, 1, \ldots H$,

$$P_n(\lambda) = \sum_{i=0}^{n} \frac{(n+i)!}{(2i)!(n-i)!} \lambda^i,$$

and that for $n = 1, 2, \ldots H - j$,

$$R_n(\lambda) = \sum_{i=1}^{n} \frac{(n+i-1)!}{(2i-1)!(n-i)!} \lambda^i.$$

Proposition 3. *The optimal choice for* w_{0j} *that solves* (23) *is given by*:

$$w_{0j} = \frac{1 + \sum_{i=j+1}^{H} R_{i-j}(\lambda)}{\sum_{i=0}^{H} P_i(\lambda)} = \frac{P_{H-j}(\lambda)}{\sum_{i=0}^{H} P_i(\lambda)}, \tag{27}$$

which simplifies to:

$$w_{0j} = \frac{\sum_{i=0}^{H-j} \frac{(H-j+i)!}{(2i)!(H-j-i)!} \lambda^i}{\sum_{i=0}^{H} \frac{(H+i-1)!}{(2i+1)!(H-i)!} \lambda^i}.$$

Proof. From Proposition 2 we can rewrite the convexity constraint as follows:

$$1 = \sum_{i=0}^{H} w_{ij} = \sum_{i=0}^{H} P_i(\lambda)w_{0j} - \sum_{i=j+1}^{H} R_{i-j}(\lambda).$$

We can now use this to express w_{0j} in terms of $P_i(\lambda)$ and $R_i(\lambda)$, as given in the proposition. We simplify the expression for w_{0j} by substituting the following for $R_i(\lambda)$:

$$\sum_{i=1}^{n} R_i(\lambda) = \sum_{i=1}^{n} \frac{(n+i)!}{(2i)!(n-i)!} \lambda^i = P_n(\lambda) - 1,$$

which is found by an induction argument. Similarly, we can simplify (27) by noting that

$$\sum_{i=0}^{n} P_i(\lambda) = \sum_{i=0}^{n} \frac{(n+i+1)!}{(2i+1)!(n-i)!} \lambda^i. \quad \square$$

Having found the optimal choice of w_{0j}, we obtain the remaining weights by iteratively solving (25). We see immediately from Proposition 3 that for positive λ, w_{0j} is positive; we can similarly show that w_{Hj} is positive. From these facts, we can obtain the following proposition by examining the first differences for the optimal weights.

Proposition 4. *The optimal weights* w_{ij} *are positive, increasing, and strictly convex over the range* $i = 0, 1, \ldots j$. *The optimal weights* w_{ij} *are positive, decreasing, and strictly convex over the range* $i = j, j + 1, \ldots H$.

Proposition 5. *The matrix of optimal weights is symmetric about the off-diagonal, i.e.,* $w_{ij} = w_{H-j,H-i}$.

Proof. This can be shown by substitution of (27) into (26). $\square$

Proposition 6. *The optimal weights are such that* $w_{ij} = w_{H-i,H-j}$.

Proof. Since the optimal weights satisfy the convexity constraint, we can substitute the convexity constraint into (25) and rewrite, after some rearrangement, as:

$$w_{i-1,j} = w_{ij} + \lambda(w_{ij} + \cdots + w_{Hj} - (1 - u_{i-1,j})) \quad \text{for } i = 1, \ldots H. \tag{28}$$

From (28), by a similar development as used to find (26), we can express the weights as linear functions of w_{Hj}:

$$w_{ij} = P_{H-i}(\lambda)w_{Hj} \qquad \text{for } i = j, \ldots H, \tag{29a}$$

$$w_{ij} = P_{H-i}(\lambda)w_{Hj} - R_{j-i}(\lambda) \quad \text{for } i = 0, 1, \ldots j - 1. \tag{29b}$$

In order for the weights to sum to one, we then find that:

$$w_{Hj} = \frac{P_j(\lambda)}{\sum_{i=0}^{H} P_i(\lambda)}. \tag{30}$$

From (29) and (30), we establish the result. $\square$

Proposition 7. *The matrix of optimal weights is symmetric about the diagonal*; *i.e.,* $w_{ij} = w_{ji}$.

Proof. This follows immediately from Propositions 5 and 6. $\square$

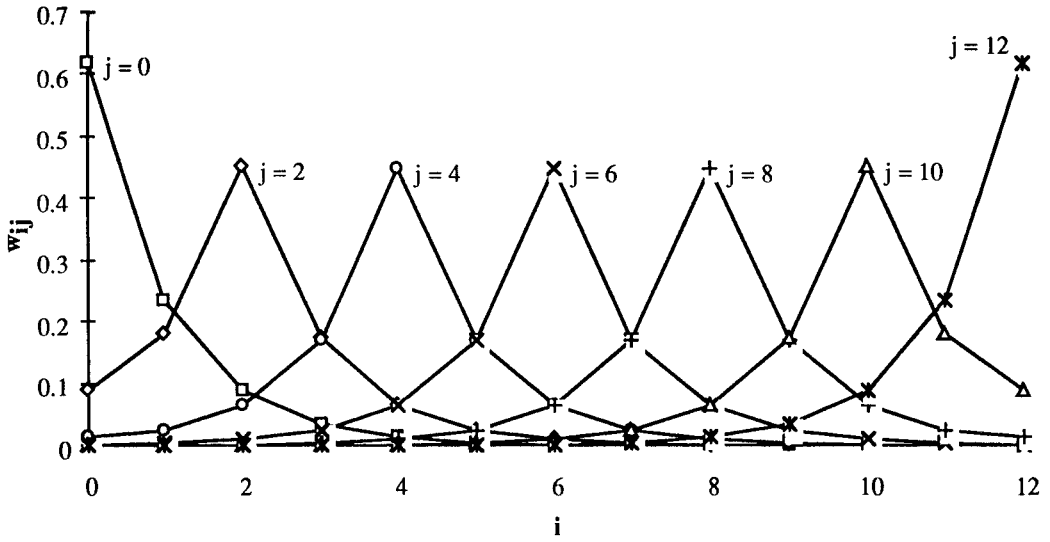


Figure 1. Optimal weights for $\lambda = 1$ and various j.

Figure 1 shows the form of the optimal weights for various values of j for $\lambda = 1$ and $H = 12$. Table I lists the actual values for the optimal weights. From the table we observe that the matrix of optimal weights is symmetric about both diagonals, as stated in the propositions above. Furthermore, for a fixed index j, the weights increase geometrically to a maximum at w_{jj}, and then decay geometrically over the rest of the column.

Figures 2 and 3 show the form of the optimal weights for $\lambda = 4$ and $\lambda = 0.25$ at $H = 12$. Intuitively, we would expect that as λ increases to ∞, w_{jj} goes to 1 and w_{ij} goes to 0 for $i \neq j$ (no production smoothing), and as λ decreases to 0, w_{ij} goes to $1/(H + 1)$ (maximum production smoothing). At $\lambda = 4$ and $\lambda = 0.25$ we already begin to observe this behavior.

Proposition 8. *The optimal objective value for the Lagrangian function in* (23) *is given by* $L_j(\lambda) = w_{jj}\sigma_j^2$ *for* $j = 0, 1, \ldots H$.

Our proof of Proposition 8 involves quite a bit of unattractive and nonintuitive algebra. (See Kletter 1994 for the details.) The basic structure of the proof is as follows: we rewrite the right-hand side of (23) strictly in terms of w_{0j} and λ for a given j by repeatedly applying (26) and factoring out σ_j^2. We then show that this expression equals w_{jj}, where w_{jj} is also expressed in terms of w_{0j} and λ. This is achieved by replacing w_{0j} with the expression given in (27), expressing all terms as polynomials in λ, and then manipulating the binomial coefficients until they are shown to be equal.

The value of Proposition 8 is that it provides a relatively quick way to evaluate the objective function of the Lagrangians, namely (21b) and (23). Also, we show next how to get a good approximation of w_{jj}, which will then yield an analytic expression for the objective function of the Lagrangian.

Suppose we define the first difference $\Delta w_{ij} = w_{ij} - w_{i-1,j}$; we can use (25) to express Δw_{ij} by:

$$\Delta w_{ij} = \Delta w_{i-1,j} + \lambda w_{i-1,j} \quad \text{for } i = 1, 2, \ldots H \text{ and } i \neq j + 1,$$

$$\Delta w_{j+1,j} = \Delta w_{jj} + \lambda w_{jj} - \lambda.$$

Table I
Optimal Weights for $\lambda = 1$, $H = 12$

	j												
	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.6180	0.2361	0.0902	0.0344	0.0132	0.0050	0.0019	0.0007	0.0003	1.1E-04	4.1E-05	1.6E-05	8.2E-06
1	0.2361	0.4721	0.1803	0.0689	0.0263	0.0101	0.0038	0.0015	0.0006	0.0002	8.2E-05	3.3E-05	1.6E-05
2	0.0902	0.1803	0.4508	0.1722	0.0658	0.0251	0.0096	0.0037	0.0014	0.0005	0.0002	8.2E-05	4.1E-05
3	0.0344	0.0689	0.1722	0.4477	0.1710	0.0653	0.0250	0.0095	0.0036	0.0014	0.0005	0.0002	1.1E-04
4	0.0132	0.0263	0.0658	0.1710	0.4473	0.1709	0.0653	0.0249	0.0095	0.0036	0.0014	0.0006	0.0003
5	0.0050	0.0101	0.0251	0.0653	0.1709	0.4472	0.1708	0.0653	0.0249	0.0095	0.0037	0.0015	0.0007
i 6	0.0019	0.0038	0.0096	0.0250	0.0653	0.1708	0.4472	0.1708	0.0653	0.0250	0.0096	0.0038	0.0019
7	0.0007	0.0015	0.0037	0.0095	0.0249	0.0653	0.1708	0.4472	0.1709	0.0653	0.0251	0.0101	0.0050
8	0.0003	0.0006	0.0014	0.0036	0.0095	0.0249	0.0653	0.1709	0.4473	0.1710	0.0658	0.0263	0.0132
9	1.1E-04	0.0002	0.0005	0.0014	0.0036	0.0095	0.0250	0.0653	0.1710	0.4477	0.1722	0.0689	0.0344
10	4.1E-05	8.2E-05	0.0002	0.0005	0.0014	0.0037	0.0096	0.0251	0.0658	0.1722	0.4508	0.1803	0.0902
11	1.6E-05	3.3E-05	8.2E-05	0.0002	0.0006	0.0015	0.0038	0.0101	0.0263	0.0689	0.1803	0.4721	0.2361
12	8.2E-06	1.6E-05	4.1E-05	1.1E-04	0.0003	0.0007	0.0019	0.0050	0.0132	0.0344	0.0902	0.2361	0.6180

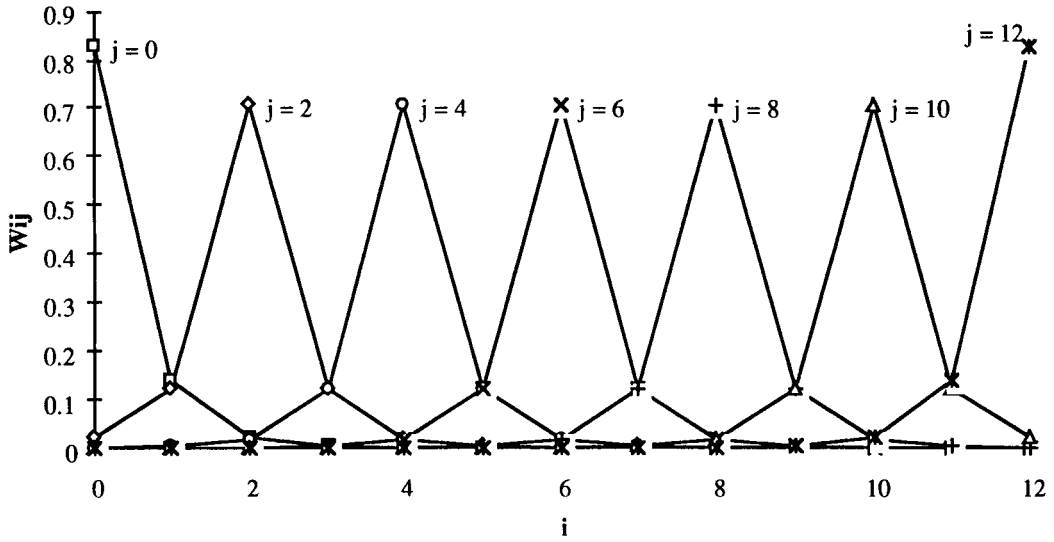


Figure 2. Optimal weights for $\lambda = 4$ and various j.

To get an approximate solution to these first difference equations, suppose we look at a limiting case where we allow both H and j to grow. In effect, we let the range be $i = \ldots -2, -1, 0, 1, 2, \ldots$, except for $i = j + 1$. Then in the limit, the solution to these difference equations is:

$$w_{j+k,j} = w_{j-k,j} = \alpha[(1 - \alpha)/(1 + \alpha)]^k \quad \text{for } k = 0, 1, 2, \ldots, \quad (31)$$

where $\alpha = \sqrt{\lambda/(\lambda + 4)}$.

Furthermore, this solution satisfies the convexity constraint over the weights. From (31) we see that in the limit:

- the optimal weights are symmetric about w_{jj};
- the optimal weights decline geometrically on either side of w_{jj};
- the value of the maximum weight w_{jj} is independent of j; and
- the maximum weight w_{jj} is a simple monotonic function of λ, that approaches 1 as λ increases.

We can see from Figure 1 and Table I that for $\lambda = 1$, the optimal weights already begin to approach the limit at $H = 12$. In particular, we observe that, except at the end points $j = 0$ and $j = H$, $w_{jj} \approx \alpha = \sqrt{\lambda/(\lambda + 4)} = \sqrt{1/5} \approx 0.4472$ and $w_{j+1,j} = w_{j-1,j} = \alpha[(1 - \alpha)/(1 + \alpha)] \approx 0.1708$.

The limit provides a simple approximation to the objective function of the Lagrangian relaxation. Using Proposition 8 and (31), we find that for large values of H we can approximate (21b) by:

$$L(\lambda) = \text{Min Var}\,[F_t(t)] + \lambda\ \text{Var}\,[I_t(t)] - \lambda K^2$$

$$\approx \text{tr}(\Sigma)\sqrt{\lambda/(\lambda + 4)} - \lambda K^2.$$

This simplification is helpful for finding the value of λ that maximizes the Lagrangian, and thus solves the original optimization problem (21).

We end this section with an interesting and perhaps useful result.

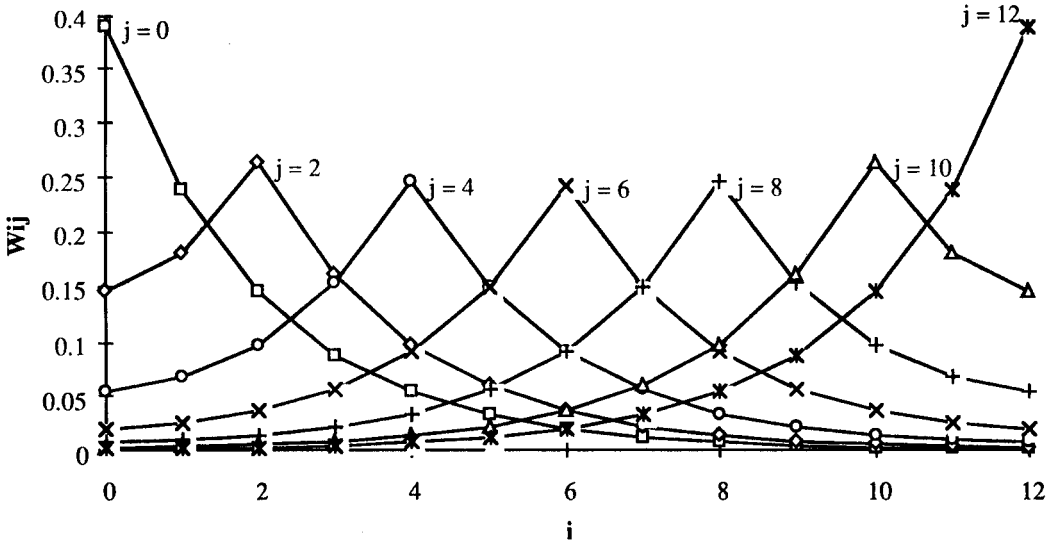


Figure 3. Optimal weights for $\lambda = 0.25$ and various j.

Proposition 9. *The optimal weight matrix is the inverse of a tridiagonal matrix* **C**, *with* $c_{00} = c_{HH} = (\lambda + 1)/\lambda$, $c_{01} = c_{10} = c_{H,H-1} = c_{H-1,H} = -1/\lambda$, *and with* $(c_{i,i-1}, c_{ii}, c_{i,i+1})$ *given by* $(-1/\lambda, (\lambda + 2)/\lambda, -1/\lambda)$ *for* $i = 1, 2, \ldots H - 1$.

Proposition 9 can be proved by construction through a series of careful matrix operations. (See Kletter 1994 for details.) Our proof simply shows that inverting the matrix **C** gives **W**, as specified in (29). This is accomplished by first factoring **C** into **LDL′**, where **L** is bidiagonal, since **C** is symmetric and tridiagonal, and then inverting to obtain $(\mathbf{LDL}')^{-1} = (\mathbf{L}')^{-1}\,\mathbf{D}^{-1}\,\mathbf{L}^{-1}$. Since the diagonal matrix **D** and the bidiagonal matrix **L** are both easily inverted, we then compute the product and simplify to show that $c_{ij}^{-1} = w_{ij}$ for all i and j.

One significance of Proposition 9 is that it makes the computation of the optimal weight matrix even easier.

3. EXTENSION TO MULTISTAGE SYSTEMS

In the previous sections we developed a single-stage model of requirements planning. We now discuss how this single-stage model can serve as a building block in modeling a general acyclic network of multiple stages.

To begin, we state the assumptions and introduce some additional notation that will be necessary for our discussion.

Assumption 1. *The production system is an acyclic network with n distinct stages, m of which produce end-items, where* $m < n$. *We index the stages so that if stage i is downstream from stage j, then* $i < j$. *In addition, the end item stages are numbered* $1, 2, \ldots m$.

Assumption 2. *The forecast processes at the end-item stages are mutually independent.*

Assumption 3. *Each downstream stage is effectively decoupled from the upstream stages, i.e., there is always adequate (raw material) inventory for a stage to make its production starts. This is an approximation that is likely to be reasonable if each stage operates with an inventory policy in which stockouts are rare.*

Assumption 4. *Each stage operates according to the assumptions for the single-stage model.*

Namely, let $\underline{f}_i$, $i = 1, \ldots n$, be the forecast vector for each stage i; for simplicity, we will omit the subscript t in this section. Note that for $i = 1, \ldots m$, $\underline{f}_i$ is an exogenous random vector, whereas for $i = m + 1, \ldots n$, $\underline{f}_i$ will be a derived forecast. Let $\underline{F}_i$, $i = 1, \ldots n$, be the output plan for each stage i. Thus, by Assumption 4, there is a weight matrix $\mathbf{W}_i$, and $\underline{\Delta F}_i = \mathbf{W}_i\,\underline{\Delta f}_i$ for each stage i.

To link the requirements of a downstream stage to an upstream stage, we need to model the production starts or releases into each stage. We assume that in each period each stage i, $i = 1, 2, \ldots n$, must translate its planned production outputs $\underline{F}_i$ into a plan of production starts, call it $\underline{G}_i$, over some planning horizon.

Assumption 5. *At each stage, we model production starts as a linear system of production outputs:* $\underline{G}_i = \mathbf{A}_i\underline{F}_i$ *for some matrix* $\mathbf{A}_i$.

We can set $\mathbf{A}_i$ to model a variety of real-world considerations as well as production policies. For instance, we might use the matrix $\mathbf{A}_i$ to model production leadtimes, where production starts are just the production outputs offset by the leadtime, to model yield factors within the production stage (e.g., need to start 1.2 units to get output 1.0), or to model the fact that production starts occur on a different time scale (biweekly rather than weekly) from the production outputs. We can also model a constant work-in-process policy where production starts for the period exactly equal production outputs. Indeed, in this way, for general multistage systems we can use this general approach to compare push policies—where starts equal planned output L periods from now—with pull policies, where starts "replace" the outputs produced in the current period.

Assumption 6. *At each stage we know how many units of input are required for one unit of output. Without loss of generality, we assume that one unit of input is required for one unit of output at each stage.*

The single-stage model that we wish to use as a building block takes as input a dynamic forecast process of the requirements for the stage. We now show that, given the assumptions above, the forecast process at each stage in the multistage network satisfies the assumptions of the single-stage model. In particular, we show the following proposition:

Proposition 10. *At each stage i, the forecast revision* $\underline{\Delta f}_i$ *can be expressed as a linear combination of* $\underline{\Delta f}_1, \ldots \underline{\Delta f}_m$: $\underline{\Delta f}_i = \sum_{j=1}^{m} M_{ij}\,\underline{\Delta f}_j$ *for some matrices* M_{ij}. *By Assumption 2, this implies that* $\underline{\Delta f}_i$ *is an i.i.d. random vector.*

We will demonstrate this proposition by an induction argument. The proposition is true by assumption for the end-item stages $1, \ldots m$. Suppose this proposition is true for stages $i = 1, \ldots j - 1$; we will now show that it is true for $\underline{\Delta f}_j$. Let S_j be the index set of immediate successors to stage j. The forecast process for outputs of an upstream stage $j > m$ is

$$\underline{f}_j = \sum_{k \in S_j} \underline{G}_k.$$

Accordingly, we can write

$$\underline{\Delta f}_j = \sum_{k \in S_j} \underline{\Delta G}_k.$$

We note by Assumption 6 that $\underline{\Delta G}_k = \mathbf{A}_k\,\underline{\Delta F}_k = \mathbf{A}_k\,\mathbf{W}_k\,\underline{\Delta f}_k$, and by the induction hypothesis that each $\underline{\Delta f}_k$ is a linear combination of $\underline{\Delta f}_1, \ldots \underline{\Delta f}_m$. Thus, we can see that each $\underline{\Delta G}_k$ for $k \in S_j$ is a linear combination of $\underline{\Delta f}_1, \ldots \underline{\Delta f}_m$, and

hence so is $\underline{\Delta f}_j$. This completes the induction argument, showing that each $\underline{\Delta f}_j$ is an i.i.d. random vector. □

We have thus shown that at each stage we have preserved the essential requirement that the forecast revisions are i.i.d. random vectors, and thus, that the assumptions for the forecast process of the single-stage model are satisfied at each stage in the multistage network. *This is an important result because it means that we can now model an acyclic multistage system by just replicating the single-stage model.* In this sense, the single-stage model serves as a building block.

4. CASE STUDY

In this section we describe an industrial application of the Dynamic Requirements Planning (DRP) model from a thesis internship performed by one coauthor (Hetzel) at the Eastman Kodak Company. The internship was conducted as part of MIT's Leaders for Manufacturing Program and ran from June 1992 to December 1992. (See Hetzel 1993 for more details on the application.)

The general charge for the thesis was to investigate cycle time reduction within the context of the film manufacturing processes at Kodak. As part of the internship, Hetzel joined an internal supply chain optimization team that was investigating opportunities for better coordination over a specific supply chain, including issues of cycle time and inventory reduction. One open issue facing the team was that of strategic inventory placement: how much inventory was needed, and where should it be placed across a multistage supply chain. Hetzel identified this as an opportunity to apply the DRP model, and the team agreed that it was an appropriate tool for their task of strategic inventory placement. The only alternative considered was to develop a simulation: since the DRP model was already available from the authors in a software package, developing a simulation would have required extensive additional work.

The goal of the supply chain analysis was to determine the optimal safety stock levels between each stage in the film making supply chain. The underlying concept is that looking at one stage of the supply chain in isolation is inherently suboptimal. All the stages in the supply chain are interconnected by information flows. In short, the inventory and production policies that are best for one stage may not be optimal for the supply chain as a whole.

In the case study, the team was able to address this situation by using the DRP model to consider all stages in the supply chain. Their recommendations challenged the conventional targets and performance measures for individual divisions (stages). For example, an upstream stage, roll coating, faced a corporate-wide mandate to lower inventories. However, by using the DRP model, the team discovered that roll coating needed to increase inventories to provide the desired service to the next stage. When roll coating holds sufficient inventory to provide a high level of service, downstream stages can hold less, resulting in a net savings for the corporation. Overall, the analysis determined that inventories for the products of the case study could be reduced by 20%. This example highlights the importance of considering the entire supply chain when setting inventory and production policies.

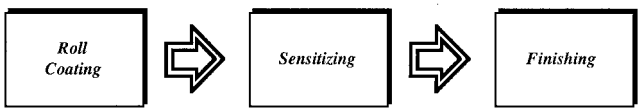


Figure 4. Simplified version of film manufacturing supply chain.

The rest of this section will describe the supply chain for the case study, provide the results from the DRP model, and comment on implementation issues.

4.1. Supply Chain for Case Study

In Figure 4 we give a simplified version of the process for film making. Roll coating transforms raw chemicals into a roll of film base. Sensitizing coats the film base with a silver halide emulsion. Then finishing cuts and packages the sensitized rolls into finished products. The structure of this supply chain has three interesting characteristics. First, the number of items grows dramatically from stage to stage; one film base might result in 5 to 10 different sensitized rolls, which might lead to a hundred or more finished goods. Second, there is a rapid growth in the value of the product due to added material (e.g., silver) and nature of the processes. Third, there is a gradual decrease in the leadtimes across the supply chain.

For the case study, the supply chain optimization team focused on a single film base (called a support). That single base becomes three different sensitized film codes because it can be coated with three different emulsions. The three film codes can be finished (slit, chopped, and packaged) into 24 different finished good items. Figure 5 illustrates the supply chain for the case study. This particular product "tree" was chosen because it is high volume, it has relatively few end items (24 total), and it represents a "typical product" that the team felt would make a useful pilot program.

It is important to note that the case study does establish arbitrary bounds on the supply chain. The case study starts with the creation of a film base in roll coating and excludes the upstream raw material stages such as chemical, gelatin, and polymer production. The case study ends with the finishing process and arrival at the Central Distribution Center, and ignores the rest of the distribution system. Besides being bounded at both ends, the case study's supply chain is also simplified. In reality, the sensitizing and finishing stages have materials flowing into them such as emulsion and packaging components. Even though these materials require inventory management, they are assumed to be available with 100% service, and were not explicitly incorporated into the model.

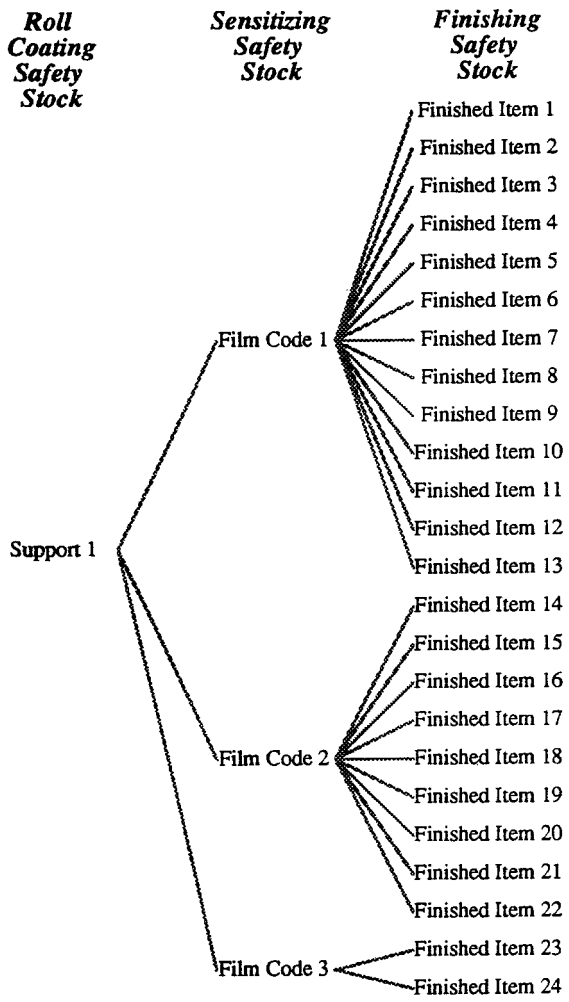


Figure 5. Supply chain analysis case study.

4.2. Data Collection

Parameterizing the DRP model required an extensive data collection effort. For each item in the chain, the team gathered data on the item's leadtime, unit cost, inventory holding cost, manufacturing frequency, and desired service level. For each end item, they needed the planning horizon, the average demand level, and a time history of the forecast process.

The leadtime and manufacturing frequency were modeled through the weight matrix (**W**). Since the team did not consider production smoothing, in the absence of leadtime and production frequency considerations, the weight matrix is simply the identity matrix. A leadtime of L periods is then captured by forcing the first L rows of **W** to be zero. To represent a production frequency of once every two weeks, **W** would then be modified so that every other row was zero. It should be noted that this method of capturing production frequency is only an approximation.

From the forecast histories, the team estimated the diagonal elements of the covariance matrix (Σ) for the forecast revision process $\underline{\Delta f}_t$; the off-diagonal elements were assumed to be zero. Associated with each "branch" linking different stages they calculated a historical "goes into factor" to capture any yield loss or conversion factors. This information was used to construct the matrix **A**, as described in Section 3.

A side benefit of applying the DRP model was that the data collection effort identified some potential issues along the supply chain. For example, in the course of reviewing the forecast data, the team discovered that the forecasts varied in a systematic way that led to a reevaluation of the forecasting process. In addition, collecting data enhanced supply chain communication and allowed the team to resolve a discrepancy in the annual planned volumes between two of the stages.

4.3. Results

The team used the DRP model to develop a base case recommendation on inventory placements. The 24 finished items were grouped into nine product aggregates, where the product aggregates shared common production processes and had similar demand histories. Service levels were set at 95% for each stage. The weight matrices were not optimized and were set to reflect each stage's leadtime and manufacturing frequency. In order to reflect Kodak's current scheduling systems, there was no production smoothing across weeks.

The DRP model showed the potential to lower inventory across the case study product "tree" by 20%, as shown in Figure 6. Note that, in general, inventories can be pushed upstream where they are in a strategic position because: (1) the inventory is common to the greatest number of finished end items desired by the customer, and (2) the inventory is at its lowest value added and thus at its lowest carrying cost. In fact, the inventory levels of roll coating's "Support 1" actually need to increase to provide savings for the supply chain as a whole.

The definition of "inventory" as it is used in this results section is important. The inventory changes and the comparisons in Figure 6 represent average inventories. Average inventory for each item includes the safety stock calculated by the DRP model, plus the cycle stock due to production batching, plus the pipeline stock from transport needs.

Besides the required safety stocks, the DRP model also provided information on the variance of the production requirements at each stage of the supply chain. The supply chain optimization team used this variance to determine the "surge" production capability needed for any stage. For instance, they might set the surge capability to be the production level that would cover the production requirements 95% (1.645 standard deviations above the mean, assuming normal forecast errors) of the time.

4.4. Validation

Before the DRP model recommendations could be implemented, the team needed to develop confidence in the results. Therefore, multiple scenarios were run to test the

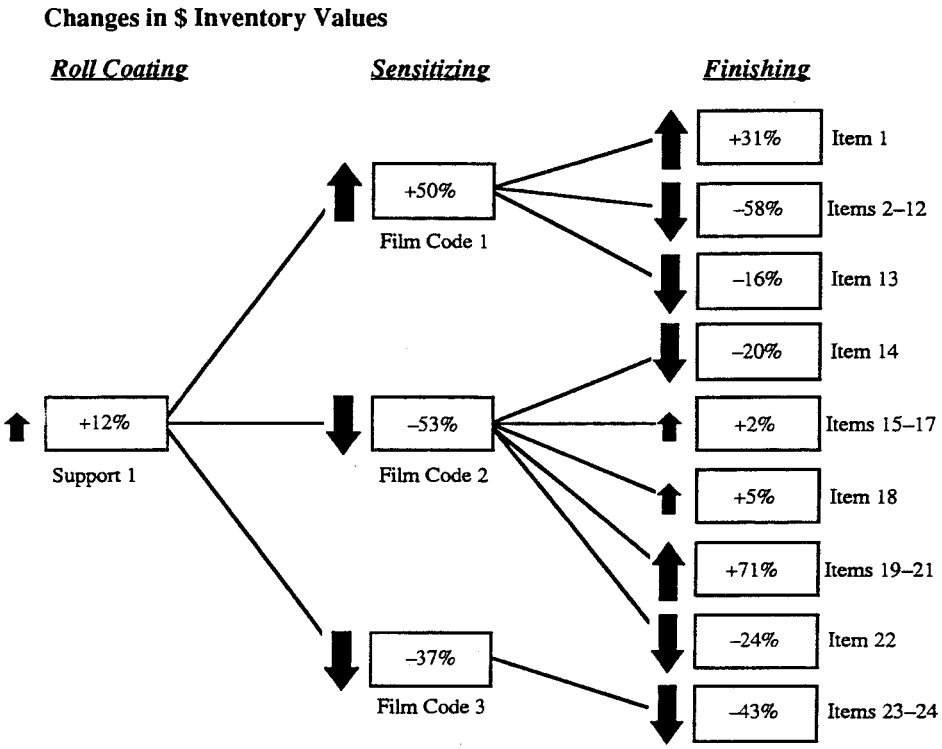


Figure 6. Results from the supply chain analysis—strategic inventory placement. Note: Excludes emulsion and chemicals inventories. Excludes regional distribution center (RDC) inventories. Includes all WIP, cycle, and in-transit stocks. Finished goods inventories are gross CDC averages. Service levels are at 95% for all stages. There is no production smoothing.

sensitivity to various parameters, including the service levels, the leadtimes, and the size (variance) of the forecast errors. (See Hetzel 1993 for details.)

The main barrier that the team had to overcome was understanding how the DRP model works. This was accomplished by exercising the model for different scenarios, especially conservative ones; by displaying all the input data and its sources for validation; by keeping the model (relatively) simple, e.g., assuming a diagonal covariance matrix and limiting the size of the explosion; by comparing the model results versus current inventory levels; and by acknowledging the model's shortcomings. Finally, a key success factor was that the model was implemented on a personal computer, provided a graphical interface for representing and visualizing the supply chain, and provided an almost instantaneous response. In addition to the analytic model, a Monte Carlo simulation was used that simply worked through the mechanics of the analytic model for a randomly generated demand stream, reporting on performance measures of interest. The simulation allowed assessment of model assumptions, thereby validating the analysis. For example, constraints were added to the simulation that enforced production capacities and prohibited production from beginning if no raw material was onhand.

No model is perfect, and no description of a model is complete without a list of shortcomings. The supply chain optimization team identified three weaknesses: (1) the DRP model does not account for lead time variability, (2) it assumes stationary average demand over time, and (3) it cannot accommodate a large product explosion. Whereas the first two are inherent assumptions for the model, the latter concern was due to a limitation in the software that could be easily overcome. However, we expect that in most practical situations a team should probably not be working at any greater level of detail than the case study, say, less than 25 items. Keeping the model at an aggregate level both reinforces the fundamental guiding principles and also makes implementation simpler.

4.5. Implementation

Once all of the supply chain analysis requirements were complete, the supply chain optimization team added local intelligence about specific customers and manufacturing issues for each item that could not be captured by the model. After reaching an understanding about how all of the model's proposed changes would impact the supply chain, the team decided to implement a pilot program, with the intention of moving to the more aggressive "base case" if there were no service problems.

The pilot program only involved the inventories of three items. The plan raised the one roll coating item's inventory by 20%, and it lowered two sensitizing items' inventories, each by 60%. The plan was implemented in early 1993. The savings were captured in the 1993 Annual Operating Plan for the case study's line of business. As of the end of April 1993, not a single end-customer order had been

missed on the pilot product due to stockouts or inventory shortages. The team then implemented the remaining recommendations over the course of 1993.

5. CONCLUSIONS

This paper presents a new model of the requirements planning process. We first describe in detail how to model a single production-inventory stage as a linear system, and provide the analysis for determining performance measures on production smoothness, production stability, and inventory requirements. We also show how to optimize the tradeoff between production smoothness and inventory for a single stage.

To model a multistage system, we can use the single-stage model as a building block. The structure of the single-stage model makes it very easy to link single-stage models together to represent the multistage system. In particular, each single-stage model takes as input a forecast of demand requirements and converts this forecast into a production plan. In the context of a network of production stages, the production plan from a downstream stage acts as the demand forecast for an upstream stage. In this way, we can cascade the single-stage models to model a multistage system.

We also report on an application of the model within the context of a supply chain study. The DRP model was used as a tool to help determine inventory placement across a multistage supply chain. This illustration provides some evidence of the value of taking a corporate-wide view by optimizing the supply chain rather than suboptimizing each of the pieces.

One outgrowth from the case study is a better understanding of industry needs, and where the DRP model is weak. Based on this experience, as well as observations from industry, we identify the following research topics.

- *Nonstationary demand.* A stationary demand process is not an accurate model for the demand experienced by many products. Common nonstationary effects include seasonal effects, end-of-quarter or end-of-year effects (the "hockey stick"), and short-product life cycles. Some of these nonstationarities get masked when products are aggregated into families or product groups. Nevertheless, an important enhancement to the model would be to capture, in some way, nonstationary demand processes.

- *Service-Level Assumptions.* In extending the single-stage model to a multistage setting, we assume that there will be sufficient inventory to decouple the stages. In effect, we assume that the service levels will be set to assure a high level of service, and in the model analysis, we ignore the downstream consequences of an upstream stockout; i.e., starvation of inputs. These assumptions raise two questions. One is, what are the consequences of ignoring the internal stockouts, and the second is, what should the internal service levels be. Graves (1988a) provides some justification for these assumptions in a related setting. And simulation tests that we have done confirm that ignoring the internal stockouts in the analysis, when service levels are high, does not distort the results of the model. But the issue remains as to how to set the service levels. The literature on multiechelon distribution systems (e.g., Jackson 1988, Schwarz 1989, Graves 1995) suggests that, from a system perspective, it often may be better to have low levels of internal service.

- *Guidelines for Consolidating Stages.* On a related note, we conjecture that, in some instances, the best policy may be to remove the inventory between an upstream and downstream stage, and thus consolidate these stages for planning purposes (Simpson 1958). Rather than have two stages separated by an inventory buffer, we would have one (combined) stage, albeit with a longer leadtime. Within a multistage system, depending on the leadtimes and holding costs, it may be optimal to consolidate some of the stages. We expect it would be helpful to have guidelines for determining what stages are good candidates for consolidation.

- *Multistage Optimization.* The paper describes the optimization of the tradeoff between capacity and inventory in a single stage for a diagonal covariance matrix: It would be interesting to explore how this development extends to nondiagonal covariance matrices, as well as to a multistage system. In particular, we would like to develop guidelines for setting the weight matrix $\mathbf{W}$ for each stage. Furthermore, one could explore how to choose among alternative production release policies, such as pull versus push, in a multistage setting.

- *Production Assumptions.* The model has a highly-simplified model of the production process. The model sets the production outputs, and these outputs are translated into production starts (e.g., by a leadtime offset). With this model, we can represent fixed lead times, yield loss factors, batch setup frequencies, as well as uncertainty that can be modeled as an additive factor. Nevertheless, there are issues as to the validity or appropriateness of this representation and the sensitivity of the model results to these assumptions. It would certainly be useful to have a richer model of the production process. For instance, it would be useful to capture the nonlinear congestion effects due to multiple items competing for a shared resource.

ACKNOWLEDGMENT

We wish to thank Chris Athaide for his contributions in the initial stages of this research, the IBM Thomas J. Watson Research Center and the NSF Strategic Manufacturing Initiative for financial support for this research, MIT's Leaders for Manufacturing Program for its support and resources to complete and apply this research, and the referees for their helpful and constructive comments on an earlier draft.

REFERENCES

BAKER, K. R. 1993. Requirements Planning. In *Handbooks in Operations Research and Management Science, Vol. 4, Logistics of Production and Inventory.* S. C. Graves, A. H. Rinnooy Kan and P. H. Zipkin (eds.), North-Holland, Amsterdam.

FINE, C. H. AND S. C. GRAVES. 1989. A Tactical Planning Model for Manufacturing Subcomponents in Mainframe Computers. *J. Manuf. and Opns. Mgmt.* **2,** 1, 4–34.

GRAVES, S. C., H. C. MEAL, S. DASU, AND Y. QIU. 1986. Two-Stage Production Planning in a Dynamic Environment. In *Multi-Stage Production Planning and Inventory Control.* S. Axsäter, C. Schneeweiss and E. Silver, (eds.), Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, **266,** 9–43.

GRAVES, S. C. 1986. A Tactical Planning Model for a Job Shop. *Opns. Res.* **34,** 4, 522–533.

GRAVES, S. C. 1988a. Safety Stocks in Manufacturing Systems. *J. Manuf. and Opns. Mgmt.* **1,** 1, 67–101.

GRAVES, S. C. 1988b. Determining the Spares and Staffing Levels for a Repair Depot. *J. Manuf. and Opns. Mgmt.* **1,** 2, 227–241.

GRAVES, S. C. 1988c. Extensions to a Tactical Planning Model for a Job Shop. *Proceedings of the 27th IEEE Conference on Decision and Control*, Austin, Texas, December.

GRAVES, S. C. 1996. A Multiechelon Inventory Model with Fixed Replenishment Intervals. *Mgmt. Sci.* **42,** 1–18.

HEATH, D. C. AND P. L. JACKSON. 1994. Modeling the Evolution of Demand Forecasts with Application to Safety Stock Analysis in Production/Distribution Systems. *IIE Trans.* **26,** 3, 17–30.

HETZEL, W. B. 1993. Cycle Time Reduction and Strategic Inventory Placement Across a Multistage Process. MIT Master's Thesis.

JACKSON, P. L. 1988. Stock Allocation in a Two-Echelon Distribution System or 'What to Do Until Your Ship Comes In'. *Mgmt. Sci.* **34,** 7, 880–895.

KARMARKAR, U. S. 1993. Manufacturing Lead Times, Order Release and Capacity Loading. In *Handbooks in Operations Research and Management Science, Vol. 4, Logistics of Production and Inventory.* S. C. Graves, A. H. Rinnooy Kan and P. H. Zipkin (eds.), North-Holland, Amsterdam.

KLETTER, D. B. 1994. Proofs of P8 and P9. Technical Appendix.

LEE, H. L. AND C. BILLINGTON. 1993. Material Management in Decentralized Supply Chains. *Opns. Res.* **41,** 5, 835–847.

SCHWARZ, L. B. 1989. A Model for Assessing the Value of Warehouse Risk-Pooling: Risk-Pooling over Outside-Supplier Leadtimes. *Mgmt. Sci.* **35,** 828–842.

SIMPSON, K. F. 1958. In-Process Inventories. *Opns. Res.* **6,** 863–873.

DEVELOPMENT OF A RAPID-RESPONSE SUPPLY CHAIN AT CATERPILLAR

UDAY RAO, ALAN SCHELLER-WOLF, and SRIDHAR TAYUR

Graduate School of Industrial Administration, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213-3890
urao@gizmo.gsia.cmu.edu ● awolf@andrew.cmu.edu ● stayur@cyrus.andrew.cmu.edu

(Received August 1998; revision received April 1999; accepted August 1999)

As part of its growth strategy, Caterpillar Inc. is launching a new P2000 product line of "compact" construction equipment and worktools. In anticipation of this, they asked the authors to construct and analyze potential P2000 supply chain configurations. Using decomposition and results from network flow theory, inventory theory, and simulation theory, we were able to provide solutions to this problem for different supply chain scenarios provided by Caterpillar. Novel features of our model include expedited deliveries, partial backlogging of orders, and realized sales that are responsive to service. Caterpillar made their decision regarding the P2000 supply chain based on our recommendations.

1. INTRODUCTION

We describe an operations research application supporting the design and deployment of a distribution logistics system for a new product line at Caterpillar Inc. (Cat). After decomposing the problem, we apply network flow techniques, recent results from inventory theory, and simulation based optimization (Infinitesimal Perturbation Analysis or IPA; see Glasserman and Tayur 1995) to arrive at a solution. In addition to such standard features as multiple-echelons, capacity constraints, uncertain demand, lead times, and multiple products, our problem also has the following novel features:

1. Dealers in Caterpillar's distribution network can order from dual suppliers. There is a low cost *regular* alternative and a high-speed, *expedited* supplier. We determine the optimal replenishment paths for each (dealer, product) pair using a deterministic minimization of cost or time over the supply network.
2. The magnitude of captured demand is sensitive to service response time. In each time period (day, week), the number of lost sales depends on the customer service provided. A certain percentage of new customers renege if not immediately served, while a different percentage of waiting customers are lost if forced to continue to wait.

To our knowledge, this is the first time a problem of this scope and complexity has been solved in this manner. In particular, the use of IPA to establish inventory levels in an industrial problem of this magnitude appears to be unprecedented.

In this paper, we describe the development of an optimization engine for use in designing Caterpillar's supply chain. We also detail the collection of data for the engine, provide the results of our optimization, and report on a sensitivity analysis of our output. Specifically, after describing the problem environment in §2, we provide details of the modeling and analysis in §§3 and 4, respectively. We report our results in §5, after which we present some concluding remarks in §6.

Throughout the paper, numerical data is disguised to maintain confidentiality.

2. THE P2000 SUPPLY CHAIN

To exploit anticipated growth in the small construction industry, Caterpillar Inc., the world's leading producer of construction and mining equipment, decided to introduce a new "compact" product line, the P2000, starting in 1999. This decision has been widely reported in the business and industry news media, with articles appearing in publications such as *The Financial Times* (Feb. 12, 1998) and *Business Week* (March 9, 1998).

One reason for the media's interest is that the P2000 not only represents a new product line, but also a new strategy for Caterpillar's construction equipment division. Caterpillar's traditional product line consists of large, low-volume, high-margin, customized machines costing $500,000 or more. Cat is a well-known leader in this market, with few large competitors (such as Komatsu Ltd.). The P2000 family encompasses smaller, medium- to high-volume, standardized products selling for as little as $20,000 per

Subject classifications: Professional: OR/MS implementation. Inventory/production: applications. Industries: machinery.
Area of review: OR PRACTICE.

machine. Specifically, the P2000 product line consists of several models of three different *machines* including a Skid-Steer Loader ("SSL"), a Compact Wheel Loader ("CWL"), and a Mini-Hydraulic Excavator ("MHE"), as well as some 40 *worktools* (such as buckets, fork sets, and grapples). Designed for use with one or more particular machines such as a Skid-Steer Loader, worktools can be sold as attachments to both competitor's machines and Caterpillar's. Worktools thus provide a means to enter the market independent of P2000 machine sales. This compact product segment currently has many entrenched market leaders including BobCat (trademark of the Melroe Company, a unit of Ingersoll-Rand Company), Deere & Company, and Case Corporation.

Both strategic and operational considerations motivated a careful analysis of the P2000 supply chain before its deployment. Cat feared that the P2000 family might not fit well in their current large equipment supply chain. They wanted to develop a network for the P2000 that could maximize profits, capture market share, and provide flexibility. They contacted the authors for assistance in determining a configuration which could fulfill these goals. We examined two study years: an initial year, 2000, and, four years later, in 2004. Although the product launch is in 1999, this first year is considered a "ramp-up," making the year 2000 demand more suitable for supply chain design and analysis. The two study years differ in the volume of forecast demand, price and cost parameters, and in routing restrictions. Compared to the year 2000, the routing restrictions generally were relaxed in year 2004, because Caterpillar expected to develop new processing capabilities.

The P2000 products are sourced, manufactured, and assembled in approximately 20 locations throughout North America and Europe. Sanford (North Carolina) and Leicester (England) are key production centers for machines. Worktools come into Caterpillar's supply chain from source-locations in the United Kingdom, the United States, Mexico, Sweden, Germany, and Finland. Each worktool has a single source of supply. In North America, P2000 products will be sold by a network of 190 Caterpillar dealers serving 58 districts in the United States and Canada.

2.1. P2000 Strategy

Specific concerns motivated Caterpillar's focus on their P2000 supply chain. The international nature of the chain, coupled with the weight of the equipment, created the potential for large lead times and shipping costs. Caterpillar previously had made the decision not to compete on price. Rather, in keeping with their core philosophy, quality and service were areas in which Cat would differentiate the P2000. Thus, long lead times were particularly worrisome. The dealer surveys reinforced this concern, implying that Caterpillar's future products would be highly substitutable with those of the competition—primarily Bobcat.

Therefore, Caterpillar believed it crucial that they capture customer demand for their P2000 products as soon as the demand materialized. By not forcing potential customers to wait for delivery, Cat would establish a reputation for product availability. This would not only generate demand for the P2000, but also allow Cat to steal demand for their competitors' (substitutable) products. Cat wanted to identify both a minimum cost channel for a product in the supply chain and an additional channel for expedited delivery. The expedited channel, likely one with a higher cost, would be used if dealer inventory levels dropped precipitously. This was the genesis of the dual supply modes within the supply chain.

To address the objective of maximizing profit subject to capturing a satisfactory portion of market demand, we constructed our model so that poor service (product availability) led to lower sales. We felt that Cat would have considered accepting a lower profit, higher inventory solution (for year 2000) to gain "market presence" for their products. To this end, we were prepared to develop a tradeoff curve between year 2000 profit and market penetration.

2.2. The Worktool Problem

The nature of the P2000 line dictates that final manufacturing and testing of some worktools take place at specific nodes of the supply chain. This constrains certain worktools to pass through selected processing facilities. Similarly, the presence of international import/export facilities such as bonded warehouses (which permit Caterpillar to forgo paying duties while storing their products in transit) requires certain items from overseas to pass through selected customs locations. With these factors in mind, Caterpillar determined that they could use up to seven additional transshipment locations—intermediary nodes between the source and the dealer—in North America, in addition to direct shipment (DS) of worktools from sources to dealers.

The seven possible transshipment locations for worktools in the United States are grouped into two disjoint sets: three *Tool Facilities* (TFs, which had not yet been constructed) and four *Parts Distribution Centers* (PDCs, which already were handling Caterpillar products). The Tool Facilities potentially would be located in Sanford (North Carolina), Laredo (Texas), and Indianapolis (Indiana). The Parts Distribution Centers are located in Morton (Illinois), Miami (Florida), Denver (Colorado), and York (Pennsylvania). In addition, other transshipment locations are included in the supply network. For instance, there is a UK tool facility in Leicester, a UK PDC in Desford, and a European tool facility in Belgium (with a PDC in Grimbergen) that feed worktools made in Europe to North America. However, we were not given the option of excluding these nodes from the supply chain.

There were four primary options for the North American worktool supply chain:

1. Use of all TFs and PDCs;
2. Use of PDCs only;
3. Use of TFs only; and
4. Use of neither TFs or PDCs, thus allowing only direct shipment (DS).

Secondary options included using the PDCs with one or two supporting TFs, or vice-versa. For example, one might add the Sanford Tool Facility to the PDCs to perform certain manufacturing or quality check operations within the supply chain.

For each of these four primary scenarios, we were asked to determine:

1. Supply path(s) from each worktool's source to every dealer region in Caterpillar's network;
2. Inventory levels and ordering policies at all points along these paths;
3. Revenues, costs, and profits, and their breakdown by product, geographical regions, and nodes (these costs *excluded* the fixed cost of constructing the tool facilities); and
4. The expected percentage of demand captured.

Our analysis showed that the optimal supply chain configuration comprised the PDCs and the Sanford TF. This configuration yields an estimated profit several million dollars higher than the TF-Only or DS-Only options, while capturing virtually all of the potential P2000 demand. These comparisons are particularly salient because political considerations prompted Caterpillar to initially favor the TF-Only and DS-Only scenarios over inclusion of the PDCs.

2.3. The Machine Problem

The nodes in the distribution network for machines were previously determined by Caterpillar. Thus, for this fixed network, the machine problem requires finding inventory levels that maximize profits while capturing no less than a specified percentage of the customer demand. This is equivalent to evaluating a single worktool scenario. Therefore, we will concentrate on the worktool problem in this paper, although we provide illustrations of some of the results for machines in §5.

The extant network for machines included two manufacturing plants and five North American storage facilities. The Sanford plant manufactured only SSL machines, and the Leicester, UK plant was responsible for CWL and MHE machines. The five storage facilities, used exclusively for machines, were: (1) Houston (Texas), Savannah (Georgia), and Harrisburg (Pennsylvania), which served the U.S. market; and (2) the bonded warehouses at Portland (Oregon) and Harrisburg, which served the rest of North America (primarily Canada).

3. MODELING THE P2000 PROBLEM

Many papers have addressed aspects of material flow management, but few have considered modeling entire supply chains. We approach Caterpillar's problem in a spirit similar to that of Lee and Billington (1993) and Feigin (1998). Lee and Billington describe their experience with a decentralized DeskJet printer supply chain at Hewlett Packard. They point out some of the challenges in modeling supply networks, and take advantage of various approximations to model a single site in the network. Feigin (1998) also uses approximations in his analysis of the trade-off between service levels and inventory investment in large supply chains. See Tayur et al. (1998) for a compilation of recent advances in supply chain management.

Finding optimal solutions to the individual components of Caterpillar's problem, such as the material routing or inventory replenishment subsystem, is in itself extremely difficult. The *deterministic* version of the routing problem, a min-cost, multicommodity, network flow problem with nonconvex costs, is the subject of a parallel work by Keskinocak et al. (1998). This problem is comparable, from a complexity viewpoint, to the transportation routing problem for next day, second day, and deferred delivery of packages considered by Barnhart and Schneur (1996). Separately, Scheller-Wolf and Tayur (1998) consider the determination of optimal policies for the inventory problem with expedited orders. They use IPA to find such optimal levels within the class of order-up-to (or base-stock) policies. Because the P2000 supply chain problem includes subproblems of these types, we believe its exact solution is unobtainable using currently available methodologies.

To make Caterpillar's problem tractable, while maintaining the model's validity, we reduced its scope in a variety of ways. This permitted us to arrive at a good solution in a reasonable amount of time. We could then conduct a sensitivity analysis on the robustness of the solution to changes in model parameters.

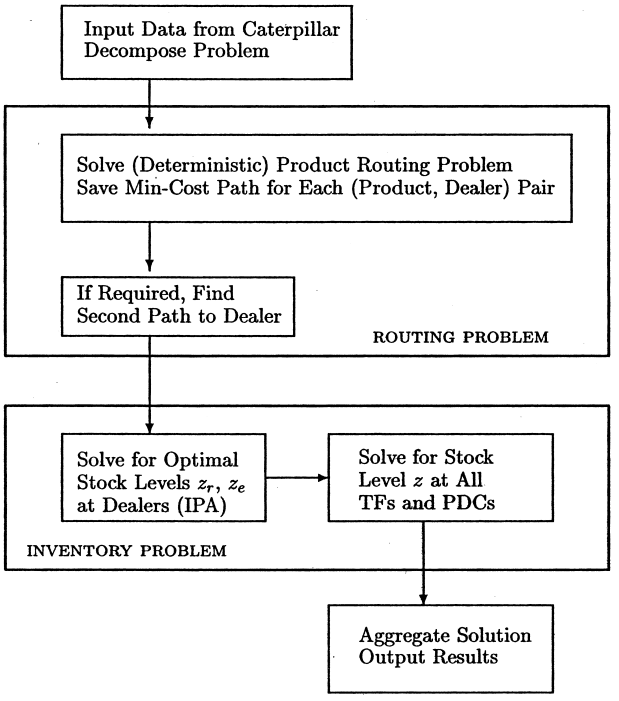
3.1. Model Assumptions and Justification

Our primary assumptions are listed below.

Problem Decomposition: Figure 1 summarizes our problem modeling and solution procedure. We decompose the problem into a network routing problem (§3.2) and a stochastic inventory problem (§3.3). The routing problem ignores safety stock levels throughout the network. Consequently, it may overlook some inventory risk pooling opportunities offered by overlapping routes for the same worktool and different destinations. However, for Caterpillar's problem, the relative cost of inventory is small compared to the transportation costs (see Figure 5). Significantly, without this decomposition, Caterpillar's problem falls in the realm of stochastic nonlinear integer programming (SNLIP; Horst and Tuy 1993, Birge and Louveaux 1997). Efficient approaches for SNLIP problems of this scale currently are not available.

Network Decomposition: Decomposing the network into dealer nodes and transshipment nodes enables us to solve the resulting subproblems as single-stage inventory systems. The accuracy of known approximations (Glasserman 1997, Tayur 1992), and the need to quickly evaluate multiple products over different scenarios, motivated this decision. If required, we could have modeled the entire network as one multiechelon inventory system using techniques similar to those described in §3.3. Glasserman and Tayur (1995) provide the theoretical base for such a multiechelon model.

Figure 1. Flow chart outlining problem decomposition and solution steps.



Dealer Aggregation: We solve the problem for one "typical" aggregate dealer within 21 regions in North America, rather than by individually considering all 190 dealers. Caterpillar accepted this because:

1. The dealers were roughly uniformly distributed within each region.

2. The sourcing and transportation costs up to each region would likely dominate the variation in costs between dealers within a region.

3. Any increase in cost resolution resulting from a more detailed model probably would be voided by the comparatively less exact estimates of individual dealer demand and logistics cost.

The use of the more detailed network structure would pose no theoretical problems for our methodology, but it would increase computational times.

Disaggregation of Sourcing: Caterpillar uses a unique source for each worktool.

Decomposition by Products: As a "base model," we decouple the distribution of different products or product families throughout the network—worktools from each other and from machines. We then solve the routing and inventory problem for each machine and worktool independently, aggregating the results. Caterpillar agreed to this based on the fact that worktools and machines were to initially use disparate distribution networks. This decoupling disregards the possible dependencies between demands for particular products.

In our "refined model," we consider situations in which worktools and machines could share transportation. Worktools may use the flatbed transportation normally used for the machines at rates, times, and capacities different from the normal closed-van mode for worktools. Section 5.1.1 describes the results of this generalization in more detail.

Demand Modeling: We model demand in one of two ways. If the estimated mean daily demand for a product is less than one unit, we approximate it with a Bernoulli random variable. Otherwise, daily demand is modeled using a truncated Normal. The data Caterpillar provided did not suggest any specific demand distribution. We conducted preliminary tests with alternative demand distributions, including the uniform and exponential. These did not change the overall optimal network configuration. We model the distribution of transship node demand with a normal random variable, because this demand is an aggregate over the dealers supplied by this transshipment node according to lot-for-lot ordering policies. Caterpillar agreed that our demand models were acceptable.

We treat demand for different products at different dealer districts as uncorrelated and time stationary. This was acceptable to Caterpillar. If they had desired (and been able to provide data on correlation/seasonality), we could have incorporated this in our simulation based optimization model. See Kapuscinski and Tayur (1998).

Lost Sales: Based on information provided by Caterpillar dealers, we use a two-parameter model for customer impatience (see §3.3.1). We adopt this model because the data indicates that customers fall into two categories: those who leave immediately and those who are willing to wait a fixed amount of time. Had the data indicated more segmented behavior, additional parameters could have been estimated and used.

Preliminary experimentation (borne out by our final results) indicated that the optimal percentage of lost sales is small. This is a consequence of the fact that the inventory cost rate for a worktool is significantly smaller than the unit profit. This leads us to believe that alternative methods of modeling lost sales (e.g., stochastic parameters, discrete parameters) would lead to qualitatively similar results.

Continuity: For simulation purposes, inventory was approximated by a continuous variable. Due to the observed unimodal nature of the model's profits, this continuity assumption was relaxed by searching the adjacent integral values after arriving at an optimal inventory level.

In summary, this paper presents a general methodology applied to a specific problem at Caterpillar. At the core of this methodology are a network routing problem and simulation based recursions (Appendix A). This methodology remains valid for many of the model enhancements mentioned above, such as a more detailed dealer network with correlated demand among products and regions. We developed our specific model in an iterative fashion based on periodic interaction with Caterpillar. As new data arrived and new features were needed, we updated the model. Our final result is a recommendation to Caterpillar regarding the configuration of their supply chain. To the extent possible, we investigated alternatives to our assumptions,

consistently finding that they did not affect the final recommendation. (For example, see the paragraph preceding §5.1.1.)

Our technique's greatest value lies in its ability to provide a good solution and perform what-if analyses while incorporating uncertainty over a large and complex problem. The question of whether our specific model, or even Caterpillar's data, is an accurate representation of the problem is a valid one. In the absence of comparable models and solution techniques we are unable to answer this question conclusively. Caterpillar is satisfied with our efforts; they are currently implementing a supply chain configuration based upon our work.

3.2. The Product Routing Model

For each product and dealer combination, we model the supply chain as a collection of nodes (sources, dealers, and transshipment points) and edges (connecting the nodes). Each edge has an appropriate lead time and cost component: overseas shipment on freighters at container rates, shipment within North America by closed vans or flatbed trucks at either truckload (TL), or less-than-truckload (LTL) rates and times. Trucking rates depend on the source and destination, and the product's weight and volume. Likewise, each node has times and costs associated with it. Inventory costs accrue at varying rates for different locations and products throughout the network, as do handling times and costs. Certain nodes are precluded from holding any inventory—we treat them as instantaneous transshipment points.

To accomplish the dual objectives of the supply chain—maximum profit and a high service level—up to two paths are found for each product and dealer node within the network. The first is the minimum cost or *regular* path. The second has the smallest lead time from the dealer to the next level up in the supply chain. This next level up is linked to the source node by a min-cost path. Together, these two links form the *expedited* path. When the last link in the minimum-cost path also has the shortest delivery lead time for direct shipment from any transshipment point to the dealer, only this single dominating path is used. The majority of shipments of a product are assumed to flow along the regular path. Worktools would utilize the expedited link in situations where unexpectedly large demand had caused dealer inventory to drop below a specified level. The inventory optimization portion of the algorithm determines this level.

3.3. The Inventory Model

After decomposing the model by products, we decompose the inventory system for each product into two subsystems: the dealers and the transshipment nodes. This decomposition implicitly assumes that service levels at transship points will be high enough to avoid stock-out occurrences. We describe experiments below, based on Glasserman and Tayur (1995), used to test this assumption. They validated this decomposition.

Our experiments compared the performance of a two-stage system under our decomposition with a two-stage system globally optimized using IPA. For a variety of cost, service, and demand parameters, the approximation tended to decrease inventory levels at the lower echelon. This decrease in lower echelon inventory does not adversely affect customer service in Caterpillar's problem. In fact, the selected inventory levels under the decomposition attain a near 100-percent customer service level. Upper echelon inventory levels could be higher or lower than optimal under the decomposition, but these stock levels always were very close under both systems, implying similar service levels to the lower echelon.

By virtue of this decomposition, we are able to model the dealer subsystem and the transshipment subsystem as decoupled single-stage inventory systems. Both systems have cost functions, lead times, demand functions, and service measures. We measure the dealers' customer service using captured demand (fraction of satisfied customer orders). The metric for the transshipment nodes is the probability of not stocking out when downstream nodes place orders. For the Markovian demand model with one replenishment path, or dual replenishment paths having lead times differing by no more than one period (a day in our case), an order-up-to policy is optimal. This latter result was first proved by Fukuda (1964). For supply chains where this is not the case, order-up-to policies, though not necessarily optimal, have the important advantage of being simple to implement. Based on this fact, Caterpillar decided that order-up-to policies would be appropriate for the P2000.

3.3.1. Dealer Nodes. The dealer nodes face a complex stochastic problem with dual replenishment paths. We use IPA to find the optimal inventory parameters within the class of order-up-to policies, as in Scheller-Wolf and Tayur (1998). This IPA procedure yields either one or two parameters for each item and dealer location—two in the case where the regular and expedited paths do not coincide, and one otherwise. These parameters specify the profit-maximizing levels at which Cat should maintain the *inventory position* (*IP*). By definition, *IP* equals the inventory on hand plus what is on order from the supplier, less what is on backorder to customers. Hence, Cat can change the *IP* value by changing the order quantity. When the inventory position, *IP*, drops below the upper parameter, a regular order is placed to increase *IP* to that level. If *IP* drops below the lower parameter, an expedited order is placed to bring the inventory position up to this lower level quickly, and then a regular order is placed to bring it up to the upper level. The rationale behind this procedure is simple; only when unusually low inventory levels endanger the satisfaction of customer demand is it worthwhile to pay the extra cost to use the expedited channel.

We determine inventory levels that will maximize expected profit. We do this rather than minimizing total expected cost because our captured demand, and thus sales

Figure 2. Sample dealer survey and response.

WORKTOOLS

Simple / Fabricated (buckets, forks, grapple)

How long is a customer willing to wait to purchase?

_85_ % now or never _5_ % 1 day _5_ % 2-4 days _5_ % 5+ days.

How long is a customer willing to wait to rent?

_95_ % now or never _3_ % 1 day _2_ % 2-4 days _0_ % 5+ days.

Complex / Hydromechanical (shear, hammer, power rake)

How long is a customer willing to wait to purchase?

_60_ % now or never _20_ % 1 day _10_ % 2-4 days _10_ % 5+ days.

How long is a customer willing to wait to rent?

_85_ % now or never _10_ % 1 day _5_ % 2-4 days _0_ % 5+ days.

COMPACT MACHINES

What percent sold will be a "standard" versus a "special ordered" configuration?

_90_ % standard _10_ % special.

Standard Configurations

How long is a customer willing to wait to purchase?

_50_ % now or never _20_ % 1 day _10_ % 2-4 days _10_ % 5-10 days _10_ % 11+ days.

How long is a customer willing to wait to rent?

_85_ % now or never _10_ % 1 day _5_ % 2-4 days _0_ % 5-10 days _0_ % 11+ days.

Special Order Configuration

How long is a customer willing to wait to purchase?

_20_ % 1 week _20_ % 2-4 weeks _60_ % 4+ weeks.

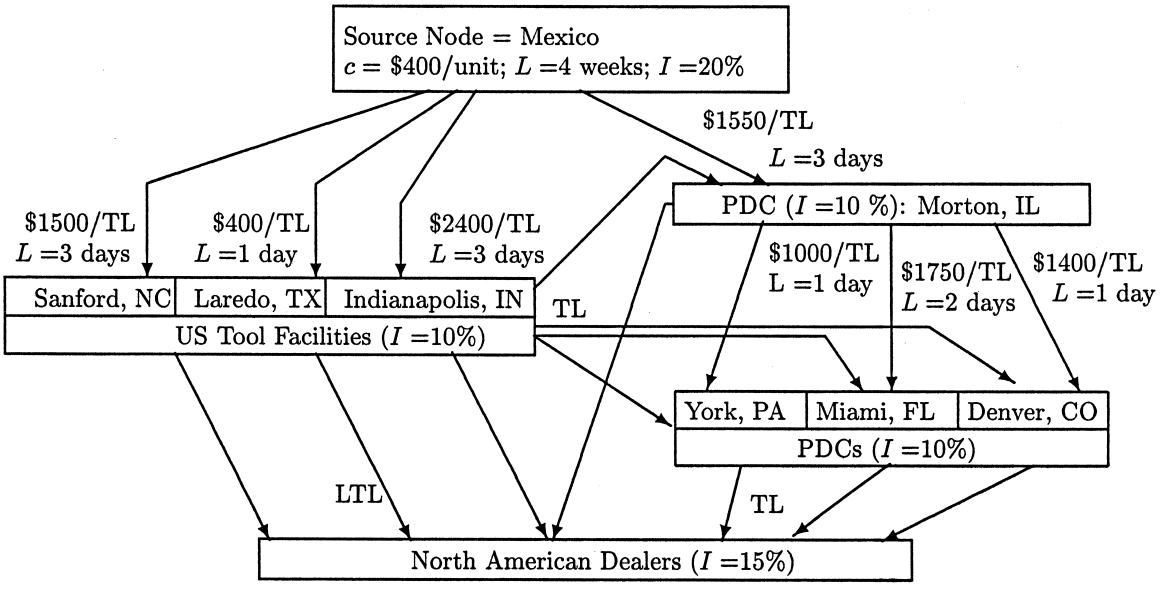
How long is a customer willing to wait to rent?

_95_ % 1 week _5_ % 2-4 weeks _0_ % 4+ weeks.

revenue, depends on the level of service provided. However, even with the objective of maximizing profit, it is conceivable that the optimal inventory parameters could permit an unacceptably large proportion of customers to be lost. Therefore, consistent with Caterpillar's strategy (§2.1), we track the customer service level resulting from our optimal parameters. If the fraction of customers lost is unacceptable to Caterpillar's management, we could incorporate additional penalty costs (beyond the lost revenue) for failing to satisfy customer demands. Making these penalty costs large forces the algorithm to find inventory levels that guarantee arbitrarily high service levels, assuming that the system has sufficient capacity at the source nodes/processing facilities. (In our problem environment, there is a limit on the maximum number of units of each product that a transship node can supply to dealers in any period.)

To efficiently model the likelihood of customers reneging, we used two parameters to incorporate data provided by Caterpillar's dealer surveys. Our parameters were based on an aggregation of dealer surveys such as the one shown in Figure 2. The first parameter captured the probability that a new customer would immediately leave should they not find the product they want in stock. The second parameter models the proportion of waiting customers,

Figure 3. Illustrative transportation data for a typical bucket worktool.



backordered in a previous period, who depart if their demand is not satisfied in the current period. Caterpillar's dealer surveys implied that a large number of customers would immediately leave if unsatisfied, while those who choose to remain undergo a more gradual rate of attrition.

Because both new and old customers renege, the sequence in which waiting customers should be satisfied becomes important. Two sequencing approaches are commonly used: First-Come-First-Served (FCFS) or Last-Come-First-Served (LCFS). Our analysis may be used with either service discipline. We decided to satisfy the more impatient customers first. For Caterpillar's problem, we used Last-Come-First-Served (LCFS) because new customers were more likely to renege than old customers. If the model predicted that a significant number of new customers would be served at the expense of those already waiting, then this LCFS assumption would have to be discussed with Caterpillar management.

3.3.2. Transshipment Nodes. For the transshipment nodes, we use aggregated demand data from the dealers and the approximation methods developed in Glasserman (1997) to compute base-stock levels, and to estimate resulting costs. These methods take into account local demand characteristics and production capacities to specify an ordering parameter that ensures a prescribed service level at each node. Once specified, this single base-stock parameter determines the ordering behavior of the node.

3.4. Problem Data

At each dealer node (and for each product),

D_t = stochastic product demand in period t, with mean μ_t, variance σ_t^2, and $cv_t = \sigma_t/\mu_t$.

L_m = delivery lead time via mode m for $m = r$, e (regular, expedited).

c_m = total unit purchase cost via mode m, including transportation costs.

p = unit selling price.

$h \equiv I \times c$ = unit holding cost, where I = interest rate and c = relevant purchase cost.

(Along each arc, a transportation mode m will be selected (§4.1). The value of c used for computing holding costs is the c_m value corresponding to the selected mode m.)

β_0 = fraction of unsatisfied new customer demand in a period that is immediately lost.

β_1 = fraction of unsatisfied old customer demand in a period that is lost.

At each transshipment node (PDC or TF), a target service level, δ, and a capacity limit, C, are specified, along with holding cost rate, h, unit cost c, and lead time L. The capacity limit, provided by Caterpillar, specifies the maximum number of units of each product that a node can obtain from the supplier in any period.

In addition, as illustrated in Figure 3, for a typical worktool, we have the following:

Product Information: Product name, whether it is a worktool or machine, unit source cost, dealer net selling price, weight and volume, source node ID, and any restrictions on paths from source to dealer. For instance, a typical bucket might have a source cost of $400/unit, a selling price of $500/unit, weight of 450 lbs/unit, with volume 32 × 66 × 25 cubic inches, be sourced from "CMSA" in Mexico, and have the following restrictions (based on quality checks, additional work requirements, and current capabilities at different nodes): No direct shipment from CMSA to dealers permitted in year 2000. In year 2000, CMSA will ship all buckets either to a U.S. Tool Facility or to the Morton PDC for processing before shipment to any other

Table 1. Sample TL transportation cost data ($/TL).

	Sanford	Laredo	Indianapolis	Morton	York	Miami	Denver	Dealer
CMSA	1,500	500	2,500	2,000	∞	∞	∞	∞
Sanford	0	∞	∞	1,000	600	1,400	2,200	1,000
Laredo	∞	0	∞	2,100	3,000	2,500	1,800	4,000
Indianapolis	∞	∞	0	600	500	2,000	1,700	1,700
Morton	∞	∞	∞	0	750	1,750	1,200	2,200
Other PDCs	∞	∞	∞	∞	∞	∞	∞	†

Note: †400 $/TL for York, PA and ∞ for other PDCs (excluding Morton).

location. In 2004, CMSA will be able to ship directly to any location.

In general, tool facilities can ship worktools to each other, to PDCs, and to dealers; but PDCs can only ship to other PDCs, or directly to dealers. As compared with Figure 3, routing networks for worktools and machines sourced from Europe tend to have more nodes and higher lead times.

Transportation Data: Feasible routes, TL, and LTL rates and times (via containers, closed vans, and flatbeds, as applicable), along with whether specified rates are charged per unit, by weight or volume. A TL transportation cost table for a typical bucket worktool is shown in Table 1. The entries specify the dollar cost per truckload at closed van rates over the subnetwork consisting of the source node + tool facilities + PDCs + a typical dealer. LTL transportation cost tables are similar, except that the entries are specified in the units of $/CWT, that is, dollars per "Cent WeighT" (100 lbs). When applicable, similar tables are available for container rates and flatbed TL & LTL rates.

We compute transportation costs per unit of product using TL & LTL rates and either the product's cent weight or volume, whichever is more restrictive. For example, if the LTL rate for Laredo to a dealer is $13.5/CWT, then the LTL transportation cost for a 400-pound worktool is $13.5 \times 400/100 = $54/unit. Now suppose a maximum of one-hundred worktools can be shipped in one truckload, based on permissible weight or volume per TL. Then, if the TL rate from Laredo to this dealer is $4,000/TL, the TL transportation cost is $40/unit. If both LTL and TL modes are available, we select the lower cost option, in this case, TL. Note that availability of the TL option does not necessitate shipment of each product in full truckloads. Essentially, Cat permits the TL option between locations where the total demand volume, summed over all products, is expected to be large enough that the sum total of orders for different products almost always will use up the entire capacity of a truck. (This is an assumption implicit in the transportation data we were given.) Cat also provided tables with LTL and TL transportation *time* data for each pair of nodes in the network. Note that, even if there was only a single mode of transport for each link, there would be multiple paths from source to dealer with different (cost, time) attributes.

Node Information: Minimum order quantities, capacity limits, processing costs and delays, inventory carrying charges, desired service levels, and a list of which products and destinations the node could serve. For example, for a bucket worktool at the Sanford TF node, the minimum order quantity is one; the production capacity is 800 units/week; storage capacity is ∞; the processing cost is $50/unit in 2000 (reduced, presumably due to learning and product/process redesign, to $20/unit in 2004); the processing delay is one day; inventory carrying charges are based on an effective interest rate of $I = 10\%$; and the desired service level (probability of not stocking out) is 95%. Sanford can ship buckets to all other locations in the network.

Demand Data: Mean and variance of demand in 2000 and 2004, for each product at each dealer. We use this data to specify a distribution for simulation of daily demand. For example, estimated year 2000 demand for a bucket worktool at one dealer is 550 units. This translates to a daily demand of 1.757 (based on 313 working days per year), which we model as a normal random variable with mean $\mu = 1.757$ and standard deviation $\sigma = 0.5\mu = 0.8786$ (corresponding to coefficient of variation, cv, of 0.5). Our solution approach is not restricted to a particular value of cv; we use 0.5 as a representative value that was considered reasonable by Caterpillar. Note that, for normal demand, a $cv > 0.35$ generates a significant amount of negative demand, necessitating truncation. Hence, in this case, we used equations from Johnson et al. (1994) to suitably update the demand parameters fed into the truncated normal generator. (In addition, we always confirmed, empirically, that the mean of the generated truncated normal demand was equal to the input mean, μ.) At a different dealer, the forecast average annual demand is 78 units, yielding a mean daily demand of $\mu = 0.249 < 1$. We model this demand using the Bernoulli distribution with probability of nonzero demand in any day set at $p_d = 0.249$. Thus, daily demand is either zero or one, with mean 0.249 and standard deviation $\sqrt{p_d(1 - p_d)} = 0.4326$, resulting in a suitably higher coefficient of variation of 1.736.

Customer Patience Parameters: β_0 and β_1, respectively, the proportion of unsatisfied customers who renege immediately and in each period thereafter. Dealer surveys (see Figure 2) established these parameters. Typically, we use $\beta_0 = 0.4$ to 0.75 and $\beta_1 = 0.15$; more service sensitive regions such as the Northeastern U.S. have higher values of β_0. The dealer surveys implied worktools should have a greater β_0 than machines because Caterpillar's worktools were considered substitutable with those of their competitors. Preliminary sensitivity analysis also was conducted on different β values to confirm that small deviations from the chosen β did not significantly affect system performance.

Furthermore, the IPA derivative estimates of performance measures with respect to β also were computed during the simulation to assess if these derivative values were unacceptably large.

4. ANALYSIS AND IMPLEMENTATION

4.1. The Product Routing Model

For each product, we determine the lowest cost path from the source to each dealer by solving a deterministic network problem. Let E^m be the set of arcs that permit use of transport mode m, $\mathcal{S}$ be the unique source node for the product, D denote the set of dealer nodes with mean daily demand d_j for $j \in D$, and T be the set of transship nodes. Let x_{ij}^m be the flow from node i to node j using mode m, and let c_{ij}^m denote the corresponding cost of a unit flow. In our model, $c_{ij}^m = a_{ij}^m + n_j + I_i L_{ij}^m y_i$, where a_{ij}^m is the unit transport cost from i to j by mode m with corresponding lead time L_{ij}^m, n_j is the unit node processing cost at j, I_i is the inventory carrying cost rate at node i, and y_i is the minimum total unit product cost from $\mathcal{S}$ to i. Then the product routing problem can be formulated as:

$$\min_{x_{ij}^m \geq 0} \sum_{i,j,m} c_{ij}^m x_{ij}^m$$

$$\text{s.t.} \quad \sum_m \sum_{j:(\mathcal{S},j)\in E^m} x_{\mathcal{S}j}^m = \sum_{j\in D} d_j,$$

$$\sum_{(i,j)\in E^m} x_{ij}^m - \sum_{(j,k)\in E^m} x_{jk}^m = 0 \quad \text{for all } j \in T,$$

$$\sum_m \sum_{i:(i,j)\in E^m} x_{ij}^m = d_j \quad \text{for all } j \in D.$$

In the absence of arc capacity constraints, the optimal extreme point solution to the above linear programming problem has the following characteristics: (1) On each arc (i, j), we use, at most, one of the modes, $m^* = \text{argmin}_m c_{ij}^m$, corresponding to the lowest cost mode with lead time of $L_{ij}^{m^*}$. (2) There is, at most, one positive incoming flow into each node, corresponding to the min cost path from $\mathcal{S}$ to the node. (3) The arcs with positive flows define a spanning tree rooted at $\mathcal{S}$, with leaves at each dealer in D; this defines a unique path from $\mathcal{S}$ to each dealer. Based on these observations, the formulation can be simplified by replacing the x_{ij}^m flow variables on arc (i, j) with one x_{ij} corresponding to the lowest cost mode. Further, without loss of generality, each positive demand d_j may be replaced by 1, because the lowest cost path for flow into dealer j will remain unchanged. Thus, once the arc costs c_{ij} are specified (based on the lowest cost mode), the problem is reduced to finding the lowest cost path in the network from $\mathcal{S}$ to each dealer $j \in D$. This is facilitated by LP duality.

Let E denote the union, over all modes m, of arcs in E^m. Then the dual of our product routing problem is:

$$\max_{y_j \geq 0} \Big\{ \sum_{j\in D} y_j \Big| y_j \leq y_i + c_{ij} \quad \text{for all arcs } (i, j) \in E; \; y_{\mathcal{S}} = 0 \Big\},$$

where, as defined earlier, y_j is the minimum total sourcing and transportation plus pipeline inventory cost of moving one worktool from the supplier $\mathcal{S}$ to node j. This dual may be solved efficiently using Dijsktra's shortest path algorithm (refer to Lawler (1976) for relevant theory on network optimization). The only difference between the product routing dual and a standard shortest path problem is that the arc lengths are not constant in our dual; the pipeline inventory portion of c_{ij} depends on the value of the decision variable y_i. This does not pose any computational difficulties because, in our implementation of Dijsktra's algorithm, we process the nodes in a specific order obtained using a topological sort (Aho et al. 1983) of the underlying directed acyclic graph. That is, we renumber the nodes of the supply chain so that, for every arc from i to j, the index of i is smaller than the index of j. If nodes are processed in increasing order of their index, the absence of directed cycles in our supply network allows us to completely determine y_i before c_{ij} is calculated, eliminating any potential difficulties.

The model above sets each dealer's regular supply node equal to the immediate predecessor of the dealer in the min-cost path from the source to the dealer. The lead time for regular deliveries is the time for shipment from this immediate predecessor node to the dealer, under the assumption that the predecessor carries sufficient inventory to provide a high level of service. For cases in which the predecessor node is allowed to carry no inventory, the delivery lead time observed by the dealer is increased appropriately.

We determine inventory levels at transship points after the ordering policies at all dealers are specified. Consequently, the use of supply nodes by different dealers is accounted for when transshipment inventory levels are set.

4.2. The Inventory Model

We use separate models, which are both stochastic, for dealer nodes and transshipment nodes. This is motivated by the difference in service level definitions at dealer nodes (which face response-sensitive customer demand) and transshipment nodes (where demand comes from captive dealers). Thus, the dealer model must incorporate the possibility of lost demand, whereas the transshipment node model just backlogs excess demand.

4.2.1. The Dealer Model. Data from the routing model serves as input to the dealer inventory model, which maximizes the total expected profit, where profit equals revenue minus regular and expedited sourcing costs minus on-hand inventory carrying costs. The unit sourcing cost from node i is y_i plus additional transportation, pipeline inventory, and node processing costs incurred between node i and the dealer. The dealer's regular supply node is determined by the product routing model. The transshipment node, i, with the smallest lead time for direct material flow to the dealer is the supply node for expedited deliveries.

The variables and features of the inventory model include:

- I_{t-1} = inventory *level* at end of period $t-1$ = on-hand inventory − backlog.
- X_τ^m = order placed in period $\tau < t$, for delivery via mode m, for $m = r$ (regular), or $m = e$ (expedited). We store past orders for $\tau = t-1, \ldots, t-L_m$. All orders that have been placed but not yet delivered contribute to the in-transit or pipeline inventory.
- $P_{t-1} = \sum_{m=r,e} \sum_{\tau=t-L_m}^{t-1} X_\tau^m$ = total pipeline inventory at end of period $t-1$.
- $IP_{t-1} = I_{t-1} + P_{t-1}$ = inventory *position* at end of period $t-1$.
- $R_t = X_{t-L_r}^r + X_{t-L_e}^e$ = receipts in period t.

The sequence of actions in period t is:

Step 1. Determine beginning inventory level, I_{t-1}, and pipeline inventories, X^r, X^e.

Step 2. Receive delivery of relevant pipeline inventory, R_t.

Step 3. Observe demand D_t.

Step 4. Satisfy as much demand as possible from on-hand inventory; a portion of unfilled demand is lost.

Step 5. Place new replenishment orders, X_t^e and X_t^r.

Step 6. Update profit.

We elaborate on Steps 4 and 5 below:

Step 4. Inventory allocation: We use inventory to satisfy the most impatient demand first. For our data set, $\beta_0 > \beta_1$, so we satisfy new demand before demand from previous periods is satisfied (LCFS).
Lost demand in period t: Let $x^+ = \max(0, x)$ and $x^- = (-x)^+$. Then

$$\mathcal{L}_t^0 = \beta_0 (D_t - I_{t-1}^+ - R_t)^+ \text{ and}$$
$$\mathcal{L}_t^1 = \beta_1 (I_{t-1}^- - (R_t - D_t)^+)^+, \quad (1)$$

where $\mathcal{L}_t^0$ and $\mathcal{L}_t^1$ denote, respectively, the portion of unfilled new demand and waiting customer orders in period t that are lost. (If FCFS were used, $\mathcal{L}_t^0 = \beta_0 (D_t - (I_{t-1} + R_t)^+)^+$ and $\mathcal{L}_t^1 = \beta_1 (I_{t-1}^- - R_t)^+$.) With $\mathcal{L}_t = \mathcal{L}_t^0 + \mathcal{L}_t^1$, the ending inventory level is

$$I_t = I_{t-1} + R_t - D_t + \mathcal{L}_t. \quad (2)$$

For example, if $\beta_0 = 0.6$, $\beta_1 = 0.15$, $I_{t-1} = -20$, $R_t = 40$, and $D_t = 50$, then $\mathcal{L}_t^0 = 0.6(50 + 0 - 40)^+ = 6$ and $\mathcal{L}_t^1 = 0.15(20 - 0)^+ = 3$. Thus, as new demand exceeds receipts by ten, six of these ten units will be lost in addition to three of the backlogged 20 units of past demand. Total lost demand is $\mathcal{L}_t = 9$. On the other hand, if $R_t = 60$ in the above example, then $\mathcal{L}_t^0 = 0$ (no new demand is lost) and $\mathcal{L}_t^1 = 1.5$.

Step 5. Order-up-to policy for replenishment: Because $c_r \leq c_e$ and $L_r \geq L_e$, it follows that $z_r \geq z_e$. Thus, the inventory position after order placement will always be $IP_t = z_r$. If the inventory position prior to ordering is $IP_{t-} = IP_{t-1} - D_t + \mathcal{L}_t$, the expedited order quantity which orders-up-to z_e is

$$X_t^e = (z_e - IP_{t-})^+ = (z_e - z_r + D_t - \mathcal{L}_t)^+. \quad (3)$$

Assuming $z_r \geq z_e$, the regular order quantity is

$$X_t^r = \min(z_r - z_e, (z_r - IP_{t-})^+) = (D_t - \mathcal{L}_t - X_t^e)^+. \quad (4)$$

Sales in period t is

$$\begin{aligned} S_t &= \min(I_{t-1}^+ + R_t, D_t - \mathcal{L}_t + I_{t-1}^-) \\ &\overset{x^+ = x + x^-}{=} \min(I_{t-1} + R_t, D_t - \mathcal{L}_t) + I_{t-1}^- \\ &\overset{(2)}{=} \min(I_t, 0) + D_t - \mathcal{L}_t + I_{t-1}^- \\ &= I_{t-1}^- - I_t^- + D_t - \mathcal{L}_t. \end{aligned}$$

Period t profit is

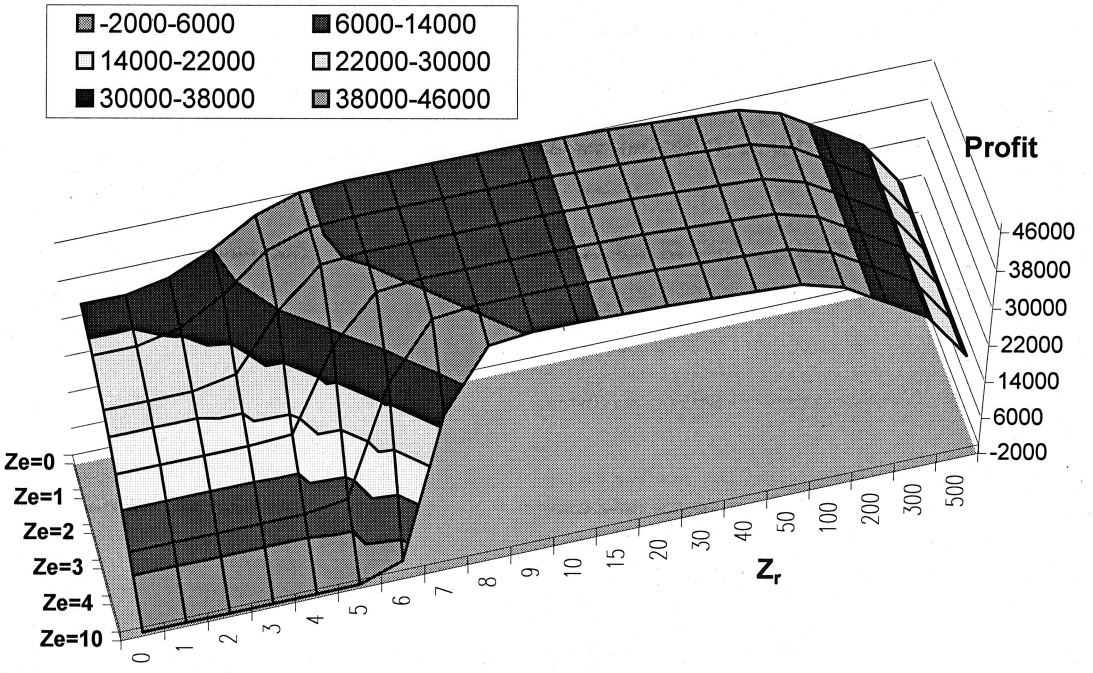
$$\pi_t = pS_t - hI_t^+ - c_e X_t^e - c_r X_t^r. \quad (5)$$

We also measure the long-run fraction of demand that is lost and the fraction of demand satisfied using the regular and expedited modes, respectively. We solve the dealer inventory model by selecting a starting value of (z_r, z_e) and generating a set of k demand scenarios. From these scenarios, we compute the estimated expected profit $\pi_t(z_r, z_e)$ and the IPA derivative estimates $d\pi_t/dz_e$ and $d\pi_t/dz_r$ according to their recursions (shown in Appendix A). We then use a subgradient-based search to find the optimal value of (z_r, z_e). Because we are using simulation, a proof of joint concavity of the profit function with respect to the parameters is desirable. This is difficult to prove for our problem, and, therefore, is the subject of parallel work. If there were no lost sales, concavity would be relatively straightforward to show using induction on Equations (1)–(5). Our computer experiments, illustrated in Figure 4, indicate that the profit function is concave when the base-stock levels are high (which eliminates lost sales). While always unimodal over $z_r \geq z_e$, the profit does fail to be jointly concave at low (z_e, z_r) values, where significant numbers of customers are lost.

Based on Figure 4 (which we generated for several problem parameters) and the concavity of profit in the absence of a loss function, we believe that, for the high service levels we use, the profit function is likely unimodal in the region of interest. Assuming the dealer profit function is, in fact, unimodal, our IPA procedure converges to the optimal values of z_r and z_e. Refer to Scheller-Wolf and Tayur (1998) for further details.

Prior to embarking on our IPA optimization, we conducted the standard practice of removing initialization bias and checking for "steady-state," based on pilot runs. As expected, simulation estimates of the optimal expected profit become more accurate as the number of simulation iterations (demand scenarios) increases. However, these profit estimates differed by no more than 0.2% over a range of simulation iterations between $k = 1{,}000$ and $k = 10{,}000$, while simulation run times increased by an order

Figure 4. SSL GP-bucket year 2000 profit function for a typical dealer.



of magnitude. Therefore, to strike a balance between computational time and accuracy, we generated 10,000 demand scenarios for the last two simulation runs, just before termination of the search for optimal (z_r, z_e), and used $k = 3{,}000$ demand scenarios during the IPA search.

4.2.2. The Transshipment Node Model. Using output from the product routing and dealer inventory models, we determine the fraction of customer demand satisfied at each dealer using the regular and expedited modes. By aggregating these product flows over all dealers we determine the mean, μ, and variance, σ^2, of daily demand for each product at each transshipment node. Given these parameters, the capacity C, desired service level δ, and the lead time L that the transshipment node faces for delivery from its source, the base-stock level at a transshipment node is set to

$$z = \left[(L+1)\mu + \frac{\sigma^2}{2(C-\mu)} \right] + \Phi^{-1}(\delta)\left[(L+1)\sigma^2 + \left(\frac{\sigma^2}{2(C-\mu)} \right)^2 \right]^{1/2}. \quad (6)$$

The first bracketed term accounts for the mean demand over the lead time and the mean shortfall, while the second term corresponds to safety stock (incorporating demand and shortfall variability). See Glasserman (1997) and Tayur (1992) for details. Because demand at each transshipment node is the sum of many demands originating at different dealers, we assumed that the cumulative distribution function (CDF) of demand at the transshipment node during its delivery lead time could be approximated by a normal distribution $\Phi(\cdot)$. We denote the inverse of the CDF by $\Phi^{-1}(\cdot)$. This demand model ignores correlation between demands at different transshipment points. If correlation effects or deviations from the normal distribution are significant, we could use a more accurate, but computationally intensive simulation model similar to the dealer model in §4.2.1.

Inventory holding costs at each transshipment node are estimated as $h[z - \mu(L + 0.5)]$. This approximation is fairly standard (see, for instance, formula (5-1) in Hadley and Whitin 1963). Costs over different transship nodes are aggregated to obtain estimated total costs for this subsystem. This cost then is subtracted from the simulation-based estimate of dealer profit for each product. This yields the product's contribution to expected profit. Total system profit is the sum of the profit for all of the products.

5. RESULTS

For each product, solution of the routing problem took just a few seconds on a Sparc20 workstation. This yields the min-cost (regular) path from the source to each dealer in the network. For expedited deliveries, we identified the transshipment location that had the shortest delivery time to the dealer. This was instantaneous. The average runtime to compute the optimal inventory levels (z_r, z_e) for each product was just under 40 seconds per dealer. Calculation of inventory levels at the transshipment nodes using Equation (6) was instantaneous. We considered 21 dealer districts, so complete analysis of one product over all dealers took approximately 21×40 seconds or 14 minutes. Cat provided comprehensive data for 21 products over two years (2000 and 2004). Consequently, one run over all

Table 2. Percentage of optimal profit across different scenarios.

Year	TF & PDC	PDC-Only	TF-Only	DS-Only
2000	100.00	96.76	88.99	77.58
2004	100.00	97.98	89.38	81.23

products, dealers, and study years took approximately 9.8 hours. For each scenario tested, we typically ran our experiments in under five hours on two computers working in parallel on independent sets of products or study years.

5.1. General Results

We considered several scenarios consisting of sets of permissible transshipment nodes. The four primary scenarios were TF & PDC, PDC-Only, TF-Only, and DS-Only, for which we generated solutions that were optimal within the class of order-up-to policies for each of these networks. A solution maximizes expected profit by setting order-up-to levels for each of the products, locations, and both the regular and expedited modes. These inventory levels ensure that a sufficiently high proportion of customers are served. Our reported results include:

1. Optimal total profit for each scenario.
2. The relative profit from machines and worktools.
3. The geographical distribution of profit percentages and the contribution of each product.
4. The market capture percentage (100% − lost sales%).
5. The breakdown of cost components (source cost, node costs, transportation costs, pipeline, and safety stock costs).
6. Optimal transportation modes for each link in the supply chain.
7. Product delivery lead times.
8. The effect of demand volume changes from 2000 to 2004.

As noted earlier, scenario costs exclude the fixed costs associated with construction of the tool facilities. Therefore, it comes as no surprise that the greatest expected profit is attained by using the entire network, as shown in Table 2. From the table, we observe that TFs will benefit less from the increased demand volume in 2004. This may be explained by noting that the linear transportation cost component dominates the sublinear inventory cost component for the TF-Only scenario.

Table 2 shows that inclusion of Tool Facilities (TFs) adds about 2–3% to profit from worktools, as compared to using the PDC-Only scenario, which is on the order of several million dollars. Nevertheless, this was outweighed by the estimated costs of building and operating the tool facilities. Therefore, Caterpillar decided to implement the use of the PDC-Only alternative (plus the Sanford TF due to routing constraints) along with direct shipments. Our model indicates that this supply chain configuration will capture almost 100% of the demand. This negligible lost sales is a consequence of the fact that the incremental inventory holding costs are significantly smaller than the lost profit due to shortages. This leads to large dealer inventories primarily supported by the regular mode, with the expedited mode used only occasionally during high demand periods. This behavior can be seen in Figure 4, which illustrates how profit changes with z_e and z_r for a representative worktool. For this example, the optimal levels are $z_r = 6, z_e = 0$, but many higher values of z_r, z_e (e.g., $z_r = 6, z_e = 1$) result in near-optimal profit with little or no expediting. However, using $z_r = 3$ instead of the optimal $z_r = 6$ reduces profit by more than 18%, primarily due to a decrease in customer service from 100 to 81 percent.

Since demand forecasts form an integral part of our optimization, we assessed the sensitivity of our recommendations to variations in input demand data. We studied performance measures (profit, inventory, and service levels) for each scenario at four distinct mean demand levels: (1) Caterpillar's forecast μ, (2) $0.8 \times \mu$, (3) $1.2 \times \mu$, and (4) $U(0.8, 1.2) \times \mu$, where $U(a, b)$ denotes a uniform random variate between a and b and each product's mean demand is multiplied by a different realization of $U(0.8, 1.2)$. In all cases, the relative profitability of the different scenarios remained remarkably insensitive to changes in mean demand. We chose 0.8 and 1.2 after discussions with Caterpillar. Caterpillar felt that a demand greater than 1.2 times the forecasted mean was unrealistic, given the aggressive nature of their target. On the downside, if the demand was less than 0.8 of the forecast, then a strategic decision would be made on price and advertising that would lead to a new analysis of the situation. Furthermore, if deemed necessary by Caterpillar, we were prepared to rerun the model for several values of mean demand to better estimate the robustness of our recommendations.

We also note that each mean demand is an aggregation of mean demands for different products/dealers (e.g., the product type we call fork sets is actually comprised of several distinct, but similar, fork sets; each dealer district is an aggregation of several dealers). It is widely accepted that such aggregate forecasts are likely to be more reliable, and so a ± 20 percent variation on the mean, in general, may be an acceptable range of analysis.

5.1.1. The Refined Model. Because the PDCs performed well for worktools, Caterpillar decided to consider an alternative scenario in which PDCs (and TFs) also were permitted to act as transshipment points for machines. Their hypothesis was that, if machines were allowed to flow through the PDCs and TFs, the worktools would benefit from lower transportation costs resulting from cheaper modes of transport normally available only to machines. This necessitated the generation of new data that specified TL transportation costs and times via closed vans *and* flatbeds, charged at new rates for worktools. Preliminary analysis showed that this would *not* increase total profits. The added node costs incurred for processing machines overwhelm any savings from combined worktool and machine

transportation. In addition, while PDCs were already well equipped to handle worktools, this was not the case for machines. Hence, Caterpillar decided to use the results of the base model, which decoupled the distribution of machines and worktools.

5.2. Revenues, Costs, and Profits of First Model

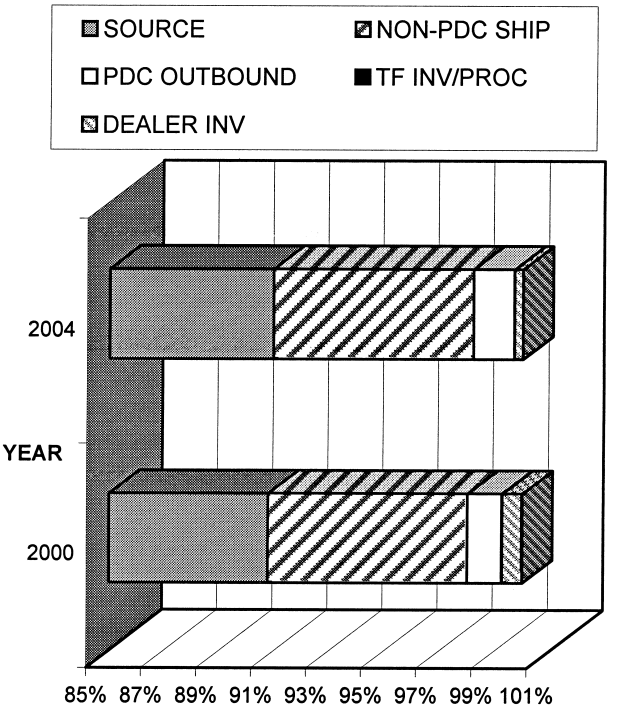
The remainder of this section illustrates output analyses conducted to provide a better understanding of the results of our study. Caterpillar was interested in the breakdown of costs by components such as source costs, transportation costs (separated by PDC and non-PDC costs), pipeline inventory costs, and on-hand inventory costs. This information served a variety of purposes. For example, Cat was considering different contracts with dealers, some of which included Caterpillar's ownership of "consignment" inventory and/or a manufacturer buy-back option. Thus Cat might be responsible for a portion of the dealer inventory costs. In addition, the PDC costs and profits were required to estimate the appropriate transfer payment to the PDC logistics group, should they be included in the supply chain. Consequently, these expected costs were categorized as:

1. Source (cost charged by external supplier);
2. Non-PDC Ship (non-PDC pipeline inventory and transportation costs);
3. PDC Outbound (costs incurred after some PDC took receipt and control of material);
4. TF Inv/Proc (on-hand inventory and node processing costs at TFs); and
5. Dealer Inv (costs of on-hand inventory at dealers).

We illustrate this cost breakdown in Figure 5 for the PDCs plus Sanford TF scenario. The bulk (90%) of costs are source costs; non-PDC pipeline inventory and transportation costs form a substantial portion (approximately 7%) of the remaining 10%; PDC-Outbound is between 1% and 1.5%; and dealer inventory costs are under 1% of total costs. This comparatively low dealer inventory cost comes despite the high service levels and the use of longer lead times (more regular shipments than expedited). The TF cost portion is negligible ($<0.02\%$) because the Sanford TF is not used for most worktools, but is required for processing a select few. Together with Figure 7, Figure 5 shows that inventory and costs shift from the dealer to the pipeline, with an increase in demand volume and additional new routes in 2004. The increased demand volume does not require much increased floor space or inventory investment at the dealer; however, it does require the capability to handle larger volumes throughout the distribution network.

Our model's output also includes regional profits. Figure 6 shows that the top six districts account for more than 60 percent of total profit from projected year 2004 worktool sales in North America. Only 18 dealer districts are shown instead of 21 because "Other Canada" actually consists of four districts including Northeast (Newfoundland and Nova Scotia) and Southeast (Montreal and Toronto) Canada. Similar graphs illustrate the breakdown by volume of material flow through distribution centers, as well as the total profit generated by each worktool. These can be used to prioritize operations by product and/or region, via an ABC type analysis.

Figure 5. Worktool cost breakdown by percentage.



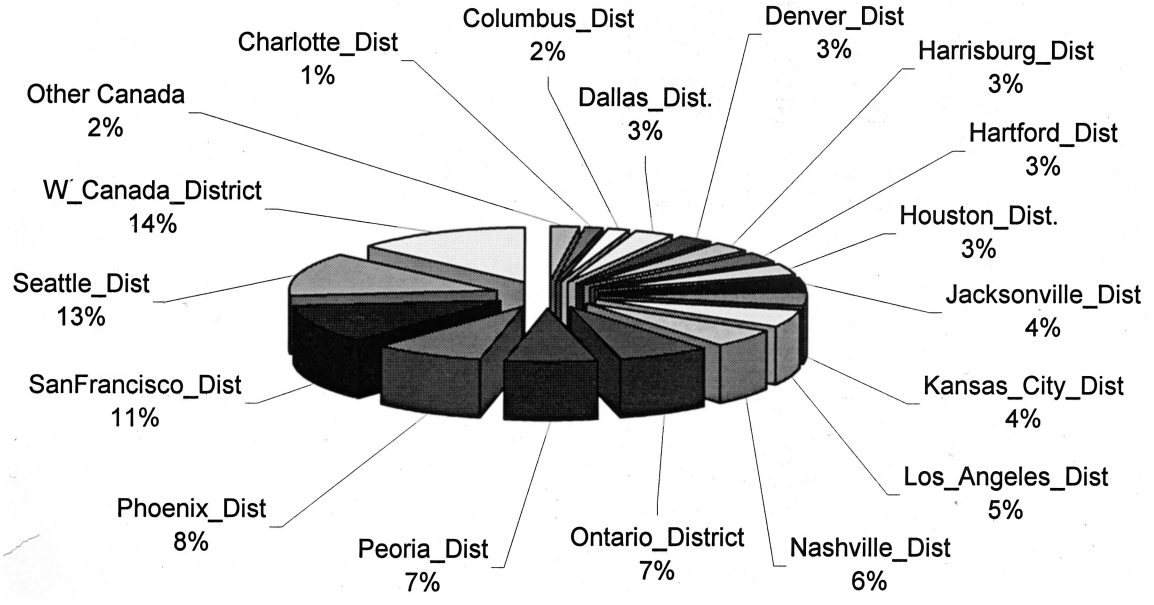
Note: Ninety percent of the total cost is due to manufacturing (termed "source" cost), approximately 7% of cost is due to transportation and pipeline inventory prior to arrival at PDCs, 1.5% of cost is due to transportation and inventory at the PDC and in transit from PDCs to dealers, and the remainder is made up in inventory holding costs at the dealers. The inventory carrying costs at the TF are negligible.

5.3. Inventory Levels

We now focus on the location and magnitude of inventory within the supply chain, and illustrate our observations using the TF & PDC scenario. Figure 7 demonstrates that most of the inventory is in transit, the transshipment nodes carry very little inventory, and the dealers carry most of the on-hand inventory (cycle stock and safety stock). In year 2000, expected total demand volume for SSL tools is about four times the demand volume for CWL/MHE tools. However, SSL tools have less relative demand variability, so their safety stock is not proportionately larger than that for CWL/MHE tools. Similar numbers hold for 2004.

We also studied the breakdown of each of these inventories by product. For year 2000, the detailed breakdown of average dealer on-hand inventory of 811 SSL worktools and 320 CWL/MHE worktools is shown in Table 3. Because Caterpillar was considering different contracts with

Figure 6. Year 2004 worktool profits by region.

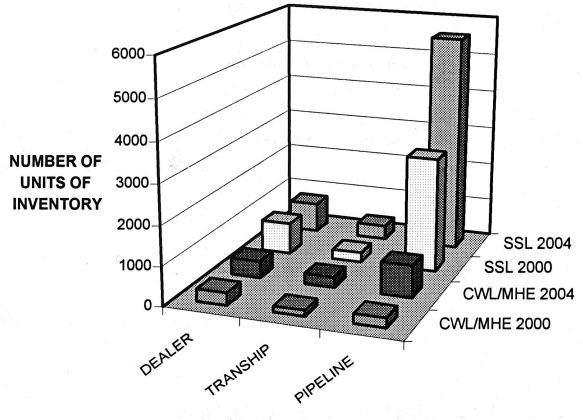


dealers, as mentioned above, these inventory breakdowns were of considerable importance. They also provide a first estimate of which products could most benefit from improved logistics. For example, there are more SSL GP buckets in inventory on average than the combined total of all of the CWL and MHE worktools that were studied.

A similar analysis of pipeline inventory (say, in terms of use of TL and LTL rates and expedited vs. regular shipments) over the network is possible from our results, but is not included. We also graphed the worktool and machine inventory by nondealer locations. The latter is shown in Figure 8, which illustrates output from our analysis of the supply chain for machines.

To summarize, our model evaluates different supply chain configurations along the dimensions of profit, captured demand, inventory parameters, and transportation mode usage. From these evaluations we obtain optimal inventory routes and levels corresponding to order-up-to policies with dual supply modes. These evaluations proved to be robust with respect to changes in system parameters such as demand intensity, transportation options, and customer impatience.

Figure 7. Worktool inventory location within supply chain.



	DEALER	TRANSHIP	PIPELINE
CWL/MHE 2000	320	158	247
CWL/MHE 2004	435	275	824
SSL 2000	811	252	2924
SSL 2004	766	358	5590

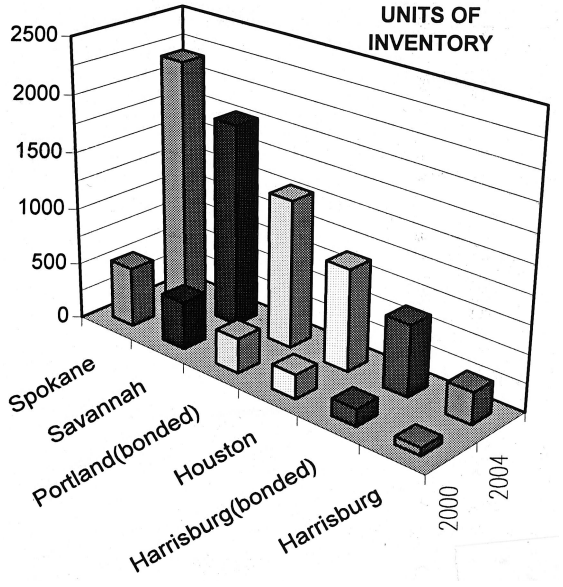
6. CONCLUDING REMARKS

In this paper, we develop an integrated model to analyze different supply chain configurations for Caterpillar's new line of compact construction equipment, the P2000 series. We use decomposition techniques, network optimization theory, inventory modeling, and simulation theory. The novel features of our model include dual modes of supply for dealer replenishments and net customer demand that is responsive to speed of service. We were able to make recommendations to Cat on the effects of different factors on profits. In the last quarter of 1998, Caterpillar launched

Table 3. Dealer inventory by worktool type 2000.

SSL Tool	Units	CWL/MHE Tool	Units
Brooms	64	Brooms	20
GP Bucket	338	GP Bucket	67
MP Bucket	58	MP Bucket	60
Grapples	63	Grapples	64
Hammers	63	Hammers	59
Augers	67	Light Material	17
Derivatives	84	Special Dump	33
Fork Sets	74		
Total	811	Total	320

Figure 8. CWL machine volumes.



its P2000 line, supporting it with the supply chain—Part Distribution Centers (PDCs) plus Sanford Tool Facility—recommended by our analysis.

Without our analysis, Caterpillar likely would have implemented either the Direct Shipment-Only (DS-Only) or the Tool Facilities-Only (TF-Only) option. Internal considerations caused Cat to question the value of inclusion of PDCs into the network. Disregarding fixed costs of TF construction, the annual benefit of our solution over TF-Only is roughly eight percent of the maximal expected profit, which is several million dollars. This comparison does not capture the full benefit of our project. It assumes that Caterpillar would have used the optimal inventory in its implementation of the TF-Only option, which we, in fact, specify. This is significant because our IPA optimization indicates that choosing the correct inventory ordering parameters can be vital. In our problem, setting the levels too low increases lost sales and requires a greater use of expediting, resulting in significantly lower profits, as shown in Figure 4. The actual benefit of our project thus is likely to be significantly greater than eight percent.

The work reported in this paper has concentrated on the North American market. A similar analysis applies to Caterpillar's European market.

APPENDIX A. IPA DERIVATIVE RECURSIONS

The derivative recursions for $d\pi/d(z_e)$ and $d\pi/d(z_r)$ (used by the gradient-based search for optimal z_e and z_r) are:

$$\frac{d\pi_t}{dz_e} = p\mathbf{1}\{S_t > 0\}\frac{dS_t}{dz_e} - h\mathbf{1}\{I_t > 0\}\frac{dI_t}{dz_e} - c_e\mathbf{1}\{X_t^e > 0\}\frac{dX_t^e}{dz_e} - c_r\mathbf{1}\{X_t^r > 0\}\frac{dX_t^r}{dz_e}. \quad \text{(A1)}$$

$$\frac{dS_t}{dz_e} = -\mathbf{1}\{I_{t-1} < 0\}\frac{dI_{t-1}}{dz_e} + \mathbf{1}\{I_t < 0\}\frac{dI_t}{dz_e} - \mathbf{1}\{\mathcal{L}_t > 0\}\frac{d\mathcal{L}_t}{dz_e}. \quad \text{(A2)}$$

$$\frac{dI_t}{dz_e} = \frac{dI_{t-1}}{dz_e} + \frac{dR_t}{dz_e} + \frac{d\mathcal{L}_t}{dz_e},$$

$$\text{with } \frac{dR_t}{dz_e} = \frac{dX_{t-L_e}^e}{dz_e} + \frac{dX_{t-L_r}^r}{dz_e}$$

(known from previous iterations). (A3)

$$\frac{dX_t^r}{dz_e} = -\frac{d\mathcal{L}_t}{dz_e} - \frac{dX_t^e}{dz_e}$$

(information from (A4) and (A5) will be used in future iterations.) (A4)

$$\frac{dX_t^e}{dz_e} = \mathbf{1}\{z_e > IP_{t-}\}\left(1 - \frac{dIP_{t-}}{dz_e}\right) = \mathbf{1}\{z_e > z_r - D_t + \mathcal{L}_t\}\left(1 - \frac{d\mathcal{L}_t}{dz_e}\right). \quad \text{(A5)}$$

$$\frac{d\mathcal{L}_t}{dz_e} = \frac{d\mathcal{L}_t^0}{dz_e} + \frac{d\mathcal{L}_t^1}{dz_e}, \text{ where } \frac{d\mathcal{L}_t^0}{dz_e} = \beta_0\left(-\mathbf{1}\{I_{t-1} > 0\}\frac{dI_{t-1}}{dz_e} - \frac{dR_t}{dz_e}\right) \cdot \mathbf{1}\{D_t - I_{t-1} - R_t > 0\} \text{ and } \frac{d\mathcal{L}_t^1}{dz_e} = \beta_1\left(-\mathbf{1}\{I_{t-1} < 0\}\frac{dI_{t-1}}{dz_e} - \mathbf{1}\{R_t > D_t\}\frac{dR_t}{dz_e}\right) \cdot \mathbf{1}\{\mathcal{L}_t^1 > 0\}. \quad \text{(A6)}$$

The order of derivative computation is in the reverse order of the listing above. We only track derivatives of I_t and X_t^k (which yield all other required derivatives of R_t, $\mathcal{L}_t$, etc.). Derivative recursions w.r.t. z_r are identical in form, except for item (A5) above, which becomes

$$\frac{dX_t^e}{dz_r} = \mathbf{1}\{z_e > z_r - D_t + \mathcal{L}_t\}\left(-1 - \frac{d\mathcal{L}_t}{dz_r}\right).$$

The validity of these simulation-based derivative estimates follows from arguments similar to Glasserman and Tayur (1995).

ACKNOWLEDGMENTS

We thank Dr. George Cusack at Caterpillar, Inc., Peoria, for providing us with the opportunity to work on this problem, and for his insightful comments. All numbers contained in this article have been disguised in accordance with a contractual agreement. We also thank the associate editor and two anonymous referees for many suggestions that greatly improved the content and presentation of this paper.

REFERENCES

Aho, A., J. Hopcroft, J. Ullman. 1983. *Data Structures and Algorithms*. Addison-Wesley, Reading, MA.

Barnhart, C., R. R. Schneur. 1996. Air network design for express shipment service. *Oper. Res.* **44**(6) 852–863.

Birge, J., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer-Verlag, Berlin.

Feigin, G. 1998. Inventory planning in large supply chains. S. Tayur, R. Ganeshan, M. Magazine, eds. *Quantitative Models for Supply Chain Management*. Chapter 24, Kluwer Academic Publishers, Boston, MA, 761–787.

Fukuda, Y. 1964. Optimal policy for the inventory problem with negotiable leadtime. *Management Sci.* **12** 690–708.

Glasserman, P. 1997. Bounds and asymptotics for planning critical safety stock. *Oper. Res.* **45**(2) 244–257.

——, S. Tayur. 1995. Sensitivity analysis for base-stock level sin multiechelon production-inventory systems. *Management Sci.* **41**(2) 263–281.

Hadley, G., T. M. Whitin. 1963. *Analysis of Inventory Systems*. Prentice Hall, Englewood Cliffs, NJ.

Horst, R., H. Tuy. 1993. *Global Optimization*. 2nd Ed., Springer-Verlag, Berlin.

Johnson, N. L., S. Kotz, N. Balakrishnan. 1994. *Continuous Univariate Distributions*. Wiley Series in Probability and Mathematical Statistics, New York.

Kapuscinski, R., S. Tayur. 1998. A capacitated production- inventory model with periodic demand. *Oper. Res.* **46**(6) 899–911.

Keskinocak, P., J. Swaminathan, S. Tayur. 1998. Solution of a multi-commodity network flow problem with non-convex costs. Manuscript in preparation.

Lawler, E. 1976. *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart and Winston, New York.

Lee, H., C. Billington. 1993. Material management in decentralized supply chains. *Oper. Res.* **41**(5) 835–847.

Marsh, P. 1998. Caterpillar digs into compact market. *The Financial Times*, *U.S. Edition*, February 12.

Scheller-Wolf, A., S. Tayur. 1998. Optimal policies for dual supplier contracts with order bands. Working Paper. GSIA, Carnegie Mellon University, Pittsburgh, PA.

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

——, R. Ganeshan, M. Magazine. 1998. *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers, Boston, MA.

Weimer, D. 1998. Strategies: a new cat on the hot seat. *Business Week*, March 9, 56–61.

Other Links of Interest

Home Page

www.informs.org

Membership

www.informs.org/General/Memben2.html

Publications

www.informs.org/Pubs/

Meetings

www.informs.org/Conf/

Public Relations

www.informs.org/Press/