The impact of endogenous demand on push–pull production systems

Paulo Gonçalves,[a]* Jim Hines[b] and John Sterman[b]

Paulo Gonçalves is Assistant Professor in the Management Science department at the University of Miami's School of Business Administration. He holds a PhD in Operations Management and System Dynamics from MIT Sloan School of Management. His current research investigates customer responses to supply chain dynamics; in particular, he focuses on over-ordering dynamics due to competition for scarce supplies.

Jim Hines' work in *Modeling at Conversation Speed* combines "molecules" and eigenanalysis to increase the speed of creating and analyzing models by a factor of ten. The ultimate goal is to change the nature of managerial conversation. Jim consults in SD and teaches at Worcester Polytechnic Institute and Brown University.

John Sterman is the Jay W. Forrester Professor of Management and Director of the System Dynamics Group at the MIT Sloan School of Management

Abstract

Though often analyzed separately, supply chain instability and customer demand interact through product availability. We investigate the feedback between supply chain performance and demand variability in a model grounded in a first-hand study of the hybrid push–pull production system used by a major semiconductor manufacturer. While customers' response to variable service levels represents an important concern in industry, with sizeable impacts on company profitability, previous models exploring supply chain instability do not account for it. This research incorporates two effects of customer responses to availability. The *sales effect* captures the negative feedback whereby product shortages cause customers to seek alternative sources of supply, reducing demand and easing the shortage. The *production effect* captures the delayed impact of changes in demand on the manufacturer's production decisions: lower demand leads to reduced production, prolonging shortages that depress demand, a destabilizing positive feedback. We show how the *sales* and *production effects* interact to destabilize the supply chain and lower average performance. Supply chain models that assume exogenous demand may therefore underestimate the amplification in the chain and the value of inventory buffers. In addition, incorporating the customer response leads to different inventory and utilization policies from those in use by the company. The model yields insights into the costs of lean inventory and responsive utilization policies in the context of hybrid production systems and endogenous demand. Copyright © 2005 John Wiley & Sons, Ltd.

Syst. Dyn. Rev. **21**, 187–216, (2005)

Introduction

Despite the supply chain revolution of the past decade, companies in diverse industries such as computers, electronics, autos, toys, seeds, and pharmaceuticals still struggle with production and shipment delays. An important effect of such delays, generated by supply chain glitches and instability, is reduced shareholder value. Hendricks and Singhal (2003) show that an abnormal decrease of over 10% in shareholder value is caused by part shortages, order changes by customers, and production ramp-up and roll-out problems, among others. In addition, it has been long recognized that the bullwhip effect tends to amplify the instability in orders as one moves upstream in a supply chain, potentially making upstream companies, such as semiconductor manufacturers, more prone to supply chain glitches. For instance, due to part shortages, Boeing had to stop production of its 747 airplane (for almost a month) and

[a] University of Miami, School of Business Administration, Management Science Department, KE 404, Coral Gables, FL 33124, U.S.A.
[b] MIT, Sloan School of Management, E53-309, Cambridge, MA 02142, U.S.A.
* Correspondence to: Paulo Gonçalves. E-mail: paulog@miami.edu

System Dynamics Review Vol. 21, No. 3, (Fall 2005): 187–216
Published online in Wiley InterScience
(www.interscience.wiley.com). DOI: 10.1002/sdr.318

Received April 2004
Accepted January 2005

delay the final assembly of the 737, leading to "more late deliveries, higher costs, upset customers and depressed profits" (Holmes 1997). Intel Corporation has consistently struggled with part shortages, high variability in demand, and order changes and cancellations by customers. In November 1999, facing shortages of Pentium III processors, Intel planned to introduce a new fabrication facility in the following year. In late 2000, blaming order cancellations by large customers and economic slowdown, Intel warned that its revenues would fall short of projections and that sales would be flat for the quarter (Gaither 2001).

The challenge of demand variability, instability, and order amplification is complicated by long production delays. In semiconductors, long throughput times (approximately 13 weeks) affect the ability of manufacturers to maintain adequate inventory levels in the face of demand variability. When customer demand varies, factory managers must adjust capacity utilization to maintain adequate service levels while avoiding excess inventory. However, the combination of variability in demand and long fabrication delays often leads to alternating periods of scarce and excess supply. The resulting supply variability can feed back to customer demand and profitability as a company's inability to meet demand leads to lost sales, eroded reputation, and decreased goodwill. The interactions of supply chain instability and customer response raise several interesting questions: What is the impact of endogenous demand on supply chain variability? What is the impact of supply chain variability on customer response? What policies can Intel and other companies implement to stabilize their supply chains?

To address these questions, this research builds and analyzes a model of a semiconductor supply chain in which customer demand responds to product availability. Based on a year-long, in-depth field study of Intel's supply chain, the model captures the material flows of production and the customers' response to the manufacturer's service level. In particular, it incorporates two effects of customer response. First, the *sales effect* captures the negative feedback whereby product shortages cause customers to seek alternate sources of supply, reducing demand easing the shortage. That is, a change in demand feeds back to mitigate the impact of the initial disturbance. Second, the *production effect* captures the delayed impact of changes in demand on the manufacturer's production decisions through a positive feedback loop. If demand falls, manufacturers reduce demand forecasts and capacity utilization to avoid excess inventory. After the production delay, lower production leads to lower inventory and service levels, causing a further drop in customer demand. The delayed *production effect* generates a reaction that reinforces the impact of the original disturbance.

We show that endogenous customer response to availability leads to greater supply chain instability compared to models in which customer demand is treated as exogenous. Therefore, supply chain instability models that assume exogenous demand may underestimate the amplification in demand and the value of inventory buffers. Moreover, treating demand endogenously leads to different inventory and utilization policies from those currently in use by the firm. In particular, the supplier should maintain higher safety levels in assembly work-in-process (WIP) and finished goods inventory (FGI); and *reduce* the responsiveness of utilization to changes in customer demand caused by inadequate service levels. Based on the costs associated with lost sales and holding inventory at assembly and finished goods, we derive a recommendation for the optimal location and quantity of safety stocks. The policy heuristic provides a sharp reduction in supply chain instability and minimizes the impact of lost sales. The model analyzed in this paper gives insights into the costs of lean inventory strategies and responsive utilization policies in the context of hybrid production systems and endogenous demand.

The next section reviews the relevant literature. The third section discusses the research site and the following section presents the model. The fifth section introduces the simulation results, analyzes the model, and derives policies for supply chain stabilization. We conclude with discussion of our main results, managerial and theoretical implications, and directions for future research.

Literature review

Supply chain instability and the influence of inventory level on demand have attracted the attention of researchers and practitioners in different fields, such as economics, system dynamics, and operations management. In economics, research on supply chain instability dates back to Thomas Mitchell's (1924) descriptions of the mechanisms through which retailers caught short of supply increased their orders to suppliers. In system dynamics, the study of supply chain instability helped lay out the foundations necessary to create the field (Forrester 1958, 1961). Subsequent research explored applications in diverse areas including interactions between the supply chain and labor force (Mass 1975); the performance of Material Requirements Planning (MRP) systems (Morecroft 1980); laboratory study of people's ability to manage complex systems such as the supply chain in the Beer Game (Sterman 1989a, 1989b); decision-making under varying levels of feedback complexity (Diehl and Sterman 1995); and the impact of business cycles on capital equipment supply chains (Anderson and Fine 1999). In addition, models in the system dynamics tradition often incorporate the feedback of inventory availability on customer demand, like Forrester's (1968) "market growth" model and Graham's (1977) model investigating the impact of adding a minor loop to oscillatory systems. While this paper emphasizes the combination of endogenous customer response with supply chain instability, the major contribution to the system dynamics literature is our investigation of the impact of the two in hybrid push–pull production systems.

In operations management, research investigating supply chain instability typically assumes exogenous customer demand, and studies exploring the influence of inventory level on customer demand do not consider multiple-stage supply chains. Examples of the former include Lee *et al.*'s (1997a, 1997b) models of demand signal processing, rationing, order batching, and price variations; Baganha and Cohen's (1998) hierarchical model; Graves' (1999) single-item inventory system with non-stationary demand; and Chen *et al.*'s (2000) model with a demand forecasting technique and order lead time. Examples of the latter include Dana and Petruzzi's (2001) extended newsvendor model where customers choose between the company and an outside supplier; Gans' (1999a, 199b) dynamic model of individual consumer behavior, where consumers update their prior beliefs about the company after each contact; and Hall and Porteus' (2000) model where the expected service level is a function of firm capacity and firms compete by investing in capacity to service customers. Our research fills a gap in the operations management literature by exploring both the effect of endogenous customer responses and supply chain instability. Our results extend Dana and Petruzzi's (2001) result to the case of a multistage supply chain with production delays, showing that when a company accounts for the effect of inventory availability on demand it is optimal to hold more safety stock.

Research site

The results draw on a year-long, in-depth analysis of Intel's supply chains. Intel is the technology leader in microprocessor manufacturing. Among many firsts, Intel was the first to produce 0.13-micron technology, allowing it to double the size of the processor's cache memory while reducing die size by over 30%. Such improvements resulted in faster microprocessors and increased number of chips manufactured per wafer. The company was also the first to transition from 200 mm to 300 mm wafers, leading to higher chip production efficiency. To manage the variability in product line, production, and demand, Intel employs about 1,500 planners who address short- and long-term production decisions, using sophisticated systems and detailed guidelines directing decisions. Model development entailed interviewing planners with diverse decision scopes and responsibilities to understand the decision-making processes in Intel's production system. In addition, the research team interviewed managers in diverse areas of the corporation, such as operations, supply chain management, information technology, demand forecasting, marketing and sales. In total, we conducted almost 100 semi-structured interviews both through site visits and weekly conference calls. The research also involved reviewing Intel's logs detailing guidelines for decision-making, and collecting related quantitative and qualitative data. The former included time-series data on quarterly capacity, utilization, wafer starts, shipments, forecasts, service level, and market share. The latter included managers' decision heuristics, company's guidelines and incentives, and information dependencies among business areas. These data helped us establish the assumptions used in the model that captures Intel's semiconductor manufacturing.

Model assumptions

Semiconductor manufacturing is commonly divided into two major phases: fabrication and assembly. The first phase (fabrication) takes place in a wafer fabrication facility, or Fab. It takes 200 mm–300 mm polished disk-shaped silicon substrates (wafers) as inputs and through a complicated sequence of steps transforms them into fabricated wafers, composed of hundreds of -inch square integrated circuits (ICs or dies).[1] A vertical cross-section of an integrated circuit reveals a number of layers formed during the fabrication process. Lower layers, produced at the "front-end" of the fabrication process, include the critical electrical components (e.g., transistors, capacitors). Upper layers, produced at the "back-end" of the fabrication process, connect the electrical components to form circuits. In addition, fabrication is characterized by a re-entrant flow process, i.e., the same equipment perform multiple steps at different stages of fabrication (such as photolithography, etching, thin films, diffusion).

In the second phase of manufacturing (assembly) wafers are cut into dies and stored in assembly die inventory (ADI) warehouses, collocated with assembly/test plants. The dies receive a protective package protecting the integrated circuit from the environment and allowing the attachment of metal connectors. The completed microprocessors (or chips) are then tested to ensure operability. Upon passing the tests, the chips can be stored in finished goods warehouses. The model proposed represents the manufacturing process by a three-stage supply chain consisting of fabrication, assembly, and distribution (Figure 1).

In addition, microprocessor production takes place in a hybrid push–pull production system, combining a push system at upstream stages and a pull system at the downstream stages. Therefore, fabrication is characterized by a push production system: long-term demand forecasts, updated weekly, and adjustments from fabrication and assembly WIP serve as the basis for the desired wafer production rate, or wafer starts. In contrast, assembly/testing and distribution operate as a pull system, with shipments based on current customer orders.

Four main assumptions based on the fieldwork drive the behavior of the model.[2] The first three assumptions address managers' decisions regarding (a) push and pull production, (b) capacity utilization, and (c) demand forecasting. These assumptions reflect Intel managers' locally rational heuristics to control their systems. While they are not optimal, they reflect heuristics managers use to make their everyday decisions and evolved because they are locally adapted

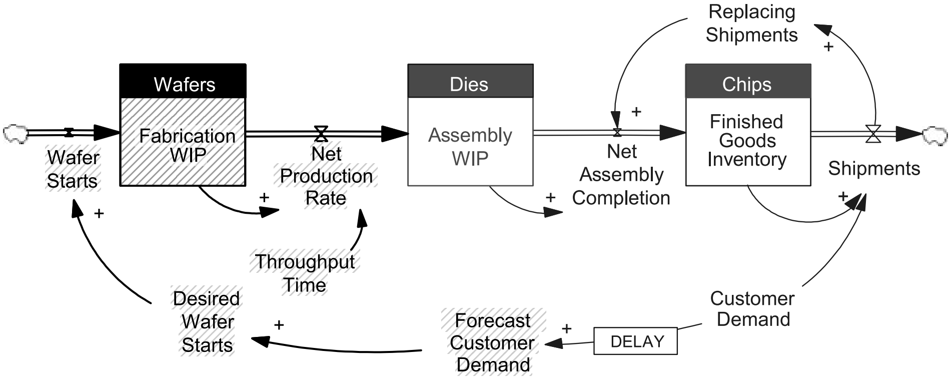


Fig. 1. Hybrid push-pull production system for semiconductors. Thick lines and patterned background refer to a push system, indicating that the upstream production process operates as a push. Thin lines and clear background refer to a pull system, indicating that the downstream stages operate as a pull

to conditions in the company and its Fabs. The fourth assumption captures customer response to inventory availability.

Push and pull production decisions

Production decisions ultimately depend on customer demand. Current demand drives shipments and assembly completions; long-term demand forecasts influence production starts. All incoming orders are logged by Intel's information system and tracked until they are shipped to customers or cancelled. If the microprocessors are available in FGI, orders can be filled immediately. Therefore, incoming customer orders "pull" the available microprocessors from FGI. In turn, replenishment of FGI shipped to customers "pulls" microprocessors from assembly. If the microprocessors are not available in FGI, backlogged orders "pull" parts directly from the ADI. Since the parts have to be assembled, filling orders from ADI increases the delivery delay experienced by customers and reduces the flow of shipments below customer orders as ADI inventory and assembly capacity limit shipments.

FGI AND ASSEMBLY PULL To model the pull characteristic of assembly and finished goods, we must capture their dependency on current demand. Actual shipments (S) from FGI are given by the minimum of the desired (pull) and feasible (push) shipment rate. By design, shipments operate in a pull mode, with shipments being determined by the desired rate; however, if not enough FGI is available the system will ship out only what is possible.

$$S(t) = \mathrm{MIN}(S^*(t), S_{\mathrm{MAX}}(t)) \tag{1}$$

Desired shipments depend on the ratio of backlog (B) and the desired delivery delay (DD^*). Feasible shipments depend on the stock of FGI and the minimum order processing time (τ_{OP}); a first-order process is assumed for simplicity.

$$S^*(t) = B(t)/\mathrm{DD}^* \tag{2}$$

$$S_{\mathrm{MAX}}(t) = \mathrm{FGI}(t)/\tau_{\mathrm{OP}} \tag{3}$$

While shipments (S) deplete FGI, the net assembly completion rate (A_{N}) replenishes it. The product of gross assembly completion rate (A_{G}) and the unit yield (Y_{U}), i.e., the fraction of good chips per assembled die, define the net assembly completion rate (A_{N}). The gross assembly completion rate (A_{G}) is given by the minimum of the desired (a pull signal) or the feasible (a push signal) gross assembly completion rate.

$$\dot{\mathrm{FGI}}(t) = Y_{\mathrm{U}} A_{\mathrm{G}}(t) - S(t) \tag{4}$$

$$A_{\mathrm{G}}(t) = \mathrm{MIN}(\mathrm{Push}A_{\mathrm{G}}(t), \mathrm{Pull}A_{\mathrm{G}}(t)) \tag{5}$$

By design, assembly operates in a pull mode, with gross assembly output being determined by the desired gross rate. However, if not enough WIP is available the system can complete only what is feasible. The feasible gross assembly rate is determined by the availability of assembly WIP (AWIP) and the assembly time (τ_{A}); for simplicity a first-order delay is used. The desired gross assembly rate ($A^*_{\mathrm{G}} = \mathrm{Pull}A_{\mathrm{G}}$) is determined by the desired net assembly rate (A^*_{N}) adjusted by the unit yield (Y_{U}).

$$\mathrm{Pull}\,A_{\mathrm{G}}(t) = A^*_{\mathrm{N}}(t)/Y_{\mathrm{U}} \tag{6}$$

$$\mathrm{Push}\,A_{\mathrm{G}}(t) = \mathrm{AWIP}(t)/\tau_{\mathrm{A}} \tag{7}$$

The determination of desired net assembly rate (A^*_{N}) by division planners begins with recent shipments (ES), a proxy for current demand that is more stable and reliable than orders. Desired net chips out is then adjusted above or below recent shipments to close any gap between target and actual FGI and to eliminate excess backlog.

$$A^*_{\mathrm{N}}(t) = \mathrm{MAX}\left(0,\ \mathrm{ES}(t) + \frac{\mathrm{FGI}^*(t) - \mathrm{FGI}(t)}{\tau_{\mathrm{FGI}}} + \frac{B(t) - B^*(t)}{\tau_{\mathrm{B}}}\right) \tag{8}$$

where *FGI** and *FGI* are target and actual finished goods inventory, B^* and B are target and actual backlog, and τ_{FGI} and τ_{B} are the adjustment times for the elimination of gaps between them.

Fabrication push The wafers produced in the fabrication process are pushed into the assembly die inventory (ADI), where they are stored until orders for specific products pull them from ADI into assembly and distribution. While the gross assembly completion rate (A_{G}) depletes AWIP, the die completion rate (D_{I}) replenishes it.

$$\dot{\mathrm{AWIP}}(t) = D_{\mathrm{I}}(t) - A_{\mathrm{G}}(t) \tag{9}$$

The die completion rate (D_{I}), measured in dies/month, is given by the gross fabrication rate (F_{G}), measured in wafers/month, adjusted by the number of dies per wafer (DPW); the die yield (Y_D), i.e., the fraction of good die per wafer; and the line yield (Y_L), i.e., the fraction of good fabricated wafers. The gross fabrication rate (F_G) is determined by the availability of fabrication WIP (FWIP) and the fabrication time (τ_F); for simplicity a first-order delay and a constant fabrication time are used.[3]

$$D_I(t) = F_G(t) \cdot \mathrm{DPW} \cdot Y_D \cdot Y_L \tag{10}$$

$$F_G(t) = \mathrm{FWIP}(t)/\tau_F \tag{11}$$

While the gross fabrication rate (F_G) depletes fabrication WIP (FWIP), wafer starts (WS) replenish it. The decision on actual production rate, wafer starts (WS), is based directly on the desired wafer starts (WS*).

$$\mathrm{F\dot{W}IP}(t) = \mathrm{WS}(t) - F_G(t) \tag{12}$$

Fab planners determine the desired wafer starts considering the desired die inflow (D_I^*) requested by assembly/test plants and an adjustment for fabrication work-in-process. The latter is based on managers' heuristic to maintain fabrication WIP (FWIP) at a desired level (FWIP*). Equation 13 shows the fabrication planners' heuristic for managing wafer starts.

$$\mathrm{WS}^*(t) = \mathrm{MAX}\left(0, \frac{D_I^*(t)}{\mathrm{DPW} \cdot Y_D \cdot Y_L} + \frac{\mathrm{FWIP}^*(t) - \mathrm{FWIP}(t)}{\tau_{\mathrm{FWIP}}}\right) \tag{13}$$

where τ_{FWIP} is the fabrication WIP correction time; and the non-negativity constraint prevents negative production targets.

The desired die inflow rate (D_I^*) depends on long-term demand forecasts (ED) and an adjustment from assembly WIP. The assembly WIP adjustment component reflects assembly planners' goal to replenish assembly WIP when the current level is below the target to correct the discrepancy over time (τ_{AWIP}). Equation 14 shows the division planners' heuristic for managing the desired die inflow (D_I^*), incorporating information on long-term demand forecast (ED).[4] Division planners provide information on the desired die inflow to Fab planners, allowing them to set production starts.

$$D_I^*(t) = \mathrm{MAX}\left(0, \mathrm{ED}(t)/Y_U + \frac{\mathrm{AWIP}^*(t) - \mathrm{AWIP}(t)}{\tau_{\mathrm{AWIP}}}\right) \tag{14}$$

where τ_{AWIP} is the assembly WIP correction time; and the non-negativity constraint prevents negative die inflow rates.

Capacity utilization

To set the capacity utilization (CU) of their Fabs, managers consider the desired production rate and the available capacity. Capacity utilization is a nonlinear function of the ratio of desired wafer starts (WS*) and available capacity (K) operating at the normal capacity utilization level ($\mathrm{CU_N}$).[5]

$$\mathrm{CU}(t) = f_{\mathrm{U}}\left(\frac{WS^{*}(t)}{K \cdot \mathrm{CU_N}}\right) \tag{15}$$

When desired production (WS*) equals the normal capacity utilized, capacity utilization is set at the normal operating point (90%), allowing all desired production to be met. The remaining 10% slack capacity is often used for engineering purposes (process improvement and development runs) as well as to accommodate manufacturing instability. When desired production is large relative to normal capacity utilized, Fab managers increase utilization, therefore reducing the capacity that is available to engineering. The opposite takes place when desired production falls below normal capacity utilization. If the function lay on the 45° reference line, utilization would vary enough to ensure that wafer starts always equaled desired starts exactly (subject to the capacity constraint). Field study showed, however, that the utilization function characterizing actual wafer start decisions lies above the 45° reference line and has a flatter slope at the normal operating point. Fab managers seek to avoid shutdown and prefer to keep their Fab running even when desired starts fall below normal, preferring instead to build inventory; similarly, they increase output less than enough to meet desired starts fully when desired starts exceed normal output so as to maintain some room for engineering purposes and to avoid yield problems. A concave function where $f_{\mathrm{U}} \geq 0, f'_{\mathrm{U}} > 0, f''_{\mathrm{U}} < 0, f_{\mathrm{U1}}(0) = 0, f_{\mathrm{U}}(1) = \mathrm{CU_{Norm}}$, $f_{\mathrm{U}}(2) = \mathrm{CU_{Max}}$, captures the response of Fab managers to variations in desired wafer starts relative to capacity (see the specification in Figure 8 for an example).

While the general shape of the function (f_{U}) is plausible, the slope of the function around the normal operating point and the normal utilization fraction play an important role in model behavior. Data for estimating such parameters are both proprietary and Fab specific. Therefore, we provide sensitivity analysis (see later section) over a broad range of plausible parameters for capacity utilization functions and investigate the impact of these parameters on model behavior.

Demand forecasting

The marketing organization is responsible for demand forecasting at Intel. As in many firms, marketing generates an initial demand forecast for microprocessors based on estimates of customer demand from different geographic regions and customer types (for another example in the semiconductor industry, see Sterman 2000, pp. 449–462). A process known as "Judged Demand" is then used to go from the initial to a final forecast. The "Judged Demand" process receives its name due to the judgment and subjective adjustments involved in elaborating the forecast. First, macroeconomic indicators are incorporated to adapt the initial estimates based on the total available market for personal and business computers. Second, an "executive adder" process often adjusts the aggregate forecast upwards to reflect the optimistic goals and aspirations of company executives. Finally, the marketing group "filters" (i.e., smoothes) the demand estimates from different regions to account for local incentives. With respect to regional information, according to a platform manager in marketing: "Customer numbers get rolled up, aggregated, and judged with a set of assumptions that may or may not be correct; and customer-level insight, when provided, gets watered down." In particular, when demand for certain products is high, regional warehouse managers tend to increase their orders to ensure that they will be able to meet demand, the familiar "phantom ordering" generated as different customers compete for larger slices of what they perceive to be a shrinking pie (Forrester 1961; Sterman 2000, pp. 743–755; Gonçalves 2003). In contrast, when demand is low, regional managers have the tendency to decrease orders to make sure they are not stuck with undesired inventory. Therefore, marketing "filters" the forecast to come up with its final forecast. Analysis of the forecast data confirmed this: when compared, regional forecasts were more variable than marketing ones.

At Intel, the demand forecast incorporates a trend component to account for the exponential growth in semiconductor sales. Because we focus on the interplay between customer response and supply chain instability, we explore a de-trended demand signal. Therefore, we model the demand forecast (ED) as a first-order exponential smooth of actual orders (D)—in practice obtained from the aggregation of regional orders—updated over a period of one month (τ_{DAdj}), the frequency with which marketing updates their forecasts.

$$\dot{\mathrm{ED}}(t) = \frac{D(t) - \mathrm{ED}(t)}{\tau_{\mathrm{DAdj}}} \qquad (16)$$

For simplicity, we do not take into consideration the random macroeconomic factors that may influence the demand forecast and the executive adder process, making the *a fortiori* assumption that marketing is able to filter out the noise and bias caused by these processes.[6]

Customer response

In this model we capture customers' response to supply availability, measured by the fraction of orders filled (FoF). Customers respond to a low fulfillment fraction by seeking alternative sources of supply; as they succeed, their orders drop. Intel's attractiveness to suppliers (A_{I}) is a non-linear function of customers' perception of supplier delivery reliability (PFoF). In turn, customers' perception of delivery reliability (PFoF) adjusts from the actual delivery reliability—fractional orders filled (FoF)—with a third-order Erlang lag (λ), with an average time constant of six months. The third-order Erlang distribution captures the plausible distribution of responses by OEMs. At the instant of a decrease in the service level, all OEMs will still perceive the supplier as reliable, and there will be no shifts to alternative sources of supply. Therefore, the immediate response of the distributed lag should be zero. If service level remains low, however, some customers will change their perceptions about supplier reliability and seek other suppliers. The distribution of OEMs' reactions eventually peak, and then decrease, reaching zero after a sufficient time has elapsed. The delay captures the time required for OEMs to perceive changes in availability, to determine that the changes are not temporary and warrant a search for alternative sources, and to close deals with those alternative sources. For simplicity, we assume that competitors maintain constant delivery performance (i.e., a constant attractiveness (A_{C}) over time). This assumption allows us to measure changes in system behavior due to customers' reactions only due to changes in supplier conditions; relaxing it is a promising direction for future work.

The non-linear function (f_{A}), characterizing Intel's attractiveness (A_{I}), is a logistic curve (the specification in Figure 9 provides an example.) Attractiveness varies on a scale going from zero to one ($0 \leq A_{L\mathrm{Min}} < A_{L\mathrm{Max}} \leq 1$). A logistic curve captures customers' mild response to small changes in supply availability, and more significant responses to large changes in supply availability.

$$A_{\mathrm{I}}(t) = f_{\mathrm{A}}(\mathrm{PFoF}(t)) \tag{17}$$

While the logistic shape of the function is plausible, the model behavior depends heavily on the slope of the function and the minimum value. At the same time, the data for estimating such parameters are not reliable or easily available. Here too we provide sensitivity analysis (see later section) over a broad range of plausible parameters for the function governing customer responses and investigate the impact of these parameters on model behavior.

Feedback structure

The Intel managers' heuristics (push–pull production, capacity utilization, demand forecast, and customer response decisions) close the feedback loops shown in Figure 2.

Balancing loop *FGI Pull* (B1) describes the company's pull system operating at the finished goods inventory (FGI) level. An increase in backlog (due to additional orders) increases desired shipments, boosting shipments and reducing the backlog—if there is sufficient FGI. When the availability of FGI is limited, the negative loop *FGI Availability* (B2) limits shipments. With FGI constraining shipments, the pull system cannot operate at the FGI level. However, the system can still pull from assembly WIP. The balancing loop *Assembly Pull* (B3) is analogous to the *FGI Pull* loop, but pulls chips out of the assembly WIP, which takes longer. Therefore, when FGI is constrained, the system still operates as pull but with longer fulfillment delay. The actual assembly completion rate is also adjusted by two other loops: a balancing loop that corrects the levels of finished goods inventory (*Adjust FGI*—B4) and a reinforcing loop (*Replenishment*—R1) that replenishes all shipments from

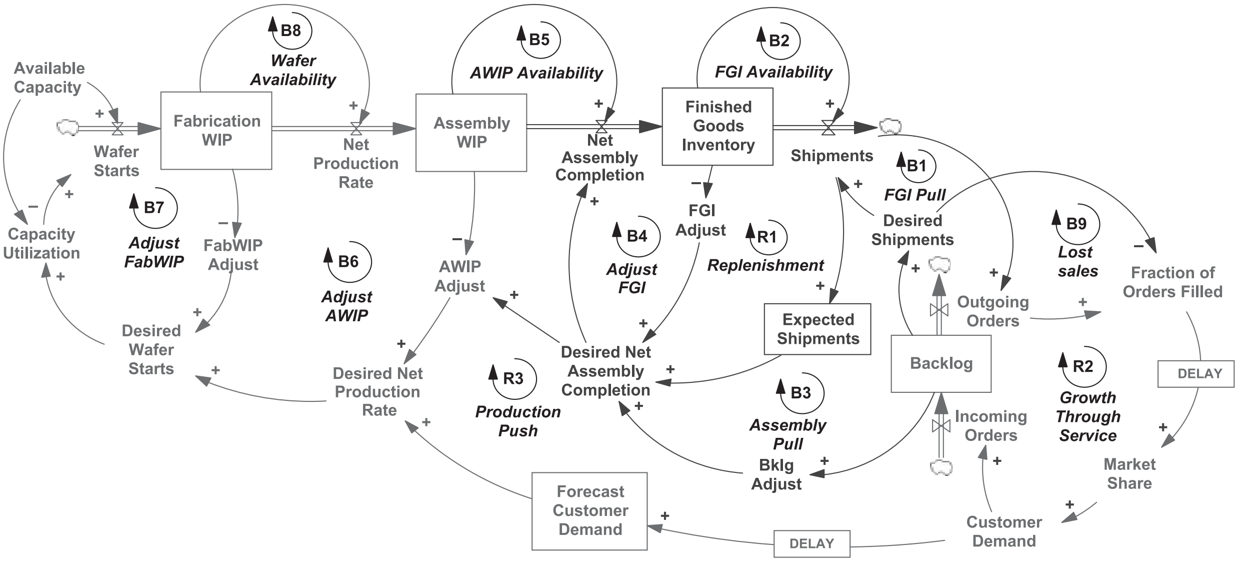


Fig. 2. Supply–demand feedback process for Intel's hybrid production system

FGI. The system can pull from assembly as long as the assembly WIP inventory is sufficiently high. When the availability of assembly WIP decreases, the first-order control for assembly—balancing loop *AWIP Availability* (B5)—prevents AWIP from going negative. If assembly WIP constrains net assembly, the system cannot even pull from AWIP, and the entire system reverts to push.

Production at the upstream stage is based on long-term demand forecasts, information on the desired assembly completion rate, and adjustments due to inventory corrections in assembly (*Adjust AWIP*—B6) and fabrication (*Adjust FWIP*—B7). The push part of the production system is determined by the first-order control for fabrication—balancing loop *Wafer Availability* (B8). In terms of customers' response, the reinforcing loop *Growth Through Service* (R2) describes the ability of the company to grow its market share as it is able to meet customer demand. In contrast, the balancing loop—*Lost Sales* (B9)—describes the inverse dynamics. As customer demand grows, the company's ability to maintain its service level (fraction of orders filled) decreases, reducing its ability to retain customers. If the company cannot adequately fill customer orders, it will lose market share to competitors. Finally, the feedback from the company's supply chain to customer demand is described in the reinforcing loop—*Production Push* (R3), which captures the long delays associated with production and customer reactions. If demand falls, manufacturers reduce demand forecasts and capacity utilization to avoid excess inventory. After the production delay, lower production leads to lower inventory and service levels, causing a further drop in customer demand. These feedback processes are capable of generating the dynamic behavior observed in the company and replicated in the model.

Model analysis and results

The model constitutes a ninth-order non-linear differential equation system. Since the equations are highly non-linear it is not possible to obtain closed-form solutions. Hence, we simulate the model to gain intuition on its behavior. While the parameters chosen for the base case (Table 1) reflect Intel's manufacturing system, the values are disguised to maintain confidentiality.

Table 1. Base case parameters

Parameter	Definition	Value	Units
D	Customer demand	5.0	Million units/month
MS	Initial market segment share	80	%
DPW	Number of die per wafer	200	Die/wafer
CU_{N}	Normal capacity utilization	90	%
Y_{L}	Line yield: Fraction of good wafers per total	90	%
Y_{D}	Die yield: Fraction of good die per wafer	90	%
Y_{U}	Unit yield: Fraction of good chips per good die	95	%
K	Available capacity	28.9	'000 wafers/month

For a given customer demand (D), the equilibrium capacity (K) required to meet that demand can be computed from the normal capacity utilization and yields. The formula for equilibrium capacity (K) is given by: $K = \frac{D \cdot \mathrm{MS}}{\mathrm{CU}_{\mathrm{N}} \cdot \mathrm{DPW} \cdot Y_{\mathrm{D}} \cdot Y_{\mathrm{L}} \cdot Y_{\mathrm{U}}}$

Figure 3 shows the behavior of backlogs and finished goods inventory coverage for three simulation runs. The model is initialized in equilibrium with constant industry demand. As mentioned earlier, while semiconductor demand has been growing exponentially for decades, we focus on a de-trended demand signal because we are interested only in issues regarding the stability of the system. Interactions of growth with supply chain stability are left for future research. In equilibrium the hybrid push–pull system functions as intended: the supplier meets its target delivery delay, fills 100% of incoming orders, and maintains the desired quantities of FGI, AWIP, and wafers. From equilibrium we introduce a demand pulse by increasing customer demand by 5% and then 20% respectively for a single month at the end of the first simulated year. The demand shocks increase the backlog (Figure 3a). As planners observe the increase in demand and backlogs, they quickly realize the need to raise production, and desired wafer starts rise (Figure 3b). Since capacity is fixed in the short run, managers' must raise capacity utilization

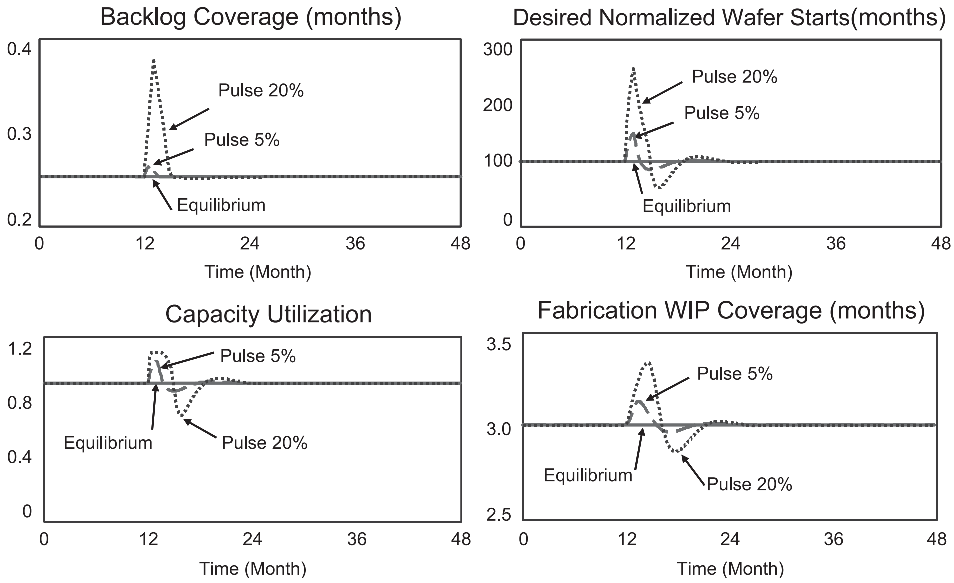


Fig. 3. (a) Backlog coverage, (b) desired wafer starts, (c) capacity utilization and (d) fabrication coverage for simulated scenarios

(Figure 3c) to increase wafer starts. Fab managers quickly adjust utilization to its maximum. Higher utilization increases the level of fabrication (Figure 3d). After fabrication and assembly delays, finished goods eventually become available to meet the demand.

In both cases, the system immediately responds to the surge in backlog by increasing shipments (not shown) and pulling more chips from FGI. In the case of the 5% increase, the depletion in FGI (Figure 4b) is insufficient to constrain shipments. Here, while the demand shock creates some supply chain instability (Figures 3d and 4a), safety stocks in FGI and AWIP allow the system to operate as desired, i.e. as a hybrid push–pull system. Despite the shock, the company is capable of meeting its target delivery delay and filling 100% of its incoming orders (Figure 4c).

The 20% shock, however, produces a different outcome. For a large enough shock, the safety stocks in FGI and AWIP are not capable of maintaining the system in desired operation mode. Here, the system behaves as a pure–push system, reacting to demand changes with a much longer production delay. The depletion in FGI constrains shipments, limiting the fraction of orders filled (Figure 4c). With FGI constraining shipments, the pull system cannot operate at the FGI level. The system compensates for the lack of FGI, however, and pulls chips from assembly WIP, increasing assembly rate. As the availability of assembly WIP decreases, it eventually constrains assembly. Now, the system can’t pull from AWIP, so it reverts to push.

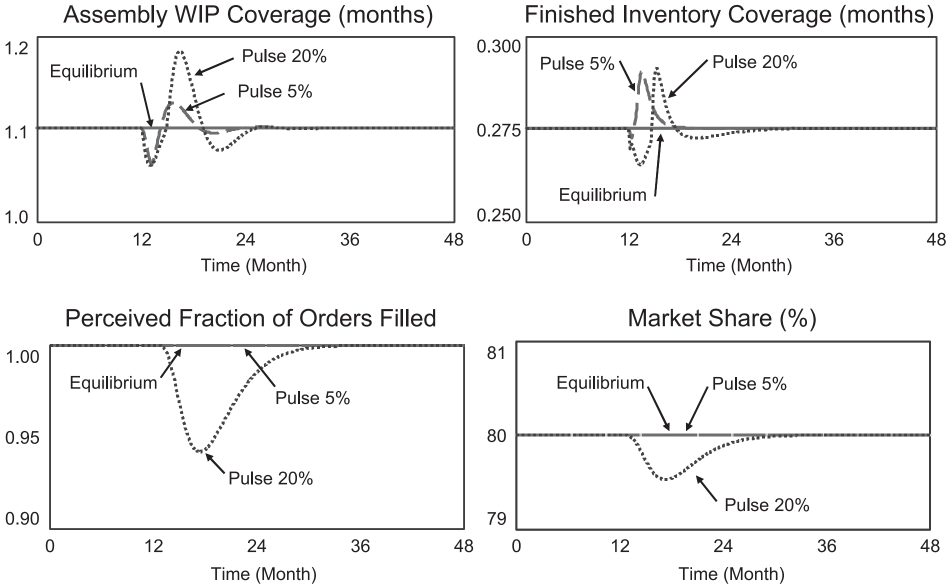


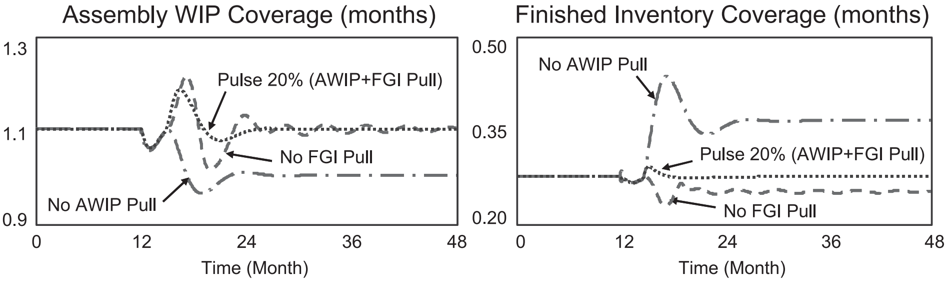
Fig. 4. (a) Assembly WIP and (b) finished inventory coverage, (c) perceived fraction of orders filled and (d) market share for simulated scenarios

In push mode, the supplier is unable to meet all customer orders. Customers (i.e., OEMs) perceive the drop in service level (Figure 4c) after a delay (accounting for decision-making and reporting delays in information systems) and seek alternative sources of supply. Market share decreases (Figure 4d) and the drop in orders contributes to the reduction in backlog coverage. Ultimately, the company's inability to meet customer demand results in reduced market share, offsetting the impact of the original increase in demand. As customer demand continues to decrease, it eventually equals the volume of shipments, allowing the backlog coverage (Figure 3a) to stop increasing and the fraction of orders filled to stop declining. Even after additional FGI becomes available, market share continues to decrease due to the delay in customers' perception. Customers' response to inventory availability feeds back to the supplier decision on production. Capacity utilization (Figure 3c) drops as the supplier reacts to declining demand. The decrease in utilization lowers the level of fabrication, assembly WIP, and FGI coverage. When customers finally perceive improved company performance, orders increase and market share rises. With time, orders increase past shipments, leading to an increase in backlog. Once again shipments are not sufficient to meet customer demand and the fraction of orders filled decreases. The 20% shock in demand generates an oscillatory response that decays as some of the excess demand is lost and the supplier closes any remaining demand gap with capacity utilization above normal. Unlike the 5% pulse, when the system is subjected to a large enough shock, the interaction of the firm's locally rational decision rules for shipments and capacity utilization with the market's response to product availability results in a lightly damped oscillation depressing firm's market share.

Impact of pull system

To obtain additional insight into the causes of oscillation, we first consider how the balancing *FGI Pull* (B1) and *Assembly Pull* (B3) loops influence system behavior. The pull from FGI allows the company to close the gap between desired shipments and actual shipments by running down backlog. Naturally, this loop can only operate while there is sufficient finished goods inventory to allow shipments to take place. The ability of the FGI pull loop to operate so effectively is due to the short time constant (1 week) associated with the desired shipment rate. When that loop is off—shipments are a function of the level of FGI—the system operates as a push system, and the oscillatory behavior of the system increases (Figure 5). Similarly, the balancing *Assembly pull* loop pulls inventory from assembly to allow the *FGI Pull* loop to operate as desired. This loop can only operate while there is sufficient assembly inventory to allow assembly completions to take place. It also has a short time constant (1 month) determined by the time to adjust backlog. Turning off the pull from assembly reduces system stability as it restricts the ability of the FGI pull loop to operate effectively. Figure 5 shows the effect of turning off the FGI and assembly pull loops compared to the base case.

Fig. 5. FGI and AWIP pull loops turned off compared to the 20% pulse in demand



Impact of endogenous demand

The impact of endogenous demand on system behavior offers further insight into the causes of oscillation. To make demand exogenous, we knock out the *customer response* loop by setting the time to perceive the fraction of orders filled to a very large number ($\tau_{\mathrm{FoF}} = \infty$). This change breaks the feedbacks from the fraction of orders filled (FoF) to customer demand (D), making demand exogenous while reducing the system to a production and backlog response to a change in demand. The production loop establishes "wafer starts" based on expected demand and inventory adjustments throughout the chain. Notice first that if inventory levels in fabrication, assembly, and finished goods are fully visible to managers and they use the same time constant for inventory adjustments, the system could be reduced to an effectively first-order system (Graham 1977). The behavior would then be a smooth increase in production to meet the additional pulse in demand, followed by a smooth decline.

The structure of the production loop, however, is different. First, while the system has full visibility, FGI is only used to set the desired level of assembly inventory, instead of also being used to set the rate of assembly outflow. Second, expected demand, smoothed with a long time constant, is used to set production, but expected shipments, smoothed with a short time constant, are used to inform the adjustment needs for desired inventory. Finally, the balancing FGI pull and assembly pull loops introduce additional complexity to the production process. While expected shipments adjust FGI it also influences, with expected demand, capacity utilization. Hence a drop in shipments due to low FGI availability sends a spurious signal to production that additional output is not needed exactly when the opposite is desired. The resulting behavior is damped oscillation of production in the manufacturing supply chain (Figure 6). By increasing the time adjustment for the inventory corrections or smoothing the expected demand over a longer time constant, we can dampen the oscillations.

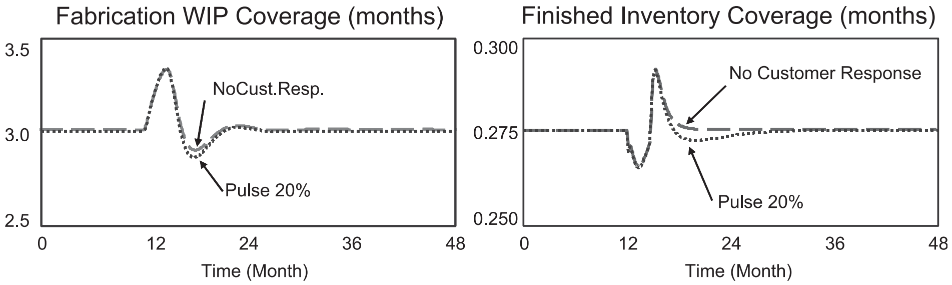


Fig. 6. Production response to demand change compared to the 20% pulse in demand.

The interaction of customer response with the rest of the system amplifies the oscillatory behavior of production. As demand increases, the company increases production as its immediate ability to meet demand decreases. With time, customers perceive that service levels are decreasing. After the manufacturing delay, the additional finished goods allow the company to meet a greater fraction of demand than it would otherwise. Just as more finished goods become available, however, the delayed responses from customers reduce orders. As the manufacturer finds itself with more finished goods inventory and reduced demand, Fab managers reduce capacity utilization, limiting the company's ability to meet future demand. The system eventually converges, however, the interaction of customer response and production amplifies supply chain instability.

While endogenous demand amplifies supply chain instability, more importantly it affects the company's policies regarding capacity utilization and inventory control. Consider first the inventory policy. Since the system reaction to a change in demand is more stable when demand is assumed exogenous, a tight inventory policy, with reduced levels of safety stock, is likely capable of providing a high service level, without incurring the additional inventory costs. Therefore, if demand is assumed exogenous, a lean inventory policy will likely lead to lower costs and be preferred to a safety stock policy. However, more unstable systems, as in the case of endogenous demand, require larger inventory buffers to provide the same level of service. In this case, the costs associated with lost sales may surpass those of holding safety inventory. Therefore, inventory buffers may be preferred when demand is assumed endogenous.

Now consider the capacity utilization policy. When demand is endogenous, inventory availability affects demand. Inventory shortages that may constrain shipments decrease service level and customer demand, sending a spurious signal that additional output is not needed. Since the decrease in demand is caused by an inventory shortage, additional output is highly desirable. An unresponsive capacity utilization policy that does not lower production level due to a decrease in demand will likely provide a higher service level when demand is endogenous. In contrast, when demand is exogenous, inventory availability does not affect demand. A responsive capacity utilization policy allows the company to prevent the accumulation of excess inventory during periods of low demand. Therefore, when demand is assumed exogenous, it is likely that a responsive capacity utilization policy be recommended.

To explore the impact of the demand assumptions on inventory and utilization we consider a simple cost structure, where total costs (TC) are the sum of inventory holding costs in assembly WIP (HC_{AWIP}) and finished goods (HC_{FGI}) and lost sales cost (LS_C). (For simplicity, we do not account for holding costs in fabrication.) Holding costs in each stage are given by the product of the inventory volume in each stage (e.g., AWIP and FGI) and the respective unit inventory holding costs ($\beta\phi$ and $\delta\phi$, that is, a fraction of the unit finished goods cost ϕ). Lost sales cost is the product of a factor (α) of unit finished goods cost (ϕ) and the amount of lost sales, given by the difference between the initial market segment share (MS_0) and the actual (MS_t).

$$HC_{AWIP} = AWIP \cdot \beta \cdot \phi \tag{18}$$

$$HC_{FGI} = FGI \cdot \delta \cdot \phi \tag{19}$$

$$LS_C = (MS_0 - MS_t) \cdot \alpha \cdot \phi \tag{20}$$

The criterion to evaluate the best policies is the comparison of net present value of cumulative discounted costs (CDC), with a discount rate (r).[7]

$$\mathrm{CDC} = \int_0^\infty e^{-rt}\{[(\mathrm{MS}_0 - \mathrm{MS}_t)\cdot\alpha + \mathrm{AWIP}\cdot\beta + \mathrm{FGI}\cdot\delta]\cdot\phi\}\mathrm{d}t \qquad (21)$$

Figure 7 shows that typical policy prescriptions of adopting lean inventory and responsive utilization policies are in fact reversed when demand is made endogenous. Figure 7(a, b) shows net present value of the costs associated with endogenous and exogenous demand for each inventory policy.[8] When demand is endogenous, a safety stock policy leads to lower costs than a lean inventory policy, suggesting that lost sales costs outweigh holding costs. Since holding inventory now prevents losing sales later, higher discount rates place a higher burden on holding costs, thereby reducing the benefits associated with lower lost sales. Figure 7(c, d) shows net present value of the costs associated with endogenous and exogenous demand under the capacity utilization policies.[9] An unresponsive capacity utilization policy allows the company to build inventory during periods of low demand caused by poor inventory availability, when demand is endogenous. Since the benefit of an unresponsive policy comes from building inventory, if the company already adopts a safety stock policy the benefits of an unresponsive utilization policy may be reduced. Therefore, inventory buffers and an unresponsive capacity utilization policy yield lower costs with endogenous demand.

The conclusions above reflect the behavior of the system for the original set of costs. Next, we changed the relation between holding and lost sales costs to

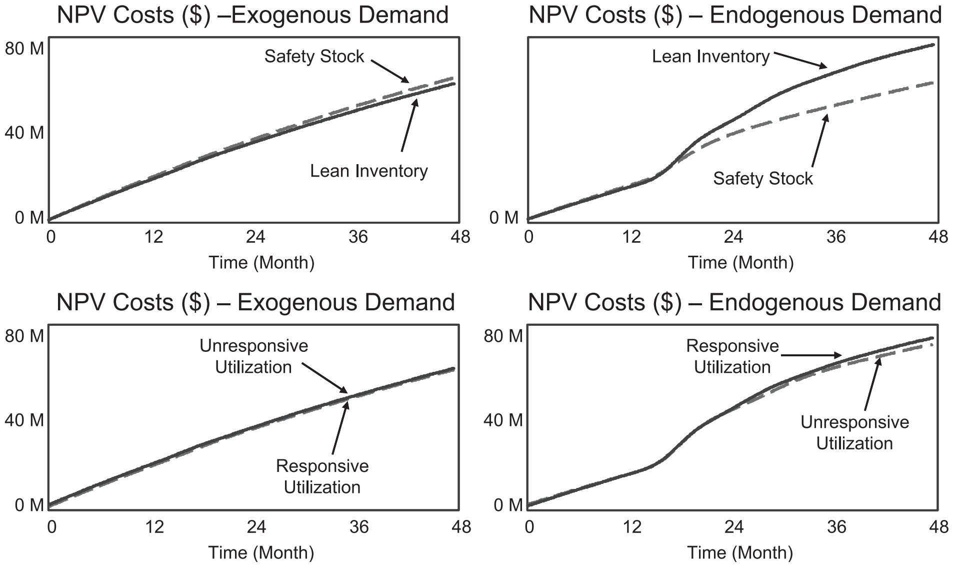


Fig. 7. (a)–(d) Impact of endogenous and exogenous demand on inventory and utilization policies

explore the impact that it may have on our conclusions. The model is simulated 2,500 times with independently randomly selected parameter values from a uniform distribution with ranges specified in Table 2.

Table 2. Range values for different cost parameters

Parameter	Symbol	Units	Min.	Base	Max.
Fractional unit FGI holding cost	δ	1/month	0.005	0.01	0.2
Ratio fractional unit AWIP to FGI cost	β/δ	dmnl	0.125	0.25	0.75
Fractional unit lost sales cost	α	dmnl	0.5	1	5

Note: The simulation is run with a unit finished good cost $\phi = 50$ \$/unit and a discount rate $r = 0.01$/month; dmnl = dimensionless.

Table 3 presents mean, median, standard deviation and confidence intervals (50%, 90%, and 95%) statistics for the net present value of cumulative discounted costs for utilization and inventory policies when customers do and do not respond to inventory availability. Statistics are evaluated at the end of the simulation (at time $t = 48$ months.)

Table 3. Utilization and inventory policy outputs for different cost parameters

NPV costs (million $): exogenous demand

Policy	Mean	Median	SD	50% CI	90% CI	95% CI	100% CI	Median savings
Safety stock (SS)	533	477	374	224, 783	29, 1241	21, 1381	14.2, 1548	
Lean inventory (LI)	511	457	358	215, 750	28, 1189	20, 1323	13.6, 1483	4.2%
Responsive utilization (RU)	511	457	358	214, 750	28, 1187	20, 1321	13.7, 1481	0.2%
Unresponsive utilization (UU)	512	458	359	215, 752	28, 1192	20, 1326	13.7, 1487	

NPV costs (million $): endogenous demand

Policy	Mean	Median	SD	50% CI	90% CI	95% CI	100% CI	Median savings
Safety stock (SS)	602	549	374	294, 851	97, 1317	68, 1429	27, 1670	26.3%
Lean inventory (LI)	790	745	383	499, 1043	234, 1499	170, 1617	65, 1987	
Responsive utilization (RU)	769	722	378	480, 1019	224, 1469	161, 1589	61, 1947	
Unresponsive utilization (UU)	731	683	374	437, 979	200, 1431	143, 1549	54, 1884	5.4%

The previous conclusions hold for a range of holding and lost sales costs. Adopting lean inventory and responsive utilization policies leads to lower costs when demand is assumed exogenous. However, when that is not the case, safety inventory buffers and unresponsive capacity utilization policy yield lower costs. The greatest savings potential take place when demand is assumed endogenous. When the policies are tested independently, savings of 26% come from the adoption of a safety stock policy and 5% come from an unresponsive utilization policy. When we explored the combined effectiveness of the inventory and utilization policies however, we verified that the responsive and unresponsive policies tested lead to similar results. That is, there was a balance between the costs of holding additional inventory obtained with the unresponsive utilization policy and the benefit of reduction in the costs of lost sales. Therefore, when demand is assumed endogenous, a safety stock policy is recommended due to its high savings potential.

Sensitivity analysis

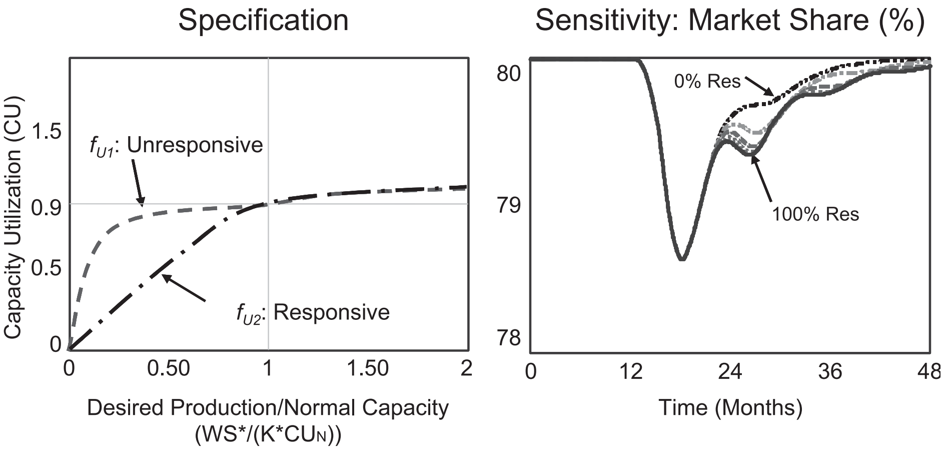
Model behavior is highly sensitive to the assumptions embedded in the capacity utilization (f_U) and customer response (f_A) functions. In particular, the model is sensitive to (1) the slopes of the non-linear functions f_U and f_A, (2) the maximum capacity utilization, and (3) the minimum of the customer demand response. The sensitivity analysis follows a common procedure to obtain its results. We represent each non-linear function—capacity utilization and customer response—as a linear combination of two polar cases, capturing extreme assumptions. By varying the weight in the linear combination it is possible to obtain a range of behaviors in the model.

Sensitivity to capacity utilization Consider two extreme types of Fab managers reacting to desired production: a responsive and an unresponsive manager. Both managers respond to high desired production by increasing capacity utilization. They respond differently, however, to low desired production volume. When desired production is low, the unresponsive manager, characterized by function (f_{U1}), prefers to keep the Fab running and build up inventory levels down the chain rather than slowing the line or shutting down. In the limit, of course, utilization must fall to zero as desired production falls to zero. The unresponsive policy is shown in Figure 8; utilization has a flat slope near the normal operating region. In contrast, a responsive manager, characterized by function (f_{U2}), responds aggressively to decreases in desired production by cutting utilization in proportion to the decline in desired production, avoiding the buildup of inventory and making the unneeded capacity available for process improvement, test runs or preventive maintenance (Figure 8). Intermediate cases are obtained from the linear combination of the two extremes; the base case sets $w_1 = 0.5$.

$$\mathrm{CU} = w_1 f_{U1} + (1 - w_1)\, f_{U2};\ w_1 \in [0,1] \tag{22}$$

Figure 8 shows market share for different specifications of capacity utilization function. System variability increases with managers' responsiveness to changes in desired production. It appears intuitive that more responsiveness should enhance stability by preventing the accumulation of excess inventory during periods of low demand. The opposite is observed, however, because demand is endogenous: by cutting production aggressively when demand is perceived

Fig. 8. Market share sensitivity to capacity utilization specification



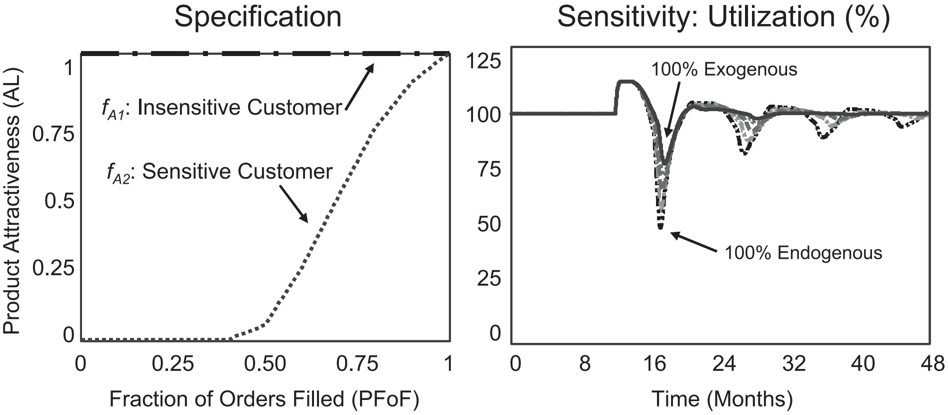
to be low, the firm ensures that inventory will be even less available, driving market share still lower. The unresponsive policy pushes product into the supply chain, improving availability and bringing customers back to the firm more quickly. Moreover, shorter delays in customer response and forecasting accentuate the impact of responsiveness on market share. Therefore, distribution channels with more responsive customers (e.g., online sales) are particularly vulnerable to a responsive capacity utilization policy.

Sensitivity to customer response Now consider the two extreme cases of customer responses. An insensitive customer base, characterized by function (f_{A1}), does not respond to changes in the perceived service level. The insensitive customer response function around the operating point (1,1) is flat. This extreme case cuts the feedback from the supplier's service level to customer demand. In contrast, a sensitive customer base, characterized by function (f_{A2}), responds aggressively to changes in perceived service. The slope of the sensitive customer response function around the operating point is steep. Sufficiently low perceived service levels can reduce product attractiveness to the minimum possible level. A general customer response function is obtained from the linear combination of the two polar cases (f_{A1} and f_{A2}); in the base case $w_2 = 0.5$.

$$\mathrm{CR} = w_2 f_{A1} + (1 - w_2) f_{A2};\ w_2 \in [0,1] \tag{23}$$

Figure 9 shows the results for different degrees of customer response. System variability increases as customers become more responsive to product availability. This result is expected. When demand is exogenous the production loop operates independently. By adding the customer response balancing loop, we have noted that oscillations in production increase. It is also sensible to expect that a more sensitive customer base, reacting with a perception delay

Fig. 9. Utilization (% of CU_N) sensitivity to customer response specification



to inventory availability, will introduce more variability in demand and consequently to production. Customer response sensitivity shows that supply chain instability with exogenous demand is much smaller than the instability with endogenous demand. That is, capturing the feedback of product availability with customer demand amplifies supply chain instability. This result suggests that models that adopt exogenous demand underestimate the instability in supply chains, which undermines associated policy prescriptions.

In “Impact of pull system”, above, we learned that the assembly and FGI pull loops are capable of stabilizing the system, but that only takes place when sufficient inventory is available. Therefore, an important policy to improve system stability is to maintain safety stocks in both assembly and FGI. The next section investigates the optimal levels of assembly and FGI safety stocks.

Optimal safety stock location analysis

The desired level of safety stock in assembly and finished goods is based on a control heuristic that optimizes the trade-off between the cost of holding inventory and the cost of lost sales. Using the same cost structure defined above (“Impact of endogenous demand”), the criterion to evaluate the optimal level of safety stocks is minimization of net present value of cumulative discounted costs (CDC), with a discount rate (r). The fraction of inventory volume destined as safety stock in each stage is given by a percentage (p_{AWIP}, p_{FGI}) of the initial volume in that stage (AWIP, FGI). The optimal safety stocks are the values of the safety stock percentage at each stage that minimizes the net present value of CDC over the simulation period. Demand is specified as the sum of the current level and an auto-correlated (pink) noise term with a standard deviation (σ) of 5% (representative of demand variations for Intel).[10]

We investigate the optimal volume of safety stocks for four different levels of lost sales cost (α) and four ratios of unit inventory holding costs in assembly

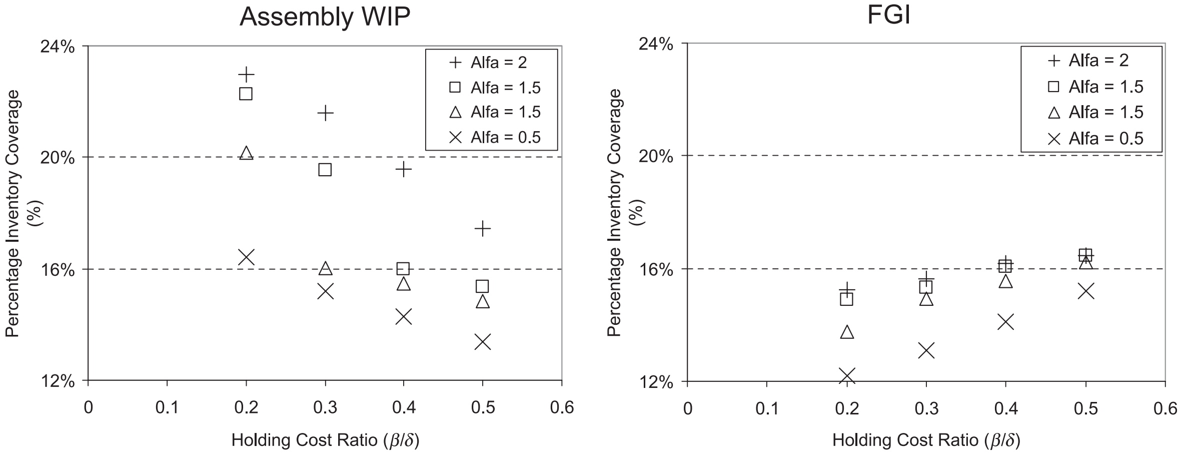


Fig. 10. Optimal percentage safety stocks in (a) assembly WIP and (b) FGI

and finished goods (β/δ).[11] Figure 10 shows the results of the optimization runs for (a) the percentage of the total volume in assembly (p_{AWIP}) and (b) the percentage of the total volume in finished goods (p_{FGI}). As expected, optimal safety stocks in both stages increase with the lost sales cost (α). Likewise, the allocation of safety stock between assembly and finished goods is highly dependent on the ratio of holding costs in the two stages. Lower holding cost ratios benefit the allocation of safety stock in assembly. In addition, for large enough holding cost ratios, the manufacturer holds more safety stock in FGI as the cost of lost sales increases. This result makes intuitive sense as higher safety stocks in FGI allows the company to meet demand with a shorter response time.

Furthermore, as the throughput time in assembly (τ_{A}) is much longer than in finished goods (τ_{OP}), the *same* percentage of safety stock volume in assembly and finished goods will translate into a *higher* safety stock coverage in assembly. For instance, as the throughput time in assembly (τ_A) equals 4 weeks compared to 1 week in finished goods (τ_{OP}), a 15% safety stock in assembly and finished goods translates into a safety stock coverage of 0.6 weeks in assembly and 0.15 weeks in finished goods. Hence, the same dollar investment in safety stocks yield higher inventory coverage in assembly. This result, while intuitive, has a direct impact on the market share that the company can retain and the resulting instability of the supply chain. Comparing the impact of the same dollar amount in safety stocks, the higher inventory coverage in assembly has a more stabilizing effect in the supply chain variability, which results in shorter delays, fewer orders expedited, and more satisfied customers. While counter to the intuition that safety stocks should be placed in FGI, due to the increased responsiveness to demand, performance is improved by keeping the safety stock in assembly. Therefore, the heuristic of maintaining additional safety stock in assembly can help stabilize the system operation and improve service

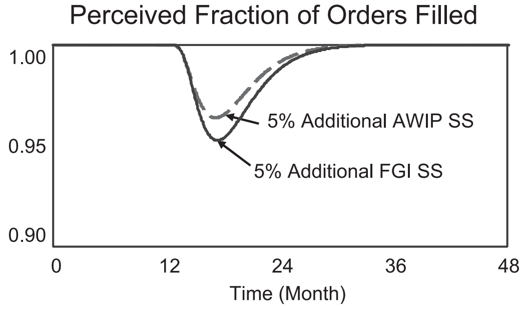


Fig. 11. Performance comparison with equal safety stock investments in AWIP and FGI

level. Figure 11 compares the performance of a 5% additional safety stock policy in assembly WIP and FGI from the base case run.[12]

The holding costs of inventory are easily measured, highly salient, and unambiguous. In contrast, the costs of lost sales are hard to assess, and demand variability is easily explained away as resulting from factors outside the firm. Thus it is likely that managers will underestimate the costs of poor delivery performance (i.e., lost sales, reputation losses, customer inconvenience, etc.). If the manufacturer underestimates the cost of lost sales, it will also underestimate the optimal safety stocks to be maintained in assembly and finished goods. In addition, not only will holding too little inventory lead to higher costs than necessary, but doing so alters the dynamics of the system by reducing the stability of the supply chain. This instability has probably many other costs that aren't accounted for. Other operational costs include shorter run lengths and smaller batch sizes, more frequent set-ups and changeovers, higher error and rework rates, and more idle time between lots and between set-ups.

Discussion and directions for future research

This paper addressed the causes of oscillatory behavior in capacity utilization at a semiconductor manufacturer and the role of endogenous customer demand in influencing the company's production and service level. The modeling effort drew on extensive fieldwork, including direct observation, collection of archival materials and data, and structured and semi-structured interviews with managers at Intel. The paper contributes to our understanding of the role that customer response has on increasing demand amplification across supply chains by exploring the mechanisms through which endogenous customer demand interacts with managers' production heuristics. The interaction of the *sales* and *production effects* greatly amplifies the oscillatory behavior of production. This result suggests that models that adopt exogenous demand underestimate the instability in supply chains, which undermines associated policy prescriptions. Our research shows that capturing the feedback of product availability with customer demand amplifies supply chain instability, ultimately reversing traditional policy prescriptions about inventory and capacity utilization policies. Simulation runs suggest that typical policy prescriptions of lean inventory and responsive utilization policies hold only when demand is assumed exogenous. With endogenous demand, inventory buffers and an unresponsive capacity utilization policy yield lower costs.

Loop knockout analysis suggests that the balancing *FGI Pull* (B1) and *Assembly Pull* (B3) loops stabilize the system. However, since they can only operate effectively when sufficient inventory is available, the manufacturer should benefit from maintaining larger inventory buffers at assembly and finished goods. While the heuristics of keeping inventory buffers at assembly and finished goods for improving system robustness is not new, this research underscores their importance in the operation of hybrid push–pull systems. In addition, the policy analysis supports an amount of safety stocks that increases with lost sales cost and an allocation dependent on the ratio of holding costs between assembly and finished goods. Therefore, the manufacturer can effectively reduce supply chain instability and reduce the impact on lost sales, as long as it internalizes the lost sales cost.

In general, semiconductor manufacturers, as well as firms in other industries, tend to keep low inventory levels and run lean supply chains, allowing them to reduce inventory costs. This practice presents manufacturers with a strategy to avoid costs associated with inventory obsolescence in industries with short product life cycles. The mental model motivating lean inventories assumes demand variability is exogenous: in a world with unpredictable demand changes, costly FGI, and rapid technological obsolescence, keeping inventories lean minimizes the risk that the firm will be caught with excess stock if demand unexpectedly declines. However, demand is not exogenous—product availability affects demand, which then feeds back to availability. Low finished and work-in-process inventories increase the chance of stockouts in different stages in the supply chain, boosting the likelihood that the system will operate in an undesirable mode (e.g., as a push system). Considering the typical inventory management heuristics adopted by companies, like the constant adjustment of desired inventory levels to reflect current demand signals, and the potential increase in demand variability introduced by customer responses, we note that companies may underestimate the true costs associated with stockouts.

Moreover, managers' heuristics of adjusting capacity utilization to respond to variability in demand—caused by the supplier's inability to satisfy the customer—can amplify the demand variability. Because demand is endogenous, by cutting production aggressively when demand is perceived to be low, the firm ensures that inventory will be even less available, driving market share still lower. Hence, the supplier's effort to meet customer demand in the short run may actually hurt customer service in the long run. In contrast, the unresponsive policy pushes product into the supply chain, improving availability and bringing customers back to the firm more quickly. However, if the company already adopts a safety stock policy, the unresponsive and responsive utilization policies yield similar results. That is, the additional inventory resulting from an unresponsive policy reduces lost sales costs by a similar amount that it increases holding costs.

There are a number of opportunities for future research motivated by this study. First, there are many other industries (e.g., automobiles, electronics) where the effects reported here may also play an important role. Second, our study currently abstracts away from the introduction of new products over time and the characteristic demand patterns over the product life cycle, where there are often initial shortages during production ramp up, followed by a secular demand decline at the end of the product life. Optimal safety stock levels and production heuristics may change over the course of the life cycle. Moreover, our model incorporates only the response of customers due to current service level (e.g., supply reliability). However, if customers consistently experience poor delivery reliability they may choose to make other firms their primary suppliers, reducing sales for other products and for future products, and also increasing the variability of demand (Risch *et al.* 1995). Order cancellations can also be added to the model. If order cancellations occur as a result of a decrease in service level, they are likely to amplify the effects caused by lost sales, which would strengthen the results presented here. In addition, the current model does not incorporate the possible inflation of orders by customers, creating phantom demand or bubbles, when multiple OEMs hedge against supply shortages (Gonçalves 2003, Sterman 2000, Ch. 18.3). Phantom demand is important since it is likely to balance the effects of lost sales and counter the effects observed in this research. While not reported here, we incorporated the assumption and conducted a number of simulations to investigate the impact on the results. Our analysis concludes that for plausible values of inflationary ordering the main results of this paper still hold. While this study does not incorporate these important assumptions—order cancellation and order inflation—they have been addressed thoroughly by one of the authors in another study (Gonçalves 2003). The hope is that by separating the effects we can help clarify their distinction and impacts to supply chains.

Notes

1. The actual number of dies per wafer range from 100 to 1,000, depending on the chip size, which varies with its architecture—whether the chip is "logic" or "memory"—and its specific design. Each die is composed of individual devices such as transistors and memory cells.
2. A full description of the model, formulations, and assumptions can be found in Gonçalves (2003).

3. The fabrication process, represented as a single stock, aggregates a complex, multi-step set of activities within the Fab. While the higher level of aggregation is consistent with the purpose of this paper, many interesting practical and research questions related to optimizing the flow of different SKUs through a Fab would require further disaggregation.
4. While the heuristic for the desired die inflow rate does not take into consideration the adjustment from FGI explicitly, information from FGI is used to set the desired assembly WIP (AWIP*).
5. We assume the normal capacity utilization level at Intel to be equal to 90% of maximum capacity.
6. It is straightforward to add random noise to the forecast to capture the impact of these sources of error and adjustment.
7. In practice, the integration limits (from 0 to ∞) go from 0 to a finite (but large) end simulated time. End time is selected ensuring that discounting has reduced the remaining contribution to NPV to a negligible amount.
8. The lean and safety inventory policies are tested with 0% and 5% of safety stock in AWIP and FGI, respectively.
9. The comparison for capacity utilization policies are performed with 0% safety stock in AWIP and FGI to make the impact of the utilization policy more salient.
10. We use auto-correlated random demand to explore the effectiveness of the safety stock strategy under a more realistic demand signal. The same random number seed is used for all simulations, allowing each optimization run to use exactly the same realization of the random process.
11. We assume also that the parameters for unit costs of finished goods (ϕ) = 50 \$/unit, and a discount rate (r) = 0.01/month.
12. To obtain a comparable dollar amount for the same amount of safety stocks in FGI and AWIP, we set the ratio of holding costs (β/δ) to 0.25, which compensates exactly for the inventory coverage ratio between the two stages.

Acknowledgements

The authors would like to thank project managers Jim Rice, from MIT, and Mary Murphy-Hoye, from Intel, and the research assistance of Charlie Lertpattarapong. We would also like to thank the many people at Intel that have contributed to this project: Jim Kellso, David Fanger, Gordon McMillan, Karl Kempf, Daniel Mckeon, George Brown, Brian Kelly, Robert Kelly, Gene Meieran, Dean Phillips, Roberta Bailey Roberts, Nick Rose, Michael Waithe, Ann Johnson, and Jay Hopman. Work reported here was supported by the Intel Corporation. Complete documentation on the model as well as more information on the research program that generated this paper can be found at http://moya.bus.miami.edu/~pgoncalves/.

References

Anderson E, Fine C. 1999. Business cycles and productivity in capital equipment supply chains. In *Quantitative Models for Supply Chain Management*, Tayur S, Magazine M, Ganeshan R (eds). Kluwer: Norwell, MA; 381–415.

Baganha M, Cohen M. 1998. The stabilizing effect of inventory in supply chains. *Operations Research* **46**: S72–S83.

Chen F, Drezner Z, Ryan J, Simchi-Levi D. 2000. Quantifying the bullwhip effect in a simple supply chain: the impact of forecasting, lead times, and information. *Management Science* **46**(3): 436–443.

Croson R, Donohue K. 2000. *Behavioral causes of the bullwhip effect and the observed value of inventory information*. Wharton School of Business working paper, University of Pennsylvania.

Dana J, Petruzzi N. 2001. Note: the newsvendor model with endogenous demand. *Management Science* **47**(11): 1488–1497.

Diehl E, Sterman JD. 1995. Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes* **62**(2): 198–215.

Forrester JW. 1958. Industrial dynamics: a major breakthrough for decision makers. *Harvard Business Review* **36**(4): 37–66.

——. 1961. *Industrial Dynamics*. MIT Press: Cambridge, MA (now available from Pegasus Communications, Waltham, MA).

——. 1968. Market growth as influenced by capital investment. *Industrial Management Review* **9**(2): 83–105.

Gaither C. 2001. Intel beats forecast; warns of revenue shortfall. *New York Times* 17 January: C1.

Gans N. 1999a. *Customer learning and loyalty when quality is uncertain*. Working paper, OPIM Department, The Wharton School, University of Pennsylvania, Philadelphia, PA.

——. 1999b. *Customer loyalty and supply strategies for quality competition*. Working paper, OPIM Department, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Gonçalves P. 2003. Demand bubbles and phantom orders in supply chains. PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Graham A. 1977. Principles on the relationship between structure and behavior of dynamical systems. PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Graves S. 1999. A single-item inventory model for a non-stationary demand process. *Manufacturing and Service Operations Management* **1**: 50–61.

Hall J, Porteus E. 2000. Customer service competition in capacitated systems. *Manufacturing and Service Operations Management* **2**(2): 144–165.

Hendricks K, Singhal V. 2003. The effect of supply chain glitches on shareholder wealth. *Journal of Operations Management* **21**: 501–522.

Holmes S. 1997. Parts shortages delay Boeing production. *Pittsburgh Post-Gazette* 4 October: C-12.

Lee H, Padmanabhan V, Seungjin Whang. 1997a. Information distortion in a supply chain: the bullwhip effect. *Management Science* **43**(4): 546–558.

——, ——, ——. 1997b. The bullwhip effect in supply chains. *Sloan Management Review* Spring: 93–102.

Mass N. 1975. *Economic Cycles: An Analysis of Underlying Causes*. Wright-Allen Press: Cambridge, MA (now available from Pegasus Communications, Waltham, MA).

Mitchell TW. 1924. Competitive illusion as a cause of business cycles. *Quarterly Journal of Economics* **38**(4): 631–652.

Morecroft JDW. 1980. A systems perspective on material requirements planning. *Decision Sciences* **14**: 1–18.

Risch J, Troyano-Bermúdez L, Sterman J. 1995. Designing corporate strategy with system dynamics: a case study in the pulp and paper industry. *System Dynamics Review* **11**(4): 249–274.

Sterman JD. 1989a. Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science* **35**(3): 321–339.

——. 1989b. Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes* **43**(3): 301–335.

——. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*., Irwin–McGraw-Hill: Chicago, IL.