Control-Oriented Approaches to Supply Chain Management in Semiconductor Manufacturing

Karl G. Kempf, *Member, IEEE*

Abstract— **After describing the general supply chain management problem with examples from the semiconductor industry, attention is restricted to the core manufacturing problem. Using a control-oriented approach for this nonlinear stochastic combinatorial optimization problem, an outer loop for addressing the planning parts of the problem and an inner loop to manage the execution aspects are proposed. The outer loop provides a material release plan generated by a linear programming formulation (LP) and inventory safety stock targets generated by a dynamic programming formulation (DP) to the inner loop to guide execution. Portions of the nonlinearity and stochasticity inherent in the problem are addressed by the outer loop that requires iterative convergence between the LP and the DP. The inner loop is formulated from the perspective of model predictive control (MPC) and integrates optimal control and stochastic control. Initial results are presented to demonstrate the ability of the inner loop to track material release and safety stock targets while improving delivery performance in the face of both supply and demand stochasticity. A simulation module is also described that supports the other components of the system by validating their efficacy before application in the real world. This component has to address the integrated flows of materials, information, and decisions through the supply chain, and employs innovative approaches combining a number of specialized models to do so quickly and accurately.**

I. Introduction to the Problem

THE economic systems that are the focus of this paper stretch from the suppliers' suppliers to the customers' customers with sets of manufacturing facilities in between. Any high volume manufacturing company in the midst of this system plays many roles. On one hand, the company is a customer sending demand signals to upstream suppliers. To manufacture, the company needs raw materials relative to its products, production facilities and equipment including spares, and often relies on subcontractors for burst capacity. In most cases, the subcontractor relies on the same materials and equipment suppliers as the company. Furthermore, the company's competitors often rely on this same set of suppliers.

Karl G. Kempf is a member of the National Academy of Engineering, a Fellow and Director of Decision Technologies at Intel Corporation, 5000 W. Chandler Blvd. Chandler, Arizona, 85226, USA, and an Adjunct Professor of Industrial Engineering at Arizona State University (karl.g.kempf@intel.com).

On the other hand, the company is a supplier satisfying demand signals from downstream customers. In this role, the manufacturing company warehouses and transports (perhaps relying on subcontractors) a variety of products to a range of geographically disperse consumers. These include other manufactures and their subcontractors, distributors, and end users, noting that the end users may also purchase from the distributors and the distributors might also purchase from the other manufacturers. Of course, the company's competitors sell to many of the same manufacturers, distributors, and end users. The tradition of referring to this supply-demand network as simply a "supply chain" grossly understates the actual complexity.

There are a broad set of flows inherent in this supply-demand network. Materials flow from suppliers to customers increasing in value while becoming products. Revenues move in the opposite direction through the many echelons in the network. Information flows in many directions including forecasts of supply towards the customers and forecasts of demand rolling up towards the suppliers. This provides a rich set of research and development opportunities for those interested in decision and control. Beyond the scope of this paper are applications of option theory and auction theory to the multiple interactions between suppliers and customers in the network, as well as applications of forecasting theory to the multiple interdependent supply and demand parameters of interest. Demand management issues, although extremely important, will not be addressed here. Neither will the extensive set of issues related to network design.

The focus here will be on integrating optimal decision making and controlling decision execution in the pre-existing manufacturing core of a supply-demand network. This business problem is addressed from the inside out under the hypothesis that, if the existing manufacturing core is not efficiently planned and executed, the probability of realizing efficient operation of the other components of the network becomes vanishingly small. Another reason for this focus is the ever-growing desire for mass customization and instantaneous doorstep delivery in our society. Consumers have come to expect higher performance at lower prices on a year to year basis. In

competition between supply-demand networks to satisfy these desires and expectations, success or failure rests to a large degree on the agility and responsiveness of the manufacturing core.

Concrete examples will be drawn from the semiconductor industry, specifically Intel Corporation as one of the international high volume manufacturing companies that represents the manufacturing core of supply-demand networks of logic, memory, and communications products (among others). While greatly simplified for the purposes of this paper, the examples represent 10's of billions of dollars in annual sales to 100's of millions of end customers for 10's of thousands of diverse products. The most sophisticated current logic products integrate roughly 250 million transistors on a silicon die the size of an average human thumb print, and have continued to increase in complexity in accordance with Moore's law for over 30 years. The factories required to manufacture products of such sophistication current cost roughly 3 billion dollars to construct and outfit, and have continued to become more expensive with every generation of decreasing transistor feature size.

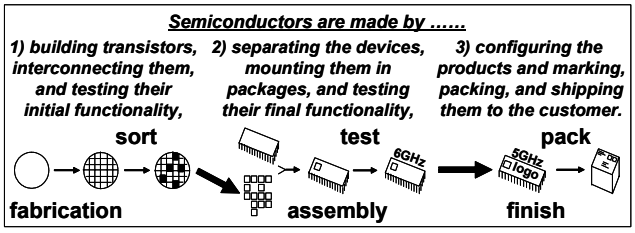

Fig. 1. The basic flow in semiconductor manufacturing.

The basic manufacturing flow is shown from left to right in Fig. 1. Transistors are built on a silicon wafer and then interconnected to form circuits in a fabrication process that might consist of 300 individual production steps and take roughly 6 weeks to complete. The resulting wafers are tested to sort working die from those that do not function, and further sort the working die into broad functional categories. Sort is necessary given the number of stochastic processes that underlie semiconductor manufacturing including random machine breakdowns that drive a distribution of throughput times (TPT) and random atomic misplacements that cause the resulting devices to work over a range of clock speeds and power consumptions including devices that do not function at all.

Sorted wafers are then passed into the assembly process that might consist of 30 individual production steps and take a week or two to complete. Here the individual die are cut from the wafers and mounted in packages to protect them and make them suitable for incorporation in other products, often being mounted on printed circuits of various types. Once packaged, they are stressed to induce infant mortality and tested again for final classification into performance categories. Stochasticity again drives a distribution of TPT as well as a distribution of end product characteristics.

Categorized product then enters the finish and pack process that involves roughly 10 processing steps that take only a few days to complete. One of the unfortunate asymmetries of semiconductor manufacturing is encountered here. Depending on the demand in the marketplace, fast devices can be configured to run more slowly, but slow devices can not be enticed to run faster. Once the final performance is configured, devices are individually labeled and packed in batches into the appropriate medium for shipment.

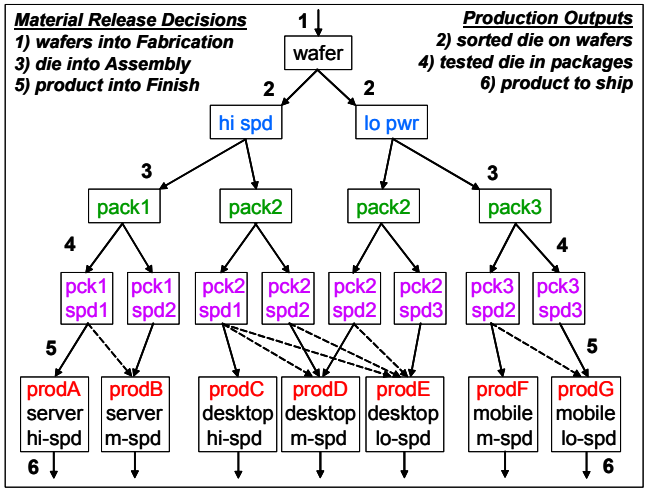

Fig. 2. Product fan-out in semiconductor manufacturing.

The product fan-out implied in Fig. 1 with the associated opportunity for delayed differentiation is shown explicitly from top to bottom in Fig. 2. Raw wafers are released into fabrication and exit covered with die exhibiting a range of properties. These are sorted into broad functional categories such as high operational speed and low power consumption. There is a correlation between speed and power with the fastest devices usually consuming the most power, and visa versa. This categorization is used in assembly to decide what die to put into which packages. In the case of microprocessors, die with the highest clock speed are placed into server packages while die with the lowest power consumption are placed into mobile packages. Both die types might be placed into desktop packages. Testing then splits the performance distributions into finer categories, usually by maximum clock speed. In finish, these categories are used to fill demand for specific products. It is sometimes the case that the multiple splittings of multiple distributions results in different production flows giving the same end product (as is seen in the middle of Fig. 2). In addition, demand for lower speed devices can be filled by configuring higher speed devices with an associated lost opportunity cost.

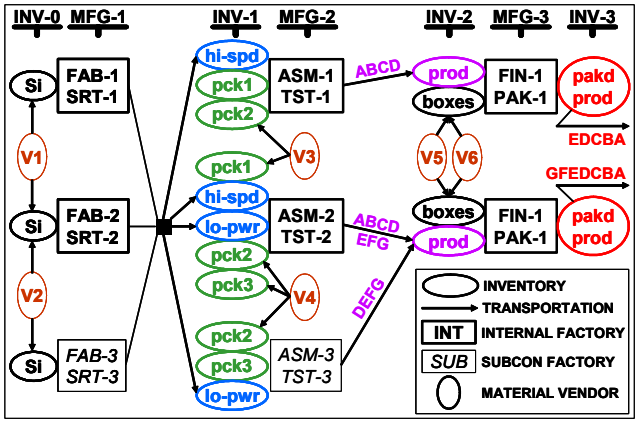

Fig. 3. Facilities topology in semiconductor manufacturing.

The network of facilities alluded to in Fig. 1 with the associated opportunity for risk sharing is shown explicitly from left to right in Fig. 3. The first few echelons show vendors of silicon wafers and transportation links supplying raw material warehouses in front of fabrication (fab) facilities. Multiple vendors and FAB/SRTs are involved to mitigate the risk of problems in any individual manufacturing facility. FAB/SRT facilities that are owned and operated by subcontractors are included as capacity buffers against variable demand. The middle few echelons show sorted die being transported to die warehouses in front of assembly (asm) facilities. In addition, vendors are supplying package warehouses through transportation links. Once again, multiple vendors and ASM/TST facilities are involved as well as subcontractors. Notice the assignment of different die types and different package types to different factories. The last few echelons show tested product being transported to finished goods warehouses in front of finish (fin) facilities as well as packing materials being supplied by vendors to pack facilities. In the example shown, multiple vendors and FIN/PAK facilities are present but no subcontractors are involved. This is a reflection, in this example, of the primary manufacturer directly managing all shipments to customers.

The core repetitive decisions that must be made during the operation of the supply-demand network are captured in these figures. Referring to Fig. 1, decisions must be made in every time period about how much of what material to release into fabrication, assembly, and finish facilities. In addition, in Fig. 2, how much material to put into which package in assembly, and how much of what material to configure into which product in finish must be decided. Supporting these production decisions are a set of inventory decisions as shown in Fig. 3. One has to do with deciding how much of which raw material to hold in front of the fabrication (wafers), assembly (packages), and finish (boxes) facilities. Another deals with deciding how much of which work in progress to hold at assembly (die) and finish (unmarked product) factories.

Previous remarks focused attention in the larger set of supply-demand network problems to those dealing with the operation of the manufacturing core. This list of relevant decisions further refines the focus on issues discussed here. Excluded are the lower level details of internal factory decisions such as the setup and maintenance of machines and the batching and prioritization of lots that must be made in all of the factories represented in Fig. 3. Details of the operations of the warehouses and transportation links are similarly not considered. At a higher level, details of the decisions for managing the market place are excluded. This includes such important decisions as product introduction strategies and the timing of special sales offers. The decision system of interest here assumes a valid demand signal from above and a capable set of facilities below.

There are a number of complications to overcome in the context of making these decisions. The most obvious from Figures 2 and 3 is the combinatorial complexity of the problem. In actual practice, across the breadth of the offerings of an international high volume manufacturer, there would be as many as 25,000 end products with the associated number of semi-finished goods, package types, and wafers. Across the globe, there would be as many as 100 factories and 500 inventory holding positions with the appropriate number of transportation links.

A second complexity is associated with Figures 2 and 3, that of supply stochasticity. Each of the factories manufacturing each of the products exhibits at least three types of variability based on different stochastic processes. The length of time it takes the raw material input to a factory to emerge as output, known as the throughput time (TPT), can be best described by a distribution. The underlying stochastic processes include the contention of lots flowing through the factory for production resources as well as the random unavailability of those resources. For example, machines experience breakdowns and operators take breaks. These TPT distributions are skewed since there are more events that can occur to slow a lot down and increase its TPT than there are events that can speed a lot up and decrease its TPT.

How much of the raw material released into the factory will emerge as output is also variable depending on a random set of unavoidable events that can occur. There are many examples of these stochastic processes. At a large scale compared with the transistors being fabricated, silicon wafers repeated heated and cooled in the production process occasionally experience thermal stress resulting in cracking with the accompanying lose of all of the die on the wafer since the wafer can not be further processed. At a

small scale, contaminates from the manufacturing process can fall onto a wafer in such a way as to short circuit transistor interconnects causing a die to malfunction and be rejected at sort.

How the output from each factory will function is another supply variability. Clock speed and power consumption are among the main functional characterizations of semiconductors and both of these can best be described with distributions. The semiconductor manufacturing process can be thought of as the arrangement of atoms to form and interconnect transistors. Controlling exactly how many atoms are added or subtracted in the individual process steps and exactly how they are positioned relative to each other involve many stochastic processes. It is the detailed outcome of all of these processes for each die that generates the functional distributions.

A third complexity is that of demand stochasticity, and it is difficult to describe the underlying random processes. Ultimately the end customer decides what products to purchase, and given the very large number of end customers and products, it is hard to forecast how much of which product will be purchased in which locations at what times because of the complex interplay of economic condition, need, and fashion. Compounding this situation is the competition in the marketplace between end product suppliers who differentiate based on form, function, price, and service (to mention but a few of the vectors). This translates into variability in the demand signal to the decision system of interest here. Orders are placed for semiconductor devices that include the specific product name and quantity as well as delivery time and place. Subsequent random events in the market result in requests to alter all of these parameters for existing orders, and the flow of materials through the supply-demand network must be altered to accommodate.

As can be seen in Figures 2 and 3, the design of the network addresses some aspects of these demand stochastics. The fan-out of products in Fig. 2 supports delayed differentiation, the final configuration being made only a few days before product released into the logistics network. The multi-trajectory flow in Fig. 3 (including multiple vendors and subcontractors) supports risk sharing. In each echelon, multiple factories are making the same product and multiple products are being made in the same factory. Between echelons, the output of each factory is tied to the input of many downstream factories just as the input of each is tied to many upstream outputs. While all of these mappings can be changed over time with the dynamics of the business, the decision system under discussion here will have to manage all of the remaining stochasticity by appropriately utilizing the product fan-out and multi-trajectory flows.

Magnifying the complexity of both the combinatorics and the supply-demand stochastics is the fact that many of the key relationships are nonlinear. Manufacturing TPT, as well as TPT variability, increase nonlinearly as the utilization of manufacturing resources increase due to congestion. The probability of stock out decreases nonlinearity as the amount of safety stock inventory increases. The well-known price elasticity curve expresses the nonlinear relationship between demand and selling price. And the history of the semiconductor industry shows a nonlinear drop in selling prices of individual products over their life cycle (currently in the range of 6 months to 3 years). Notice as well that these nonlinearities interact. A drop in price could lead to an increase in demand. An increase in demand could lead to a higher safety stock target to protect against stock out. An increase in demand and inventory increases the load and congestion in the factory leading to an increase in TPT. An increase in manufacturing TPT could lead to a less responsive system leading to an increase in stock out probability and a decrease in demand. The decision system directing the supply-demand network must recognize and deal with these interacting nonlinearities.

The final complexity in the decision problem described here has to do with the financial aspects of the problem. The primary goal in operating a supply chain is to realize a profit, and this involves a number of conflicting objectives. The most obvious is the tug of war between minimizing cost and maximizing revenues. More subtle is the balance between maximizing profits now (perhaps risking future profits) and maximizing profits in the future (perhaps delaying or foregoing current profits). Making too little of a product or delivering it late as a result of striving to minimize current or future costs too often lead to decreased revenues through delayed payments, late penalties, lost sales, and (in the worst case) lost customers. Making too much of a product or introducing it early before the market is ready as a result of trying to maximize revenue now and in the future too often lead to higher costs through inventory holding charges and increased risk of obsolesce leading to fire sales and write offs.

II. A CONTROL-ORIENTED SOLUTION APPROACH

Given an international supply-demand network that operates around the clock every day of the year, the resulting problem can be described as a continuous nonlinear stochastic combinatorial financial optimization. With the scale of international supply-demand networks, the difference between an optimal and a non-optimal solution can be worth hundreds of millions of dollars per year. For example, in the case of Intel Corporation with roughly 30B$ in annual revenue from its network, a 3 1/3% improvement in operations would result in an addition 1 B$ per annum.

The approach described here relies on splitting the problem into a strategic planning function and a tactical execution function as shown in Fig. 4. The former can be thought of as an outer loop controller that considers business goals and trends over months and quarters into the future, the latter as a companion inner loop controller that manages day to day operations providing the network with responsiveness and agility.

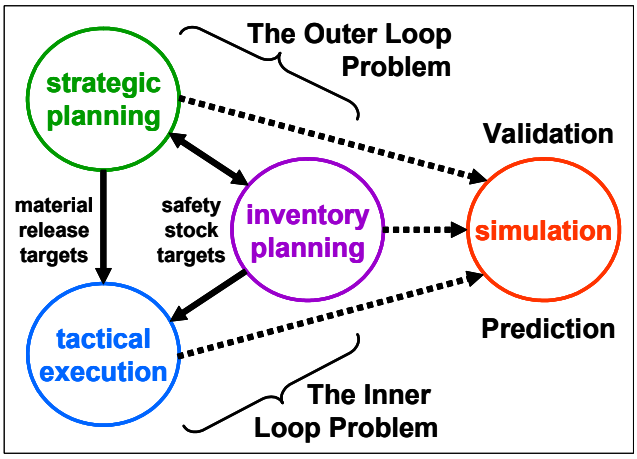

Fig. 4. Configuration of the proposed control system.

A major component of the strategic outer loop is a Linear Programming (LP) formulation that addresses the combinatorial financial optimization. The supply and demand nonlinearity and stochasticity is managed in two ways. On one hand, a major component of the tactical inner loop is a Model Predictive Control (MPC) formulation that accomplishes both feedback and feed forward control, selecting actions based on optimization of a control-relevant objective function. Stochasticity is contained by rapid measured responses as soon as deviation from the plan is recognized. On the other hand, a major component of both the outer and inner loops is an Inventory Planning formulation. The placement and sizing of safety stocks provides an additional hedge against both supply and demand stochasticity. From an outer loop perspective, inventory targets are set based on long term demand forecasts that include potentially large errors and passed to the inner loop to guide execution. From an inner loop perspective, upper and lower control limits can be placed around the targets based on recent history in the execution environment and used to guide current execution. Simulation supports all of these components by providing for the testing of policies and plans before they are implemented in the real supply-demand network. These components and interactions between them will be described in the next several sections.

A. A Strategic Planning Formulation

One solution to the strategic planning problem can be formulated as a mathematical optimization to allocate capacity to satisfy demand while minimizing costs and maximizing revenue [1], [2]. There are three major categories of inputs in this approach as shown in Fig. 5. One set of inputs specify the basic structure of the problem. This includes the material required to make any particular product as depicted in Fig. 2 and the possible manufacturing flows for products shown in Fig. 3. Supporting these descriptions are forecast future values of performance parameters including factory and transport TPTs and product yields. Financial data completes the set with the cost of manufacturing in each facility, transport costs for each link, the cost of holding inventory in each warehouse, and the average selling prices. Note that all of this data is time varying. New products are introduced and old products are discontinued with the associated modification of factory qualifications on a monthly basis. Yields and TPTs are intended to improve as products and factories mature over time. There is always external market pressure to lower selling prices and the associated internal pressure to reduce manufacturing costs. Note also that much of this data also varies by factory and product depending on the maturity of each. Different factories have different costs and performance parameters for manufacturing the same product. The same factory exhibits different performance parameters and costs for different products. Finally note that while TPTs and yields are the result of stochastic processes and are best described by distributions, only forecast means are used here.

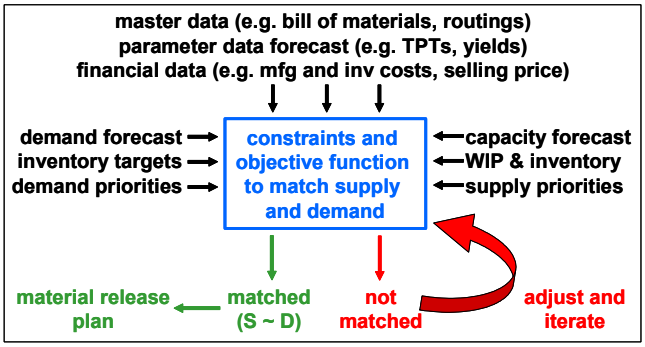

Fig. 5. A solution to the basic strategic planning problem.

The second set of inputs describes the supply scenario. Part of this description is the capacity forecast for all of the facilities in the system. This is product and facility specific and changes over time as individual facilities are modified. While there is a stochastic component to all of these capacity forecasts, again means are used. Another major part of the supply description is the current work in progress in each of the manufacturing facilities and transportation links as well as the current inventories in each warehouse. In practical applications there can be an arbitrary number of supply preferences included in the input. For example, it may be advantageous from a risk mitigation standpoint to distribute the load somewhat

evenly across the supply facilities rather than heavily loading the low cost facility while leaving the higher cost facilities under-loaded. It is often challenging to include heuristic preferences in a mathematical formulation since they are usually difficult to quantify.

The demand scenario is described in the third set of inputs. The demand forecasts by product can vary dramatically over time and are a collection of numerical estimates on a highly stochastic process rooted in the dynamics of the marketplace. Included are inventory targets intended to mitigate a portion of this demand stochasticity as well as part of the supply stochasticity. They are included on the demand side since they require capacity to be satisfied. The calculation of these targets will be the focus of a later section. Analogous to the supply scenario, an arbitrary number of demand preferences can be included. For example, a particular product or a particular customer may be deemed to have an elevated strategic importance relative to others although from a tactical financial perspective this may not be apparent. Once again, appropriate formulation is challenging but necessary to satisfy business needs.

Considering all of this information, the core formulation is based on mass balance and capacity constraints and an objective function that includes minimizing costs and maximizing revenues. The simple example formulation that follows includes only the left-most single flow leading to "prodA" from Fig. 2. From a facilities perspective, it includes only the top-most manufacturing facilities ("FAB-1 / SRT-1", "ASM-1 / TST-1", "FIN-1 / PAK-1") and the top-most product warehouses ("hi-spd", "prod", and "pakd prod") from Fig. 3. It ignores raw materials and transportation. Demand forecasts and inventory targets for the single product are included as are factory capacity, yield and TPT, initial factory WIP, and initial warehouse inventory for the single manufacturing line. The corresponding implied financial model includes manufacturing costs, inventory holding costs, penalties for missing inventory targets and demand, and the average selling price. It ignores both demand and supply priorities.

Indices:

m	manufacturing stage (i.e. MFG-2) and Sales
i	inventory location (i.e. INV-1)
p	product (intermediate or end product)
t	time

Input Variables:

$\mathrm{Cap}_{m,p,t}$	total product that can be in a mfg stage
$\mathrm{Yield}_{m,p,t}$	ave fraction of input to a mfg stage that exits
$\mathrm{TPT}_{m,p,t}$	ave number of periods to complete a mfg stage
$\mathrm{MfgCost}_{m,p,t}$	cost of manufacturing a product
$\mathrm{InvTar}_{i,p,t}$	inventory target
$\mathrm{InvCost}_{i,p,t}$	cost of holding inventory
$\mathrm{InvPen}_{i,p,t}$	penalty for not satisfying inventory target
$\mathrm{Dem}_{p,t}$	average demand
$\mathrm{SaleBen}_{p,t}$	benefit from selling the end product
$\mathrm{BackPen}_{p,t}$	penalty for not satisfying demand

Decision Variables (assume non-negativity):

$\mathrm{Rel}_{m,p,t}$	material released into a mfg stage
$\mathrm{Inv}_{i,p,t}$	inventory held at an inventory location
$\mathrm{Sales}_{p,t}$	amount of satisfied demand
$\mathrm{Backlog}_{p,t}$	amount of unsatisfied demand
$\mathrm{InvUnd}_{i,p,t}$	slack variables, for deviation from inv target
$\mathrm{InvOvr}_{i,p,t}$	

Initialization:

$\mathrm{Rel}_{m,p,t\in\{0-\mathrm{TPT}_{m,p,t},0\}}$	previous material release
$\mathrm{Inv}_{i,p,t=0}$	beginning on hand inventory
$\mathrm{Backlog}_{p,t=0}$	no initial unsatisfied demand

Capacity Constraint

$$\sum_{t'\in(t-TPT_T,t]} \mathrm{Rel}_{m,p,t'} \le \mathrm{Cap}_{m,p,t} \qquad (1)$$

$$\text{where } \mathrm{T} = t - \mathrm{TPT}_{m,p,T}$$

Inventory Mass Balance Constraint

$$\mathrm{Inv}_{i,p,t} = \mathrm{Inv}_{i,p,t-1} + \sum_{T:T=t-TPT_{m,p,T}} \left(\mathrm{Rel}_{m,p,T} \cdot \mathrm{Yield}_{m,p,T}\right) - \mathrm{Rel}_{m+1,p,t} \qquad (2)$$

Backlog Mass Balance Constraint

$$\mathrm{Backlog}_{p,t} = \mathrm{Backlog}_{p,t-1} + \mathrm{Dem}_{p,t} - \mathrm{Sales}_{p,t} \qquad (3)$$

$$\text{where } \mathrm{Sales}_{p,t} = \mathrm{Rel}_{m=Sales,p,t}$$

Inventory Target Constraint

$$\mathrm{Inv}_{i,p,t} + \mathrm{InvUnd}_{i,p,t} - \mathrm{InvOvr}_{i,p,t} = \mathrm{InvTar}_{i,p,t} \qquad (4)$$

Objective Function:

$$\max\left\{\begin{array}{l} +\sum_{p,t} \mathrm{Sales}_{p,t} \cdot \mathrm{SaleBen}_{p,t} \\ -\sum_{p,t} \mathrm{Backlog}_{p,t} \cdot \mathrm{BackPen}_{p,t} \\ -\sum_{m,p,t} \mathrm{Rel}_{m,p,t} \cdot \mathrm{MfgCost}_{m,p,t} \\ -\sum_{i,p,t} \mathrm{Inv}_{i,p,t} \cdot \mathrm{InvCost}_{i,p,t} \\ -\sum_{i,p,t} \left(\mathrm{InvUnd}_{i,p,t} + \mathrm{InvOvr}_{i,p,t}\right) \cdot \mathrm{InvPen}_{i,p,t} \end{array}\right\} \qquad (5)$$

The principal result of executing such a formulation as a Linear Program is a material release plan (or equivalently a product output plan) for all the manufacturing facilities for each time segment. Also available in the output is a transportation plan, an inventory plan including level relative to the specified targets, and a demand satisfaction plan including backlog. A profit projection is also available.

In an industrial setting, this tool is used for solidifying the response of the company to the market. It is rarely the case that initial executions provide a satisfactory match, usually leaving some demand unsatisfied and some capacity unused. In these circumstances, four of the inputs shown in Fig. 5 are manipulated. Selected demand can be moved earlier (incurring inventory holding costs), specific backlog can be authorized (incurring potential penalties), or some demand can be ignored (incurring lost revenue). Particular inventory targets can be adjusted (with attendant changes in inventory holding costs and penalties). Of course, demand and supply priorities can be modified potentially changing all facets of the plan. These manipulations reflect the practical impossibility of including all of the relevant business considerations in the formulation. Fortunately, a satisfactory strategic plan can usually be realized within a manageable number of iterations.

The principal shortcoming of this formulation is its treatment of the operational dynamics of the system. The most obvious is the disregard of the stochastic nature of the input parameter data, demand forecast, and capacity forecast. Addressing this deficiency is the focus of most of the rest of this paper. Another is the use of time periods with associated buckets of capacity to model the flow of work through manufacturing. An example of the difficulties this precipitates is the disregard of the fact that factory TPT rises nonlinearly with factory utilization. To build the plan, fixed factory TPTs are input as parameters. The plan that results specifies the amount of material to be released into those factories, which determine their work loads and, in turn, the TPTs that will be realized.

One approach to this circularity has been to iteratively use an LP formulation and a discrete event model as described by Hung and Leachman [3]. Given an initial TPT, the LP provides an initial plan. This plan is executed in simulation to determine the resulting TPT. This TPT is fed back to the LP, and iteration is continued seeking convergence. Although convergence is not guaranteed, the results are generally adequate for most real-world applications. Unfortunately, solution time of such a scheme can be quite long if the simulation model contains the detail necessary to generate accurate TPTs.

More recently an approximate but very efficient approach has been developed by Asmundsson, Rardin, and Uzsoy [4], [5] using the idea of clearing functions. Here the expected throughput of a capacitated factory in a period of time is expressed as a function of its workload in that period. Based on an outer linearization of this nonlinear clearing function, an LP is formulated that appropriately captures the dynamics of factory capacity, TPT, and work load in a manner that supports rapid execution.

B. An Inventory Planning Formulation

One popular approach to managing the inherent stochasticity in the supply and in the demand is to put safety stock in place [6], [7], [8]. Such extra inventory hedges the risk of an unexpected supply downside or demand upside. The important decisions include where to hold the safety stock as well as how much to hold, both varying over time. Consider the product and facilities topologies in Figs. 2 and 3, respectively.

Positioning completed product in the final warehouse in the manufacturing flow so that it can be shipped from stock to the customer on demand tends to maximize revenues, but requires complete product differentiation and incurs full manufacturing costs. Positioning extra undifferentiated material near the beginning of the manufacturing flow tends to minimize manufacturing costs, but requires customers to wait for the entire manufacturing TPT for their orders to be shipped.

Holding too much extra inventory as safety stock increases holding costs and risks steep discounts or write-offs as the market becomes saturated or the products becomes obsolete. It can also waste manufacturing capacity, building what ultimately turns out to be the wrong product, potentially precluding building other more appropriate revenue products. Holding too little safety stock risks stock-outs with late delivery penalties, lost revenue, and in the worst case, lost customers. It can also stress the manufacturing system with rush orders and increased congestion in the factories leading to longer TPTs for all orders and the possibilities of lower yields.

Practical complications abound. For example, safety stock requirements change through the life cycle of products. In the ramp-up phase when market penetration is of paramount importance and stock-outs can not be tolerated, high levels of safety stock might be desirable but might be difficult to attain with immature production facilities. During the middle phase of market stability, two situations are possible based on the assumption that the initial market forecasts used to put capacity in place were flawed. If forecasts were pessimistic and the network is supply constrained, building safety stocks will be difficult since all capacity is being used to fill orders. If the network is demand constrained due to optimistic forecasts, the difficulty will be in overbuilding safety stock since manufacturing personnel loath idle capacity. In the ramp-

down phase of a product's life cycle, safety stock is a liability when the focus is on moving customers to new improved products.

Furthermore, in operating supply-demand networks there is a substantial amount of inventory present and it can be difficult to identify that which is extra. In an efficient network, the majority of the material flowing through manufacturing and transportation facilities is destined to satisfy firm customer orders. The majority of the material in warehouses has a very low residency time as it moves steadily toward the market. However, some material might appear to be extra. Raw materials might have to be ordered in batches larger than can be immediately consumed. Intermediate production may arrive in a warehouse ahead of schedule due to manufacturing stochasticity in either yield or TPT. Final product for actual customer orders may be built ahead due to mismatches between demand and supply at the time the order is due. None of these can be considered safety stock which must be scheduled into production facilities in addition to that committed to firm orders.

The simple supply-demand network described in Figs. 2 and 3 is among the most difficult for which to compute safety stock positions and amounts even for a bounded time horizon [9], [10]. Multiple facilities exhibit a complex network flow with multiple products. Both supply and demand forecasts included substantial stochasticity with means that can vary over time. The flow contains multiple points at which multiple raw materials from outside vendors are injected. Fortunately the same product can be made through a number of routes, and the products incrementally differentiate along the flow providing opportunities for risk pooling.

One approach to the computation of safety stock positions and amounts is to construct a simulation that incorporates many of these complexities and then use it to search the space of possible positions and amounts for a good (but possibly not optimal) solution. Simulation speed and accuracy as well as the efficiency and effectiveness of the search control algorithm are crucial to the practicality of this approach. Glasserman and Tayur [11] have demonstrated a gradient estimation technique called infinitesimal perturbation analysis (IPA) for estimating from simulation the derivatives of inventory costs with respect to policy parameters. They have shown that they can use these derivatives to help steer the search for improved policies in multi-echelon systems with demand uncertainty related to those considered here.

Other approaches rely on representing the safety stock problem in such a way that mathematical optimization techniques can be employed. For example, Graves and Willems [12], [13] have developed a method for supply-demand networks modeled as spanning trees that captures the stochastic nature of the demand and allows the safety stock problem, given a few key simplifying assumptions, to be formulated as a deterministic optimization. The goal of this method is to place and size decoupling safety stocks that are large enough to permit downstream portions of a network to operate independently from the upstream, provided the upstream portion replenishes the external demand in a timely fashion. The simplifying assumptions include 1) bounded demand, 2) deterministic production TPTs at each stage that are independent of load (this is equivalent to assuming no capacity constraints), and 3) guaranteed service times by which each stage will satisfy its downstream demand. The first assumption is a practical one reflecting the fact that, to cover any possible demand eventuality however improbable, very large inventories would have to be positioned. Bounding demand simply means that, in extraordinary demand scenarios, the personnel operating the network would take extraordinary measures in response. The second assumption clarifies the focus on demand variability (not supply variability) and inventory target setting (not capacity allocation). The third assumption is key to the formulation. These service times for both end items and the internal stages are decision variables an optimization model used here including the possibility of setting a maximum service time for end items as required for customer satisfaction.

This inventory planner models the supply chain as a set of nodes and arcs where the nodes denote a processing function and the arcs capture the precedence relationship between nodes. While the LP formulation draws a distinction between manufacturing stages and inventory locations, here the stages are defined such that they are also potential stocking locations and can hold safety stock after processing activity has been completed. The decision variables of a stage must be bound to prevent the possibility of it holding safety stock. This safety stock optimization problem can be formulated as a mathematical program …

Indices:

i stage

Input Variables:

$MaxDem(\tau)_i$	max demand at a stage over an interval τ
$AvgDem(\tau)_i$	ave demand at a stage over an interval τ
TPT_i	ave number of periods to complete a stage
$InvCost_i$	cost of holding inventory
$MaxServ_i$	max outgoing service time at a stage

Decision Variables (assume non-negativity):

$ServOut_i$	outgoing service time at stage i
$ServIn_i$	incoming service time to stage i

Constraints:

$$ServOut_i - ServIn_i \le TPT_i \quad \text{for all nodes} \quad (6)$$

$$ServOut_i \le MaxServ_i \quad \text{for all nodes} \quad (7)$$

$$ServIn_i - ServOut_{i-1} \ge 0 \quad \text{for all arcs} \quad (8)$$

Objective Function:

$$\min_{ServOut_i} \sum_i InvCost_i [MaxDem_i(ServIn_i + TPT_i - ServOut_i) - AveDem_i(ServIn_i + TPT_i - ServOut_i)] \quad (9)$$

In this formulation, the service times at a stage are the decision variables. It is assumed that each stage quotes its downstream customers a guaranteed service time (ServOut) by which time it will satisfy demand requests. The incoming service time to a stage (ServIn) is the maximum outgoing service time that its upstream supplier stages quote. The net replenishment time τ at each stage dictates the inventory requirements at each stage to cover the demand over this time. The net replenishment time equals the outgoing service time at a stage minus the incoming service time plus the production time (TPT). The function $\text{MaxDem}(\tau)_i$ characterizes the maximum demand at each stage as a function of the net replenishment time. If a stage has a net replenishment time of t, the stage sets its base stock level equal to $\text{MaxDem}(\tau)_i$. The expected safety stock level at stage *i* will then equal the maximum demand minus the expected demand over the interval of length τ.

This formulation is solvable by dynamic programming (DP) where each stage solves a functional equation f(ServOut) or g(ServIn) depending on its orientation in the network. The DP seeks to determine the optimal set of service times that minimizes total safety stock cost while satisfying the maximum service time constraints to the final customer. At each stage, the net replenishment time, τ equals ServIn +TPT -ServOut. The constraints on service time ensure that net replenishment times are nonnegative, incoming service time is no less than the maximum outgoing service time quoted to the stage, and the outgoing service times of demand stages do not exceeds the maximum service times imposed by customers.

From a practical perspective, both of these approaches can raise philosophical questions. Both suggest the computation of inventory targets preceding the computation of the strategic plan on the grounds that the LP used for strategic planning expects inventory targets as input. The philosophical difficulty with this is rooted in the iterative business use of the LP as described previously. That iteration is motivated by the desire to play out multiple supply and demand scenarios to find the one that best suits the overall strategy of the company. From this perspective, the role of safety stock is to protect the plan that is ultimately selected from the relevant stochasticities that might disrupt it, and this can not be done a priori.

Note that both the stochastic simulation and dynamic programming approaches utilize some form of mapping between demand and production facilities during their computation. Notice also that the LP allocates specific demand to specific capacitated production facilities during its operation. This means that an iterative scheme might be appropriate. Two starting points are possible. In one, the LP is run first including heuristically set inventory targets, allocations result that are passed to the inventory computation for its first run, inventory targets result that are passed to the LP for a second run, and so on until convergence is attained. In the other, the iteration is initiated by first running the inventory computation with heuristically set allocations.

The results of initial investigations of this iterative scheme are in preparation, and while promising, have identified an additional difficult decision problem. When the capacity of the supply system is greater than the sum of the demand forecast and the safety stock computed, convergence is relatively easily identified. In circumstances when either a) supply capacity is greater than demand forecast but less than demand forecast plus safety stock, or b) supply capacity is less than demand forecast and less than the sum of the demand forecast and the safety stock computed, convergence can not be resolved until additional capacity allocation choices are made. The decision is between building for demand or building for safety stock. This is especially interesting when deciding such comparisons as demand for lower margin products and safety stock for higher margin ones, or demand for products at the end of their life cycles and safety stock for ones early in theirs.

C. A Tactical Execution Formulation

The combination of strategic planning and inventory planning addresses a large portion of the continuous nonlinear stochastic combinatorial financial optimization problem of concern here. But given the data preparation requirements for these tools, as well as the business processes necessary to modify strategic directions, it is not likely that planning at the strategic level will be practical or effective more often than once or twice per week for operational purposes. (Note that it is entirely possible that these tools will be used much more often to explore supply and / or demand scenarios in the process of considering strategic options offline.) Unfortunately the nonlinear and stochastic aspects of a continuously operating supply chain are active minute to minute, hour by hour, day after day. It is clear that responding on a timescale much shorter than weekly will result in lower supply chain costs and

improved levels of delivery performance generating higher revenues.

The approach suggested is not more rapid execution of the outer loop tools, but rather relies on decision policies based on control-theoretic concepts applied to supply-demand networks. For more than 50 years, control methodologies have been continuously improved and reduced to reliable practice in a variety of process industries [14]. Process control systems are widely used to adjust flows to maintain product levels and compositions at desired levels. This is analogous to the management goals of high volume supply-demand networks and material flows in these networks can be modeled using a fluid analogy as shown in Fig. 6. In a very general sense, the manufacturing stages are represented as long and leaky "pipes" (to include TPT and yield, respectively) with the material in the pipes correspond to production work in progress. Additional pipes represent transportation links containing work in transit. Warehouses are represented as holding "tanks" and their contents correspond to inventory. Decisions about releasing material to initiate a production process or satisfy demand are implemented by adjusting control valves. Compare the system shown in Fig. 6 with the top-most facilities in Fig. 3. As a result of using this fluid analogy, one can expect that decision policies based on process control concepts to have a large and beneficial impact on supply chain management.

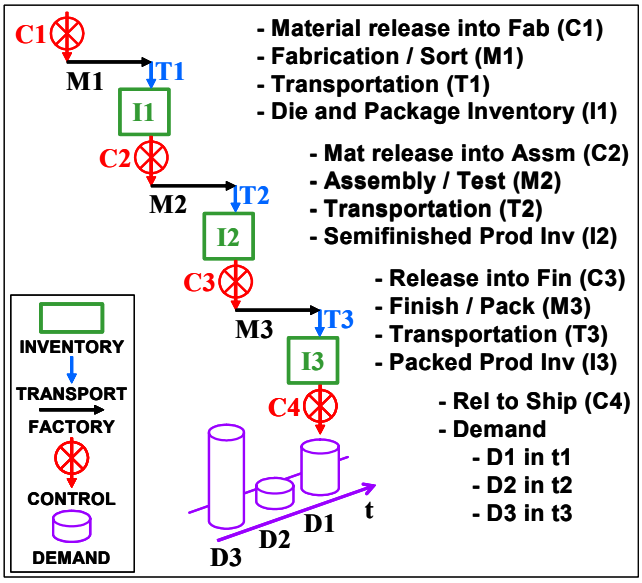

Fig. 6. A fluid analogy for semiconductor manufacturing.

In particular, Model Predictive Control (MPC) offers a combined feedback-feed forward decision framework that can be tuned to provide enhanced performance and robustness in the presence of significant supply and demand variability and forecasting error while enforcing constraints on inventory levels and production and transportation capacities. Its formulation integrates optimal control, stochastic control, multivariable control, and control of processes with dead time. MPC is arguably the most general method currently known of posing the process control problem in the time domain [15]. In addition, there are early indications that MPC is applicable to the supply-demand network problems of interest here [16]-[21].

In MPC, a system model and current and historical measurements of the process are used to predict the system behavior at future time instants. A control-relevant objective function is then optimized to calculate a sequence of future control moves that must satisfy system constraints. The first predicted control move is implemented and at the next sampling time the calculations are repeated using updated system states. This is referred to as a Moving or Receding Horizon strategy and is illustrated in Fig. 7.

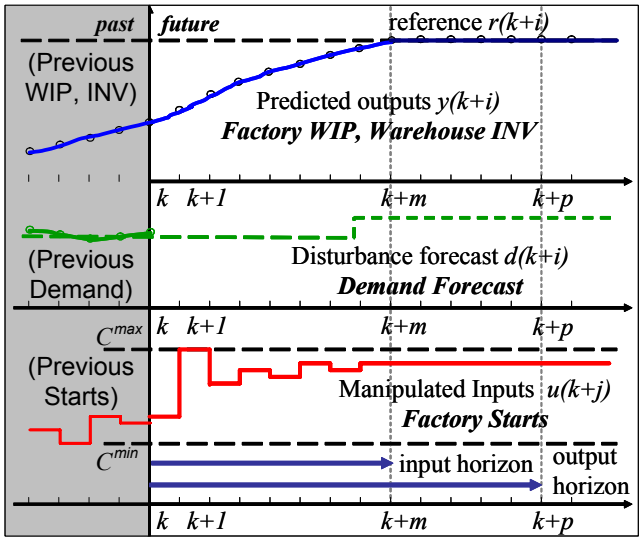

Fig. 7. The moving or receding horizon strategy.

Input variables consist of two types: manipulate variables ***u*** which can be adjusted by the controller to achieve desired operation and disturbance or exogenous variables ***d***. The starts rates for F/S (C1), A/T (C2), and F/P (C3) represent manipulated variables for the problem in Fig. 6, with suggested targets determined by the strategic planning module. Demand (D1, D2, …) in Fig. 6 is treated as an exogenous signal. This signal consists of actual demand which is only truly known in the past and for a very short time into the future, and forecasted demand which is provided to all of the components shown in Fig. 4 by a separate organization such as Sales and Marketing. As noted in Fig. 7, the demand forecast is used in the moving horizon calculation to anticipate future system behavior and plays a significant role in the starts decisions made by the MPC controller. Representing quantities of primary importance to the system, ***y*** is a vector of output variables. Outputs can be classified in terms of controlled variables

which must be maintained at some setpoint value and associated variables which may not have a fixed setpoint, but must reside between high and low limits. For the problem in Fig. 6, controlled variables consist of the three inventory levels (I1, I2, and I3) whose setpoint targets are determined by the inventory planning module. Associated variables include loads on the manufacturing nodes (M1, M2, and M3) determined on the basis of their WIP.

In MPC control, predictions of $\boldsymbol{y}$ over a horizon are computed on the basis of an internal model arising from mass conservation relationships describing the dynamics of the manufacturing, inventory, and transportation nodes. For the problem in Fig. 6, the mass conservation relationship for inventories *I* can be written as:

$$I_x(k+1) = I_x(k) + C_x Y_{M_x}(k - TPT_{M_x}) - C_{x+1}(k) \qquad (10)$$

where x can equal 1, 2, or 3. For the manufacturing nodes an expression for the work in progress *WIP* is written as:

$$WIP_{M_x}(k+1) = WIP_{M_x}(k) + C_x(k) - C_x(k - TPT_{M_x}) \qquad (11)$$

where x can equal 1, 2, or 3, and *TPT* and *Y* represent the nominal throughput time and yield for the manufacturing node, respectively, while *C* represents the manufacturing starts per time period that constitute inflow for factories and outflow streams for warehouses. These systems of equations can, in general, be organized into a discrete-time state-space model representation amenable to Model Predictive Control implementation and analysis [22].

The goal of the MPC decision policy is to seek a future profile for $\boldsymbol{u}$, the manipulated variables, that brings the system to some desired conditions consistent with the relevant constraints and the minimization of an objective function.

The ability to address constraints explicitly in the controller formulation is part of the appeal of MPC. For the problem in Fig. 6, constraints need to be imposed on the magnitudes of factory starts (12), the changes in factory starts (13), factory loads (14), and warehouse inventory levels (15).

$$0 \le C_x(k+j \mid k) \le C_x^{\max} \quad \text{x=1 to 3} \quad \text{j=1,2,…,m} \qquad (12)$$

$$\Delta C_x^{\min} \le \Delta C_x(k+j \mid k) \le C_x^{\max} \quad \text{x=1 to 3} \quad \text{j=1,2,..,m} \qquad (13)$$

$$0 \le WIP_{M_x}(k+i \mid k) \le CAP_{M_x}^{\max} \quad \text{x=1,2,3} \quad \text{i=1,2,..,p} \qquad (14)$$

$$I_x^{\min} \le I_x(k+i \mid k) \le I_x^{\max} \quad \text{x=1,2,3} \quad \text{i=1,2,..,p} \qquad (15)$$

While there is significant flexibility in the form of the objective function used in MPC, a meaningful formulation for the problem in Fig. 6 is:

$$\begin{aligned} &\min_{\Delta C(k|k)\ldots\Delta C(k+m-1|k)} J = \\ &\sum_{x=1}^{3}\sum_{j=1}^{m} Q_C\left(C_x(k+j-1 \mid k) - C_{x:tar}(k+j-1 \mid k)\right)^2 + \\ &\sum_{x=1}^{3}\sum_{j=1}^{m} Q_{\Delta C}\left(\Delta C_x(k+j-1 \mid k)\right)^2 + \\ &\sum_{x=1}^{3}\sum_{i=1}^{p} Q_I\left(I_x(k+i \mid k) - I_{x:tar}(k+i)\right)^2 \end{aligned} \qquad (16)$$

The first input target term is meant to maintain the starts close to target values for each time period over the move horizon $\boldsymbol{m}$ based on the targets calculated by the outer loop strategic planner. The second move suppression term penalizes changes in the starts over the move horizon $\boldsymbol{m}$. This term serves an important control-theoretic purpose as the primary means for achieving robustness in the controller in the face of uncertainty [15]. The third setpoint tracking term is intended to maintain inventory levels at targets specified by the outer loop inventory planner over time. These targets need not be constant and can change over the prediction horizon $\boldsymbol{p}$. The emphasis given to each one of the sub-objectives is achieved through the choice of weights *Q* that can potentially vary over the move and prediction horizons.

For an MPC system relying on linear discrete-time state-space models to describe the dynamics, with an objective function as described above, and subject to linear inequality constraints, a numerical solution is achieved via a Quadratic Program (QP). Depending on the nature of the objective function, model and constraint equations, other programming approaches (LPs) may also be utilized [23].

TABLE I
MAJOR EXPERIMENTAL INPUTS

	M1	M2	M3
Planned Starts			
(units / day)	1,025	975	960
	I1	I2	I3
Inventory Targets			
(units)	2,700	1,550	1,450
	Ave		Var
Demand			
(units/day)	950		150
(uniform distribution)			

It is suggested that MPC-based formulations are able to perform satisfactorily if properly tuned in spite of the nonlinearities and stochasticity associated with semiconductor manufacturing supply-demand networks.

This is illustrated for the representative problem described in Fig. 6. The major inputs to the controller for the experiment are shown in Table I including planned starts from the strategic planning system, inventory targets from the inventory planning system, and the demand forecast as used by both outer loop modules. The experimental model parameters for factories are shown in Table II including fixed factory capacity with stochastic TPTs and yields. Note that the TPTs are nonlinear with load for the fab/sort factory.

TABLE II
EXPERIMENTAL MODEL PARAMETERS

			M1	M2	M3
Capacity	*(concurrent items*	Max	45,000	7,500	2,500
	in factory)	Initial	33,500	5,700	1,900
TPT	*(uniform distribution)*				
		Min TPT (days)	30	5	1
	Load 0% to 70%	Ave TPT (days)	32	6	2
		Max TPT (days)	34	7	3
		Min TPT (days)	32	5	1
	Load 70% to 90%	Ave TPT (days)	35	6	2
		Max TPT (days)	38	7	3
		Min TPT (days)	35	5	1
	Load 90% to 100%	Ave TPT (days)	40	6	2
		Max TPT (days)	45	7	3
Yield	*(uniform distribution)*	Min Yield (%)	93	98	99
	(applied at factory exit)	Ave Yield (%)	95	99	99
		Max Yield (%)	97	99	100

Demand can be measured and used to make predictions, but can not be manipulated. Planned starts can be manipulated by the controller and are modified to satisfy customer demand and keep inventories as close to target as possible while keeping factory within their allowed operational parameters.

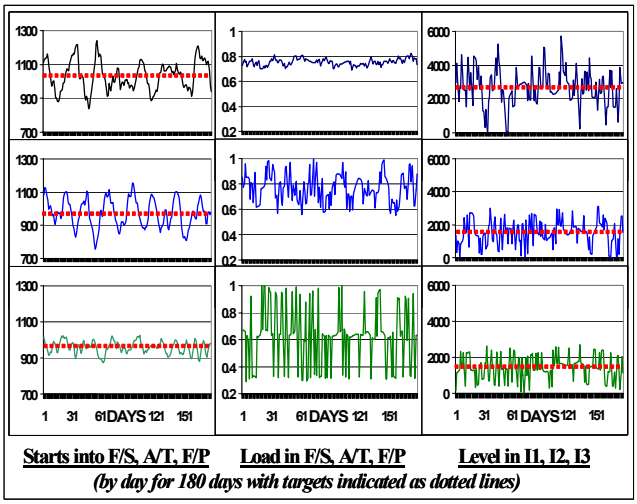

Fig. 8. The results of an MPC experiment.

The performance of the controller over a six month trial using $Q_C = 0, Q_{\Delta C} = 10, Q_I = 1$ is shown in Fig. 8. The factory starts in the left column have been adjusted relative to their targets in each time period while the loads in each factory in the middle column are appropriately managed. The load in the F/S factory is maintained at a high and stable level as desired for very expensive factories with long TPTs where "thrash" is likely to degrade performance. At the other end of the manufacturing flow, the F/P factory load varies from 30% of its maximum to 100% as expected in a low cost very short TPT factory that is tracking wildly fluctuating demand. The A/T factory in the middle is absorbing supply stochasticity as well as demand stochasticity. The inventory levels in the right column are performing a similar function. In all three cases, the inventory averages are roughly at their target levels. But in each case there are times when levels range from zero (stock-out for a few days) to brief periods at nearly double target levels. This is precisely what the safety stocks are intended to do, insulating manufacturing facilities from a large part of the variation in the system while assuring demand satisfaction. It is important to note that there was no backlog of customer orders during the six month experiment.

D. Simulation Support

Simulation can play many roles in the system described here for managing supply-demand networks. In each case, the speed with which the plan is generated, or the quality of the resulting plan, or the confidence in the plan producing the expected results (or some combination of these factors) is increased.

In the outer loop, there are multiple ways to utilize simulation. As described previously, using appropriate search control (such as Infinitesimal Perturbation Analysis [11]), a simulator can be run repeatedly to compute inventory targets. Or if inventory calculations are performed using mathematical optimization (for example, with Dynamic Programming [12], [13]), a simulator that models the stochasticity of the real environment can be run to test whether the computed targets perform as expected. A similar test could be performed on the plan that results from running the strategic planning module (the LP) alone or in tandem with the inventory planning module. This is similar in spirit to the iterative computation described previously between an LP model and a simulator to manage the nonlinearity between TPT and factory utilization [3]. The idea behind all of these activities is, if there is some question about whether or not the planning system has adequately modeled the nonlinearity and stochasticity in the execution environment, build a simulation of the execution environment and use it to evaluate the plan.

One interesting extension of this idea addresses the fragility of plans. Assume that a system has been implemented including a strategic planner and an inventory planner that iteratively generate a plan, and that an accurate simulation

of the execution environment is available. Assume that the business problem being addressed is complex enough that a number of strategies have been investigated. It is tempting to believe that, since the solution to each strategy has been generated by mathematical optimization, the plan with the best objective function value is the plan that should be passed into the execution environment. Unfortunately this overlooks the fact that little if any of the dynamics of the supply-demand network have been included in the optimization model. A more discriminating process would take each plan generated from each business strategy and subject it to multiple executions in the simulation environment. The result would be a distribution of results for each providing insight into plan fragility. The plan that appears best upon initial generation might not be the most robust under stress. Additional decision making would be required to make the selection based on such criteria as worst case, expected, and best case profit, but the plan selected would have a higher likelihood of producing the desired results.

In the inner loop, there are also a number of important uses for simulation. In developing a controller, a test bed that simulates the intended application environment is needed to tune and test the formulation and parameters. Although the controller might not deal with all of the intricacies of the real world, the simulation should to adequately test the system before it is deployed. In the Model Predictive Control approach advocated here [15], a simulation to provide repeated forward projection is an inherent part of the method. Higher fidelity projection generates a number of benefits, although as shown in the experiment presented earlier, very impressive results are produced even from an approximate simulation.

This range of uses highlights one of the inherent tradeoffs that must be weighed in applying simulation to the supply-demand network problem. Generally, the higher the accuracy required of the simulation, the longer the computation time. On one hand, a very detailed discrete event simulation could be used internally in a MPC formulation to produce very high quality results for forward projection, but the run time of the resulting controller might compromise its usefulness. On the other hand, a very abstract fluid flow simulation could be used externally to the strategic and inventory planning modules to very quickly evaluate proposed plans, but the lack of detail might compromise the desired discriminating power.

Recent efforts have focused on improving this tradeoff. Historical applications of simulation to manufacturing have included fine details of the production process as well as the equipment and the personnel to address problems of factory design and operation. Including this level of detail when considering the set of manufacturing, warehousing, and transportation entities involved in a supply-demand network would raise computation times well beyond practical limits. To address this difficulty, there have been explorations to find the simplest possible discrete event elements that can be coupled to provide a sufficiently accurate simulation for these networks [24], [25]. Another approach to modeling is the use of fluid networks as used, for example, in traffic theory [26]. Expanding this type of modeling to address the specific features important to simulating supply-demand networks has been explored recently [26]-[30].

Whatever the simulation approach used, robustly interfacing the components shown in Fig. 4 is a challenge. The simulation component is present as a way to accurately model the material flow in the system. The other components are all involved in the decision flow required to manage the network. Algorithms for modeling material flow, whether based on differential equations or discrete events, are very different from the optimization and control algorithms used to model decision flow. From one perspective, interfacing these algorithms is equivalent to modeling the data and information flow throughout the network. Recent work in this area has focused on a versatile methodology to interconnect a wide variety of simulation approaches and decision approaches based on the computing principles of model composability and system interoperability [31], [32].

III. CONCLUSION

Efficient operation of a complex supply-demand network can provide a company with a substantial competitive advantage. Unfortunately the planning problem for the manufacturing core of a network requires continuous nonlinear stochastic combinatorial financial optimization. One of the best ways to attack such a difficult problem is from a control-oriented perspective. An outer loop combining strategic planning and inventory planning can address the combinatorial and financial aspects of the problem as well as one of the major nonlinearities. An inner loop based on Model Predictive Control can take material release and inventory target plans from the outer loops (with inventory control limits from recent execution results) and provide excellent customer service over long periods of time in the presence of nonlinearities and stochasticities. Multiple simulation techniques can provide support for each of these activities. Integrating these modules as a control system promises to deliver practical solutions to this very difficult but very important economic problem.

IV. ACKNOWLEDGMENT

The author thanks John Bean of the Decision Technologies Group at Intel Corporation for the linear programming formulation, Sean Willems of the School of Management at Boston University for the dynamic programming

formulation, and Wenlin Wang of the Department of Chemical and Materials Engineering at Arizona State University for the quadratic programming formulation. The author also acknowledges Jakob Asmundsson, John Bean, Amit Devpura, Gary Godding, Michael O'Brian, Shamin Shirodkar, and Kirk Smith of Intel Corporation and Dieter Armbruster, Daniel Rivera, and Hessam Sarjoughian of Arizona State University for the stimulating interactions that have generated the approach to managing the core manufacturing component of a complex supply-demand network described here.

REFERENCES

[1] W. J. Hopp and M. L. Spearman, *Factory Physics: Foundations of Manufacturing Management*, New York: McGraw Hill, 1996, ch. 16 (Aggregate and Workforce Planning).

[2] S. Chopra and P. Meindl, *Supply Chain Management: Strategy, Planning, and Operation*, Upper Saddle River, NJ: Prentice-Hall, 2001, part 2 (Planning Demand and Supply in a Supply Chain).

[3] Y.-F. Hung and R. C. Leachman, "A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations," *IEEE Trans. on Semiconductor Manufacturing*, vol. 9, no. 2, pp. 257-269, 1996.

[4] J. Asmundsson, R. L. Rardin, and R. Uzsoy, "Tractable nonlinear capacity models for production planning part I: modeling and formulation," *Operations Research*, submitted for publication.

[5] J. Asmundsson, R. L. Rardin, and R. Uzsoy, "Tractable nonlinear capacity models for production planning part II: implementation and computational experiments," *Operations Research*, submitted for publication.

[6] H. L. Lee and C. Billington, "Managing supply chain inventories: pitfalls and opportunities," *Sloan Management Review*, vol. 33, pp. 65-73, Spring 1992.

[7] W. J. Hopp and M. L. Spearman, *Factory Physics: Foundations of Manufacturing Management*, New York: McGraw Hill, 1996, ch. 17 (Inventory Management).

[8] S. Chopra and P. Meindl, *Supply Chain Management: Strategy, Planning, and Operation*, Upper Saddle River, NJ: Prentice-Hall, 2001, part 3 (Planning and Managing Inventories in a Supply Chain).

[9] S. C. Graves, "Safety stocks in manufacturing systems," *J. Manufacturing and Operations Management*, vol. 1, pp. 67-101, 1988.

[10] P. H. Zipkin, *Foundations of Inventory Management*, New York: McGraw Hill, 2000.

[11] P. Glasserman and Sridhar Tayur, "Sensitivity analysis for base-stock levels in multiechelon production-inventory systems," *Management Science*, vol. 41, no. 2, pp.263-281, 1995.

[12] S. C. Graves and S. P. Willems, "Optimizing strategic safety stock placement in supply chains," *Manufacturing and Service Operations Management*, vol. 2, no. 1, pp. 68-83, Winter 2000. (Erratum, *M&SOM*, vol. 5, no. 2, pp. 176-177, Spring 2003.)

[13] S. Willems, "A Tutorial on Strategic Safety Stock Placement in Supply Chains," this volume.

[14] B. A. Ogunnaike and W. H. Ray, *Process Dynamics, Modeling, and Control*, New York: Oxford University Press, 1994.

[15] C. E. Garcia, D. M. Prett, and M. Morari, "Model predictive control: theory and practice - a survey," *Automatica*, vol. 25, no. 3, pp. 335-348, 1989.

[16] S. Tzafestas, G. Kapsiotis, and E. Kyriannakis, "Model-based predictive control for generalized production planning problems," *Computers in Industry*, vol. 34, no. 2, pp. 201-210, 1997.

[17] E. Perea-Lopez, B. E. Ydstie, and I. E. Grossmann, "A model predictive control strategy for supply chain optimization," *Computers and Chemical Engineering*, vol. 27, pp. 1201-1218, 2003.

[18] M. W. Braun, D. E. Rivera, M. E. Flores. W. M. Carlyle, and K. G. Kempf, "A model predictive control framework for robust management of multi-product multi-echelon demand networks," *Annual Reviews in Control (Special Issue on Enterprise Integration and Networking)*, vol. 27, no. 2, pp. 229-245, 2003.

[19] M. W. Braun, D. E. Rivera, W. M. Carlyle, and K. G. Kempf, "Application of model predictive control to robust management of multi-echelon demand networks in semiconductor manufacturing," *Simulation: Transactions of the Society for Modeling and Simulation International*, vol. 79, no. 3, pp. 139-156, March 2003.

[20] W. Wang, D.E. Rivera, and K.G. Kempf, "Centralized model predictive control strategies for inventory management in semiconductor manufacturing supply chains," in *Proc. 2003 American Control Conference (Denver),* pp. 585-590, 2003.

[21] W. Wang, D. E. Rivera, K. G. Kempf, and K. D. Smith, "A model predictive control strategy for supply chain management in semiconductor manufacturing under uncertainty," this volume.

[22] J. H. Lee, M. Morari, and C. E. Garcia, "State-space interpretations of Model Predictive Control," *Automatica*, vol. 30, no. 4, pp. 707-717, 1994.

[23] F. D. Vargas-Villamil, D.E. Rivera, and K.G. Kempf, "A hierarchical approach to production control of re-entrant semiconductor manufacturing lines," *IEEE Transactions on Control Systems Technology*, vol. 11, no. 4, pp. 578-87, July 2003.

[24] S. Shirodkar, C. Arnold, K. Kempf, and J. Fowler, "Modeling and simulating supply chains for increased performance and profitability," in *2000 Proc. Inter. Conf. Modeling and Analysis of Semiconductor Manufacturing (Tempe, AZ)*, pp. 346-352.

[25] K. Kempf, K. Knutson, J. Fowler, B. Armbruster, P. Babu, and B. Duarte, "Fast accurate simulation of physical flows in demand networks," in *2001 Proc. Semiconductor Manufacturing Operational Modeling and Simulation Symposium (Seattle)*, pp. 111-116.

[26] D. Helbing, "Traffic and related self-driven many particle systems," *Reviews of Modern Physics*, vol. 73, pp. 1067-1141, 2001.

[27] D. Marthaler, D. Armbruster, and C. Ringhofer, "A mesoscopic approach to the simulation of semiconductor supply chains," *Simulation: Transactions of the Society for Modeling and Simulation International*, vol. 79, no. 3, pp. 157-163, 2003.

[28] D. Armbruster, D. Marthaler, C. Ringhofer, K. Kempf, and T. C. Jo, "A continuum model for a re-entrant factory," *Operations Research*, submitted for publication.

[29] D. Armbruster, C. Ringhofer, and T.-C. Jo, "Continuous models for production flows," this volume.

[30] E. Lefeber, R. A. van den Berg, and J. E. Rooda, "Modeling, validation and control of manufacturing systems," this volume.

[31] G. W. Godding and K. G. Kempf, "A modular, scalable approach to modeling and analysis of semiconductor manufacturing supply chains", *Proc. IV SIMPOI/POMS (Sao Paulo)*, p. 1000-1007, 2001.

[32] G. W. Godding, H. S. Sarjoughian, and K. G. Kempf, "Semiconductor supply network simulation," in *2003 Proc. IEEE Winter Simulation Conf. (New Orleans)*, pp. 1593-1601, 2003.