Development of a Rapid-Response Supply Chain at Caterpillar

Uday Rao Alan Scheller-Wolf Sridhar Tayur

Graduate School of Industrial Administration, Carnegie Mellon University

5000 Forbes Avenue, Pittsburgh, PA 15213-3890

July 1998

Abstract

As part of its growth strategy, Caterpillar Inc. plans to launch a new P2000 product line of "compact" construction equipment and work tools. In anticipation of this, they asked the authors to construct and analyze potential P2000 supply chain configurations. Through the use of decomposition, and results from network flow theory, inventory theory, and simulation theory, we were able to provide a set of solutions to this problem, corresponding to different supply chain scenarios (provided by Caterpillar). Based on our recommendations, Caterpillar made their decision regarding the P2000 supply chain configuration.

Subject classifications: Professional: OR/MS Implementation. Inventory/Production: Applications. Industries: Machinery.

Area of review: OR Practice.

1 Introduction

We describe an application of Operations Research techniques to support 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, inventory theory, and simulation based optimization (Infinitesimal Perturbation Analysis or IPA, see Glasserman and Tayur (1995)) to arrive at a solution. Apart from some standard features such as multiple-echelons, capacity constraints, uncertain demand, lead times and multiple products, the nature of Cat's problem required our model to have the following novel features:

1. We permit entities in our distribution network to order from dual suppliers: A low cost (regular) alternative and a high speed (expedited) supplier. The determination of the optimal replenishment paths for each (dealer,product) pair is accomplished through a deterministic minimization of cost (or time) over the supply network.

2. We allow the magnitude of captured demand to be sensitive to service response time: In each time period, the fraction of lost sales depends on the customer service provided. A certain fraction of new customers are impatient and renege if not immediately served, while another fraction of waiting customers are lost if forced to wait longer.

To our knowledge, this is the first time a problem of this scope has been solved in this manner; in particular, the use of IPA to establish inventory levels in an industrial problem of this magnitude and complexity appears to be unprecedented.

Using the above techniques, we built an optimization engine for use in designing Caterpillar's supply chain. In what follows, we will describe the building of this tool, along with the methods used to collect data for input into the engine, and the results of a sensitivity analysis of output from the model. Specifically, after describing the problem environment in Section 2, we provide details of the modeling and analysis in Sections 3 and 4 respectively. The results of our efforts are reported in Section 5, after which we present some concluding remarks in Section 6. Throughout the paper, whenever we report numerical data, its values have been disguised to maintain confidentiality.

2 Caterpillar

In an effort to exploit anticipated growth in the small construction industry, Caterpillar Inc., the world's leading producer of construction and mining equipment, made the decision to introduce a new "compact" product line, the P2000, starting in 1999. This decision has been widely reported in the media, with articles appearing in *The Financial Times* (Feb. 12, 1998) and *Business Week* (Mar. 9, 1998), for example.

One reason this project has been so widely covered is that the P2000 represents not only a new product line, but also a new business 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), while the P2000 family will encompass smaller, medium to high volume, standardized products (selling for as little as $20,000 per machine). Furthermore, while Caterpillar is a well-known leader in the heavy equipment business with few large competitors (such as Komatsu Ltd.), the compact product segment has many entrenched market leaders such as BobCat (trademark of the Melroe Mfg. Company, a unit of Ingersoll-Rand), Deere & Co., and Case Corp. Adding to the interest surrounding the P2000 launch are the well-publicized work force difficulties at Caterpillar and the recent Asian financial crisis, both of which increase the uncertainty of future profitability and demand. (While it was initially hypothesized that collapsing Asian currencies would lead to a flood of cheap products from Asian rivals, it has become clear that most companies do not have the capital to launch an export attack.)

Before embarking on this new product introduction a careful analysis of the supply chain design was required; Cat knew that they needed a network which could maximize profits while also capturing market share and providing flexibility. Cat also recognized that the P2000 family likely would not fit well in their current supply chain. They contacted the authors of this study for assistance in determining a configuration which could fulfill these goals for two study years: An initial year, 2000, and four years later, in 2004. Although the product is to be launched in 1999, the first year was considered a "ramp-up" year and the demand in 2000 was considered more suitable for supply chain design and analysis. The two study years differed in the volume of forecast demand, the value of cost parameters and the routing restrictions (which were generally relaxed as new facilities and new processing capabilities were developed).

Caterpillar's new product line will consist of several models of three different *machines* (a Skid-Steer Loader, "SSL", a Compact Wheel Loader, "CWL", and a Mini-Hydraulic Excavator, "MHE") and some forty *work tools* (including items such as Buckets, Fork Sets and Grapples).

The work tools, which are designed for use with one or more particular machines (such as a Skid-Steer Loader), can be sold as attachments to *competitor's* machines as well as those produced by Caterpillar. They thus provide a means to enter the market independent of P2000 machine sales. The P2000 products are to be sourced, manufactured, and assembled in approximately twenty locations throughout North America and Europe. Key production centers for machines are Sanford, North Carolina, and Leicester, England, with specific machines assigned to each production center. Work tools come from locations in the United Kingdom, the United States, Mexico, Sweden, Germany and Finland, with each work tool having 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.

Caterpillar provided demand forecasts, by dealer district, for all of the machines and work tools for both of the study years (mean with degree of variability), as well as customer patience data gathered through surveys of their dealers. These surveys were an effort to understand the behavior of customers forced to wait for delivery of an item, and hence measure the cost of such delays through customer attrition. They implied that customers renege at a rate which depends upon how long they have been waiting, thus establishing a relationship between lost sales and level of service.

Caterpillar also furnished cost and logistical data, including source costs, source capacities, minimum order sizes, holding cost rates at the different locations, handling costs, and shipping times and rates for different combinations of transportation modes. The nodes in the distribution network were completely specified for machines. A set of potential transshipment locations (intermediary nodes between the source and the dealer) had been determined for work tools. In our initial analysis, Caterpillar stated that work tools and machines would not share transshipment nodes. This requirement was relaxed in later models, where shipment of machines and work tools could be done either separately or together, at varying capacities and rates.

2.1 P2000 Strategy

Caterpillar's focus on their supply chain configuration was motivated by specific concerns: The international nature of the P2000 supply chain, coupled with the considerable weight of the equipment, created the potential for both lead times and shipping costs to be large. Prior to the introduction of the P2000 Caterpillar had made the decision not to compete on price; rather, in keeping with their core philosophy, quality and service were targeted as areas in which Cat would distinguish its products. Long lead times were thus particularly worrisome. The dealer surveys

reinforced this concern, by implying that Cat's future products would be highly substitutable with those of the competition – primarily Bobcat.

Caterpillar thus believed it crucial that they capture a very high percentage of customer demand for Caterpillar products as soon as it materialized, and not force potential customers to wait for delivery. By establishing a reputation for product availability, Cat felt that they would not only generate demand for their products, but also possibly be able to satisfy demands for their competitors' products, if these competitors failed to keep sufficient quantities of products in stock. It therefore became important not simply to find the minimum cost channel for a product in the supply chain, but to also provide an additional channel for expedited delivery to the dealer, likely at a higher cost, should inventory levels drop precipitously. Cat did not want to underestimate the value of quickly getting inventory to dealers, and hence to new customers. This was the genesis of the dual supply modes within the supply chain.

Cat also wanted demand to be modeled in a way such that poor customer service resulted in lower effective demand (sales). This was in keeping with their goal of maximizing profit subject to capturing a satisfactory portion of market demand: For strategic reasons, Cat would consider accepting a lower profit solution (for year 2000) to gain "market presence," since the revenue benefits of market penetration were not easily quantifiable at the same level as the cost analysis. To this end, we were prepared to develop a trade-off curve between year 2000 profit and market penetration, if appropriate.

2.2 The Work Tool Problem

The complex nature of the P2000 line dictates that some final manufacturing be done within the supply chain, forcing certain work tools to pass through selected centers or sets of centers. Similarly, the presence of facilities unique to the international import/export trade, such as bonded warehouses (which permit Caterpillar to forgo paying duties while storing their products in-transit), force certain items coming from overseas to pass through selected customs locations before distribution to dealers. With these factors in mind, Caterpillar determined that they could use up to seven additional transshipment locations in North America to facilitate distribution of work tools to the dealers, in addition to direct shipment (DS) from sources.

The seven possible transshipment locations for work tools 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 were already built and in use handling other Caterpillar products). The Tool Facilities were to potentially 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, there were other transshipment locations in our supply chain that were not among the alternatives to be considered for inclusion into, or exclusion from, the network. For instance, there was a UK tool facility in Leicester, a UK PDC in Desford, and a European tool facility in Belgium (with a PDC in Grimbergen), that fed work tools made in Europe to the North American PDCs or TFs.

There were four primary options for the construction of the North American work tool supply chain:

1. Use of all TFs and PDCs.
2. Use of PDCs only.
3. Use of TFs only.
4. Use of neither TFs or PDCs; allowing only direct shipment (DS) of work tools.

Secondary options included using the PDCs with one or two supporting TFs, or vice-versa. For example, instead of using only PDCs one might add the Sanford Tool Facility to this set, due to a need to perform certain manufacturing operations or quality check procedures on selected work tools within the supply chain.

For each of these four scenarios, we were asked to determine:

1. Supply path(s) from each work tool's source to every dealer region in Caterpillar's network.
2. Inventory levels and ordering policies along all such 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.

2.3 The Machine Problem

In our initial model of the supply chain, replacement work tools and machines could be treated as independent entities that did not interact with each other. The distribution network for machines was completely fixed and included the two manufacturing plants and five North American storage facilities: The Sanford, NC, plant manufactured only SSL machines, the Leicester, UK, plant was responsible for CWL and MHE machines. The five storage facilities, used exclusively for machines, were: Houston, TX, Savannah, GA, and Harrisburg, PA, which served the US market,

and the bonded warehouses at Portland, OR, and Harrisburg, PA, which served the rest of North America.

Given this distribution network, all that was required for this portion of the problem was to determine the best routes from each machine's source to all of the dealer regions and the optimal inventory levels at each storage facility and dealer. This corresponds to a single work tool scenario. Therefore, we concentrate on the work tool problem, although we provide illustrations of some of the results for machines in Section 5. Note that we did consider an extension of the initial model in which work tools and machines were permitted to share distribution resources, thereby charging work tools lower transportation costs whenever they were shipped together with machines. Results for this model are summarized in Section 5.1.1.

3 Modeling

While there have been many papers addressing aspects of material flow management, fewer papers have considered modeling entire supply chains. We approach our problem in a spirit similar to Lee and Billington (1993) and Feigin (1998). Lee and Billington (1993) 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 develop a model for a single site in the network. Feigin (1998) uses different techniques and approximations to conduct a similar analysis on the tradeoff between service levels and inventory investment in large supply chains.

Finding optimal solutions to the individual components of our system (such as the material routing subsystem or the inventory replenishment subsystem), is in itself extremely difficult. The *deterministic* version of the routing problem, a min-cost multi-commodity network flow problem with non-convex costs, is the subject of a parallel work by Keskinocak *et al* (1998). Similarly, the determination of optimal policies for the inventory problem with expedited orders, and the use of IPA to find such optimal levels within the class of order up-to (or base stock) policies, is considered by Scheller-Wolf and Tayur (1998). Taken together, the exact solution of our problem, which includes sub-problems of these types, is likely unobtainable.

To make the 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 while also conducting 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 decomposed the problem into a network routing problem (Section 3.2) and a stochastic inventory problem (Section 3.3). The routing problem ignores safety stock levels throughout the network. Consequently, it may overlook some inventory risk pooling opportunities arising from utilizing a more expensive route. However, for our problem the relative cost of inventory is small as compared to the transportation costs (see Figure 4). More importantly, without this decomposition, the problem falls in the realm of stochastic nonlinear integer programming (Horst and Tuy (1993), Birge and Louveaux (1997), for which efficient approaches for a problem of this scale are not currently available.

Network Decomposition The network was decomposed into dealer nodes and transshipment nodes, permitting us to solve the resulting subproblems as single stage inventory systems. This decision was motivated by the accuracy of known approximations (Glasserman (1997) and Tayur (1992)) and the need to quickly evaluate multiple products over different scenarios. If required, we could have treated the entire network as one multi-echelon inventory system using techniques similar to those described in Section 3.3. The theoretical base for such a multi-echelon model is provided by Glasserman and Tayur (1995).

Dealer Aggregation We solved the problem based on one "typical" aggregate dealer within each of 21 regions in North America, rather than individually considering all 190 dealers. This was deemed acceptable, because: (i) the aggregated dealers were roughly uniformly distributed within the region. (ii) the variation in costs within a region will likely be dominated by the sourcing and transportation costs up to the region. (iii) Caterpillar decided that any increase in cost resolution resulting from using a more detailed model would probably be voided by the comparatively less exact estimates of individual dealer demand and logistics costs. The use of a more detailed network structure poses no conceptual problem with regard to our methodology. It will increase the computational times.

Disaggregation of Sourcing Caterpillar uses a unique source for each work tool.

Decomposition by Products: As a "first model", we decoupled the sourcing of different products throughout the network – work tools from each other and also from machines. This allowed us to solve the routing and inventory problem for each machine and work tool

independently, and then aggregate the results. This decoupling disregards the possible dependencies between demands for particular products; this was acceptable to Caterpillar in light of the fact that initially work tools and machines were to use disparate distribution networks.

In our "second model" we did consider situations where work tools and machines could share transportation: Work tools were permitted to use the flat-bed transportation mode normally used by the machines, at rates, times and capacities different from the normal closed-van transportation mode for work tools. Section 5.1.1 describes the results of this generalization in more detail.

Demand Modeling The demand itself was modeled in one of two ways. If the estimated mean daily demand for a product was less than one unit, daily demand was modeled as a Bernoulli random variable with the appropriate parameter. For products where the daily demand was anticipated to exceed one unit, demands were modeled as being drawn from a truncated Normal distribution, again with appropriately selected parameters. The data we were provided did not suggest any specific demand distribution and Caterpillar agreed that our demand models were acceptable. We conducted preliminary tests with alternative demand distributions, including the uniform and exponential. These did not change the overall optimal network configuration.

The mean, variance of demand at transshipment points was obtained from the mean, variance of demand at dealers served by this transship node. The distribution of transship node demand was approximated by a Normal based on the fact that transship node demand was an aggregate over several dealers. Furthermore, demand at different dealer districts and for different products was taken to be uncorrelated and time stationary. This level of analysis was acceptable to Caterpillar. If they had desired (and been able to provide data on correlation and seasonality), we could have incorporated this in our simulation based optimization model. In particular, see Kapuscinski and Tayur (1998).

Lost Sales Based on information provided by Caterpillar dealers, we used a two parameter model for customer impatience (Section 3.3.1). This model was adopted because the data indicated that customers fell into two categories: those who left immediately and those who were 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 (which was borne out by our final results) indicated that the percentage of lost sales was small. This resulted from the fact that, for Caterpillar's problem, the inventory

cost for a work tool 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 behavior for our problem data.

Continuity For simulation purposes, inventory was taken to be continuous. Due to the convex nature of the inventory model's costs, after arriving at an optimal inventory level this continuity assumption could be relaxed by searching the adjacent integral values.

In summary, this paper presents a general methodology applied to a specific problem at Caterpillar. At the core of this methodology lie simulation based recursions of the type presented in Appendix A. Many of the model enhancements mentioned above would have resulted in similar recursions for which the proposed IPA technique would have remained valid.

Our specific model was developed in an iterative fashion and involved periodic interaction with Caterpillar. As new data was provided and new features needed, the model was updated. The ultimate result is a recommendation to Caterpillar regarding the configuration of their supply chain. To the extent possible, we investigated alternatives to our assumptions and found that they did not affect the final recommendation (ref: section with $\mu \pm 20\%$ experiment). In its incorporation of uncertainty over such a large and complex problem, we believe our technique provides both an adequate solution and a tool for what-if analyses. The question of whether our specific model or even Caterpillar's data is an accurate representation of the real problem to be faced by Cat is a valid one. In the absence of comparable models and solution techniques, we are unable to answer this conclusively, except to state that Caterpillar is currently implementing a solution based upon our recommendations and sensitivity analyses.

Details of uur model are presented in the following section.

3.2 The Product Routing Model

After these reductions, for each product and dealer combination we modeled the supply chain as a collection of nodes (for sources, dealers and transshipment points) and edges (connecting the nodes). Each edge had an appropriate lead-time and cost component: Overseas shipment was done on freighters at container rates, while shipment within North America was done by closed vans or flatbed trucks, at either truckload (TL) or less-than-truckload (LTL) rates, depending on the product, source and destination, weight, and volume.

In a similar manner each node had times and costs associated with it. Inventory costs were assessed at varying rates for different locations and products throughout the network, as were

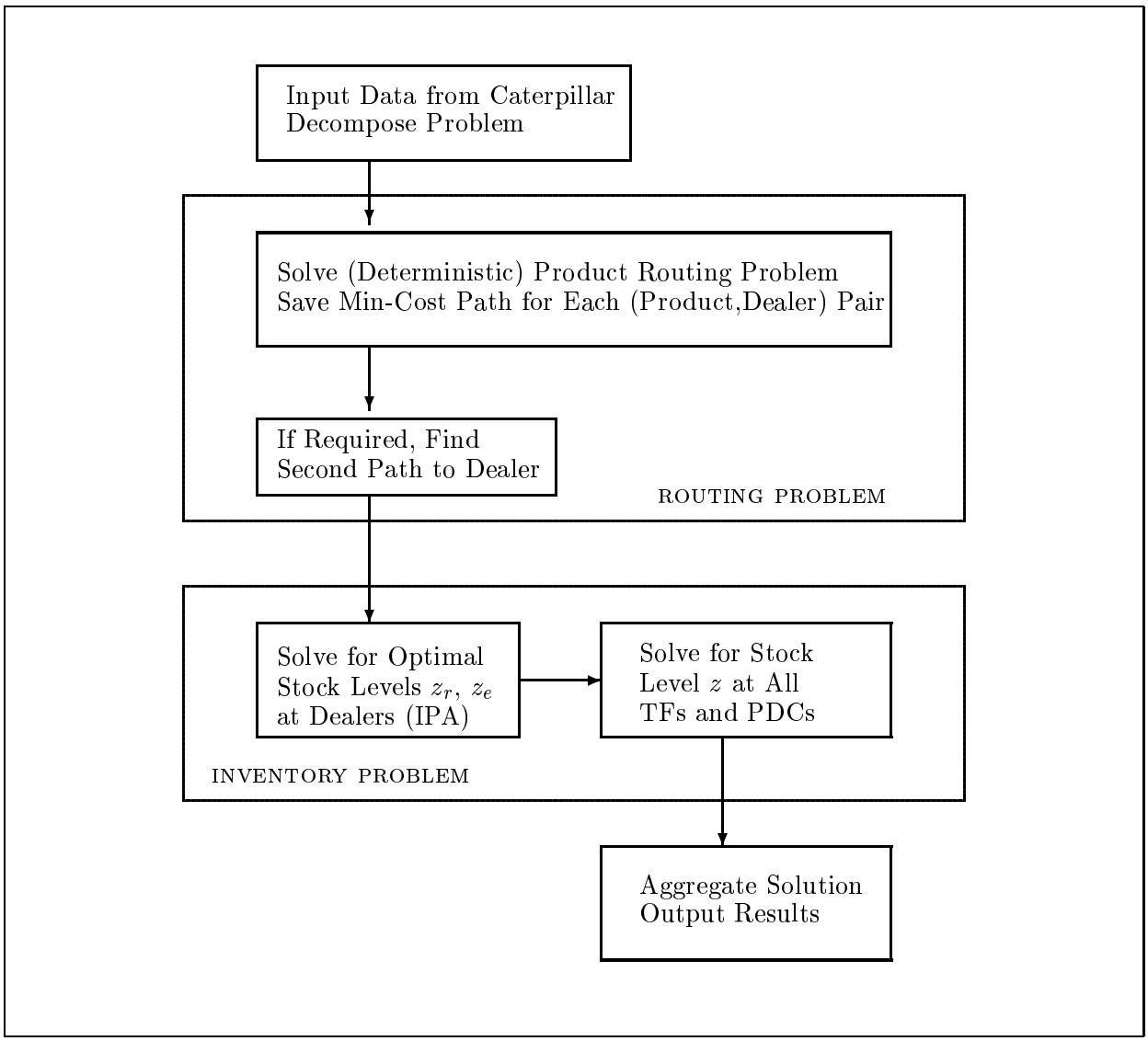


Figure 1: Flow Chart Outlining Problem Decomposition and Solution Steps

handling times and costs. Certain nodes were precluded from holding any inventory – they were treated as instantaneous transshipment points.

In order to accomplish the dual objectives of the model – minimal cost with a simultaneous high service level – for each product and dealer node of the supply chain up to two paths were found within the network. The first was the minimum cost path, while the second was the path with the smallest delivery lead time along the link from the dealer to the next level up in the supply chain. This node at the next level up was linked to the source node by a min-cost path. The two supply paths were called the *regular* and *expedited* paths, respectively. In the case that the last link in the minimum-cost path also proved to have the shortest delivery lead time for direct shipment from any transshipment point to the dealer, only this single dominating path was used. Otherwise, the regular path was intended to be used for the majority of shipments of a product, while the expedited link would be utilized in situations where intense demand had caused inventory to drop below a certain minimal level, determined by the inventory optimization portion of the algorithm.

3.3 The Inventory Model

After decomposing the model by products we further simplified the problem by decomposing the multi-echelon inventory system for each product into two subsystems, one consisting of the dealers (who face uncertain customer demand and receive regular or expedited material deliveries from suppliers) and another consisting of the source node plus transshipment nodes. This decomposition was based on the assumption that service levels at transship points would be high enough to avoid stock-out occurrences. Experiments, based on Glasserman and Tayur (1995), confirmed this assumption, implying that imposing the decomposition a priori would not result in increased costs or reduced service. These 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, these experiments showed that the approximation tended to decrease inventory levels at the lower echelon. Upper echelon inventory levels could be higher or lower under the approximation, but in either case they were very close under both systems implying similar service levels to the lower echelon. As applied to our problem, the decrease in lower echelon inventory does not adversely affect customer service since inventory costs at the dealers are small relative to the lost profit from shortages, and virtually a 100% customer service level is attained. If the above were not the case, our methodology could have been applied to the entire multi-echelon inventory system, at the expense of the increased

computational times required to jointly optimize the vector of base stock levels at all nodes in the system.

By virtue of this decomposition, we were able to model both the dealer subsystem and the transshipment subsystem as decoupled single-stage inventory systems with appropriate cost functions, capacity constraints, lead times, demand functions and service measures. (While the dealer's customer service was measured using captured demand, the transshipment nodes' service was measured as a probability of not stocking out when downstream nodes placed orders.) It is known that for the Markovian model, for the cases of a single replenishment path or dual replenishment paths having lead times differing by no more than one period - one day in our case - an order up-to policy will be 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 very simple to implement. It was thus decided that this would be an appropriate policy to adopt at Caterpillar.

3.3.1 Dealer Nodes

The dealer nodes face a complex stochastic problem with dual replenishment paths. Therefore, we used IPA to find the optimal inventory parameters within the class of order up-to policies, see Scheller-Wolf and Tayur (1998). The IPA procedure yields either one or two parameters for each item and dealer location. (Two in the case where there are two unique paths for the item, i.e., when the regular and expedited paths do not coincide, and one otherwise.) These parameters specify the levels that the *inventory position* (inventory on hand plus what is on order from the supplier, less what is on backorder to customers) should maintain to maximize expected profit. If the inventory position drops below the upper parameter, then a regular order is placed to replenish it. If it drops below the lower parameter then an expedited order is placed to bring it up to an acceptable level, and a regular order is placed to bring it up to the optimal level. The rationale behind this procedure is simple: Only when the inventory is unusually low, and dealers are in danger of failing to satisfy customer demand, is it worthwhile to pay the extra cost to secure goods along the expedited channel.

As mentioned above, the determination of the inventory levels is done in a way so as to maximize expected profit. We maximize profit instead of minimizing total expected cost since our captured demand (and thus sales revenue) depends on the level of service provided. However, even with the profit maximizing objective it is conceivable that the optimal inventory positions could permit an unacceptably large proportion, say $\geq$ 30 %, of customers to be lost. Hence,

consistent with Cat's strategy (Section 2.1), we tracked the customer service level resulting from our optimal inventory positions, and if the fraction of customers lost proved to be too high, we were prepared to incorporate an additional penalty cost (above and beyond the lost revenue) for failing to satisfy customer demands. By making this penalty cost arbitrarily large, we could force the system to find inventory levels guaranteeing service levels of any desired magnitude (assuming the system had sufficient capacity).

To model the likelihood of customers reneging if their demands were not satisfied in a period, we fit two parameters to the customer patience data provided by Caterpillar. The first captured the likelihood that a new customer would immediately leave should they not find the product they wanted in stock. The second parameter was used to model the proportion of already waiting (backordered) customers who would depart if their demand was not satisfied in the current period. By using two parameters were were able to efficiently approximate the data from the dealer surveys; this data implied that a large number of customers would immediately leave if unsatisfied, while those who choose to remain undergo a somewhat regular rate of attrition. A sample dealer survey is presented in Figure 2, along with the response of a typical dealer. Our estimation of attrition rates were based on an aggregation of such dealers. (Note that in this paper we do not consider the product rental business and only focus on purchased products).

Customer reneging raises the issue of the sequence in which waiting customers should be satisfied: First-Come-First-Served (FCFS) seems the most fair, but if customers who are already waiting are less likely to renege than newly arrived customers, from a strictly analytical viewpoint it makes more sense to satisfy new customers first. Since our analysis may be used with either service discipline, we decided to satisfy the more impatient customers first. That is, if new customers were more likely to renege than waiting customers we used Last-Come-First-Served (LCFS), otherwise we used FCFS. If the model predicted that a significant number of newly arrived customers would be served at the expense of those waiting (or vice versa), then this LCFS/FCFS assumption on the service discipline would need to be discussed further with Caterpillar management.

3.3.2 Transshipment Nodes

For the transshipment nodes we used aggregated demand data from the dealers to compute base stock levels and estimate resulting costs for each product, via the approximation methods developed in Glasserman (1997) and Tayur (1992). These methods use equation (1), from Section 4.2.2, to specify an ordering parameter which ensures a prescribed service level at each

Work Tools

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.

Figure 2: Sample Dealer Survey & Response

node, taking into account local demand characteristics and production capacities. Once this single parameter of the base stock policy is specified, the ordering behavior of the node is determined.

3.4 Problem Data

At each dealer node (and for each product),

D_t = product demand in period t, random with mean μ_t, variance σ_t^2, and $cv_t = \sigma_t/\mu_t$;
L_k = delivery lead time via mode k for $k = r, e$ (regular, expedited);
c_k = unit purchase cost via mode k;
p = unit selling price;
$h \equiv I \times c$ = unit holding cost, where I = interest rate, and c = effective purchase cost;
β_0 = fraction of unsatisfied new customer demand in a period that is immediately lost;
β_1 = fraction of other waiting customers at the end of each period whose demand 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. This capacity limit refers to maximum number of units which can be obtained from the supplier in any period. This information was provided by Caterpillar.

In addition, as illustrated in Figure 3, for a typical work tool we have the following:

1. *Product Information* such as product name, whether it was a work tool or machine, unit source cost, dealer net selling price, weight and volume, source node ID and any restrictions on paths for material flow 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 $32 \times 66 \times 25$ 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 all buckets from CMSA were to be shipped either to a US Tool Facility or to the Morton PDC for processing before shipment to other transshipment nodes or dealers. This requirement was to be relaxed in 2004 by allowing CMSA to ship directly to any of the PDCs.

   In general, tool facilities could ship work tools to each other, to PDCs and to dealers, but PDCs could only ship to other PDCs or directly to dealers. Routing networks for work tools and machines sourced from Europe tended to have more nodes and higher lead times.

2. *Transportation Data* including feasible routes, TL and LTL rates and times (via containers,

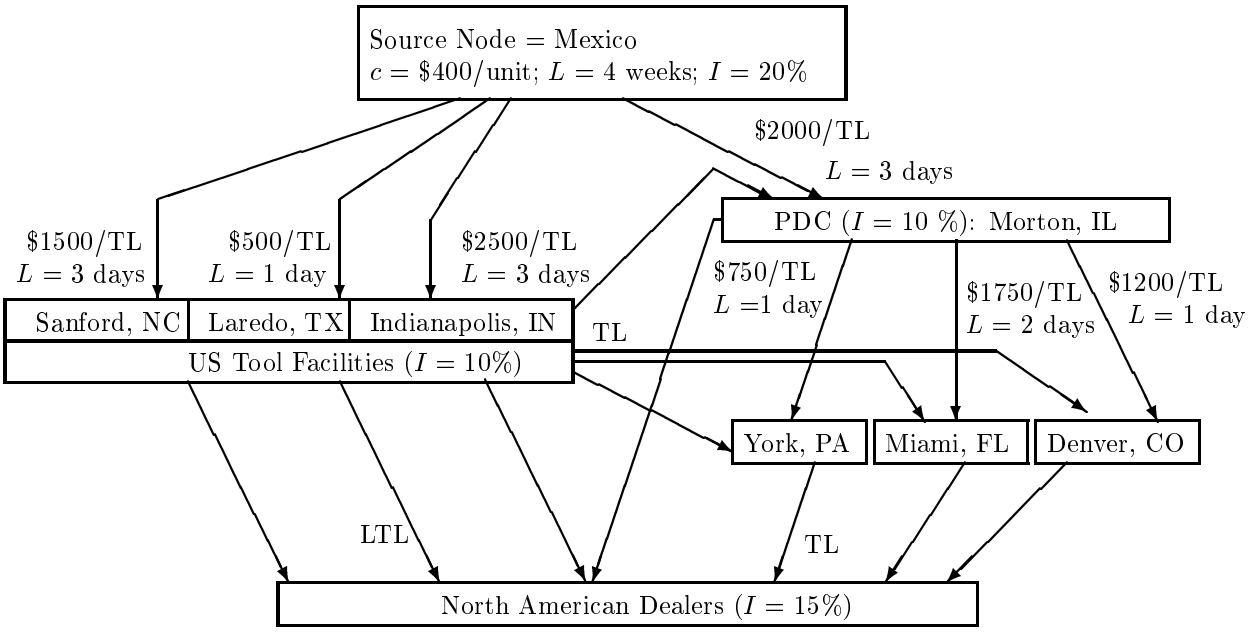


Figure 3: Illustrative transportation data for a typical "bucket" work tool (see Table 1 for all cost data)

closed vans and flat beds, as applicable), along with information on whether specified rates were charged per unit, by weight or volume. A TL transportation cost table for a typical bucket work tool is shown below in Table 1, where the entries specify the dollar cost per truck load at closed van rates, over the subnetwork consisting of the source node + tool facilities + PDCs + typical Dealer. LTL transportation cost tables were similar, except that the entries were specified in the units of \$ / CWT, that is, dollars per cent weight (100 lbs). When applicable, similar tables were available for container rates and flat bed TL & LTL rates.

Table 1: Sample TL Transportation Cost Data (\$/TL)

	Sanford	Laredo	Indianapolis	Morton	York	Miami	Denver	Dealer
CMSA	1700	600	2600	1750	∞	∞	∞	∞
Sanford	0	∞	∞	1000	600	1400	2200	1000
Laredo	∞	0	∞	2100	3000	2500	1800	4000
Indianapolis	∞	∞	0	600	500	2000	1700	1700
Morton	∞	∞	∞	0	1200	1950	1600	2200
Other PDCs	∞	∞	∞	∞	∞	∞	∞	†

† 400 \$/TL for York, PA and ∞ for other PDCs (excluding Morton).

Transportation costs per unit of product were computed using TL & LTL rates and the product's cent weight or volume, whichever was considered more restrictive. For example:

(i) If the LTL rate for Laredo to Dealer was \$13.5/CWT, then the LTL transportation cost per unit for a 400 lb work tool was $\$13.5 \times 400/100 = \$54$/unit. (ii) If a maximum of 100 work tools could fit in one truck load (based on permissible weight or volume per TL) and the TL transportation cost rate from Laredo to this Dealer was \$4000/TL, then the unit TL transportation cost would be \$40/unit. If both modes were permitted, the lower cost option (in this case TL) would be used. In addition, tables with LTL and TL transportation *time* data for each pair of nodes in the network were also specified. Note that even with a single mode of transport for each link, there are multiple paths from source to the dealer with different (cost,time) attributes.

3. *Node Information* comprised of minimum order quantities, capacity limits, processing costs and delays, inventory carrying charges, desired service level measures, and a description of what products could flow through each node to which destinations. For example, for a bucket work tool at the Sanford TF node, the minimum order quantity was 1; the production capacity was 800 units/week; storage capacity was ∞; the processing cost was \$50/unit in 2000 (which was reduced, presumably due to learning and product/process redesign, to \$20/unit in 2004); the processing delay was one day; inventory carrying charges were based on a effective interest rate of $I = 10$ %; desired service level (probability of not stocking out) was set to 95 %.

4. *Demand Data* consisted of mean and variance (or coefficient of variation) of demand, in 2000 and 2004, for each product at each dealer. From this data, a demand distribution was chosen for simulation of daily demand. For example, estimated year 2000 demand for a bucket work tool at one dealer was 550 units. This translated to a daily demand of 1.757 (based on 313 working days per year), which was modeled 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). However, at a different dealer the forecast average annual demand for buckets was only 78 units, yielding a mean daily demand of $\mu = 0.249 < 1$; demand was expected to occur relatively infrequently. Hence, this demand was modeled using the Bernoulli distribution with probability of non-zero demand in any day set at $p_d = 0.249$. Thus daily demand was thus 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.

5. *Other Data* included the customer patience parameters, β_0 and β_1 (respectively, the proportion of unsatisfied customers who renege immediately and in each period thereafter). These values were chosen based on the dealer survey (Figure 2) and preliminary sensitiv-

ity analysis of the effects of different β values on system performance. Typically, we used $\beta_0 = 0.4 - 0.75$ and $\beta_1 = 0.15$; higher values of β_0, e.g. $\beta_0 = 0.7$, were used for more service sensitive regions such as the North Eastern US. In addition, consistent with the dealer survey, a higher value of β_0 was used for work tools than for machines (since Cat's work tools were considered completely substitutable with competitors', higher customer impatience was anticipated).

Before we present our results, we discuss details of the analysis.

4 Analysis & Implementation

4.1 The Product Routing Model

We solve the product routing problem using network flow theory, in particular, Dijsktra's shortest path algorithm (refer to Lawler (1976) for relevant theory on network optimization). In this section we describe how our routing problem can be formulated as a shortest path problem with special arc lengths.

The product routing problem aims to find a lowest cost directed rooted spanning tree that defines a unique path from the source to every other node. For each product, let y_j denote the minimum sourcing, transportation plus pipeline inventory cost of moving one work tool from the supplier to node j. Let c_{ij} be the incremental cost for a unit flow from node i to node j and let x_{ij} denote this flow. Then the routing problem can be written as follows:

$$\text{Product Routing Problem:} \qquad \min_{x_{ij} \geq 0} \left\{ \sum_{i,j} c_{ij} x_{ij} \mid \sum_i x_{ij} - \sum_k x_{jk} \geq 1, \quad \text{for all nodes } \ j \right\}.$$

This ensures that a unit flow from the source can reach node j in the least costly way possible. The following dual formulation of this problem is easier to solve.

$$\text{Product Routing Dual:} \qquad \max_{y_j \geq 0} \left\{ \sum_j y_j \mid y_j \leq y_i + c_{ij}, \quad \text{for all arcs } \ i \to j \right\}.$$

The above linear program was solved using an efficient version of Dijkstra's shortest path algorithm where nodes were processed in a specific order (using a topological sort of the underlying directed acyclic graph, we can renumber nodes so that for every arc from i to j, the index of i is smaller than the index of j). This ordering of the nodes was required since the value of c_{ij} depends on y_i. Arc cost c_{ij} consists of node processing, transportation and pipeline inventory

costs. Node processing costs are part of the node data; transportation costs were computed using the lowest cost, feasible mode of transportation from i to j; pipeline inventory cost is $I_i L_j y_i$. (Recall that I_i denotes the effective interest rate at node i, L_j is the lead time from i to j corresponding to the transportation mode used, and y_i is the minimum total cost from the source to node i.)

Although c_{ij} depends on y_i, the absence of directed cycles allows us to completely determine y_i before c_{ij} was calculated. In this manner, we solve a shortest path problem in which arc lengths for $i \rightarrow j$ depend on the shortest path from the source to node i. In addition to determining the min-cost path, knowledge of the corresponding mode of transportation dictated the delivery lead time along each arc in this path.

Thus, by the above deterministic routing model, each dealer's regular supply node was set equal to the immediate predecessor of the dealer in the min-cost path from the source to the dealer node. The lead time for deliveries to the dealer was taken to be equal to the time for shipment from the immediate predecessor node (transship point) to the dealer, under the assumption that this transship point would carry sufficient inventory to provide high service. (We determine inventory levels at transship points after the ordering policies at all dealers are specified. Consequently, use of different supply nodes by dealers is accounted for when transshipment inventory levels are set.) For cases where the predecessor (transship) node was allowed to carry no inventory, the delivery lead time observed by the dealer was appropriately increased.

4.2 The Inventory Model

We use separate models for dealer nodes and transshipment nodes. This is motivated by the difference in service level definitions at dealer nodes (who face response-sensitive customer demand) and transshipment nodes (whose demand comes from captive dealers).

4.2.1 The Dealer Model

The dealer inventory model maximizes total expected profit, where dealer profit is computed as revenue minus regular sourcing costs minus expedited sourcing costs minus on-hand inventory carrying costs. The sourcing costs included transportation costs and pipeline inventory costs. The dealer's regular supply node was determined by the product routing model. The supply node for expedited deliveries was the transshipment node, i, with the smallest lead time for direct

material flow to the dealer. The unit cost for deliveries was based on y_i plus additional transportation costs and pipeline inventory costs from i to the dealer. The two (regular,expedited) supply nodes formed the input used by the inventory model corresponding to each dealer node.

The variables and features of the model include:

- I_{t-1} = the inventory *level* at end of period $t-1$ = on-hand inventory − backlog.
- X_τ^k = pipeline inventory due to an order placed in period $\tau < t$, for delivery via mode k, for $k = r$ (regular), or $k = e$ (expedited). We store past orders for $\tau = t-1, \ldots, t-L_k$.
- $P_{t-1} = \sum_{k=r,e} \sum_{\tau=t-L_k}^{t-1} X_\tau^k$ = total pipeline inventory at end of period $t-1$.
- $IP_{t-1} = I_{t-1} + P_{t-1}$ = inventory *position* at end of period $t-1$.
- Receipts in period t are $R_t = X_{t-L_r}^r + X_{t-L_e}^e$.
- The sequence of actions in period t is:
  1. Determine beginning inventory level, I_{t-1}, and pipeline inventories, X^r, X^e.
  2. Receive delivery of relevant pipeline inventory, R_t.
  3. Observe demand D_t.
  4. Satisfy as much demand as possible from on-hand inventory; a portion of unfilled demand is lost.
  5. Place new replenishment orders.
  6. Update net profit.

  Steps 4, 5, and 6 are elaborated on below:

  - Step 4: *Inventory allocation*: If $\beta_0 > \beta_1$, then we satisfy new demand before demand from previous periods is satisfied. Otherwise, a FCFS service discipline is followed. Thus we use inventory to satisfy the most impatient demands first. For our data set, $\beta_0 > \beta_1$, so we focus on the LCFS case.
  - Step 5: *Lost demand* in period t: Let $x^+ = \max(0, x)$ and $x^- = (-x)^+$. Then,

$$\mathcal{L}_t^0 = \beta_0 (D_t - I_{t-1}^+ - R_t)^+ \text{ and } \mathcal{L}_t^1 = \beta_1 (I_{t-1}^- - (R_t - D_t)^+)^+,$$

    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 is used, $\mathcal{L}_t^0 = \beta_0 (D_t - (I_{t-1} + R_t)^+)^+$ and $\mathcal{L}_t^1 = \beta_1 (I_{t-1}^- - R_t)^+$.) For $\mathcal{L}_t = \mathcal{L}_t^0 + \mathcal{L}_t^1$, the ending inventory level is $I_t =$

$I_{t-1} + R_t - D_t + \mathcal{L}_t$.

For instance, if $\beta_0 = 0.6$, $\beta_1 = 0.15$, and in period t, $I_{t-1} = -20$, $R_t = 40$, 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, when LCFS is followed and new demand exceeds receipts by 10, 6 of these 10 units will be lost in addition to 3 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 6: *Order up-to policy for replenishment*: We assume that $z_r \geq z_e$ (which follows from the fact that $c_r \leq c_e$ and $L_r \geq L_e$), thus the inventory position after order placement will always be $IP_t = z_r$. Then, if the inventory position before an order is placed is $IP_{t-} = IP_{t-1} - D_t + \mathcal{L}_t$, the expedited order quantity, which orders up to z_e is,

$$X_t^e = (z_e - IP_{t-})^+ = (z_e - z_r + D_t - \mathcal{L}_t)^+$$

and the regular order quantity is

$$X_t^r = \min(z_r - z_e, (z_r - IP_{t-})^+) = D_t - \mathcal{L}_t - X_t^e.$$

- Sales in period t, $S_t = I_{t-1}^- - I_t^- + D_t - \mathcal{L}_t$.
- Profit in period t is $\pi_t = pS_t - hI_t^+ - c_e X_t^e - c_r X_t^r$.
- Other measures we track and report include the long-run fraction of demand that is lost and the fraction of demand satisfied using the regular and expedited modes, respectively.
- The derivative recursions for $d\pi_t/dz_e$ and $d\pi_t/dz_r$, used by the gradient based search (IPA) for optimal z_e and z_r, are specified in Appendix A.

The inventory model is solved by selecting a starting value of (z_r,z_e), generating a set of k demand scenarios, computing the expected net profit $\pi_t(z_r, z_e)$, the derivative estimates $d\pi_t/dz_e$ and $d\pi_t/dz_r$ according to their recursions, and then using a sub-gradient based search to find the optimal value of (z_r,z_e). As shown in Appendix A, under a last come first served (LCFS) service discipline the profit function is jointly concave in z_r and z_e. At a 100 % service level, FCFS and LCFS are identical. This leads us to believe that for our extremely high service levels, the profit function is likely concave or at worst unimodal. This is borne out by computer experiments illustrated in Figure ??. As the dealer profit function is 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 on this simulation-based technique for determining optimal inventory levels. The usual

practice of removing initialization bias and checking for "steady-state," based on pilot runs, was conducted; the last two simulation runs (just before termination of the search for optimal (z_r, z_e)) were conducted with 10,000 demand scenarios; during the IPA search, $k = 3000$ demand scenarios were generated. This choice of k provided us with a balance between computational time and accurate estimates of expected profit (and its derivative). Preliminary experimentation showed that estimates of optimal profit differed by no more than 0.2 % over a range of simulation iterations between $k = 1000$ and $k = 10,000$, while run times increased by an order of magnitude.

4.2.2 The Transshipment Node Model

For the Transshipment nodes, 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 standard deviation, σ, of daily demand for each product at each transshipment node. Given daily demand parameters μ and σ, capacity C, and desired service level δ, the base stock level at a transshipment node was set to

$$(1) \qquad z = \left[(L+1)\mu + \frac{\sigma^2}{2(C-\mu)}\right] + \Phi^{-1}(\delta)\left[(L+1)\sigma^2 + \left(\frac{\sigma^2}{2(C-\mu)}\right)^2\right]^{1/2},$$

where $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function (CDF) of demand at the transshipment node during its delivery lead time. 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.

Since demand at each transshipment node is the sum of many demands originating at different dealers, we assumed that the CDF, $F(\cdot)$, could be approximated by a normal distribution, $\Phi(\cdot)$. However, as stated earlier, any correlation between demands at different transshipments points was neglected. If correlation effects are considered significant, a more accurate, but computationally intensive simulation model similar to the dealer model in Section 4.2.1 could be used to compute base stock levels and costs.

Inventory holding costs at each transshipment node were estimated as $h(z - \mu(L + 0.5))$, where L is the lead time along the single min-cost path from the source to the transship node. This approximation is fairly standard, see, for instance, formula (5-1) in Hadley and Whitin (1963). Costs over different transship nodes was aggregated to obtain estimated total costs for this subsystem. Our simulation based estimate of net profit for each product was obtained by

subtracting costs of the transshipment subsystem from expected profit at the dealer subsystem. Total system profit was the sum of the net 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 quickest delivery to the dealer. The average run-time for each product in the dealer inventory subsystem was just under 40 seconds per dealer. This computes optimal inventory levels (z_e, z_r) at the dealer. Calculation of inventory levels at the transshipment nodes, using equation (1), was instantaneous.

5.1 General Results

We considered 21 dealer districts, so complete analysis of one product over all dealers took approximately 21×40 seconds or 14 minutes. Comprehensive data was available for 21 products over two years (2000 and 2004), consequently one run over all products, dealers, and study years took approximately 9.8 hours. For each scenario tested, our experiments were typically run in under five hours, on two computers working in parallel on independent sets of products (or study years).

Several scenarios (consisting of a restricted set of permissible transshipment nodes) were considered, the four primary ones being TF & PDC, PDC-Only, TF-Only, and DS-Only. Solutions which were optimal within the class of order-up-to policies were generated for each of these four scenarios; these consisted of order up-to levels for each of the products, locations, and modes of shipment, which ensured that there was a sufficiently high proportion of customers served at a minimal cost. The results included:

1. Optimal total profit for each scenario.
2. The relative profit from machines and work tools.
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. Product delivery lead times.

7. The effect of demand volume changes from 2000 to 2004 on the above.

Some of these results are illustrated below.

Recall that scenario costs exclude the fixed costs associated with construction of the tool facilities. Therefore, it comes as no surprise that the optimal (highest) expected profit results are achieved by using the entire network. The profit results for the work tools for each of the two study years across the four scenarios are presented in Table 2, as a percentage of the optimal. The higher percentages in year 2004, relative to year 2000, may be attributed to increased economies of scale due to higher demand, which has a greater risk-pooling effect on inventory cost for the PDC-Only and TF-Only alternatives (where fewer facilities are used) compared with the TF & PDC alternative. TFs benefit less from increased volume since, for the TF-Only scenario, the transportation cost component dominates the inventory cost component.

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

Table 2 demonstrates that inclusion of Tool Facilities (TFs) adds about 2-3 % to net profit from work tools as compared to using the PDC-Only scenario. This is several million dollars. Given that the tool facilities were not in current use and would need to be constructed, and that our model did not include costs of building and operating tool facilities (which were expected to exceed the incremental 3 % annual profit from use of tool facilities), Caterpillar decided to implement use of PDCs (plus the Sanford TF due to routing constraints) and direct shipments. This supply chain configuration captured almost 100 % of the demand (so that lost sales were negligible). This was because the incremental inventory holding costs was significantly smaller than the lost profit due to shortages.

In addition, a high percentage of the material delivery to the dealer occurred via the regular mode, with the expedited mode used occasionally during high demand periods. The relative amounts of regular vs. expedited deliveries depend on the increase in dealer on-hand and pipeline inventory costs (required to reduce shortage penalty costs) resulting from slower regular shipments as compared to the incremental sourcing costs of expediting.

For the purpose of comparison, without our analysis, Caterpillar would have likely imple-

mented either the DS-Only or the TF-Only option. (Internal considerations caused Cat to question the value of inclusion of PDCs into the network.) A primary goal of our project was to quantify the benefits of incorporating PDCs into the P2000 network. Taking TF-Only as the base case for comparison and disregarding fixed costs of construction, the annual benefit of the PDC-Only solution is of the order of 8 % of the highest expected profit, which is several million dollars. The above comparison underestimates the benefit since it assumes that Caterpillar would have used the optimal inventory levels from our analysis in its implementation of the TF-Only option.

In order to assess the sensitivity of our recommendations to input demand data, we studied performance measures (profit, inventory, service levels) for each scenario at four distinct mean demand levels: (i) Cat's forecast μ, (ii) 0.8 $\times\mu$, (iii) 1.2 $\times$ μ, (iv) $U(0.8, 1.2) \times \mu$, where $U(a, b)$ denotes a uniform random variate between a and b. In case (iv), the mean demand for each product was multiplied by a different random variate. In all four cases, the relative profitability of the different scenarios remained remarkably insensitive to changes in mean demand.

5.1.1 The Second Model

Since the PDCs performed well for work tools, Caterpillar decided to consider an alternative scenario where PDCs (and TFs) were also permitted to act as transship points for machines. Their hypothesis was that if machines were allowed to flow through the PDCs and TFs, the work tools would benefit from lower transportation costs resulting from use of cheaper modes of flat bed transport normally available only to machines. This necessitated generation of new data which specified TL transportation costs and times via Closed Vans *and* via Flat Beds charged at new rates for work tools. Preliminary analysis showed that permitting machines to use PDCs and TFs would *not* increase total profits since the added node costs, incurred at PDCs and TFs, for processing machines would overwhelm any savings from combined work tool and machine transportation. Furthermore, while PDCs were already well equipped to handle work tools, this was not the case for machines. Hence, Caterpillar decided to stay with its initial model which decoupled the distribution of machines and work tools.

5.2 Revenues, Costs, and Profits of First Model

The remainder of this section illustrates output analysis conducted to provide Caterpillar with a better understanding of the results of our study. For instance, 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. Since the PDCs were treated as a separate profit center, these expected costs were categorized as:

"Source" (cost charged by external supplier),

"non-PDC Ship" (non-PDC pipeline inventory and transportation costs),

"PDC Outbound" (costs incurred after some PDC took receipt and control of material),

"TF Inv/Proc" (on-hand inventory and node processing costs at TFs), and

"Dealer Inv" (costs of on-hand inventory at dealers).

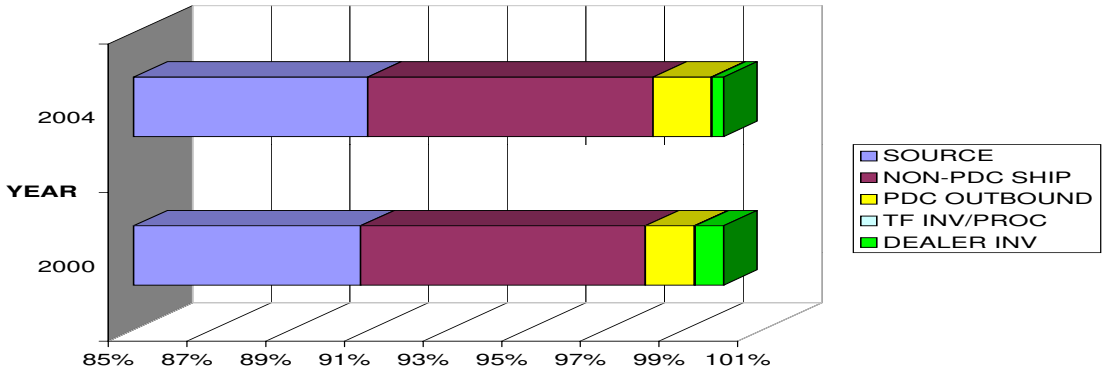


Figure 4: Work Tool Cost Breakdown By Percentage

This cost breakdown is illustrated in Figure 4 for the scenario where the PDCs plus the Sanford TF are used. We note that the bulk (90 %) of costs are source costs, that non-PDC transportation & pipeline 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, 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 %) since the Sanford TF is of limited use for the scenario considered. Together with Figure 6, this figure shows that with the increase in demand volume and additional new routes in 2004, inventory (particularly SSL work tool inventory) will shift from the dealer to the pipeline. So the increased deamnd volume does not require increased floor space or inventory investment at the dealer, rather, it requires the capability to handle larger volumes throughout the distribution network.

Regional profits could also be estimated from our model's output. For example, an ABC classification of regions (dealer districts) could be developed using the pie-chart in Figure 5,

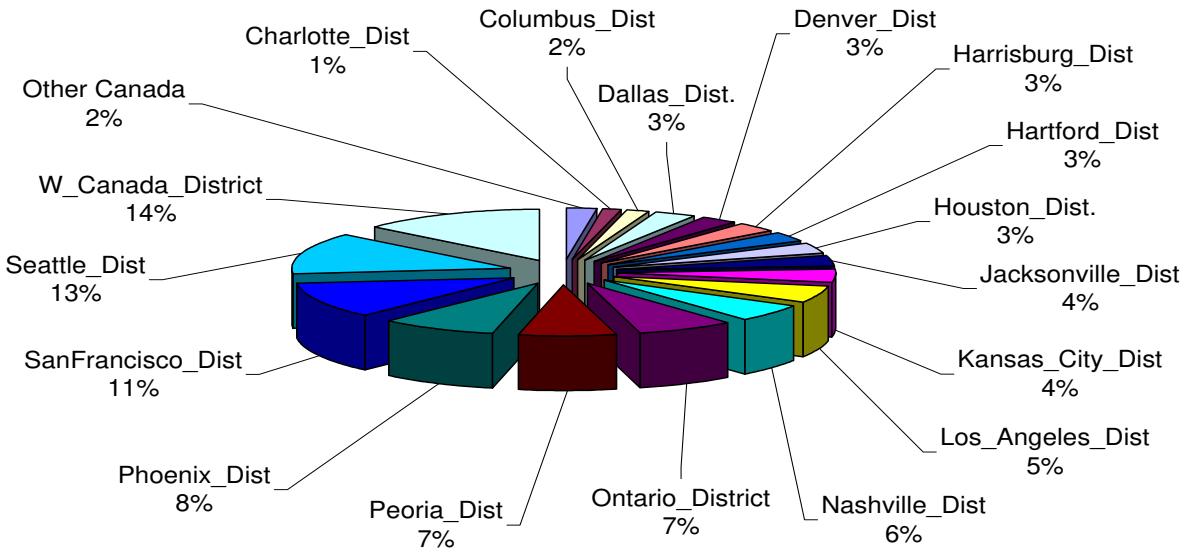


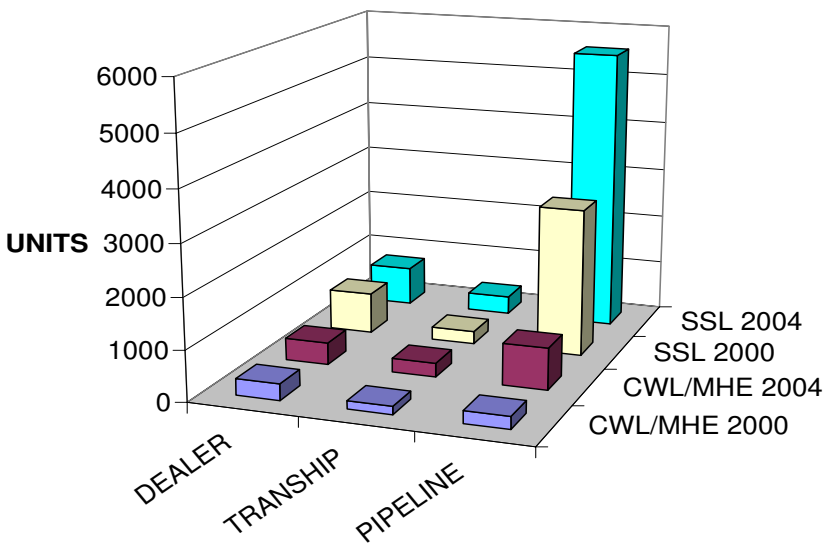
Figure 5: Year 2004 Work Tool Profits by Region

which illustrates that the top 6 districts account for over 60 % of total profit from projected year 2004 work tool sales in the North American region. Only 18 dealer districts are shown instead of 21 since "Other Canada" actually consists of four districts including North-East (Newfoundland, Nova Scotia) and South-East (Montreal, Toronto) Canada. Similar graphs illustrated the breakdown, by product volume, of material flow through distribution centers as well as the profit generated by each work tool.

5.3 Inventory Levels

In this section we focus on the location and magnitude of inventory within the supply chain, with some of the illustrated results pertaining to the PDC & TF scenario. Our analysis specifies optimal inventory levels at the dealers within the class of base stock policies. During the course of our IPA optimization we observed that choosing the correct inventory ordering parameters is critical: Setting these levels incorrectly (at lower than optimal levels for our problem) often results in significantly higher total costs. Figure ?? illustrates how profit changes with z_r for a representative worktool. Graphs for other worktools were similar.

Figure 6 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 + safety stock). There is more inventory of SSL work tools because of the higher mean demand for SSL



	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

Figure 6: Work Tool Inventory Location Within Supply Chain

work tools. 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. We also studied the breakdown of each of these inventories. For instance, for year 2000, the detailed breakdown of dealer on-hand inventory of 811 units of SSL work tools and 320 units of CWL/MHE work tools is shown in Table 3. Since Caterpillar was considering different contracts with dealers, some of which involved a manufacturer buyback option and Caterpillar's ownership of "consignment" inventory, the breakdown of these dealer inventories was of considerable importance.

Table 3: Dealer Inventory by Work Tool 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

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 in this paper. We also graphed the work tool and machine inventory by non-dealer locations. The latter is shown in Figure 7, which is output from our analysis of the supply chain for machines.

6 Concluding Remarks

In this paper we have presented our efforts to develop and use an integrated model to analyze the performance of different supply chain configurations in a real industrial setting. Our analysis was based on decomposition techniques, network optimization theory, inventory modeling, and simulation theory. The novel features of the model include dual modes of supply for dealer replenishments and net customer demand that is responsive to speed of service. Via our model, we were able to make recommendations to Cat on the effects of different factors on profits.

In last quarter of 1998, Caterpillar plans to launch its new P2000 product line of compact

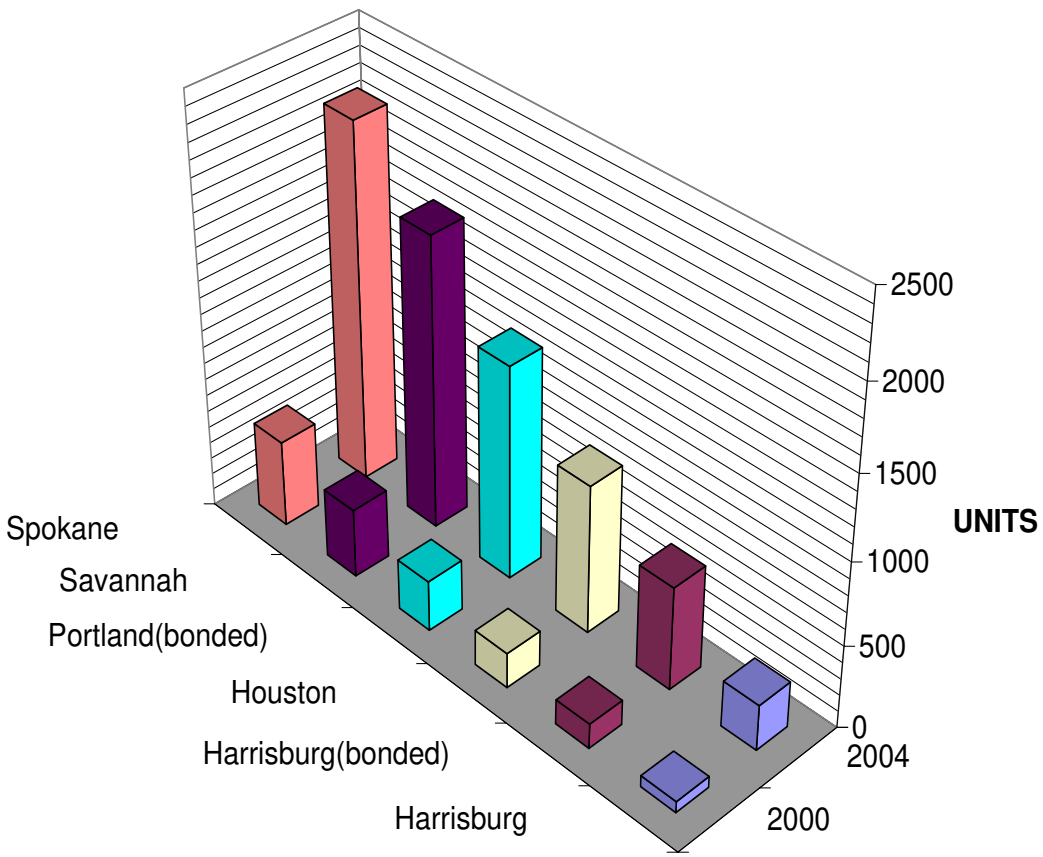


Figure 7: CWL Machine Volumes

machines and associated work tools, supporting it with the supply chain recommended by our analysis. The work reported in this paper has concentrated on the North American market; we are currently analyzing supply chain configurations to handle Caterpillar's European market.

A Appendix: 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:

1. $\frac{d\pi_t}{dz_e} = p\mathbf{1}\{S_t > 0\}\frac{dS_t}{dz_e} - h\mathbf{1}\{I_t > 0\}\frac{dI_t}{dz_e} - c_e\mathbf{1}\{X_t^e > 0\}\frac{dX_t^e}{dz_e} - c_r\mathbf{1}\{X_t^r > 0\}\frac{dX_t^r}{dz_e}$.
2. $\frac{dS_t}{dz_e} = -\mathbf{1}\{I_{t-1} < 0\}\frac{dI_{t-1}}{dz_e} + \mathbf{1}\{I_t < 0\}\frac{dI_t}{dz_e} - \mathbf{1}\{\mathcal{L}_t > 0\}\frac{d\mathcal{L}_t}{dz_e}$.
3. $\frac{dI_t}{dz_e} = \frac{dI_{t-1}}{dz_e} + \frac{dR_t}{dz_e} + \frac{d\mathcal{L}_t}{dz_e}$, 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).
4. $\frac{dX_t^r}{dz_e} = -\frac{d\mathcal{L}_t}{dz_e} - \frac{dX_t^e}{dz_e}$. (Information from 4 & 5 will be used in future iterations).
5. $\frac{dX_t^e}{dz_e} = \mathbf{1}\{z_e > IP_{t-}\}(1 - \frac{dIP_{t-}}{dz_e}) = \mathbf{1}\{z_e > z_r - D_t + \mathcal{L}_t\}(1 - \frac{d\mathcal{L}_t}{dz_e})$.
6. $\frac{d\mathcal{L}_t}{dz_e} = \frac{d\mathcal{L}_t^0}{dz_e} + \frac{d\mathcal{L}_t^1}{dz_e}$, where $\frac{d\mathcal{L}_t^0}{dz_e} = \beta_0(-\mathbf{1}\{I_{t-1} > 0\}\frac{dI_{t-1}}{dz_e} - \frac{dR_t}{dz_e})\mathbf{1}\{D_t - I_{t-1} - R_t > 0\}$ and $\frac{d\mathcal{L}_t^1}{dz_e} = \beta_1(-\mathbf{1}\{I_{t-1} < 0\}\frac{dI_{t-1}}{dz_e} - \mathbf{1}\{R_t > D_t\}\frac{dR_t}{dz_e})\mathbf{1}\{\mathcal{L}_t^1 > 0\}$.
7. The order of derivative computation is in the reverse order of the listing above. We only track derivatives of I_t, X_t^k (which yields 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 5 above, which becomes $\frac{dX_t^e}{dz_r} = \mathbf{1}\{z_e > z_r - D_t + \mathcal{L}_t\}(-1 - \frac{d\mathcal{L}_t}{dz_r})$. The validity of these simulation based derivative estimates follows from arguments similar to Glasserman and Tayur (1995).

Proof of Concavity

A Notation

We consider an inventory system with two types of orders, expedited and regular. It is well known that if the leadtimes of these two orders differ by more than one unit, the optimal policy of such a system is state-dependent, and hence may be quite complex. Therefore we reduce

the solution space to consist of only order-up-to policies. This type of policy will be explicitely defined below.

Formally we define:

n: Period index.

N: Horizon length.

I_n: On-hand inventory at start of period n.

R_n: Amount of inventory received in period n.

l_r, l_e: Lead time for regular and expedited orders, respectively. We assume $l_e < l_r$.

c_r, c_e: Unit ordering cost for regular and expedited orders, respectively. We assume $c_e > c_r$.

S_n: Amount of sales in period n, at price p.

β, γ: Proportion of new and previously waiting customers respectively who, if left unsatisfied, depart the system (lost sales) in each period. We assume $\gamma \geq \beta$; impatience grows with the length of the wait, and hence customers who arrived first are served first.

z_r, z_e: Regular and expedited order-up-to levels, respectively. In period n, one will order to bring the on-hand inventory, minus the backorders, plus the orders in route in periods $n+1, \ldots, n+l_e$ up to the level z_e, if below. An analogous process occurs for z_r.

U_n^r, U_n^e: Regular and expedited orders placed in period n, respectively. We denote by $U^r(n, n+k)$ the sum of the regular orders placed in periods n up to and including $n+k$. $U^e(n, n+k)$ is defined analogously.

$L(n, n+k)$: Amount of demand lost between periods n up to and including period $n+k$. We abbreviate $L(n, n)$ as L_n.

d_n: New customer demand in period n. $\{d_n : n \geq 0\}$ are assumed to form a stationary and iid sequence. We call the cumulative demand between periods n up to and including $n+k$ $d(n, n+k)$.

$\mathcal{B}_n$: The set of periods between $n - l_r$ and n in which an expedited order is placed which arrives prior to period n. Formally, $\mathcal{B}_n \stackrel{\text{def}}{=} \{x | n - l_r + l_e < x + l_e < n, U_x^e > 0\}$. We define $\hat{\mathcal{B}_n}$ as a similar function, where $x + l_e$ may be equal to n: $\hat{\mathcal{B}_n} \stackrel{\text{def}}{=} \{x | n - l_r + l_e < x + l_e \leq n, U_x^e > 0\}$.

At the start of period n, orders are placed to bring the *estimated inventory position* in period $n + l_e$ up to z_e and in period $n + l_r$ up to z_r. By estimated inventory position we mean the amount of inventory on hand, minus customers on back-order, plus the amount of inventory which will arrive by $n + l_e$ or $n + l_r$, respectively. Thus in deciding the expedited order amount, we ignore orders already placed which will arrive after $n + l_e$. This is what makes our policy sub-optimal. This is only one possible policy; we will explore another in section D.

After orders are placed, a shipment of R_n units is received, and demand d_n is revealed. Waiting customers are satisfied first, after which, if any inventory remains, new customers are satisfied as much as possible. Then if any new or previously waiting customers remain, they are lost in proportions β and γ respectively. Any remaining unsatisfied customers or inventory is carried to the next period.

B Results

It can be seen from the definition of the problem,

$$I_{n+1} = I_n + R_n - d_n + L_n. \tag{2}$$

As we have FCFS service, the loss function is:

$$L_n = \beta((I_n + R_n)^+ - d_n)^- + \gamma(I_n + R_n)^-, \tag{3}$$

where $^-$ denotes the negative part of the expression, and $^+$ the positive part. The profit in any period is

$$\pi_n = p_n S_n - h I_{n+1}^+ - c_r U_r - c_e U_e. \tag{4}$$

Lemma B.1 *1. If an expedited order was placed in period $n - l_e$, then*

$$I_n + R_n = z_e - d(n - l_e, n - 1) + L(n - l_e, n - 1).$$

2. Otherwise,

$$I_n + R_n = z_r - \sum_{n - l_r \le i < n | i \notin \mathcal{B}_n} (d_i - L_i) - \sum_{i \in \mathcal{B}_n} U_i^r.$$

Proof :

1. If an expedited order arrives in period n, this order was placed in period $n - l_e$ to bring $I_{n-l_e} + R(n - l_e, n)$ up to z_e. In other words:

$$I_{n-l_e} + R(n - l_e, n) = z_e. \tag{5}$$

Using (2) repeatedly, we find:

$$I_n = I_{n-l_e} + R(n - l_e, n - 1) - d(n - l_e, n - 1) + L(n - l_e, n - 1). \tag{6}$$

Substituting (5) into (6) yields:

$$I_n = z_e - R_n - d(n - l_e, n - 1) + L(n - l_e, n - 1),$$

from which the result follows.

2. (a) If $\mathcal{B}_n = \{\}$, one would arrive at a formula similar to that above:

$$I_n = z_r - R_n - d(n - l_r, n - 1) + L(n - l_r, n - 1), \tag{7}$$

from which the result is immediate.

(b) If $\mathcal{B}_n \neq \{\}$, then (7) must be modified by adding the expedited orders which were placed in periods which are in $\mathcal{B}_n$. These units were not accounted for when the order up to z_r was placed, but they will have arrived by period n. Starting from this modified (7):

$$\begin{aligned} I_n &= z_r - R_n - d(n - l_r, n - 1) + L(n - l_r, n - 1) + \sum_{i \in \mathcal{B}_n} U_i^e \\ &= z_r - R_n - d(n - l_r, n - 1) + L(n - l_r, n - 1) + \sum_{i \in \mathcal{B}_n} (d_i - U_i^r - L_i) \\ &= z_r - R_n - \sum_{n - l_r \leq i < n | i \notin \mathcal{B}_n} (d_i - L_i) - \sum_{i \in \mathcal{B}_n} U_i^r, \end{aligned}$$

from which the result is immediate. The second equality above uses the fact that the total order in a period will equal the number of units demanded less the lost sales; $U_i^r + U_i^e = d_i - L_i$.

■

Theorem B.1 *Along any sample path ω, for all $0 \leq n \leq N < 1$:*

1. *I_n is jointly convex in (z_e, z_r).*

2. *$I_n + R_n$ is a jointly convex function of (z_e, z_r).*

3. *L_n is jointly convex in (z_e, z_r).*

4. *U_n^r is a jointly concave function of (z_e, z_r).*

Proof : We assume the hypothesis is true for all $i \leq n-1$.

1. This follows from (2) and our induction assumption.

2. Assume that we increase z_e to $z_e + \epsilon_e$ and z_r to $z_r + \epsilon_r$, where $\epsilon_r < \epsilon_e$. (The case where $\epsilon_e = 0$ is straightforward.) Furthermore, we choose the (ϵ_e, ϵ_r) terms in such a way that if we placed an expedited order in period n with the original order-up-to parameters, we will do so along the path with the increased parameters as well. Similarly, if the inventory position is strictly greater than z_e along the initial sample path, we will not place an expedited order along the altered path. This is possible as I_n and R_n are continuous with respect to (z_e, z_r). Incrementing in this manner is sufficient for our proof, as need only show convexity (or concavity) in an infinitesimal neighborhood around every point.

   We look at three cases, with the understanding that if in any period, prior to ordering, the inventory position is exactly equal to z_e, we say that we place an expedited order for zero units in this period.

   (a) Along both paths an expedited order (possibly zero) was placed in period $n - l_e$. Then after incrementing, using Lemma B.1, and denoting by $\Delta(\cdot)$ the value of a function along the incremented path minus the value along the original path:

   $$\Delta(I_n + R_n) = \epsilon_e + \Delta(L(n - l_e, n-1)).$$

   Using the induction assumption, this is jointly convex in the parameters.

   (b) Along both paths a regular order was placed in period $n - l_r$, but an expedited order was not placed in period $n - l_e$. There are two cases here.

      i. $\mathcal{B}_n^{New} = \mathcal{B}_n^{Old}$, where the superscripts denote the changed and original sample paths, respectively. In this case:

      $$\Delta(I_n + R_n) = \epsilon_r + \sum_{n-l_r \leq i < n | i \notin \mathcal{B}_n} \Delta(L_i) - \sum_{i \in \mathcal{B}_n} \Delta(U_i^r).$$

      From our induction assumption, the result is immediate.

ii. $\mathcal{B}_n^{New} \supset \mathcal{B}_n^{Old}$, strictly; in other words, there are certain periods where we placed expedited orders along the altered path, in which we did not along the original. Then if we partition the periods between $n - l_r$ and $n - 1$ into three sets, $\mathcal{B}_n^B, \mathcal{B}_n^N,\ \mathcal{B}_n^0$, consisting of the elements which are in both $\mathcal{B}_n^{New}$ and $\mathcal{B}_n^{Old}$, only the former, and in neither:

$$\Delta(I_n + R_n) = \epsilon_r + \sum_{i \in \mathcal{B}_n^0} \Delta(L_i) - \sum_{i \in \mathcal{B}_n^B} \Delta(U_i^r) + \sum_{i \in \mathcal{B}_n^N} (d_i - L_i^{Old} - U_i^{r,New}),$$

where L_i^{Old} denotes the loss in period i along the unchanged sample path, and $U_i^{r,New}$ the regular order placed in period i along the incremented path.
Looking at the final sum above, if $i \in \mathcal{B}_n^N$, then an expedited order could have been placed along the original sample path which would have arrived prior to n, but was not. Therefore for all such i, $U_i^r = d_i - L_i^{Old}$. Hence the above expression reduces to:

$$\Delta(I_n + R_n) = \epsilon_r + \sum_{i \in \mathcal{B}_n^0} \Delta(L_i) - \sum_{i \in \mathcal{B}_n^B} \Delta(U_i^r) - \sum_{i \in \mathcal{B}_n^N} \Delta(U_i^r).$$

(c) An expedited order is placed in period $n - l_e$ along the initial sample path, but not along the incremented one. As we assume $\epsilon_r < \epsilon_e$, this is not possible, as I_{n+1} increases at rate one with respect to $I_n + R_n$, and $I_n + R_n$ increase at rate no greater than one with z_e and z_r. Therefore the inventory position in period $n - l_e$ along the incremented sample path can be no more than that along the initial path plus $\epsilon_r \leq z_e + \epsilon_r < z_e + \epsilon_e$. The last two inequalities are due to our assumptions that we order along the path with the unchanged parameters and $\epsilon_r < \epsilon_e$.

3. This follows from (3), and the above results, noting that we are acting under FCFS service and $\gamma \geq \beta$. Waiting customers will be driven down to zero exclusively when there are both new and old customers left unsatisfied; once they reach zero new customers will be served. Hence $\gamma \geq \beta$ is required.

4. Assume we increase z_e and z_r in a manner similar to part 2.

   We again look at three cases:

   (a) Along both paths an expedited order (possibly zero) is placed in period n. Then along both paths:

   $$U_n^r = z_r + \epsilon_r - z_e - \epsilon_e.$$

   This is jointly concave in the parameters.

(b) Along both paths a regular order was placed in period n, but an expedited order was not placed. Then along both paths:

$$U_n^r = d_n - L_n.$$

(c) This is not possible, due to our assumptions on the increments of the parameters.

As the results are trivially true for any fixed initial values, the result is proved. ■

Theorem B.2 *For the given system, assuming we may salvage goods for their ordering cost, the total expected finite horizon cost is concave in* (z_e, z_r).

Proof : Starting from (4), we may rewrite the total profit along any sample path as:

$$\pi = (p - c_e) \sum_{n=0}^{N} (d_n - L_n) - h \sum_{n=1}^{N+1} I_n^+ + (c_e - c_r) \sum_{n=0}^{N} U_n^r.$$

Using the fact that by assumption $p > c_e > c_r$, Theorem B.1, and the fact that the total expected cost of any policy is the expectation along all sample paths, which preserves concavity, the result follows. ■

Theorem B.3 *For the given system, the total expected infinite horizon average cost is jointly concave in* (z_e, z_r).

Proof : Using Theorem B.2, the fact that salvage value becomes insignificant as horizon length grows, and the fact that a series of concave functions converges to a concave function, the result is proved. ■

C Alternate Policy

In this section we examine a policy where when determining the expedited and regular orders in period n we examine the *total* inventory position, the on-hand inventory, minus what is on back-order, plus *all* items on order. This differs from the above in that when considering whether to place an expedited order, the count of items on order includes those which will arrive *after*

$n+l_e$. Thus orders placed in period n bring the expected inventory position $I_n+R(n+1,n+l_r)$ to z_e and then z_r, respectively.

Under this policy Lemma B.1 becomes

Lemma C.1 *Define τ_n as follows:*

$$\tau_n \stackrel{\text{def}}{=} \max\{x+l_e | x \in \hat{\mathcal{B}}_n\}.$$

In other words, $\tau_n \leq n$ is the index of the last period an expedited order arrives which was placed after the regular order arriving in period n was placed. Then,

1. *If $\tau_n \neq \{\}$:*

$$I_n + R_n = z_e - d(\tau_n - l_e, n-1) + L(\tau_n - l_e, n-1) - U^r(n - l_r + 1, \tau_n - l_e - 1).$$

2. *Otherwise, if $\tau_n = \{\}$:*

$$I_n + R_n = z_r - d(n - l_r, n-1) + L(n - l_r, n-1).$$

Proof : We only illustrate the first case.

The order in period $\tau_n - l_e$ brought the inventory position up to z_e:

$$I_{\tau_n - l_e} + R(\tau_n - l_e, \tau_n - l_e + l_r - 1) = z_e.$$

As the orders which are to arrive after τ_n must all be regular, and letting $n-k \stackrel{\text{def}}{=} \tau_n - l_e$:

$$I_{n-k} + R(n-k, n) + R(n+1, n-k+l_r-1) = I_{n-k} + R(n-k, n) + U^r(n - l_r + 1, n-k-1) = z_e.$$

Using (6) and substituting:

$$I_n = z_e - R_n - d(n-k, n-1) + L(n-k, n-1) + U^r(n - l_r + 1, n-k-1).$$

■

Our results then follow as before, noting that for all n, $\tau_n^{New} \geq \tau_n^{Old}$. This fact will be needed in the proof of Theorem B.1.

D Final Comments

- Extension to the case of order capacities is immediate. We modify our definition of ϵ_e and ϵ_r, so that along the sample path an order which prior to the increment did not hit the capacity barrier does not along the modified path.

- Serving arriving customers before waiting customers leads to a loss function which may not be convex, thus making a proof for this case appear impossible.

Acknowledgements: We thank Dr. George Cusack at Caterpillar, Inc., Peoria, for providing us the opportunity to work on this problem, and for his insightful comments. All numbers contained in this article have been disguised in accordance with contractual agreement.

References

J. Birge and F. Louveaux 1997. Introduction to Stochastic Programming. Springer-Verlag, Berlin. Y. Fukuda. 1964. Optimal Policy for the Inventory Problem with Negotiable Leadtime, *Management Science*, **12**, 690-708.

G. Feigin. 1998. Inventory Planning in Large Supply Chains. In *Quantitative Models for Supply Chain Management*. S. Tayur, R. Ganeshan and M. Magazine (eds.). Ch.24, 761-787. Kluwer Academic Publishers, Boston.

P. Glasserman. 1997. Bounds and Asymptotics for Planning Critical Safety Stock, *Operations Research*, **45**, 2, 244-257.

P. Glasserman and S. Tayur. 1995. Sensitivity Analysis for Base-stock levels in Multi-echelon Production-inventory Systems, *Management Science*, **41**, 2, 263-281.

G. Hadley and T.M. Whitin. 1963. Analysis of Inventory Systems. Prentice Hall, Inc., New Jersey.

Horst and Tuy. 1993. Global Optimization. 2nd Edn. Springer-Verlag, Berlin.

R. Kapuscinski and S. Tayur. 1998. Optimal Policies and Simulation-Based Optimization for Capacitated Production Inventory Systems. In *Quantitative Models for Supply Chain Management*. S. Tayur, R. Ganeshan and M. Magazine (eds.). Ch.2, 7-40. Kluwer Academic Publishers, Boston.

P. Keskinocak, J. Swaminathan and S. Tayur. 1998. Solution of A Multi-Commodity Network Flow Problem with Non-Convex Costs, Manuscript in Preparation.

E. Lawler 1976. Combinatorial Optimization: Networks and Matroids. Holt, Rinehart and Winston, New York.

H. Lee and C. Billington. 1993. Material Management in Decentralized Supply Chains, *Operations Research*, **41**, 5, 835-847.

P. Marsh. 1998. Caterpillar Digs Into Compact Market, *The Financial Times, US Edition*, February 12.

A. Scheller-Wolf and S. Tayur. 1998. Optimal Policies for Dual Supplier Contracts with Order Bands, *Submitted*.

S. Tayur. 1992. Computing the Optimal Policy for Capacitated Inventory Models, *Comm. Statist. – Stochastic Models*, **9**, 585-598.

D. Weimer. 1998. Strategies: A New Cat on the Hot Seat, *Business Week*, 56-61, March 9.