Managing Variability in the Semiconductor Supply Chain

by

Jeffrey D. Johnson
Bachelor of Science in Industrial & Systems Engineering,
University of Florida (2001)

Submitted to the Engineering Systems Division in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Engineering Systems

at the Massachusetts Institute of Technology
February 2005

Signature of Author______________________________
Engineering Systems Division
January 10, 2005

Certified by______________
Stephen C. Graves
Abraham J. Siegel Professor of Management Science and Engineering Systems
Thesis Supervisor

Accepted by______________________________
Richard de Neufville
Professor of Engineering Systems
Chair, Engineering Systems Division Education Committee

This page is intentionally left blank.

Managing Variability in the Semiconductor Supply Chain
by
Jeffrey D. Johnson

Submitted to the Engineering Systems Division on January 10, 2005 in partial fulfillment of the Requirements for the Degree of

Master of Science in Engineering Systems

Abstract

Within the semiconductor industry, the variability in both supply and demand is quite high; this uncertainty makes supply chain planning very difficult. We analyze the current tools and processes at a large semiconductor manufacturing company and then propose a framework for improvement based on hierarchical production planning. We present an appropriate decomposition for this specific planning problem and illustrate some limitations of traditional inventory models. New safety stock equations are developed for this planning problem based on a simple analysis using the basic ideas from probability theory. We also devise a new method to determine lead times that more accurately captures the actual lead time seen in the supply chain. Finally, an algorithm is developed to determine appropriate inventory levels and production allocation. These ideas, when used together, provide a powerful framework to properly manage supply chains in highly stochastic environments.

Thesis Supervisor: Stephen C. Graves
Abraham J. Siegel Professor of Management Science and Engineering Systems

This page is intentionally left blank.

Acknowledgements

I am grateful to my advisor, Steve Graves, for his patience and time. His advice and insight provided me many opportunities to learn and grow professionally. Additionally, I would like to thank the resources made available to me through the Singapore-MIT Alliance.

My work could not have succeeded without the help of many people at the Intel Corporation. Specifically, Karl Kempf and Dennis Arnow provided countless amounts of advice, suggestions, and ideas that helped strengthen this work. I am grateful to have had the opportunity to work with both of them.

I would also like to thank Sean Willems for taking time to meet with me to discuss my research. His support and advice was invaluable in this experience.

Also, I wish to thank all of my colleagues at MIT. It has been an amazing and humbling experience to be surrounded by the best and brightest people in the world.

Finally, I want to thank my wife Sarah. Her support and encouragement has been amazing and I could not have succeeded in this endeavor without her.

This page is intentionally left blank.

Table of Contents

List of Figures

List of Equations

1 Introduction and Overview

This thesis examines several important issues in the management of the supply chain for a semiconductor company. Many of the results easily extend to other industries; however, this research was motivated by work with a large semiconductor corporation. We open with a background on the industry and the manufacturing process. Then, we diagnose the current supply chain / production planning system. Next, we discuss a framework for improvement followed by several specific improvements supported by the framework. In particular, we evaluate some generally accepted inventory equations, show shortcomings in their application in this industry, and propose new methods to evaluate inventory targets. Additionally, we propose a new way to measure lead time that more accurately reflects the actual lead time seen in the supply chain. We also propose a solution method to solve the production planning problem given the semiconductor binning characteristics.

1.1 Industry Background

The semiconductor industry is characterized by very short product lifecycles, rapid technological change, and very high fixed costs. To support these claims, we provide the following facts. In 2004, a new wafer fabrication facility cost over $2 billion, which clearly supports the notion of high fixed costs[1]. To support the claims regarding product lifecycles and technological change, we cite Moore's Law, which states that the transistor density on integrated circuits doubles every 18-24 months[2]. It is widely held

---

[1] http://www.semiconductorfabtech.com/industry.news/0006/20.08.shtml
[2] http://www.intel.com/labs/eml/index.htm

that this rapid technological advancement described by Moore's Law will hold for the foreseeable future of the industry. Performance pressures in all aspects of the business will only increase in the future and finding every area of competitive advantage has become essential for growth companies.

Several companies have identified supply chain management and inventory optimization as areas to explore in terms of deriving an advantage. Bain and Co. report that 85 percent of senior executives say improving supply chain performance is a top priority (see Cook and Hagey 2003). This recognition has led to significant investment in making operations a source of competitive advantage for many companies.

We now continue with an overview of the semiconductor manufacturing process.

1.2 Manufacturing Process Overview

In this section, we provide an overview of the manufacturing process for semiconductor products, specifically microprocessors. The figure below summarizes the high level stages in the manufacturing process. Note that the actual process is much more complicated; however, it is well beyond the scope of this thesis.

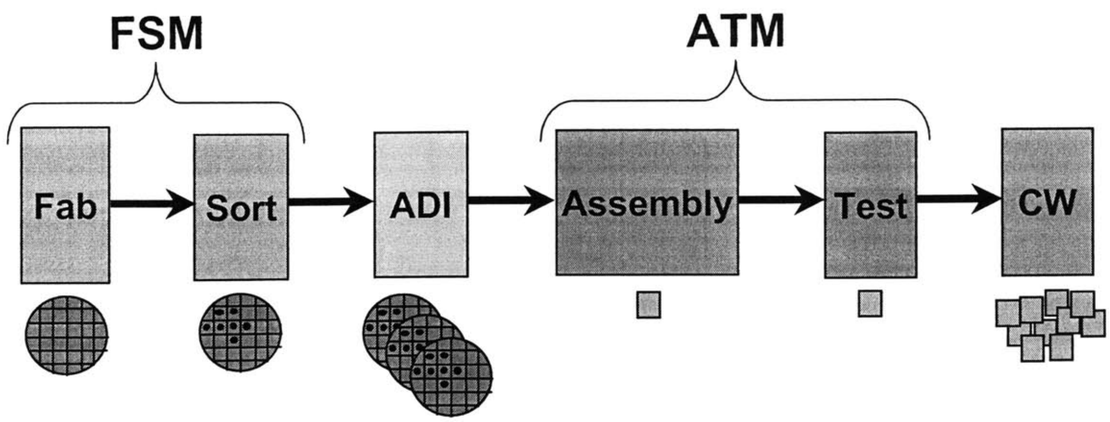


Figure 1: Semiconductor Manufacturing Overview

The process begins in fab sort manufacturing (FSM). Fab is the term used by the industry to describe a fabrication facility or the place where raw silicon is turned into actual devices (e.g. microprocessor chip or memory chip) on a wafer. Generally, there are hundreds of individual operations that take place in the fab. The details are unimportant in terms of this thesis and will be omitted. These facilities are generally located in higher cost countries that have highly trained personnel and support infrastructure. At the fab, electrical components are formed in layers or stages using a process that takes two to three months to complete (which represents a fairly lengthy throughput time).

Once the wafers finish the 'fab' portion of the process, they are sent to the sorting function where they are probed for functionality and performance in a process the company calls die level cherry picking (DLCP). Here, the individual devices (or die) are either given a code that corresponds to its predicated final characteristics (like power consumption and speed) or marked as defective. This marks the completion of the FSM portion of the manufacturing process. The wafer is then sent into assembly die inventory

(ADI) where it waits for final assembly and testing. ADI is a major inventory location in the internal supply chain.

After ADI, the process continues in assembly test manufacturing (ATM). ATM sites are generally located in lower cost regions as they are less technologically intensive as fab processes while also requiring a greater amount of manual labor. ATM is responsible for turning die from ADI into finished goods. The process begins with the ATM site taking wafers from ADI, sawing them into individual die and attaching them to the appropriate piece parts (or packages). The decision made at this point is very important as die sitting in ADI can be put into many different packages that can result in different end products. These assembled units are then sent through testing machines where they are tested and binned. Binning is the process that determines the speed at which the processor will run. The binning process occurs by placing the product in a machine that determines the natural speed (or "natural bin"). At this point, decision makers can "downbin" a product whereby its speed is lowered by blowing fuses in the chip. The natural bin is determined by the product design and events that occur during the manufacturing process while downbinning is a decision that planners can make in order to properly align supply and demand. The final product is then sent to the component warehouse (CW) where it awaits shipment to the end customer. End customers typically include large OEMs (e.g. Dell, Gateway, etc.), smaller OEMs (e.g. Toshiba, Acer, etc.), and distributors (e.g. Arrow, Avnet, etc.) The ATM portion of the process usually takes two to three weeks (much shorter compared to the FSM throughout time).

Many details were left out of the description above; however, they are not necessary for the development of this thesis. For an approachable overview on the details of the manufacturing process, see Quirk and Serda (2001).

1.3 Organization of Thesis

In this chapter, we have described the semiconductor industry and manufacturing process as well as giving insight as to why supply chain management is important for the future.

The remainder of this thesis is organized as follows. The next chapter describes the planning details at the particular semiconductor manufacturing company examined, provides background on variability and inventory, and concludes by describing this system as a so-called "engineering system." Chapter three provides an analysis of the current system while chapter four highlights a framework for improving the current system. Chapter five details some specific improvements based on the framework, including new inventory models and a new factory allocation algorithm.

We finally conclude in chapter six with some closing remarks and recommendations for future work.

2 Description of Process

We begin this chapter by discussing the naming conventions and current tools used to manage the supply chain. Then, we describe the important issues in understanding the key concerns behind managing the supply chain. Next, we discuss the impact of variability and finally conclude by examining this supply chain as an "engineering system."

2.1 Semiconductor Supply Chain –General Description

In this section, we describe the details around the information needed to manage this system. Properly understanding and capturing appropriate data regarding the flow of products through the internal supply chain is quite important in this context and further aids in understanding the manufacturing process. Many pieces of critical information are needed to "map" a product from beginning to end. This mapping provides a way to track products as they move through the manufacturing steps as well as a way to plan future products.

We now describe the flow of products through the internal supply chain in terms of their naming conventions or product mapping. Recall that products begin in the fab; these wafers are given *base product* names at this stage. We do not use any real product names or numbers from the motivating company or any other company in this thesis.

Once a wafer has left the fab, it goes into the sort process and the resulting die are given a *sort name* that depends on the assigned DLCP category. For example, there may be three categories for a particular wafer type (let's call them A, B, and C). Category A might be assigned to die that are expected to be very fast at the expense of consuming higher than average power while category B might be assigned to die that consume much lower power and thus are expected to run a bit slower. Category C may be the catch-all category for those that do not fall into the other two. This is a one-to-many relationship as one base product name usually feeds three to six sort names. We note that some products may not be assigned DLCP categories. In this case, the die are still checked for functionality and all those that pass are assigned the same category name.

Products wait in ADI as sort names; once they are pulled from inventory, they are assigned *level 3 names* that signify the package type and test program they will run on. This is a key decision point in the production process and is another one-to-many relationship as one sort name can feed many different level 3 names.

These level 3 products go through the assembly/test process and are "binned" to their final speed. Binning is the process by which the product is fused to run at its final speed; at this time products are given an *MM number* based on several characteristics (including the newly determined product speed and other characteristics, like package type, that were determined when the level 3 name was assigned). End customers place orders at the MM level. This is again a one-to-many relationship; however this time we have a form of "recombination" in the products. Even though each level 3 has a one-to-many relationship with MMs, different level 3's can feed the same MM. The figure below provides a clear illustration of the mapping scheme.

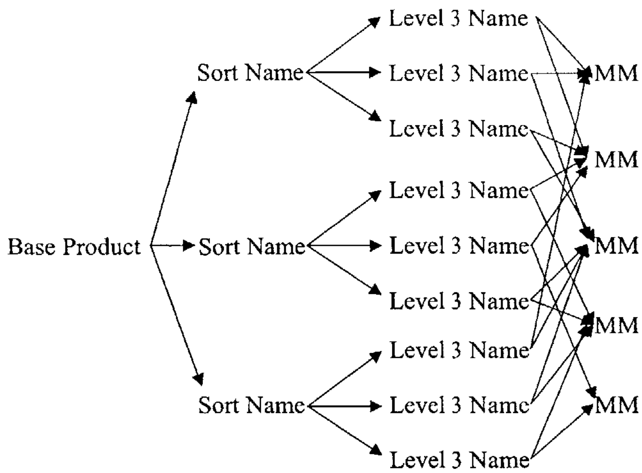
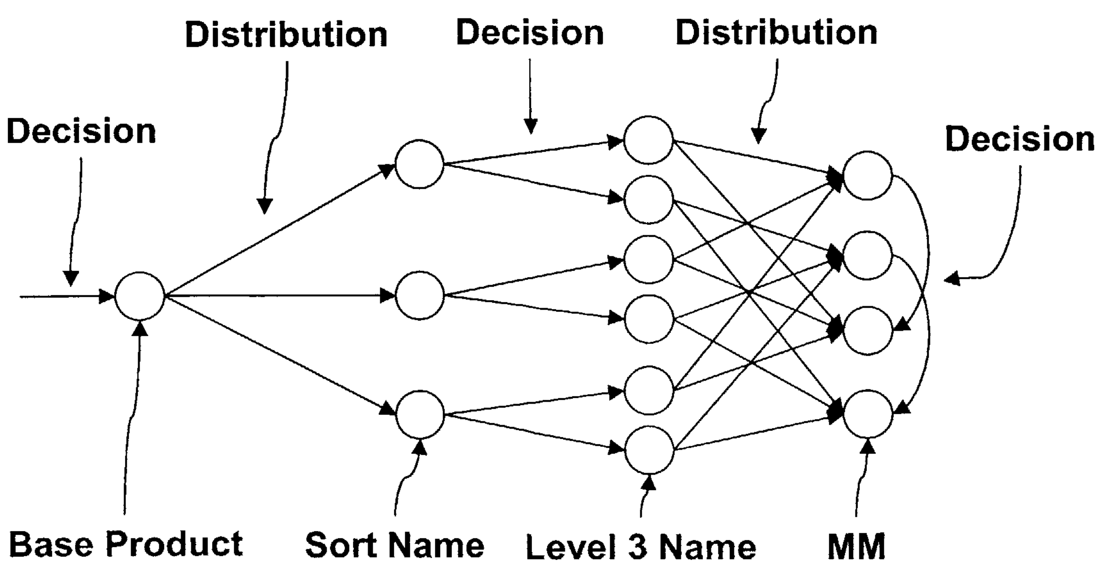


Figure 2: Product Mapping Scheme

A common source of confusion in the manufacturing process is where decisions are made versus where a probabilistic distribution governs the output from a process step.



The figure below provides a summary.

Figure 3: Decisions vs. Distributions

The figure shows that decisions are made on base products and level 3 names while probabilistic distributions govern the output as sort names and MMs. The last decision we see represents the downbinning process whereby you can reduce the speed of the final product from its natural bin by blowing appropriate fuses.

Obviously, the kind of level 3 products you have available depends on which type of wafer you started (and which sort names result from those wafers). An example of how the same sort name could be given different level 3 names is if the products were placed into different packages. Specifically, the DLCP distribution governs which sort names will come from a specific base product and the bin split distribution governs which MMs will come from a specific level 3 name.

2.2 *Variability Discussion*

High levels of variability are commonplace in many industries, including the semiconductor industry. We think of variability in terms of the two-by-two matrix below.

	Supply	Demand
Time	S-T	D-T
Quantity	S-Q	D-Q

Figure 4: Variability Matrix

It shows that we are interested in both supply and demand variability in terms of time and quantity uncertainties. Examples of each type of variability follow:

- supply/time uncertainty - the variability of factory throughput times,
- supply/quantity uncertainty – the variability of product yields,

- demand/time uncertainty – uncertainty regarding when customers want products delivered,
- demand/quantity uncertainty – uncertainty involving which products customers want and how many units of each a customer wants.

These uncertainties cause manufacturing firms to hold inventory in their supply chain. It is important to fully understand the uncertainties you face in order to properly manage the risk.

A summary of all the important parameters for the semiconductor supply chain follows below with definitions of each and a categorization into an uncertainty group. We note that a data analysis of these different sources of variability is provided in Levesque (2004).

- FSM TPT = throughput time for wafers in the fab. The data is in terms of base product names. FSM TPT falls into supply-time uncertainty.
- GDPW = good die per wafer. This represents how many die on the wafer are actually functional. The data is in terms of base product names. GDPW falls into supply-quantity uncertainty.
- DLCP = die level cherry picking. DLCP provides a predictor of future performance of each chip. The data is in terms of both base product names and sort names. DLCP falls into supply-quantity uncertainty.
- ATM TPT = throughput time for die in the assembly/test site. The data is in terms of level 3 names. ATM TPT falls into supply-time uncertainty.
- ATM yield = yield of products at the assembly/test site. The data is in terms of level 3 names. ATM yield falls into supply-quantity uncertainty.

- Bin splits = provide information regarding the speeds at which each level 3 is predicted to test out. The data is in terms of both level 3 names and MMs. Bin splits fall into supply-quantity uncertainty.
- Demand = end customer demand. Customers order by MM number and currently forecasts are made by MM number. Demand falls into both demand-time and demand-quantity uncertainty.

A typical engineer's approach to variability generally takes the following four steps: (1) identify the uncertainty, (2) quantify its impact, (3) reduce what you can, and (4) manage what's left. We refer the reader to Levesque (2004) for details regarding steps one and two. This thesis focuses on methods to manage the remaining variability.

Properly managing variability is not always intuitive. The following example was created to educate managers on the importance of variability and why properly understanding it is both useful and profitable. It is included in this thesis to illustrate how a simple example can have a profound impact on understanding. This example helped managers in several organizations understand why properly comprehending uncertainty is so important.

2.2.1 Variability Example

Let's assume a company manufactures two products, called Product A and Product B. Also, assume we have a one period model, the total production capacity is 100 units, and we have the following demand and financial information (note we assume that the demand is normally distributed).

Demand				
Product A mean =	55			
Product A st dev =	3			
Product B mean =	55			
Product B st dev =	20			
Revenue			**Costs**	
Product A revenue =	$ 100		*Product A cost =*	$ 30
Product B revenue =	$ 105		*Product B cost =*	$ 30

Figure 5: Variability Example Data

We see that Product B has the same average demand as Product A; however, there are two important differences between the products. Product B has higher variability in its demand forecast (as measured by standard deviation) than Product A, but it also has a higher margin than Product A (the difference is $5).

We use the standard inventory result that our inventory level is calculated according to the equation

$$InventoryLevel = \mu + z\sigma$$

Equation 1: Inventory Equation - Demand Variability

where μ is the average demand, σ is the standard deviation of demand and z corresponds to the required service level. See Nahimas (2001) for details on this equation and a detailed discussion of service levels. Note we will provide a basic overview of this equation later in this thesis.

Given our data set and assuming a 95% service level (thus z=1.645), the company should produce 55+(1.645*3) = 59.9 units of Product A and 55+(1.645*20) = 87.9 units of Product B. Unfortunately, the company's capacity is only 100 units and they can not provide this kind of service level. The question becomes how much of each product should the company produce in order to maximize profit.

We now turn to evaluating different strategies for dealing with this problem. Three different ideas will be examined in terms of expected profit. The expected profit will be calculated for each strategy and important questions will be discussed regarding the results.

- Idea #1 – Allocate *equal* capacity to each product

	Prod	~E[Profit]
Product A	50	$ 3,493
Product B	50	$ 3,318
Total	100	$ 6,811

Figure 6: Expected Profit - Strategy 1

We see this results in an expected profit of $6,811. The first important question is given that the expected profit per unit of Product A is $70 and the company produced 50 units, why is the expected profit not $70*50 = $3,500?

The answer lies in the fact that even though 50 units are produced, there's no guarantee that all 50 will be sold. A probabilistic model takes this possibility into account. Since there's positive probability that the company will sell less than 50 units, their expected profit cannot be $3,500, it must be less.

- Idea #2 – Allocate *more* capacity to the *higher margin* product

	Prod	~E[Profit]
Product A	47	$ 3,271
Product B	53	$ 3,447
Total	100	$ 6,718

Figure 7: Expected Profit - Strategy 2

We see this results in an expected profit of $6,718; this is nearly $100 less than the equal allocation policy. The important question here is why did the expected profit go down? For many managers, this result is counterintuitive.

The reason it worked out this way is because the higher margin enjoyed by Product B was not enough to counteract the variability that Product B sees. Thinking in terms of certainty of dollars is helpful here. Each unit of Product B sold results in $5 more than Product A, but the certainty of selling that product and actually getting that dollar is much lower with Product B. Stated another way, there's not a 1:1 relationship between production and sales.

- Idea #3 – Allocate *more* capacity to the *less variable* product

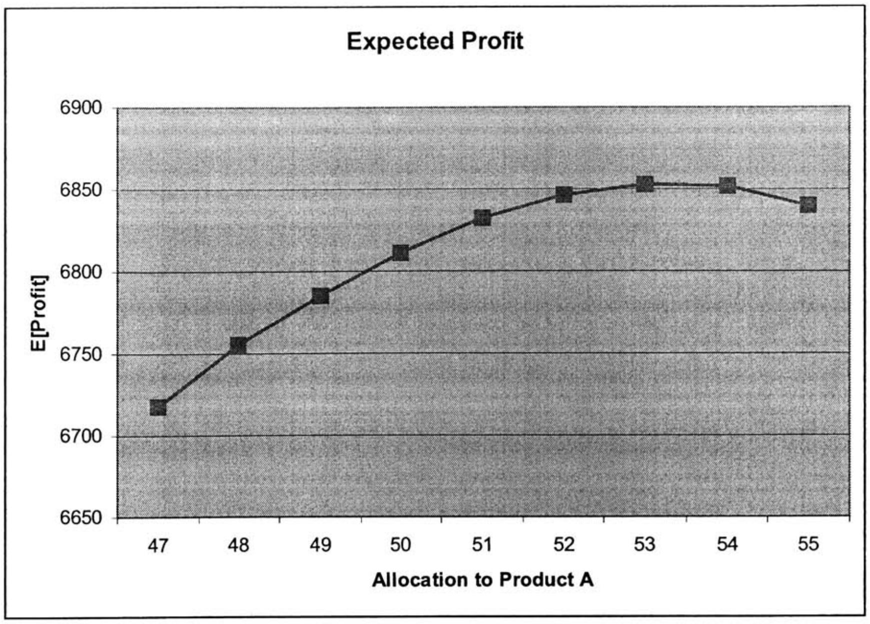
	Prod	~E[Profit]
Product A	53	$ 3,676
Product B	47	$ 3,176
Total	100	$ 6,853

Figure 8: Expected Profit - Strategy 3

We see this results in the highest expected profit yet at $6,853. Why is this the highest profit yet?

This is because the variability in Product B is too great to be overcome by the increase in profit. Under this strategy, the company is better off because it's producing more of the product that's more certain to be sold.

An obvious last question is what strategy results in the maximum profit for this company? The answer is shown in the graph and table below.



A	B	E[Profit]
47	53	$ 6,718
48	52	$ 6,756
49	51	$ 6,785
50	50	$ 6,811
51	49	$ 6,832
52	48	$ 6,846
53	47	$ 6,853
54	46	$ 6,852
55	45	$ 6,840

Figure 9: Expected Profit Graph and Table

2.3 Supply Chain Characterization

Now that we have introduced the manufacturing process, product naming conventions, and given an introduction to variability, we will begin to discuss some of the details within this supply chain. The semiconductor industry has some very specific characteristics that make their supply chain planning problem difficult. We now turn to identifying and discussing some of these important characteristics.

1. Lead time issues

   The lead time in the fab (FSM) is usually between 8-12 weeks while the lead time in assembly/test (ATM) is usually between 2-4 weeks. Thus we see that ATM lead time << FSM lead time.

2. Zero cancellation window

    Competition has dictated that customers can cancel or modify orders until they physically leave the company's shipment dock. Thus customers are constantly pushing out or pulling in orders with no repercussions.

3. Well-behaved supply parameters

    The manufacturing process is well controlled and understood. Thus, estimates of the probability distributions are available for supply parameters (e.g. yield). Usually, normality holds and fairly good estimates for mean and variance are available. Future estimates are more difficult due to technology uncertainty, but not impossible.

4. Highly variable demand parameters

    The demand parameters are much harder to characterize than supply.

5. Binning & mapping complexities

    We have discussed the ideas of binning and mapping already in this thesis. This refers to the combinatorial explosion of products as you move through the manufacturing process. Below is a figure that summarizes the main ideas.

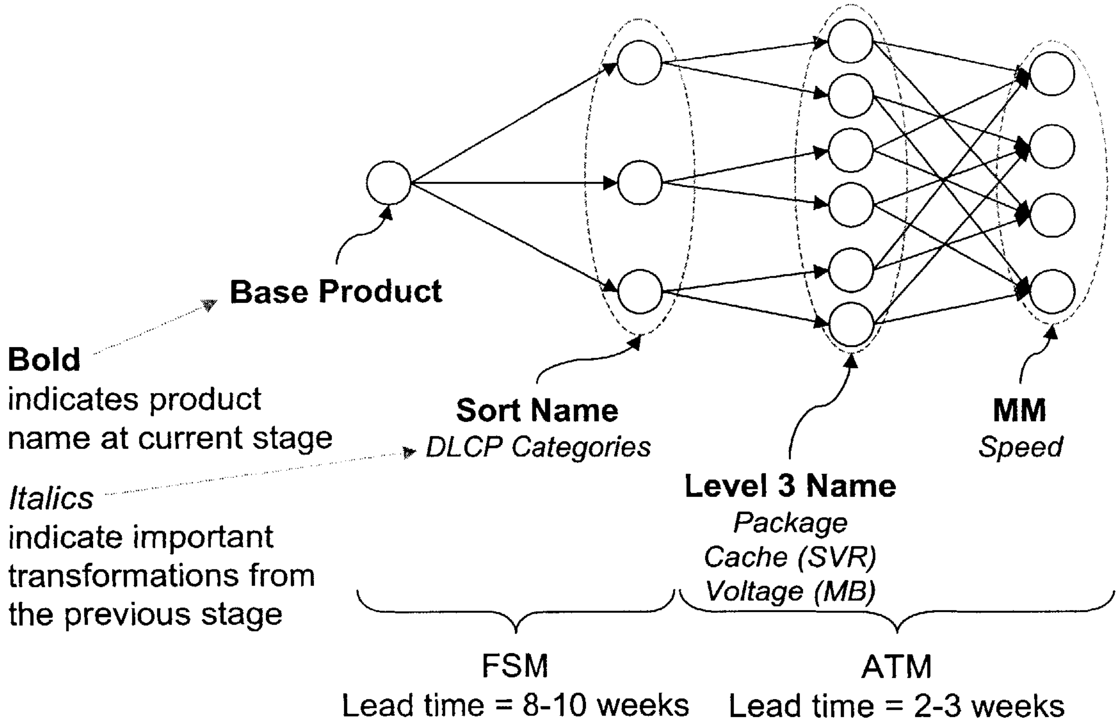


Figure 10: Binning and Mapping Complexity

2.4 Specific Supply Chain Description

We now turn our attention to the supply chain under consideration in this thesis. We examine planning inventory and production beginning at FSM, through ATM to the end customer. Distribution and warehousing details are not considered specifically in this work. We use a generic components warehouse (CW) to represent the aggregate final finished good inventory.

The supply chain considered in this thesis is an abstraction of the real life situation. This abstraction is pervasive within the industry. We model one inventory location at the end of FSM (called ADI) and a second at the end of ATM (called CW). The following figure summarizes the supply chain under consideration.

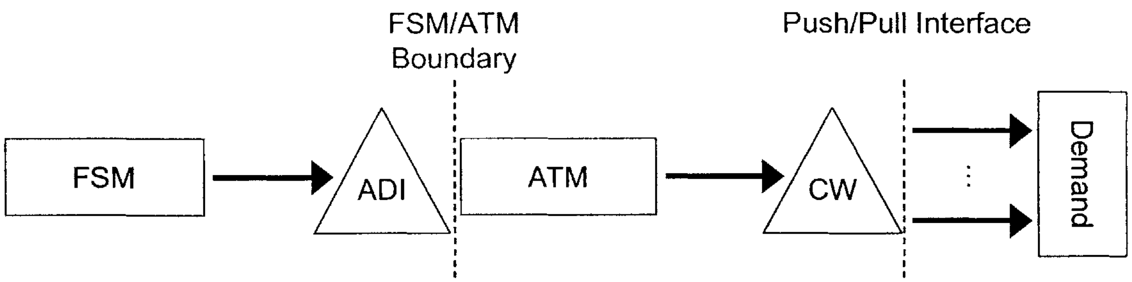


Figure 11: Supply Chain

We recognize that ADI serves to decouple FSM and ATM. Thus ADI serves to buffer ATM from uncertainties that arise in the fab. CW serves to decouple ATM from customer demand. It is important to note that due to the zero cancellation window, customers can (and do) change orders right up to the ship date. This means that actual orders are not known until the product is shipped.

The implication of the zero cancellation window is the placement of the push/pull interface (or inventory/order interface or push/pull boundary) after the CW inventory location. This interface represents the change from make-to-stock to make-to-order (see Hopp and Spearman 2004 for details). In this specific application, since customers can modify orders up until shipment, this interface is placed after products are placed into finished goods inventory (or CW).

2.5 Semiconductor Supply Chain as an Engineering System

In this section, we examine this system in terms of the definition of an engineering system. We begin with several important definitions regarding engineering systems; these are all taken from the MIT ESD Internal Symposium overview paper:

Engineering System - a system designed by humans having some purpose; large scale and complex engineering system, which are of most interest to the Engineering Systems Division, will have a management or social dimension as well as a technical one.

Complex system - a system with components and interconnections, interactions, or interdependencies that are difficult to describe, understand, predict, manage, design, or change.

Large scale systems - systems that are large in scale and/or scope; such systems have a large number of components; as a result large scale physical systems will be distributed over a region that is large relative to its smallest components.

Given that we have already described the semiconductor manufacturing process and the high-level planning issues, we are in a position to claim the semiconductor supply chain planning system is indeed an engineering system (ES).

This supply chain system was designed by humans for the purpose of fulfilling customer requests for products, thus the first part of the ES definition is satisfied.

The next part of the definition says the system must be complex and large scale. We will next connect the semiconductor supply chain planning system to the definitions given for complex and large scale.

This system obviously has many interactions that are difficult to manage and predict. Given the many uncertainties in the system (yield, DLCP, bin splits, demand), the definition for a complex system is clearly satisfied.

Now, given that a company may simultaneously manufacture 10-15 base products that give rise to a number of sort names in the hundreds that can yield several hundred (potentially thousands) level 3 names, which finally produces several hundred (potentially thousands) MMs. This combinatorial tree represents the kind of large scale scope of the planning problem. Combine this with the fact that the manufacturing sites and inventory locations are distributed throughout the entire world and clearly the definition for a large scale system is satisfied.

The final part of an ES is not only having technical complexity, but social or managerial dimensions as well. This supply chain is distributed throughout the world. Decision makers exist in the US, Europe, and Asia and have to coordinate constantly. This distributed decision structure makes for tremendous social and managerial complexity. For example, some stakeholders keep rather unorthodox hours in order to keep up communication. Cultural differences also play a role in this system. For example, the planning groups in Malaysia and Costa Rica have very different styles and philosophies, but must come together under one process to manage the supply chain effectively.

Another aspect of social and management complexity comes in the form of working with customers. These companies each have their own cultures and you must respond to that to keep them satisfied. A one-size fits all method would not work as different customers have different value drivers and these must be comprehended.

Given the above analysis, we conclude that the semiconductor supply chain planning system is an engineering system.

3 Diagnosis of Current Planning System

In this chapter, we provide a detailed explanation of the planning problem and the methods that are currently used to manage the decisions.

3.1 Concepts of the Planning System

We now look to continue our analysis of the supply chain in terms of identifying how the different issues discussed in the previous section impact how the supply chain is managed. Generally these problems can be decomposed based on answering two key questions:

1. Can we still change capacity in a meaningful way?
2. Can we still change our allocation in a meaningful way?

If the answer to the first question is yes, then we're dealing with long range planning (generally over 1 year in the future) which is out of scope for this thesis. If the answer to the first question is no (i.e. capacity is more or less fixed), then we move on to the second question. If the answer to the second question is yes, then we're making "strategic" decisions while if it's no we're making "tactical" decisions. There are many different terms used for these distinctions, we will proceed using the definitions above. We concern ourselves in this thesis with strategic decisions.

Given these definitions and the problem structure at hand, it is reasonable to say strategic decisions involve planning to uncertain demand. This is due to many things, including the zero cancellation window issue. Thus, it is also reasonable to say that tactical decisions involve planning when the demand is known. Again the push/pull interface is the place that separates these two modes of operation. See the figure below for a graphical description of these issues.

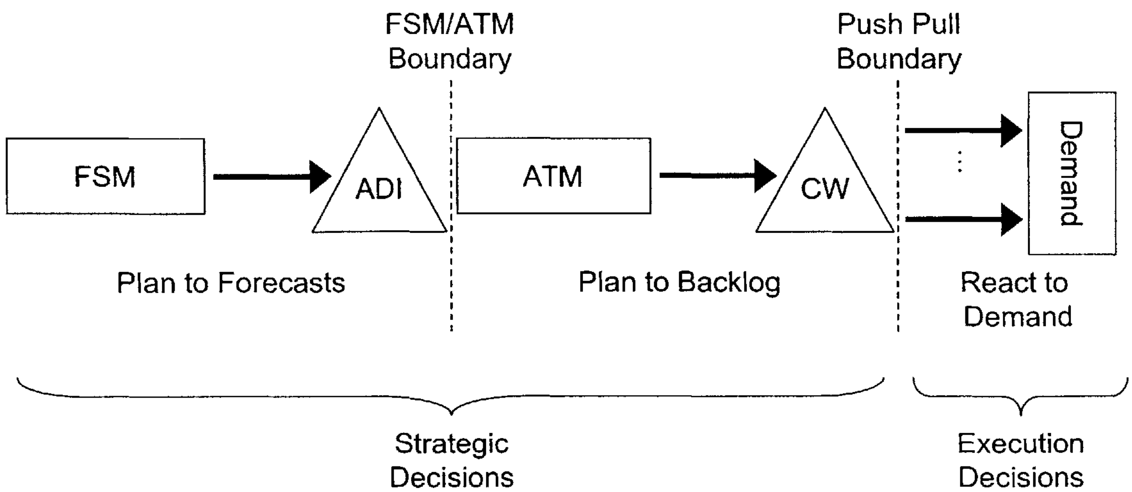


Figure 12: Strategic vs. Execution Decisions

The figure further breaks strategic decisions into two sub-sections, the first being 'plan to forecasts' and the second being 'plan to backlog.' We begin by defining the terms forecast and backlog. A forecast is defined as estimate of future customer requirements. Forecasts are made by the marketing organization based on historical sales, macroeconomic conditions, and price elasticity. Backlog is defined as customer order on the books that have not yet shipped. Given what is known about orders in this industry, it is clear that backlog is a form of a forecast. Since customers can modify orders up until shipment, their orders give only an indication of their actual intentions. In this context once an order is firm and cannot change, it is called demand.

Currently, many companies use forecasts to plan FSM production. Thus their marketing organization provides numbers that are used to generate strategic FSM production plans. Since the lead times are generally two to three months, this is a reasonable procedure. ATM production plans are generally built using backlog information. The assumption is that backlog is the best predictor of actual customer requirements given the short ATM lead time (recall that this lead time is only two to four weeks). See Levesque (2004) for a detailed discussion of these issues and a detailed data analysis.

In the execution space, the idea is to react to actual customer demand. Obviously demand is defined here as the actual customer requirement. Many different definitions for the term demand exist and here we define it as actual (known) customer requirements. Reacting to demand includes various logistics issues, including deciding which warehouse to ship the product from, deciding the mode of transportation, etc.

3.2 Details on the Current Planning Tools & Processes

We now briefly describe the current tools and business process to plan production at this specific company. The process begins with demand forecast generation by the marketing organization. Currently, point estimates are generated with no indication of variability in the forecast. These forecasts, in addition to all other necessary supply data, then are used by the FSM allocation linear programming tool to allocate fab (wafer) capacity.

Once wafers are allocated, product planners request die from ATM through a high-level analysis (i.e. without considering specific ATM site capacity or detailed product characteristics). This is accomplished using a detailed set of MS Excel™ spreadsheets. The stakeholders for this specific step are spread throughout the world. Once this high-level request for die from ATM is generated, it goes to the ATM planning community for analysis based on the current fab plan and current conditions at the ATM sites. Upon completion of this analysis, ATM provides a response to the product planners on their initial request. Again we emphasize that this is a high level plan; processor speed has not been considered as of right now and the analysis has been at the virtual ATM level (i.e. not broken down by specific sites).

The next step requires product planners to request specific speeds based on their allocation of high level die. This time, a request is sent to each individual ATM site. Again, the ATM planners analyze the request and return a response to the product planners on their actual allocation. The above steps take about two weeks to complete, starting from the FSM wafer allocation through the detailed ATM response.

There are many different computer systems involved in these planning processes. These details only complicate things further and in the interests of keeping this work as general as possible, specific ERP, planning, and data management systems will not be examined.

We illustrate this process through the use of a figure. It shows when these models are run relative to the manufacturing process.

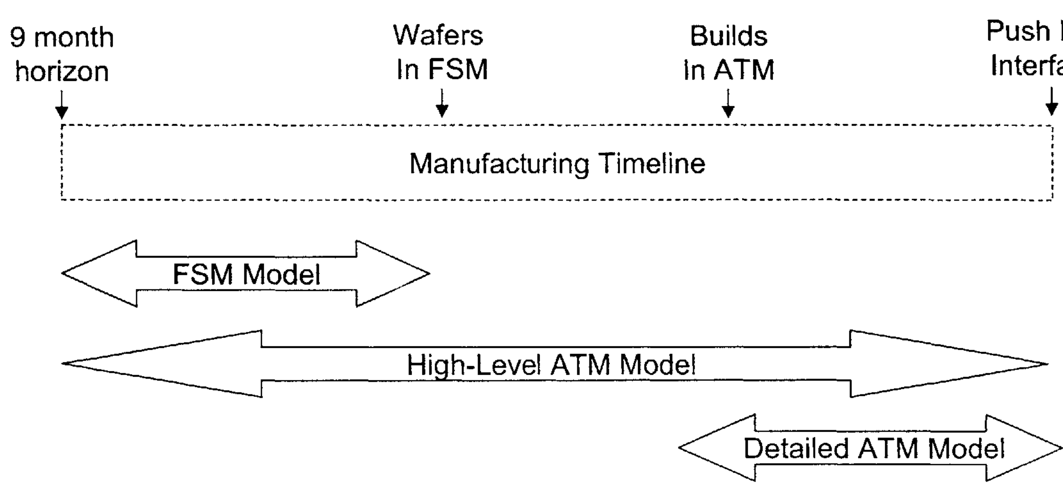


Figure 13: Planning Tool Timing

The FSM model has a 9 month horizon. Thus, it plans products from the current week through week 39. This model is complete when the wafers go into the fab. This is why, in the figure above, the arrow for the FSM model ends at the time when wafers enter in the fab. We assume another system takes over the WIP management (which is outside the scope of this thesis). The high-level ATM model also works on a 9 month horizon. The key here is that this high-level model is completed at the virtual factory level (meaning that all sites are viewed as one entity in the analysis). We see this model runs all the way through product completion. Immediately before ATM begins working on products, a detailed plan is worked out at each ATM site using an iterative procedure between the high-level ATM model and the detailed ATM model.

We see that this process is very involved and cumbersome. It takes a very long time to drive change and is fairly basic in its logic. Inventory targets are set according to heuristic judgments and much of the planning is done with very little in terms of decision support tools (i.e. many decision are made with the aid of only a simple spreadsheet). There is both good and bad to this. The good is that the process remains somewhat flexible and able to adapt quickly to changes in the marketplace. However, the down side is that there is very little control of the system and the opportunity to bring analytics into the process definitely exists.

Currently, the business organization is also very tied to its cumbersome computer systems. At several points in the above process, planners will print out data from one system and manually enter it into another. Just when the automation group gets a computerized process to do this data transformation, another one comes up due to the ever-changing business climate and needs.

The main strength in the current planning system is the flexibility inherent with not being tied to even more complex computer systems. Even though manual transformations are cumbersome, it is still possible for the job to get done. Another strength is that the system does not require advanced mathematical skill sets or large investments in training. Although it is complex, it is fairly straightforward.

The main weaknesses of the system are a lack of coordination between organizations and a lack of analytics to improve decision-making capability. We hope through this thesis to provide models and methods to improve the analytics available to the planning community at the company.

4 Framework for Improvement

In this chapter, we provide a framework for improving the supply chain planning system. Once the framework is established, we can begin to suggest improvements to the current system.

4.1 Hierarchical Production Planning Framework

We use the concept of hierarchical production planning (HPP) as a basic framework to analyze this supply chain and recommend improvements to the management system. Hax and Meal (1975) describe the HPP framework in the following way:

> Optimal decisions at an aggregate level (planning) provide constraints for detailed decision making (scheduling)

Based on this definition, we see that HPP is, in essence, a decomposition technique. The key is to properly determine what are the aggregate levels and what are the detailed levels.

Graves (1982) provides a nice discussion on the difference between the HPP approach and the more traditional, or *monolithic,* approach. We summarize the key points on this topic in the Graves paper as follows.

A monolithic approach attempts to formulate the entire planning problem in one formulation while the HPP approach breaks the problem into a hierarchy of sub problems. The key advantage of the monolithic approach is that you have one model that solves the entire problem. The drawbacks are the sheer size of the resulting model, significant data requirements, and difficult stakeholder management since all parties are at the mercy of one model.

The key advantages of the HPP approach represent the drawbacks of the monolithic approach. First, with HPP, the computational burden is dramatically reduced due to the decomposition. Thus, it is possible to solve problems that were intractable before. Second, with HPP, the data burden is reduced as aggregated data is required as opposed to detailed data. Finally, the HPP decomposition can be done to mirror a company's decision making structure.

Thus, we propose thinking about managing the supply chain under examination in this thesis using this HPP framework. It makes sense to break this problem apart given the steps in the process as well as the fact the decision makers are distributed around the work in different organizations. We will see that making optimal decisions for FSM capacity allocation provides a constraint on the lower-level, more detailed ATM decisions. This is the essence of the HPP framework and we will use this framework as a basis to propose new ways to determine wafer start plans and inventory targets.

4.2 HPP Applied to this Planning Problem

The easiest way to understand HPP is through an example. Our next step is to apply the HPP idea to the specific semiconductor supply chain planning problem under discussion here. We break the problem into three key decisions (using the HPP framework), they are:

1. Capacity Decisions
2. FSM Allocation
3. ATM Allocation

We see that overall the capacity decisions create constraints on what the FSM and ATM factories can produce. This is why it is considered the first step in the hierarchy. Capacity decisions are generally made using a long-range planning process. This particular process is out of the scope of this thesis and the capacity decisions are assumed to have been made (although minor adjustments are allowed, for example, buying additional low-cost burn-in boards for ATM).

Notice that the second and third decisions are allocating the FSM and ATM capacity. Thus, we have effectively decoupled the two processes. The company currently operates using this simplification (although the details presented later in this thesis are new), and the HPP framework supports doing so. The idea is that solving the FSM problem creates a die supply that can then be used by the ATM model. Thus solving the FSM problem creates a constraint for the ATM problem, specifically the amount of product that will be available to go into ATM.

We now turn our attention to the specific problem of determining wafer starts in FSM. Solving this FSM problem consists of the following five key steps:

1. Determine product modeling (aggregation) based on critical economic drivers
2. Determine CW inventory targets for these aggregated, finish good products
3. Solve for die (i.e. how many die of each DLCP category are required to meet the inventory targets determined in step 2)
4. Determine ADI inventory targets based on die requirements

5. Solve for wafer starts (i.e. how many wafers are needed to meet the inventory targets determined in step 4)

Step 1 (of the five steps presented above) represents where you abstract the real life setting into a model. In reality, there are countless products in the supply chain. Products can be tracked down to the lot number; however, planning wafers by using such detailed data is cumbersome and unnecessary. Thus, step 1 is where you determine which parameters are important enough to include in the model and what level of granularity is required.

Proper modeling for products should depend on the company's strategy. We provide two examples to clarify what we mean by modeling. Assume a set of five finished goods in a single product family exists. Let us further assume that the products are substitutable in the marketplace. Thus, for the purposes of planning wafers, we can add their demands into one aggregate value. This represents a product family strategy (where enough wafers are started to meet the demand for an entire product family, thus the products are assumed to be completely substitutable). In order to implement this, we can model all five products as one since their demand is substitutable.

Now assume that there exists a fast processor that has a poor bin split (meaning that you get very few of these fast processors per wafer). Further assume that this product is strategically important to the company and they want to meet the demand regardless of its impact to other products. This is an example of a bin split chasing strategy (where extra wafers may have to be started in order to get more of a fast product if the bin split is poor at the expense of making large quantities of slower products). In order to implement this, we must model the bin split to the fast product separately from the remaining products. Thus, in this case we represent all of the finished goods by two planning items with one for the fast processor and the other for all of the rest. The key idea here is to only model what is important for the decision being made. We call this the *simplicity principle*.

Step 1 is perhaps the most difficult of them all as different people have different opinions on what represents a critical economic driver for a product. The idea is that you start wafers in the fab and thus we need to understand what drives us to start wafers. For example, if we know that the wafer produces a very small amount of the fastest speed chips, but the demand for those is expected to be quite high and important, then it is necessary to model that trade-off (as in the bin split chasing case).

Many people want to place every bit of detail into a model; however, this does not help and in fact may hurt. The details make the models much larger and harder to solve. Also, when large amounts of uncertainty are in the representative data, vast quantities of it make understanding the output very difficult and also makes what-if analysis nearly impossible. Additionally, the extra details can suggest a level of precision that does not exist in reality.

Once the proper product modeling has been determined for the model (i.e. will an entire family be represented by one aggregate product, two products, etc.), the next step is to determine a CW inventory target (step 2). In this step, you are figuring out how much finished goods inventory you need (in CW) to buffer against demand and ATM variability. Once this CW inventory target is determined, you can determine the die requirements (step 3). In this step, you are calculating how many die it will take to adequately meet the inventory requirements determined in step 2.

In step 4, you are determining the inventory targets for ADI products. This is accomplished by building safety stock on top of the die requirements determined in step 3. This is done to buffer against FSM variability. Once the safety stock requirement for die is determined, you can finally solve for wafer starts (step 5).

Solving for ATM is a similar procedure to that described above for FSM. The key difference is that there will be less steps and the product modeling will probably be more detailed as ATM has a much shorter lead time and is the last major stage before the product goes to the customer. A summary of the steps for ATM is as follows:

1. Determine product modeling (aggregation) based on critical economic drivers
2. Determine CW inventory targets for these aggregated, finish good products
3. Solve for die allocation (based on the targets generated in step 2)

Again, we determine the appropriate product modeling based on critical economic trade-offs. This time the details are much more important as stated in the previous paragraph. Then, based on this model, inventory targets are created and an allocation scheme is determined.

We see that the ideas of HPP provide a very clear framework for solving the semiconductor planning problem. In the next chapter, we use these ideas to tackle three specific problems. The first is determining inventory targets, the second is how lead times are calculated, and the third is a method for FSM allocation.

5 Detailed Problems & Solutions

In this chapter, we begin by examining methods for setting inventory targets. Afterwards, we use these equations to develop a production allocation model that accurately captures the necessary binning and mapping complexities that exist within the semiconductor industry and at this particular company.

In an effort to make this thesis self-contained and complete, we will derive and explain some elementary results as experience has shown the author that many people in industry have a hard time fully grasping what the equations really mean.

5.1 Inventory Target Background

One purpose of a planning system is to keep the WIP close to a specified target. This target is set in order to maximize the chance that a customer will have their order filled when it is desired. At many companies, this inventory target represents the amount of inventory they desire to have in finished goods inventory (or CW in our context) and is set using a simple weeks-of-inventory (WOI) model. This WOI model has a very simple equation that generates the inventory target:

$$InvTar = WOI * \mu_D \text{ where } \mu_D \text{ is the mean weekly demand forecast}$$

Equation 2: WOI Equation

Note that when setting this target, one only needs to worry about two parameters. The first parameter is WOI or the required number of weeks of inventory desired to buffer against variability. The second parameter is the mean weekly demand forecast and is usually provided to supply chain managers by the marketing/forecasting divisions. Two obvious and important questions are

1) How do the values for WOI get set?

2) How accurate is the demand forecast?

In practice, WOI values are generally set according to management gut feeling. This gut feeling generally comes from years of experience in the industry and with the products.

The one and only certainty regarding demand forecasts is that they are uncertain. The so-called three fundamental principles of forecasting are

1) The forecast is always wrong

2) The longer the horizon, the worse the forecast

3) The more granular the forecast, the worse the forecast

We intend to propose a model that will better account for the variability in forecasting and give managers more than just intuition to set targets. The WOI model says inventory is a function of demand and an arbitrarily set value for WOI.

A logical next step using the standard results from inventory theory is to set inventory as a function of demand, service level and the variability that exists within the system. Thus, our goal is to establish appropriate equations for setting inventory levels as to account for the important uncertainties.

We begin this discussion by reviewing some of the basic results of inventory theory. In the development of this thesis, we assume the system works as follows. Assume there exists a finished goods inventory and it is replenished by starting wafers in the fab. There is a finite lead time and the system is managed using a continuous (Q,R) policy.

Let's assume that we have a single item with known (deterministic) lead time, L, and stationary demand with mean μ_D and variance σ_D^2 . For simplicity, we also assume we have a one period model that operates as follows. We begin with no initial inventory and then a production decision is made. Once this decision is made, demand and yield are realized for that period. Given the production quantity and the realized yield and demand, a new inventory level can be determined by multiplying the production quantity by the yield, adding this to the current inventory level, and finally subtracting the current demand. Then, the next production decision is made given the current inventory level. The inventory target represents the amount of finished goods inventory required in the system.

For our simple one period model (assuming we start with no inventory), in order to cover the *average* lead time demand, you would need to produce $\mu_D L$ units. Of course, we must actually produce more than this quantity; otherwise, we would miss orders approximately half of the time due to the variability under the normality assumption. We introduce safety stock into the system to cover situations when the demand is higher than average. We will express this safety stock using the expression $z\sqrt{L\,\sigma_D^2}$. Thus our base stock target, called B, is

$$B = \mu_D L + z\sqrt{L\,\sigma_D^2}$$

Equation 3: Demand Variability

If we assume that the demand is normally distributed, then this expression has real meaning. Suppose we want the probability of the lead time demand being larger than B (thus incurring a stockout situation) to be less than some α. We can write this as a probabilistic expression $\Pr(D_{LT} \leq B) = 1 - \alpha$ where D_{LT} is the actual demand over the lead time in a particular interval. The normality assumption allows us to set the value z in Equation 3 according to the expression $\Phi(z) = \alpha$ where $\Phi(\cdot)$ represents the cumulative distribution function for the normal distribution. For example, setting z=1.645 provides 95% probability that demand will be satisfied while z=3 provides 99.9% probability. We will refer to these coverage probabilities as the service level in this thesis. Thus, we will proceed using the idea of Type I service. For details about this derivation, service level, and multi period models, see Nahmias (2001).

Let us now assume that not only is demand a random variable, but lead time is also a random variable. We will denote the mean lead time as μ_{LT} and the variance of lead time as σ_{LT}^2. We will first state the result for the inventory target, but a detailed explanation will follow. Under the assumptions of random demand and lead time, the target is

$$B = \mu_D \mu_{LT} + z\sqrt{\mu_{LT}\sigma_D^2 + \mu_D^2\sigma_{LT}^2}$$

Equation 4: Demand and Lead Time Variability

The general form of this expression is that B equals average lead time demand plus safety factor times standard deviation. The first term represents the average demand over the expected lead time. The second term represents the required safety stock for a given service level (corresponding to a value for the safety factor z) and the standard deviation assuming stochastic demand and lead time.

We now look to explain the derivation of the mean and standard deviation in equation 4. The explanation follows closely one found in Bertsekas and Tsisiklis (2002). Let us assume that each unit of demand is an independent and identically distributed random variable with mean μ_D and variance σ_D^2 . We will denote the demand per time period as $D_1, D_2, \ldots, D_N$ where N corresponds to the lead time and it itself is an integer random variable. For the purposes of this derivation let

$Y = D_{LT} = D_1 + D_2 + \ldots + D_N = \sum_{i=1}^{N} D_i$. We first look to establish the first term. Let us condition on the random variable being equal to a specific n. Thus,

$$
\begin{aligned}
E[Y|N = n] &= E[D_1 + \ldots + D_N | N = n] \\
&= E[D_1 + \ldots + D_n | N = n] \\
&= E[D_1 + \ldots + D_n] \\
&= n\mu_D
\end{aligned}
$$

Since the above argument is valid for all positive integer n, we have

$E[Y|N] = N\mu_D$. Note that N is a random variable in this expression. Finally, the law of iterated expectations allows us to write

$E[Y] = E[E[Y|N]] = E[\mu_D N] = \mu_D E[N] = \mu_D \mu_{LT}$.

We now turn our attention to the safety stock expression. We will focus on the variance term as the safety factor is a simple multiplier that is not important to this derivation. Again conditioning on N we can write

$$
\begin{aligned}
\mathrm{var}(Y|N = n) &= \mathrm{var}(D_1 + \ldots + D_N | N = n) \\
&= \mathrm{var}(D_1 + \ldots + D_n) \\
&= n\sigma_D^2
\end{aligned}
$$

This time using the law of total variance as defined in Bertsekas and Tsisiklis (2002), we have

$$
\begin{aligned}
\mathrm{var}(Y) &= E[\mathrm{var}(Y|N)] + \mathrm{var}(E[Y|N]) \\
&= E[N\sigma_D^2] + \mathrm{var}(N\mu_D) \\
&= E[N]\sigma_D^2 + \mu_D^2\,\mathrm{var}(N) \\
&= \mu_{LT}\sigma_D^2 + \mu_D^2\sigma_{LT}^2
\end{aligned}
$$

Thus, we have established the results in Equation 4. Note the following assumptions are implicit in the equation:

- Lead times are integer values
- Lead time and demand are independent (for a given product)
- Lead time and demand have stationary distributions
- The demands are added together over the lead time (thus the mean lead time is not squared in the variance term as would happen if you assumed that we were multiplying by the lead time, this is very different)

The previous two expressions are well-known equations in inventory theory.

We now present an equation proposed in Levesque (2004) for handling demand, lead time, and yield variability. This equation is

$$B = \frac{\mu_D \mu_{LT}}{\mu_Y} + z\sqrt{\mu_{LT}\sigma_D^2 + \mu_D^2\sigma_{LT}^2 + \frac{\mu_D \mu_L}{\mu_Y}\sigma_Y^2}$$

Equation 5: Initial Demand, Lead Time, Yield Equation

Equation 5 above assumes that the yield variability is independent between all products during the lead time, thus allowing the variability to "cancel" itself out over the production run. This is evident in the last term involving the variance of yield; it assumes that each unit of production is independent in terms of yield (i.e. the data required is detailed yield estimates by individual lot). Stated another way, this equation assumes that the variability is seen by each item in the production run and thus some will have higher and lower yield values and the impact of the variability is diminished. However, this is not true in practice as the data is usually summarized by week and thus variability data represents the average variability over a week's worth of production (i.e. the yield values are not independent for each unit of demand). We intend to improve on the ideas in this expression, as it does not accurately capture the way yield information is reported and used by the planning community at the company of interest. We illustrate these difficulties in the following example.

5.1.1 Inventory Target Example

We now provide an example that motivates our work to develop new equations for setting inventory targets in this supply chain context. Assume the following set of data (from Levesque 2004) for demand (D), lead time (LT), and yield (Y):

$$\mu_D = 1{,}802{,}529 \qquad \sigma_D = 475{,}246$$
$$\mu_{LT} = 2.10 \qquad \sigma_{LT} = 1.58$$
$$\mu_Y = 0.993 \qquad \sigma_Y = 0.0258$$

The units for demand are the number of processors per week, the units for lead time are weeks, and the units for yield are percent. Given the above data set, we use the second term of equations 3, 4, and 5 to determine a safety stock target (using a z value of 1.645). The results of using the data set above with the three equations are given in the figure below:

Variability Type	***Safety Stock***	***Increase***
Demand (Eq 3)	1,005,497	
Demand & LeadTime (Eq 4)	4,277,920	3,272,423
Demand & LeadTime & Yield (Eq 5)	4,277,920	0.0006

Figure 14: Safety Stock Calculations

We see that demand variability causes the system to hold just over 1,000,000 units of safety stock. When lead time variability is added, the inventory requirement jumps by over 3 million units to approximately 4.27 million units. Finally, notice that when yield is added, the inventory requirement does not change by a measurable amount.

There are two things that jump out of these results. First is the fact that even though lead times are variable, the amount of safety stock required is much higher than many company experts would imagine. Content experts hold this claim because nearly all of these products are produced in a high volume manufacturing process, thus you really don't have to buffer against entire weeks worth of output not being realized.

The second issue that we recognize in these results is the fact that yields have no impact on the safety stock. This illustrates the issue with equation 5 given our assumption that the data is the aggregate yield for a product over an entire week. The variability for the yield over the week is moderately high, but equation 5 does not indicate any additional safety stock is needed. This is again because equation 5 assumes that each individual unit of demand is independent and thus the variability in effect cancels itself out. However, the data provided does not warrant such an assumption as the yield values represent the average seen by an entire week's worth of production (not individual units of production).

Given the fact that we do not have an appropriate expression for handling yield variability, we derive an alternate model in the following section.

5.2 Demand and Yield Variability

Given some of the difficulties discussed in the previous section, we will now propose a new set of equations to account for variability. We begin our analysis by investigating demand and yield variability together for one product in a one period model. We assume the system operates as follows. There is no initial inventory in the system. Next, a production decision is made. Then, the demand and yield are realized. The production quantity is multiplied by the yield amount to determine the supply. Finally, we subtract the demand from the supply to determine the new inventory level.

In the second period, you begin with an inventory level (where negative inventory represents backlogged demand that is not lost). Then, another production decision is made, followed by the realization of demand and yield. This time the supply is the inventory level plus the production quantity multiplied by the yield. Finally, we subtract the supply from the demand to determine the new inventory level. This process repeats itself over a specified horizon.

Let us begin with the figure below that shows the above system. Production is in terms of wafers while all of the other nodes are in terms of finished goods. We see that supply comes from both inventory and production. Also note that in this context, the yield value represents the yield for all wafers seen in that specific time period (independent of demand). Both yield and demand values are independently, identically distributed per time period.

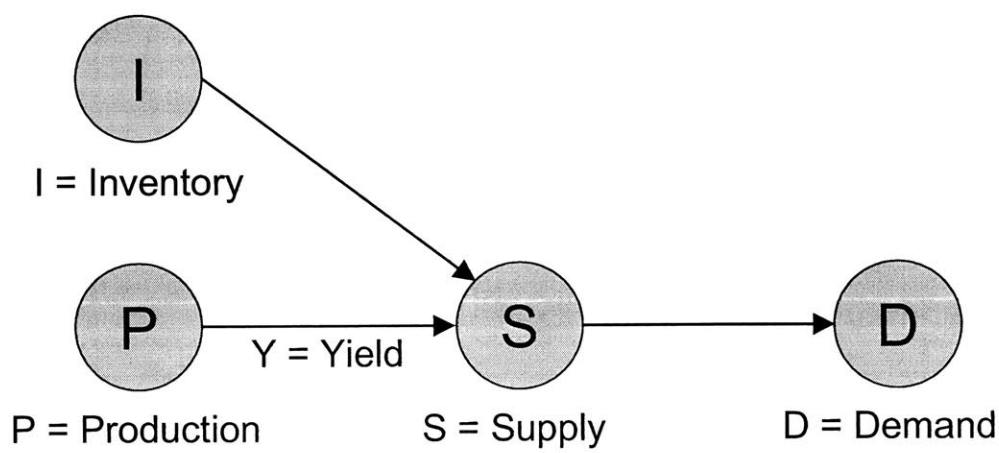


Figure 14: Supply and Demand Nodes

Let us assume that demand and yield are general random variables, we operate using the single period model described above, and that we have zero inventory in the system. Given these assumptions, we can define the required supply as the following expression

$$S = \frac{D}{Y}.$$

Clearly, if demand and yield were known, deterministic values, then the required supply to meet the demand given the yield would be determined using the equation above. For a concrete example, if demand were 100 units and yield were 50%, then the required supply would be 100/0.5 = 200 units.

We see that S is a ratio of two random variables. We will assume that Y > 0 to ensure that S is defined.

Next, we write expressions for the expectation and variance of S where σ_{DY} is the covariance between demand and yield.

$$E[S] = E\left[\frac{D}{Y}\right] \approx \frac{\mu_D}{\mu_Y}\left[1 + \left[\frac{\sigma_Y}{\mu_Y}\right]^2 - \frac{\sigma_{DY}}{\mu_D \mu_Y}\right]$$

$$Var(S) = Var\left(\frac{D}{Y}\right) \approx \left[\frac{\mu_D}{\mu_Y}\right]^2 \left[\left[\frac{\sigma_D}{\mu_D}\right]^2 + \left[\frac{\sigma_Y}{\mu_Y}\right]^2 - 2\frac{\sigma_{DY}}{\mu_D \mu_Y}\right]$$

Equation 6: Expected Value and Variance of Supply

The derivations for both the expectation and variance are based on first-order Taylor series expansions. Casella and Berger (1990) provide a clear explanation of the result.

If we assume that demand and yield are independent, then the equations simplify to

$$E[S] \approx \frac{\mu_D}{\mu_Y}\left[1+\left[\frac{\sigma_Y}{\mu_Y}\right]^2\right]$$

$$Var(S) \approx \left[\frac{\mu_D}{\mu_Y}\right]^2\left[\left[\frac{\sigma_D}{\mu_D}\right]^2+\left[\frac{\sigma_Y}{\mu_Y}\right]^2\right]$$

Equation 7: Expected Value and Variance of Supply Under Independence

We will later assume that demand and yield are normally distributed; however, the expressions above are general for any distribution provided that Y>0. There is an exact result for the ratio of two normally distributed random variables. Hinkley's 1969 paper, "On the Ratio of Two Correlated Normal Random Variables," discusses the derivation of the exact result. We make note of it here for completeness, but will use the approximate results above in this thesis.

We now turn our attention to using these equations to develop an appropriate safety stock equation. This analysis parallels the development of the traditional safety stock equations given earlier in this thesis. We assume normality for demand and yield.

Let $\mu_S \equiv E[S]$ and $\sigma_S^2 \equiv Var(S)$ where S is the usual supply value. Given these definitions, we can write the equation for supply requirements as $\mu_S + z\sqrt{\sigma_S^2}$. The safety stock expression is clearly $z\sqrt{\sigma_S^2} = z\sqrt{\left[\frac{\mu_D}{\mu_Y}\right]^2\left[\left[\frac{\sigma_D}{\mu_D}\right]^2+\left[\frac{\sigma_Y}{\mu_Y}\right]^2\right]}$.

We see that the above expression is in terms of supply. Thus, if we produce $\mu_S + z\sqrt{\sigma_S^2}$ units (assuming zero starting inventory and a one period model), we can be confident we'll meet demand according to the service level corresponding to the value used for z. To bring it in the more traditional terms of demand, we multiply by the mean yield quantity. Thus we can write the safety stock expression in terms of demand as

follows $z\mu_Y\sqrt{\sigma_S^2} = z\mu_Y\sqrt{\left[\frac{\mu_D}{\mu_Y}\right]^2\left[\left[\frac{\sigma_D}{\mu_D}\right]^2 + \left[\frac{\sigma_Y}{\mu_Y}\right]^2\right]}$.

Given that we now have an expression for safety stock (one in terms of demand and another in terms of the required supply to meet the demand), we turn our attention to using these expressions to manage the supply decision. What is meant here, is how much supply is required to buffer against the variability of demand and yield. Thus we can think of managing the system in two ways, either in terms of supply or demand.

5.2.1 Supply Management Policy

Let $SS_S \equiv z\sqrt{\sigma_S^2}$ be the safety stock requirement for the demand in terms of its supply and let $B_S \equiv \mu_S + z\sqrt{\sigma_S^2}$ be the inventory requirement.

Assuming we begin with no inventory, producing B_S in the first period (recall that we have assumed a single period model) will give us a percentage stock outs that corresponds to the service level. For example, a 5% stock out result should correspond to a 95% service level (z=1.645). This is illustrated in the simulation results that are presented shortly.

Clearly, if we have inventory, we need to reduce our starts by the amount of inventory (adjusted for yield). Thus we have the following expression for determining the required amount of production (or starts)

$$\text{Starts} = B_S - \frac{Inv}{\mu_Y}.$$

Equation 8: Supply Management Policy Starts

Thus, we have established the number of units we need to start each week in order to meet our inventory requirements in terms of B_S. While this is a valid derivation, many planning managers think in terms of reaching targets of finished goods, thus we extend these results to capture such a situation.

5.2.2 Demand Management Policy

Let $SS_D \equiv z\mu_Y\sqrt{\sigma_S^2}$ be the safety stock requirement for the demand in terms of demand and let $B_D \equiv \mu_S\mu_Y + z\mu_Y\sqrt{\sigma_S^2}$ be the inventory requirement for demand.

Assuming we begin with no inventory, producing B_D in the first period (again we have assumed a single period model) will give us a percentage stock outs that corresponds to the service level. For example, a 5% stock out result should correspond to a 95% service level (z=1.645). Again, this is illustrated in the simulation results that are presented shortly.

Thus, we have the following expression for determining the required amount of production (or starts)

$$Starts = \frac{\mu_D + SS_D - Inv}{\mu_Y}.$$

Equation 9: Demand Management Policy Starts

We see that if our inventory level drops below the target, then we must produce/order more then the mean amount. This adjustment is equal to the amount of product that you are currently under the target by (adjusted for yield). A similar argument holds if the inventory level is above the target.

5.2.3 Simulation Results

We perform a simulation study to verify that these proposed policies perform as expected. Microsoft Excel ™ was used for the analysis. The code used to perform all simulations for this thesis can be found in Appendix B. The data set consisted of

- mean and standard deviation for demand,
- mean and standard deviation for yield, and
- service level.

The model assumes zero initial supply and simulates a one quarter (13 week) cycle. Each value for yield and demand makes a draw from the appropriate distribution. Inventory is calculated as BOH + supply – demand where BOH stands for beginning on-hand inventory. Note that the supply is the production amount adjusted for yield and is in the same units as demand. Below is a sample run of the model.

Data Set & Inventory Values												
	Mean Demand	1000										
	StDev Demand	300										
	Mean Yield	0.9										
	StDev Yield	0.01										
	Starts Target @95%SL	1660.0										
	Supply Cycle Stock	1111.2										
	Demand Safety Stock	493.8										
Simulation Model												
Week	1	2	3	4	5	6	7	8	9	10	11	12
Produce	1660.00	810.99	1612.08	949.48	1669.56	630.57	371.95	1003.19	1175.06	1044.73	1765.58	1190.31
Yield	0.90	0.92	0.89	0.90	0.90	0.89	0.88	0.90	0.91	0.91	0.89	0.89
Supply	1494.60	743.70	1429.88	851.20	1501.36	561.35	328.30	899.23	1075.06	953.72	1570.50	1059.47
Demand	730.70	1464.68	833.54	1499.27	566.27	328.59	896.42	1053.90	957.76	1602.49	1052.76	1149.70
Inventory	763.91	42.92	639.26	(8.81)	926.28	1159.04	590.92	436.24	553.54	(95.23)	422.52	332.28

Figure 15: Simulation Sample

In all simulations in this thesis, the following output values are obtained. The mean and standard deviation for the production levels (denoted in the output by P), the mean and standard deviation of the inventory levels (denoted in the output by I) and the percentage of time the inventory was negative (indicating a stock out has occurred, denoted in the output by % Neg). Below we show the results of the simulation for this section.

Input	Output				
Service Level	Mean(P)	StDev(P)	Mean(I)	StDev(I)	% Neg
93%	1148.47	346.43	443.30	300.25	7.0%
95%	1153.27	352.58	494.14	300.55	5.0%
97%	1158.74	362.19	565.18	300.98	3.0%

Figure 16: Simulation Results

The results are exactly as expected. We see that if a 93% service level is entered into the system, the results show an expected 7% stock out probability. Similarly if a 95% service level is entered, a stock out occurs 5% of the time while if a 97% level is entered, a stock out occurs only 3% of the time. Also, as the service level is increased, the average production and inventory levels increase.

We ran all simulations for 100,000 iterations to ensure accuracy in the results. This was an unusually large number of iterations; however, we wanted to ensure the results were as accurate as possible. Additionally, confidence intervals are calculated for the mean production quantity and mean inventory level in all simulations using the Excel CONFIDENCE function. The results for the calculations on this data set are shown below where CI(P) represents the 95% confidence interval for the mean production level and CI(I) represents the 95% confidence interval for the mean inventory level. Thus for the 93% service level run, we can be 95% confident that the true mean production level is 1148.47 +/- 2.15 wafers and the true mean inventory level is 443.3 +/- 1.86. These are very tight confidence intervals (this is expected given the very large number of iterations performed in the simulation) and lend much credibility to the results.

Service Level	Mean(P)	StDev(P)	CI(P)	Mean(I)	StDev(I)	CI(I)
93%	1148.47	346.43	2.15	443.3	300.25	1.86
95%	1153.27	352.58	2.19	494.14	300.55	1.86
97%	1158.74	362.19	2.24	565.18	300.98	1.87

Figure 17: Simulation Confidence Intervals

5.3 Alternative Policies

We now look to reduce the variance in production levels through two possible smoothing procedures. These alternative policies are based on the idea of control limits that attempt to smooth production levels. For example, maybe it is acceptable to have a service level between 93% and 97% and to "correct" whenever we leave this range. The idea is that if we make fewer corrections to the plan, then the factory will see less thrash (or large swings in production requirements). Smoothing production is a key issue for plant managers and thus the planning community tries to accommodate. We use standard deviation of production starts as a proxy for measuring the thrash in the production facility. We hope to see the standard deviation of production starts decrease using these policies.

Specifically there are three correction policies that we will examine in this thesis. The first has already been thoroughly discussed and is to make an adjustment in every time period.

A second approach is called "Correct to the Target." Under this policy, upper and lower service limits are set. If the inventory drops below the lower limit or above the upper limit, the supply is adjusted to attempt to bring it back to the target level. This is shown in the figure below.

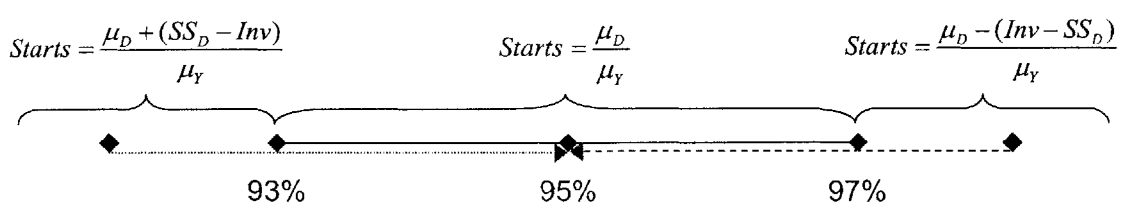


Figure 18: Correct to Target Policy

A third approach is called "Correct to the End Point." Under this policy, upper and lower service limits are again set. However, if the inventory drops below the lower limit, the supply is adjusted back to the lower limit (i.e. only to bring the system back inside the requirement). Similarly, if the inventory rises above the upper limit, the supply is adjusted back to the upper limit. Again, the figure below illustrates this policy pictorially.

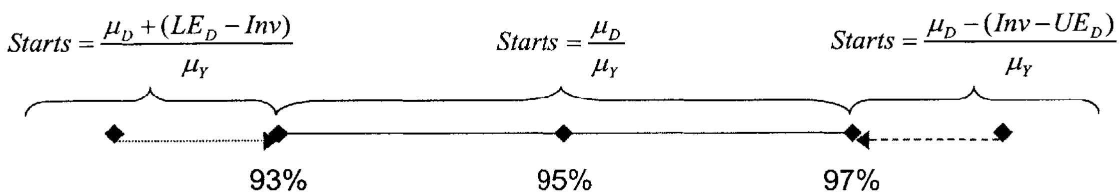


Figure 19: Correct to Endpoint Policy

The company was actively discussing the second policy while the author proposes they consider the third.

There are countless additional policies that could be examined. Another common one is "close α of the gap" where $0 \leq \alpha < 1$. We see that if $\alpha = 1$, then we see that this corresponds to the first policy examined where we fully correct in every time period. However, if $\alpha = 0.5$, we would only close 50% of the gap. We do not examine this policy in this thesis, but mention it for completeness.

5.3.1 Simulation Results

We performed a simulation study to determine how these proposed policies performed. Microsoft Excel ™ was used to perform the analysis. Again, details about all simulations performed for this thesis can be found in the appendix.

For this study, each policy used the same data set and random number draws to ensure consistency in the results. A sample of the model is shown below.

Data Set & Inventory Targets													
	Mean Demand		1000		93%	443.0							
	StDev Demand		300		95%	493.8							
	Mean Yield		0.9		97%	564.6							
	StDev Yield		0.01										
	Req Req @95%SL		1660.0										
	Cycle Stock		1111.2										
	Safety Stock		493.8										
Week	1	2	3	4	5	6	7	8	9	10	11	12	13
Randon Number Generator													
Rand-Y	0.050	0.180	0.607	0.079	0.911	0.566	0.254	0.433	0.860	0.535	0.218	0.190	0.014
Rand-D	0.879	0.132	0.332	0.173	0.463	0.894	0.081	0.147	0.073	0.425	0.989	0.484	0.425
Correct To Target													
Produce	1660.00	1531.32	754.32	963.99	812.28	1067.95	1525.15	655.61	761.98	617.91	1047.20	1878.23	1116.06
Yield	0.88	0.89	0.90	0.89	0.91	0.90	0.89	0.90	0.91	0.90	0.89	0.89	0.88
Supply	1466.64	1364.17	680.93	853.97	742.00	962.92	1362.55	588.94	694.02	556.66	934.32	1673.91	980.03
Demand	1351.04	664.87	869.63	717.43	972.10	1374.41	579.96	684.68	564.36	943.02	1682.25	987.95	943.09
Inventory	115.60	814.91	626.21	762.74	532.64	121.16	903.75	808.01	937.67	551.31	(196.62)	489.34	526.28
Outside Range - Correct To Target													
Produce	1660.00	1531.32	754.32	963.99	812.28	1111.11	1481.91	655.29	761.98	617.91	1111.11	1814.88	1111.11
Yield	0.88	0.89	0.90	0.89	0.91	0.90	0.89	0.90	0.91	0.90	0.89	0.89	0.88
Supply	1466.64	1364.17	680.93	853.97	742.00	1001.84	1323.92	588.65	694.02	556.66	991.34	1617.44	975.69
Demand	1351.04	664.87	869.63	717.43	972.10	1374.41	579.96	684.68	564.36	943.02	1682.25	987.95	943.09
Inventory	115.60	814.91	626.21	762.74	532.64	160.07	904.03	808.01	937.67	551.31	(139.60)	489.89	522.49
Outside Range - Correct To Endpoint													
Produce	1660.00	1474.93	888.84	963.58	812.27	1067.95	1390.06	789.71	762.23	617.91	1047.20	1743.14	1114.75
Yield	0.88	0.89	0.90	0.89	0.91	0.90	0.89	0.90	0.91	0.90	0.89	0.89	0.88
Supply	1466.64	1313.93	802.36	853.61	741.99	962.92	1241.87	709.40	694.25	556.66	934.32	1553.51	978.88
Demand	1351.04	664.87	869.63	717.43	972.10	1374.41	579.96	684.68	564.36	943.02	1682.25	987.95	943.09
Inventory	115.60	764.67	697.40	833.58	603.47	191.99	853.89	878.61	1008.50	622.14	(125.79)	439.77	475.56

Figure 20: Simulation Sample

Again, we ran all simulations for 100,000 iterations to insure accuracy in the results. Additionally, confidence intervals were calculated for the mean production quantity and mean inventory level for all simulations using the Excel CONFIDENCE function. The results for these tests again show very tight intervals (similar to the results found in section 5.2.3) lending credibility to the results. We refer the reader to Appendix B for the results of the confidence tests in this section.

The results for the data set with mean of demand = 1000, standard deviation of demand = 300, mean of yield = 90%, standard deviation of yield = 1% is shown below. Note that in the figure P = production starts and I = inventory level.

	Mean(P)	StDev(P)	Mean(I)	StDev(I)	% Neg
Replenish to Target (95% SL)	1152.38	352.62	494.60	300.39	4.9%
Replanish to Target if outside range (93%-97% range, 95% SL)	1152.53	352.56	496.11	300.71	4.9%
Replenish to Endpoint if outside range (93%-97% range, 95% SL)	1153.20	313.05	503.85	305.51	4.9%

Figure 21: Simulation Results - Data Set 1

We make the following important observations from the above results:

- Simulation results support the model validity as when a 95% service level is entered, the probability of a stockout is approximately 5%
- Correcting to the target provides no value (because the standard deviation of P is the same for using the range and correcting to the target and not using the range)
- Correcting to the endpoint decreases variability of production (because the standard deviation of P decreased from approximately 353 to 313)

We now present results where we modify the acceptable range. A total of four different ranges (all with target service level of 95%) were used in separate simulations and each has similar output characteristics.

	Mean(P)	StDev(P)	Mean(I)	StDev(I)	% Neg
Replenish to Target (95% SL)	1152.38	352.62	494.60	300.39	4.9%
Replanish to Target if outside range (93%-97% range, 95% SL)	1152.53	352.56	496.11	300.71	4.9%
Replenish to Endpoint if outside range (93%-97% range, 95% SL)	1153.20	313.05	503.85	305.51	4.9%
Replenish to Target (95% SL)	1154.35	351.45	492.68	299.49	5.0%
Replanish to Target if outside range (93%-99% range, 95% SL)	1156.36	350.48	514.34	303.70	4.4%
Replenish to Endpoint if outside range (93%-99% range, 95% SL)	1160.81	281.63	560.60	319.23	3.9%
Replenish to Target (95% SL)	1153.81	351.16	493.44	299.22	5.0%
Replanish to Target if outside range (91%-97% range, 95% SL)	1153.61	351.16	491.32	300.01	5.1%
Replenish to Endpoint if outside range (91%-97% range, 95% SL)	1152.96	301.63	484.33	307.95	5.8%
Replenish to Target (95% SL)	1153.12	351.56	493.60	299.51	5.0%
Replanish to Target if outside range (91%-99% range, 95% SL)	1154.82	350.70	511.66	304.36	4.6%
Replenish to Endpoint if outside range (91%-99% range, 95% SL)	1158.16	275.02	543.09	325.22	4.8%

Figure 22: Simulation Results - Data Set 1 - Four Different Ranges

We see in the results above that the "replenish to endpoint" policy reduced the variability of production while the "replenish to target" policy offered no improvement in all cases.

The results for another data set with mean of demand = 10000, standard deviation of demand = 3500, mean of yield = 95%, standard deviation of yield = 2% is shown below.

	Mean(P)	StDev(P)	Mean(I)	StDev(I)	Stockout %
Replenish to Target (95% SL)	10990.88	3893.11	5772.46	3500.248	4.9%
Replanish to Target if outside range (93%-97% range, 95% SL)	10992.46	3893.01	5789.53	3504.593	4.9%
Replenish to Endpoint if outside range (93%-97% range, 95% SL)	11000.10	3453.61	5877.98	3558.736	4.9%

Figure 23: Simulation Results - Data Set 2

Again we observe that correcting to the target provides no value while correcting to the endpoint decreases the variability of production.

5.4 *Lead Time Variability*

We now turn our attention to lead time variability. So far, our analysis has assumed a deterministic lead time (in fact, we have assumed it to be zero without loss of generality). We now wish to examine issues related to managing this system when lead time is a stochastic parameter. We examine this by proposing a new way to calculate lead time variability, which we call the cumulative flow method.

5.4.1 Cumulative Flow Method

The motivation behind the cumulative flow method is that the usual or traditional way of measuring lead time is not appropriate given this specific system. We define the traditional lead time method as one that subtracts the out date from the in date for each individual lot in the factory and then summarizes this data with a mean and variance.

During the semiconductor manufacturing process, there are several places where orders can "cross" or "jump" each other. This can occur in the semiconductor supply chain in several ways. For example, as products move through the manufacturing process they can easily be switched by moving one cart of wafers ahead of another when moving between machines. Another example is when products are put on hold for review. If a set of products is held for inspection, products that started afterwards will move ahead in the process. These types of review are quite common (given the tight specification required to product microprocessors) and the resulting impact on the lead time statistics can be quite important.

This behavior has no impact on what actually comes out of the fab; however, it can dramatically affect the summary statistics of lead time. Again, this is because lead time is traditionally calculated per unit in terms of time period out minus time period in. Thus if orders "cross," this calculation would be impacted.

In summary, we see that the lead times for individual lots can be quite variable, but some of this variability is due to order crossing. Since this order crossing does not affect the output of the factory, it is necessary to find a way to remove this when considering inputs to a stochastic inventory model.

We propose the following alternative method that removes the effects of order crossing from lead time calculations. The idea is to calculate cumulative ins and outs and base lead time on their difference. This removes the effects of order crossing and provides a much more accurate picture of the impact of lead time variability. The following example introduces the method.

Week	Input	TPT
1	10	2
2	10	3
3	10	1
4	10	2
	Average LT =	2
	StDev LT =	0.82

Week	Output
3	10
4	10
5	10
6	10

Figure 24: Traditional vs. Cumulative Flow Lead Time Calculation

We see in the above example that 10 units were started in weeks 1, 2, 3, and 4 with the factory throughput time (TPT) being 1, 2, or 3 each with probability 1/3. We see that the 10 units started in week 1 have a TPT of 2 weeks. Similarly, the 10 units started in week 2 have a TPT of 3 weeks while the week 3 starts have a TPT of 1 week and the week 4 starts have a TPT of 2 weeks. Thus, the traditional lead time method would say the average is 2 weeks with a standard deviation of 0.82 weeks.

The cumulative flow looks to see how the inputs and outputs compare. In this example, we see that the output is a steady 10 units despite the variability seen by each lot. Thus, using this method, we claim the lead time is 2 weeks with no variability.

We now show an example based on a realistic data set consisting of weekly data to further the intuition behind this idea. In the figure below, we have plotted the cumulative number of products started in a facility (per week) and the cumulative cumber of products exiting the facility (per week) against time. We note that the horizontal distance between the cumulative in and cumulative out curves represents the lead time. while the vertical distance represents the work-in-process (WIP).

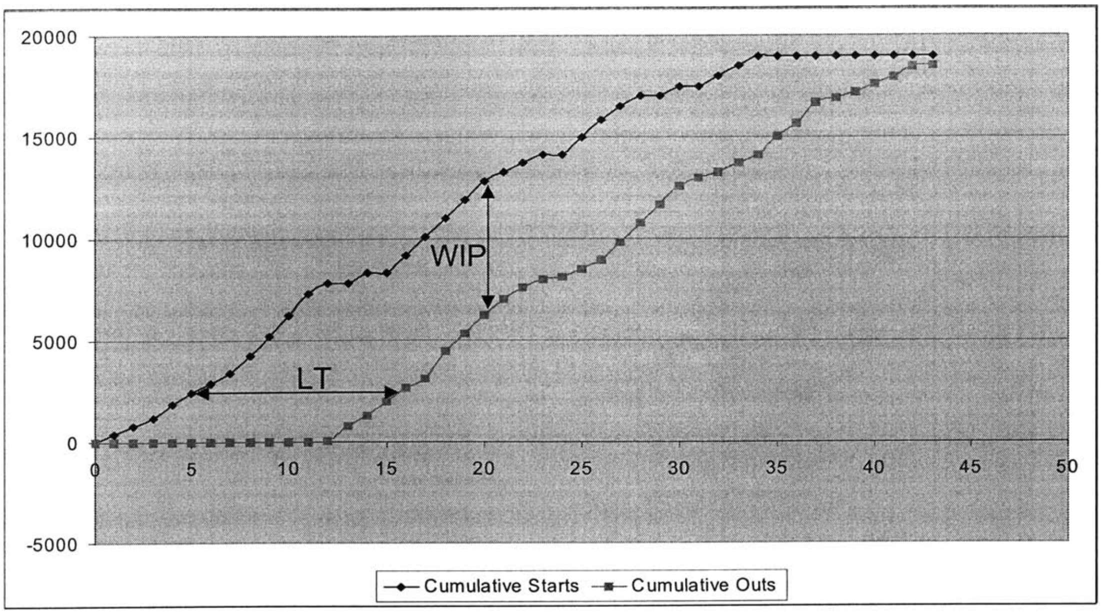


Figure 25: Cumulative Flow Lead Time Calculation Curves

If we sample the lead time at each data point on the cumulative starts line (i.e. at each large cumulative start dot), we obtain the set of lead times shown in the figure below.

Week	Starts	Outs	CumStarts	CumOuts	Obs	Lead Time
0	0	0	0	0		
1	400	0	400	0	12.46	11.46
2	400	0	800	0	12.99	10.99
3	375	0	1175	0	13.72	10.72
4	700	0	1875	0	14.82	10.82
5	550	0	2425	0	15.63	10.63
6	500	0	2925	0	16.58	10.58
7	500	0	3425	0	17.23	10.23
8	825	0	4250	0	17.83	9.83
9	1000	0	5250	0	18.87	9.87
10	1025	0	6275	0	20.06	10.06
11	1050	0	7325	0	21.51	10.51
12	500	51	7825	51.02041	22.58	10.58
13	0	757	7825	808.1633	22.58	9.58
14	500	509	8325	1317.347	24.54	10.54
15	0	679	8325	1995.918	24.54	9.54
16	900	682	9225	2677.551	26.32	10.32
17	900	430	10125	3107.143	27.32	10.32
18	900	1373	11025	4480.612	28.25	10.25
19	900	888	11925	5368.367	29.24	10.24
20	900	858	12825	6226.531	30.54	10.54
21	475	830	13300	7056.122	31.97	10.97
22	475	532	13775	7587.755	33.09	11.09
23	375	406	14150	7993.878	33.98	10.98
24	0	127	14150	8120.408		
25	850	381	15000	8501.02		
26	850	450	15850	8951.02		
27	675	859	16525	9810.204		
28	525	982	17050	10791.84		
29	0	915	17050	11707.14		
30	450	913	17500	12620.41		
31	0	378	17500	12997.96		
32	500	312	18000	13310.2		
33	500	429	18500	13738.78		
34	500	426	19000	14164.29		
35	0	912	19000	15076.53		
36	0	597	19000	15673.47		
37	0	1012	19000	16685.71		
38	0	253	19000	16938.78		
39	0	255	19000	17193.88		
40	0	426	19000	17619.39		
41	0	352	19000	17971.43		
42	0	479	19000	18450		
43	0	76.5	19000	18526.53		

Figure 26: Cumulative Flow Time Calculation - Sample Data

This time, the lead times were calculated using an interpolation method. As an example, let's calculate the lead time for the first week. We see that in week one, the cumulative starts were 400 wafers. Our next step is to determine when there was a cumulative exit of 400 wafers. We see this occurs between weeks 12 and 13. Thus we know it took between 11 and 12 weeks for the 400 wafers to exit the fab. We assume a linear relationship with time and thus calculate a weighted average to arrive at a calculated lead time of 11.46 weeks. For an exact calculation on this observation, the expression would be 11 + ((400-51.02)/(808.16-51.02)) ≈ 11.46. The following figure shows these calculated lead times plotted against the week for which they were calculated.

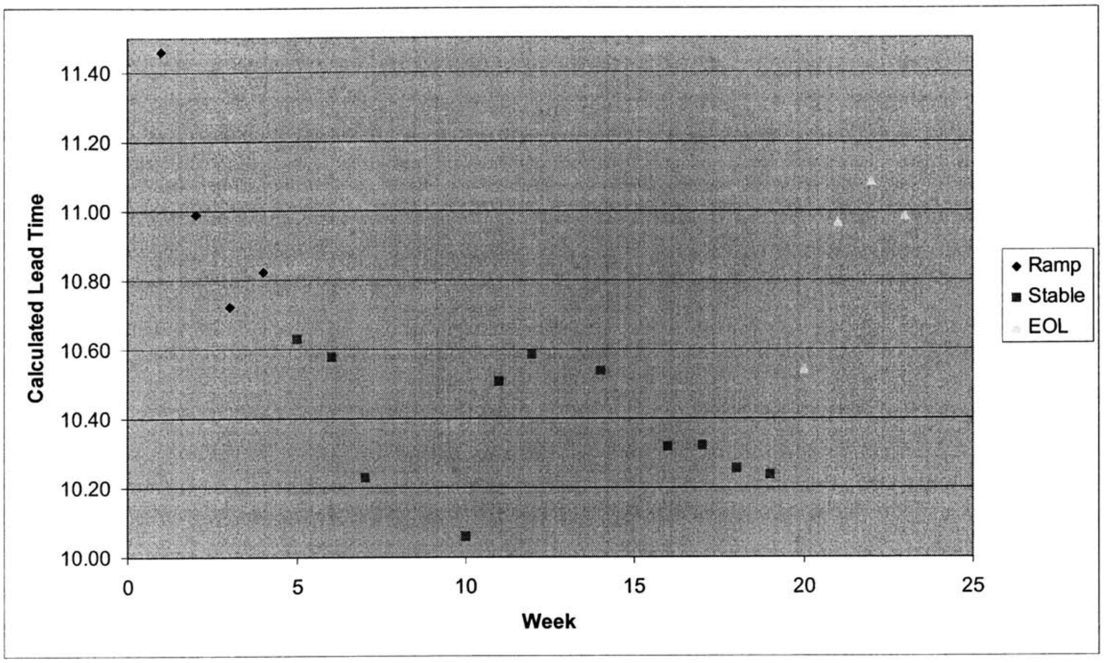


Figure 27: Cumulative Flow Time Calculation - Lifecycle Plot

In the plot, you can clearly see the effects of the product ramp and end of life. Given this method yields a more accurate picture of lead time variability than the traditional method of calculation, you could use the traditional lead time variability equation (equation 4 in this thesis) combined with equations for demand and yield derived in this thesis for a more accurate methodology for the combined problem. The above example came from actual data regarding fab inputs and output.

We now provide results where we compare calculating lead time the traditional way versus this new way. This analysis was based on daily, lot level data taken from one of the fabrication facilities. Any identifying marks have been removed to protect company confidentiality. Below are the mean and standard deviations for three different semiconductor products for both daily and weekly levels of aggregation.

	Daily Data		Weekly Data	
	Old Avg	Old StDev	Old Average	Old StDev
Product A	9.38	1.59	9.35	1.68
Product B	9.41	1.71	9.41	1.76
Product C	12.59	2.76	12.56	2.81

Figure 28: Traditional Lead Time

Note that we calculated these values by simply taking weekly data and subtracting the out week from the start week for each lot in the data set.

We will now use two different implementations of the cumulative flow method, both described above, on all three of these products. The first is a lot level computation method based on the trivial example in Figure 24 (called the sorting method). This will be used on both our daily and weekly data sets. The second implementation method is based on the example in Figure 25 (called the graphical method) and will be used only on the weekly data set for reasons to be discussed later. We will see that they provide nearly identical results.

We begin by describing the sorting method algorithm we used to implement the cumulative flow method on the daily lot data. The idea is quite simple. Assume that you have three columns in a spreadsheet where column A is the product name, column B is the start date, and column C is the end date. An example of this is shown below.

Product	In	Out
Product B	4	101
Product B	4	89
Product B	4	92
Product B	4	84
Product B	4	90
Product B	5	88
Product B	5	89
Product B	5	101
Product B	6	89

Figure 29: Sample Data - Cumulative Flow / Sorting Method

To remove the effects of order crossing, simply sort the numbers in ascending order in the "In" column and "Out" columns independently. Performing this sorting on the data in the figure above produces the result below.

Product	In	Out
Product B	4	84
Product B	4	88
Product B	4	89
Product B	4	89
Product B	4	89
Product B	5	90
Product B	5	92
Product B	5	101
Product B	6	101

Figure 30: Sorted Data - Cumulative Flow / Sorting Method

To finish the method, find the difference between the out and in date for each lot (or row). This represents the leadtime with the effects of order crossing removed. We summarize the impact on this small data set below.

	Old Method			New Method		
Product	In	Out	TPT	In	Out	TPT
Product B	4	101	97	4	84	80
Product B	4	89	85	4	88	84
Product B	4	92	88	4	89	85
Product B	4	84	80	4	89	85
Product B	4	90	86	4	89	85
Product B	5	88	83	5	90	85
Product B	5	89	84	5	92	87
Product B	5	101	96	5	101	96
Product B	6	89	83	6	101	95
		Average	86.89		Average	86.89
		StDev	5.88		StDev	5.23

Figure 31: Method Comparison

We see that the average leadtime did not change, but the standard deviation was reduced from 5.88 to 5.23. This result is expected as our new method simply reduces the impact when orders cross.

We now provide results when applying this method to the full data set for all three sample products. The data set has 1007 observations for Product A, 6390 observations for Product B, and 1807 observations for Product C. In the figure below, the results using the traditional leadtime calculation method is referred to as "Old" while the results using the sorting implementation is referred to as "New." Again, this analysis was done using daily, lot level data.

Product	Old Avg	Old StDev	New Avg	New StDev	StDev Reduction
Product A	9.38	1.59	9.38	0.80	50%
Product B	9.41	1.71	9.41	0.80	53%
Product C	12.59	2.76	12.59	1.52	45%

Figure 32: Daily Data – Sorting Implementation

In the results above, we see dramatic reductions in the standard deviations while the average lead time remains identical.

We now move on to the situation where we use data at a weekly level. Note that we are using the exact same source data; however, it has been aggregated to a weekly level before doing the analysis. To begin, we will use the sorting method implementation as described earlier. The results are shown in the figure below again where the results using the traditional leadtime calculation method is referred to as "Old" while the results using the sorting implementation is referred to as "New.".

Product	Old Avg	Old StDev	New Avg	New StDev	StDev Reduction
Product A	9.35	1.68	9.35	0.88	48%
Product B	9.41	1.76	9.41	0.92	48%
Product C	12.56	2.81	12.56	1.59	43%

Figure 33: Weekly Data - Sorting Implementation

We now show an alternative implementation method based on the graphical method given earlier in this section to calculate lead times. Again, we are still using the weekly data set.

Below is the data and leadtime calculation for a Product A aggregated into weekly buckets. A description of the data in each column follows the figure.

Name	Ins	Outs	Cuml_Ins	Cuml_Outs	TPT
Product A	5	0	5	0	10.38
Product A	21	0	26	0	10.76
Product A	8	0	34	0	10.12
Product A	33	0	67	0	10.16
Product A	5	0	72	0	9.42
Product A	5	0	77	0	8.68
Product A	29	0	106	0	9.18
Product A	34	0	140	0	9.52
Product A	16	0	156	0	9.55
Product A	16	0	172	0	9.28
Product A	22	0	194	0	8.85
Product A	48	13	242	13	8.74
Product A	33	17	275	30	9.17
Product A	20	34	295	64	9.12
Product A	21	19	316	83	8.62
Product A	10	18	326	101	7.86
Product A	15	28	341	129	7.69
Product A	35	21	376	150	8.55
Product A	24	11	400	161	8.27
Product A	47	39	447	200	8.85
Product A	26	57	473	257	9.17
Product A	31	15	504	272	9.16
Product A	24	18	528	290	8.70
Product A	36	42	564	332	9.00
Product A	27	13	591	345	10.03
Product A	30	15	621	360	9.78
Product A	40	29	661	389	9.58
Product A	29	41	690	430	9.30
Product A	28	20	718	450	9.52
Product A	26	18	744	468	9.83
Product A	26	29	770	497	10.28
Product A	14	44	784	541	9.57
Product A	30	23	814	564	9.19
Product A	32	5	846	569	8.79
Product A	32	21	878	590	8.72
Product A	22	40	900	630	10.00
Product A	33	53	933	683	10.95
Product A	23	23	956	706	10.50
Product A	37	23	993	729	10.83
Product A	14	18	1007	747	11.00
Product A	0	10	1007	757	
Product A	0	47	1007	804	
Product A	0	53	1007	857	
Product A	0	29	1007	886	
Product A	0	9	1007	895	
Product A	0	5	1007	900	
Product A	0	13	1007	913	
Product A	0	21	1007	934	
Product A	0	44	1007	978	
Product A	0	18	1007	996	
Product A	0	11	1007	1007	
Product A	0	0	1007	1007	
				Weighted Average	9.41
				Weighted StDev	0.75

Figure 34: Product A - Lead Time Calculation - Cumulative Flow / Graphical Method

The column "Ins" represents the number of wafer starts (in terms of lots) that occurred for a given week. Note that the data set is ordered; thus, values in the first row represent the first week of observation, values in the second row represent the second week, etc. The column "Outs" represents the number of wafer lots that came out of the fab in a given week. The "Cuml Ins" and "Cuml Outs" columns represent, respectively, the cumulative number of wafer starts in a given week and the cumulative number of lots that came out of the fab. The "TPT" column represents the calculated lead time using our graphical implementation with interpolation. Note that the average and standard deviation are weighted by the number of lots started in a particular week. Below is a graph of the cumulative ins and outs over time.

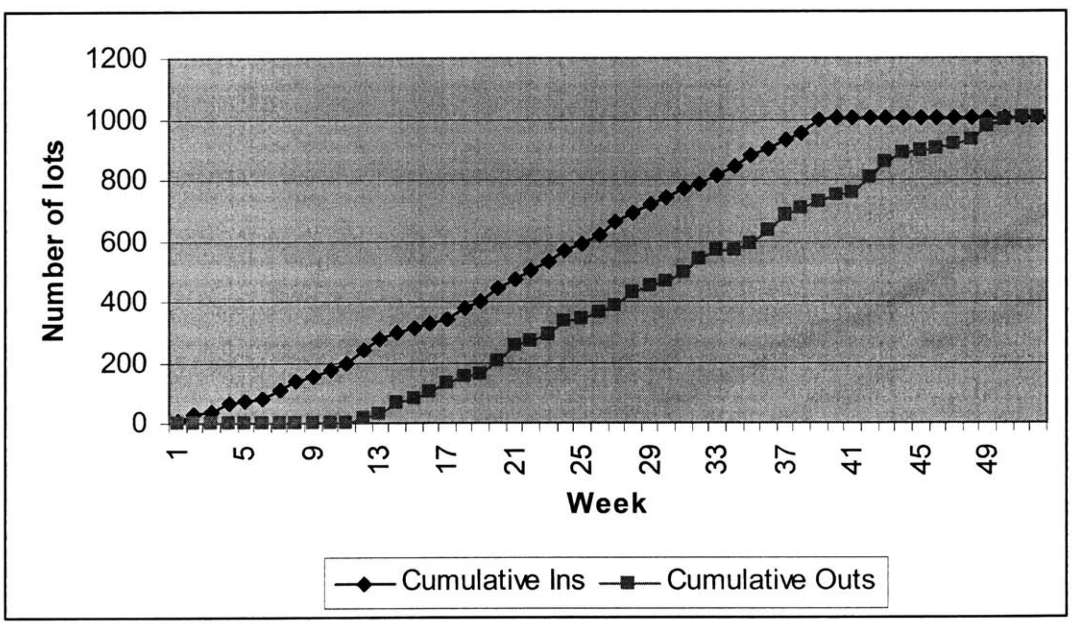


Figure 35: Product A - Cumulative Flow / Graphical Method Plot

This analysis is completed for two additional products and the results are presented on the following pages.

Name	Ins	Outs	Cuml_Ins	Cuml_Outs	TPT
Product B	83	0	83	0	11.49
Product B	58	0	141	0	11.32
Product B	55	0	196	0	11.14
Product B	66	0	262	0	11.10
Product B	55	0	317	0	12.13
Product B	34	1	351	1	11.75
Product B	18	0	369	1	11.12
Product B	36	1	405	2	11.11
Product B	78	0	483	2	11.28
Product B	77	0	560	2	11.01
Product B	88	6	648	8	10.63
Product B	78	39	726	47	10.28
Product B	74	73	800	120	10.19
Product B	91	66	891	186	10.52
Product B	90	73	981	259	10.59
Product B	88	31	1069	290	10.38
Product B	93	20	1162	310	10.05
Product B	126	55	1288	365	9.79
Product B	125	34	1413	399	9.52
Product B	167	55	1580	454	9.50
Product B	157	104	1737	558	9.50
Product B	177	144	1914	702	9.56
Product B	162	86	2076	788	9.35
Product B	156	64	2232	852	9.05
Product B	131	75	2363	927	8.65
Product B	165	91	2528	1018	8.46
Product B	157	136	2685	1154	8.33
Product B	166	170	2851	1324	8.56
Product B	153	170	3004	1494	8.66
Product B	197	171	3201	1665	9.23
Product B	178	144	3379	1809	9.36
Product B	210	187	3589	1996	9.42
Product B	177	226	3766	2222	9.33
Product B	237	217	4003	2439	9.45
Product B	222	193	4225	2632	9.44
Product B	167	160	4392	2792	9.34
Product B	158	105	4550	2897	9.30
Product B	115	163	4665	3060	8.84
Product B	177	109	4842	3169	8.67
Product B	192	139	5034	3308	8.68
Product B	215	200	5249	3508	8.82
Product B	203	192	5452	3700	8.81
Product B	211	200	5663	3900	9.03
Product B	241	230	5904	4130	9.28
Product B	200	214	6104	4344	8.98
Product B	141	141	6245	4485	8.83
Product B	80	214	6325	4699	8.45
Product B	61	213	6386	4912	7.97
Product B	4	179	6390	5091	
Product B	0	193	6390	5284	
Product B	0	208	6390	5492	
Product B	0	166	6390	5658	
Product B	0	167	6390	5825	
Product B	0	284	6390	6109	
Product B	0	163	6390	6272	
Product B	0	118	6390	6390	
				Weighted Average	9.39
				Weighted StDev	0.80

Figure 36: Product B - Lead Time Calculation - Cumulative Flow / Graphical Method

Name	Ins	Outs	Cuml_Ins	Cuml_Outs	TPT
Product C	48	0	48	0	12.82
Product C	22	0	70	0	12.33
Product C	32	0	102	0	11.96
Product C	43	0	145	0	11.84
Product C	51	0	196	0	12.90
Product C	57	0	253	0	13.26
Product C	40	0	293	0	12.85
Product C	35	0	328	0	12.64
Product C	45	0	373	0	12.65
Product C	61	0	434	0	12.64
Product C	52	3	486	3	12.32
Product C	57	1	543	4	12.00
Product C	44	21	587	25	11.88
Product C	63	28	650	53	12.82
Product C	73	51	723	104	13.57
Product C	73	49	796	153	14.07
Product C	94	34	890	187	14.82
Product C	93	10	983	197	15.29
Product C	82	38	1065	235	15.44
Product C	48	68	1113	303	15.03
Product C	27	39	1140	342	14.44
Product C	24	48	1164	390	13.80
Product C	17	69	1181	459	13.29
Product C	12	84	1193	543	13.08
Product C	7	50	1200	593	12.38
Product C	23	34	1223	627	12.10
Product C	27	28	1250	655	11.43
Product C	39	42	1289	697	10.90
Product C	46	46	1335	743	11.27
Product C	25	50	1360	793	11.88
Product C	15	41	1375	834	11.87
Product C	0	68	1375	902	10.87
Product C	3	62	1378	964	10.06
Product C	21	65	1399	1029	10.13
Product C	21	82	1420	1111	9.78
Product C	58	66	1478	1177	11.18
Product C	39	14	1517	1191	10.70
Product C	50	24	1567	1215	10.48
Product C	63	82	1630	1297	10.77
Product C	50	34	1680	1331	10.98
Product C	31	15	1711	1346	11.67
Product C	24	16	1735	1362	11.42
Product C	27	15	1762	1377	11.17
Product C	22	18	1784	1395	10.57
Product C	19	32	1803	1427	9.926
Product C	4	11	1807	1438	
Product C	0	26	1807	1464	
Product C	0	76	1807	1540	
Product C	0	56	1807	1596	
Product C	0	44	1807	1640	
Product C	0	41	1807	1681	
Product C	0	8	1807	1689	
Product C	0	33	1807	1722	
Product C	0	31	1807	1753	
Product C	0	54	1807	1807	
Product C	0	0	1807	1807	
				Weighted Average	12.61
				Weighted StDev	1.54

Figure 37: Product C - Lead Time Calculation - Cumulative Flow / Graphical Method

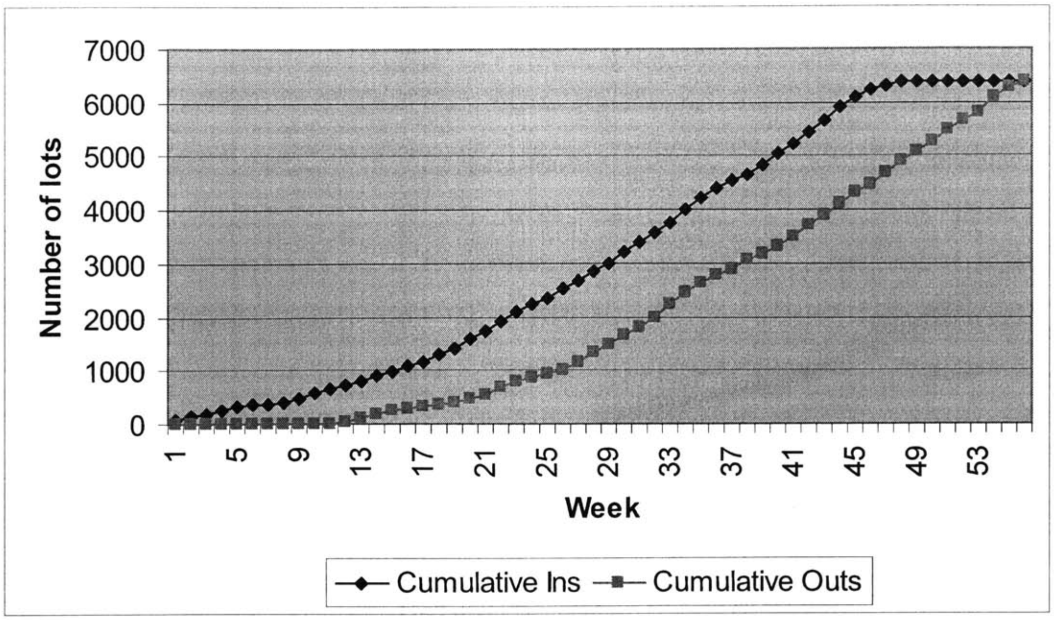


Figure 38: Product B - Cumulative Flow / Graphical Method Plot

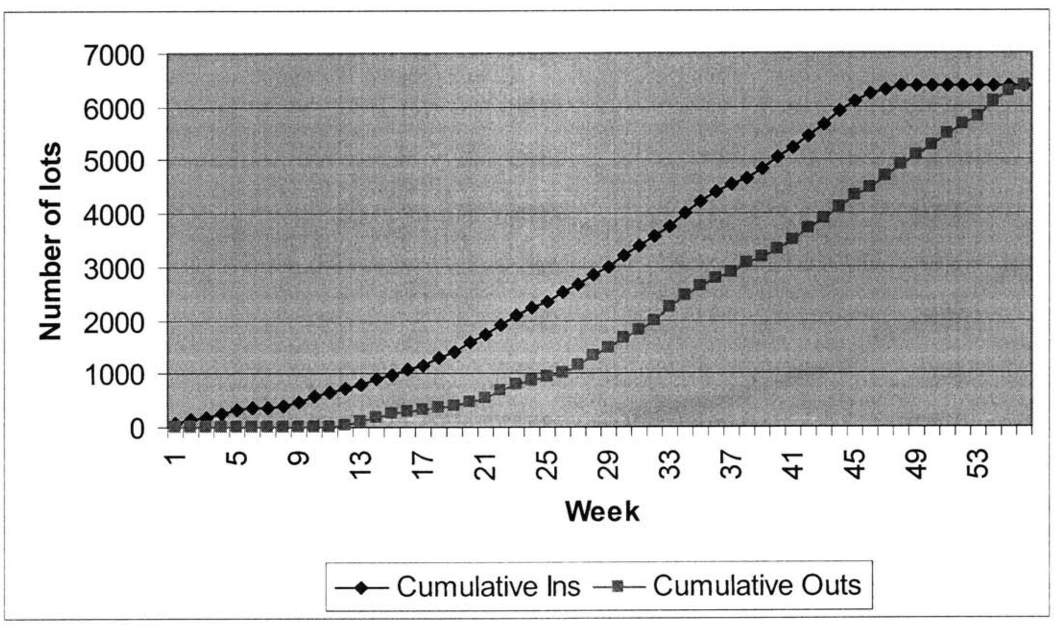


Figure 39: Product C - Cumulative Flow / Graphical Method Plot

We summarize and compare the mean and standard deviation for each product between the traditional method on weekly data (labeled "Old) and graphical method on weekly data (labeled "New").

Product	Old Avg	Old StDev	New Avg	New StDev	StDev Reduction
Product A	9.35	1.68	9.41	0.75	55%
Product B	9.41	1.76	9.39	0.80	55%
Product C	12.56	2.81	12.61	1.54	45%

Figure 40: Weekly Data – Traditional vs. Graphical Implementation

Again, we see that the average values don't change by a meaningful amount while the standard deviations are reduced significantly.

Finally, note that regardless of the data granularity and implementation method (sorting or graphical method), the results are nearly identical (see the summary below).

Data	Method	Product	Old Avg	Old StDev	New Avg	New StDev	StDev Reduction
Daily	Sorting	Product A	9.38	1.59	9.38	0.80	50%
		Product B	9.41	1.71	9.41	0.80	53%
		Product C	12.59	2.76	12.59	1.52	45%
Weekly	Graphical	Product A	9.35	1.68	9.41	0.75	55%
		Product B	9.41	1.76	9.39	0.80	55%
		Product C	12.56	2.81	12.61	1.54	45%
Weekly	Sorting	Product A	9.35	1.68	9.35	0.88	48%
		Product B	9.41	1.76	9.41	0.92	48%
		Product C	12.56	2.81	12.56	1.59	43%

Figure 41: Summary of Cumulative Flow Implementations

We point out that the graphical method should be used only when the graphs are strictly increasing over time. This is because when the function is nonincreasing, the evaluation becomes much more difficult and the interpolation method suggested above is no loner valid. Given this complication, the sorting implementation is preferred due to its wider range of applicability and ease of implementation.

This difference in lead time variability has a large impact on safety stock calculations given a stochastic lead time. Since the standard deviations have been cut down considerably, the safety stock required to operate at a given service level would be dramatically reduced.

5.5 Production Planning

Using the hierarchical production planning framework and ideas, we develop a procedure to plan production in this system. This procedure will be described by means on an example. We will assume that all of the work in determining proper modeling, grouping, etc. has been completed and we have the following representation.

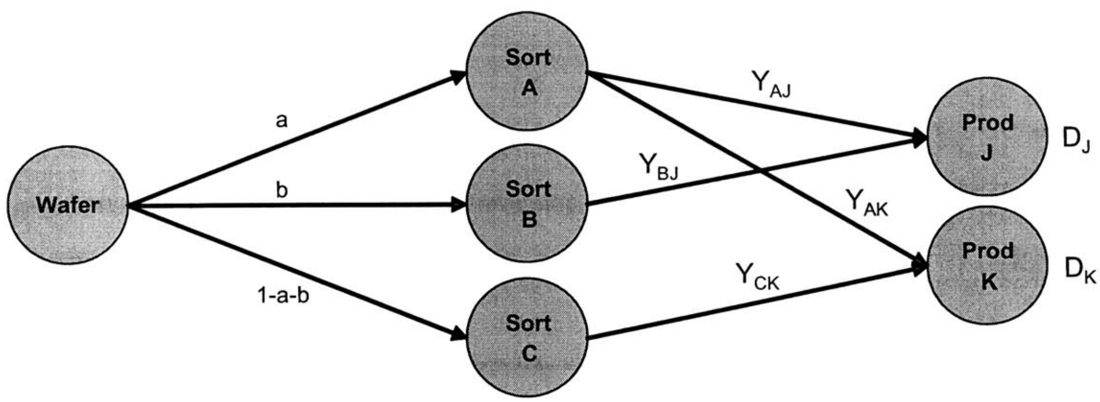


Figure 42: Product Mapping

We see that there is one wafer (called Wafer) that gives rise to three sort names (Sort A, B, C) which become one of two finished goods (Prod J, Prod K). Note that we have left out the level 3 complexities in this example. This was done to make the explanation easier; it could easily be incorporated. We will assume in this example that DLCP is deterministic while yield (denoted by Y_{AJ}, Y_{AK} , Y_{BJ} , Y_{CK}) and demand (denoted by D_J, D_K) are normal random variables with given mean and standard deviation. To make this example concrete, assume the following data set (where the units for DLCP and yield are percent and the units for demand are number of products):

DLCP (Deterministic)		Yield (Random Variable)		Demand (Random Variable)	
a=	0.2	$E[Y_{AJ}]$=	0.9	zJ=	1.645
b=	0.3	$E[Y_{AK}]$=	0.8	zK=	1.645
		$E[Y_{BJ}]$=	0.8	$E[D_J]$=	100
		$E[Y_{CK}]$=	0.95	$E[D_K]$=	150
		$StDev(Y_{AJ})$=	0.05	$StDev(D_J)$=	50
		$StDev(Y_{AK})$=	0.1	$StDev(D_K)$=	50
		$StDev(Y_{BJ})$=	0.1		
		$StDev(Y_{CK})$=	0.05		

The first column shows that the DLCP split to Sort A is 20%, to Sort B is 30%, and to Sort C is 50%. The expectation and standard deviation for yield are given in the second column while the third column has the service level (represented by the z value) for each finished product as well as the expectation and standard deviation for demand for each product. Assuming our usual one stage model, we can write the inventory target for product J using the equations derived earlier in this thesis as follows (the equation for product K is analogous):

$$\text{CW Tar}_J =$$

$$\frac{\mu_{D_J}}{\mu_{Y_J}}\left[1+\left[\frac{\sigma_{Y_J}}{\mu_{Y_J}}\right]^2\right]+z\mu_{Y_J}\sqrt{\left[\frac{\mu_{D_J}}{\mu_{Y_J}}\right]^2\left[\left[\frac{\sigma_{D_J}}{\mu_{D_J}}\right]^2+\left[\frac{\sigma_{Y_J}}{\mu_{Y_J}}\right]^2\right]}$$

where

$$\mu_{Y_J} = f_{AJ}\mu_{Y_{AJ}} + f_{BJ}\mu_{Y_{BJ}}$$

$$\sigma_{Y_J}^2 = f_{AJ}^{\,2}\sigma_{Y_{AJ}}^2 + f_{BJ}^{\,2}\sigma_{Y_{BJ}}^2$$

$$f_{AJ} = \frac{AJ}{AJ+BJ}$$

$$f_{BJ} = \frac{BJ}{AJ+BJ}$$

Note that AJ represents the number of Sort A products sent to satisfy Product J demand while BJ represents the number of Sort B products sent to satisfy Product J demand. Similarly, we define AK as the number of Sort A products sent to satisfy Product K demand and CK represents the number of Sort C products sent to satisfy Product K demand. Given the product mapping, we see the system must observe the following inventory and allocation constraints:

$$\begin{aligned}
&\text{Inventory Relationship} = \\
&\text{AJ} + \text{BJ} \geq \text{CW Tar}_J \\
&\text{AK} + \text{CK} \geq \text{CW Tar}_K
\end{aligned}$$

$$\begin{aligned}
&\text{Allocation Relationship} = \\
&\text{AJ} = \text{A} \cdot f_{AJ} \cdot Y_{AJ} \\
&\text{AK} = \text{A} \cdot f_{AK} \cdot Y_{AK} \\
&\text{BJ} = \text{B} \cdot Y_{BJ} \\
&\text{CK} = \text{C} \cdot Y_{CK}
\end{aligned}$$

The inventory relationship says that the amount of product J produced must be greater than or equal to the inventory target. An analogous expression is written for product K.

The allocation relationship says that the amount of sort name A that is allocated to either product J or product K equals the total available amount, multiplied by the weighted average of the yields. Sort names B and C are straightforward as there are no factional decisions being made.

This problem as formulated is a non-linear optimization problem. The fractional allocations represent the non-linearity. The objective function is to minimize the number of wafers subject to the inventory and allocation constraints. This example was solved using the MS Excel ™ solver and produced the following results.

Model				
Wafer	491.0769			
Sort A		98.21539		
Sort B		147.3231		
Sort C		245.5385		
AJ			88.39385	
AK			0	
BJ			117.8585	
CK			233.2615	
CwTarJ				183.7991
CwTarK				233.2615

meanYJ	0.842857
sdYJ	0.082375
meanYK	0.95
sdYK	0.05

206.2523	>=	183.7991	Inventory Constraint
233.2615	>=	233.2615	Inventory Constraint
88.39385	=	88.39385	Allocation Constraint
0	=	0	Allocation Constraint
117.8585	=	117.8585	Allocation Constraint
233.2615	=	233.2615	Allocation Constraint

DLCP (Deterministic)	
a=	0.2
b=	0.3

Yield (Random Variable)	
E[AJ]=	0.9
E[AK]=	0.8
E[BJ]=	0.8
E[CK]=	0.95
StDev(AJ)=	0.05
StDev(AK)=	0.1
StDev(BJ)=	0.1
StDev(CK)=	0.05

Demand (Random Variable)	
zJ=	1.645
zK=	1.645
E[J]=	100
E[K]=	150
StDev(J)=	50
StDev(K)=	50

Figure 43: Production Planning Model & Results

We see the results show that we need to start 491 wafers to generate enough die and finished goods to cover our inventory targets. We see the model used all of the Sort A die for Product J, this is logical for the following two reasons, (1) the yield for Sort A die on Product J is much higher then the yield when used for Product K and (2) the yield from Sort C die is much higher for Product K than the yield for Sort A die.

We see here the basic idea is to perform the following steps.

1. Determine appropriate product modeling and collect relevant data
2. Write finished goods (or CW) inventory equations
3. Determine the inventory relationships and allocation relationships
4. Solve non-linear model to determine ADI requirements and wafer start requirements

This method could provide value as it uses more accurate inventory equations and also captures the very important non-linear allocation relationship. Most models used today do not explicitly take this non-linear relationship into account. This method easily generalizes to more industrial sized problems provided a more sophisticated mathematical solver is available.

6 Conclusion

In this thesis, we have described the semiconductor supply chain, provided a framework for improvement, and given a detailed analysis for several specific problem areas.

This thesis developed out of a specific need at a company. The goal was to determine an effective way to allocate fab capacity to the worldwide factory network. This led to the creation of the allocation model described in section 5.5; however, the inventory equations were problematic in the development of this model and these issues led to the work of sections 5.2 through 5.4.

In this thesis we have (1) provided a new set of equations that captures demand and yield variability, (2) provided an analysis of two different production smoothing procedures and illustrated the one favored by the company provided no value, (3) proposed a new method to calculate lead time based on the cumulative flow of products through the factory, and (4) developed a non-linear model to determine wafer starts using the new inventory equations.

There are many areas that warrant further attention based on the work presented here, below are some key areas:

- *Managing the evolution of process and tools as the company's products and roadmaps evolve*

  The product lifecycle for companies continues to shrink and building very large, complex, and cumbersome planning systems makes it difficult to keep up with the company's ever changing products and offerings. There's a need to examine ways to build effective planning systems that evolve easily as products become more complex or as the organization changes directions and produces entirely new kinds of products.

- *Integrating analytics into everyday planning and enterprise systems*

  It is critical to find ways to build analytic tools into enterprise systems. MS Excel ™ is a wonderful example of an analytic tool that is used in businesses everyday. It is easy to use and intuitive, thus business analysts are experts in using it. Giving the supply chain analyst a similar tool, but specialized for their problems would greatly improve their ability to make good decisions.

- *Detailed investigation of the non-linear fab allocation model*

  The non-linear allocation model presented in this thesis is effective for many small problems, it remains to be seen how well it can scale to larger problems. Coding the model using an industrial strength solver (LOQO for example) would help determine how strong the method could scale as the business becomes more complex.

- *Investigation of the lead time methods and integration with the demand and yield model*

  We developed models where lead time was separated from the rest of the problem. It would certainly be worth exploring how to integrate all of the models together.

- *Methods when normality assumption is not valid*

The question of handling non-normal parameters is a difficult one. They theory breaks down quickly under this assumption, but many real-life situations follow this behavior.

7 References

Bertsekas, Dimitri P. and John N. Tsitsiklis. Introduction to Probability. Athena Scientific: Belmont, MA, 2002.

Billington, Corey, Gianpaolo Callioni, Barrett Crane, John Ruark, Julie Unruh Rapp, Trace White, Sean Willems. "Accelerating the Profitability of Hewlitt-Packard's Supply Chains," Interfaces, Vol. 34, No. 1, Jan-Feb 2004, pp. 59-72.

Bitran, Gabriel, Elizabeth Haas, Hirofumi Matsuo. "Production Planning of Style Goods with High Setup Costs and Forecast Revisions," Operations Research, Vol. 34, No. 2, Mar-Apr 1986, pp. 226-236.

Bitran, Gabriel and S. Dasu. "Ordering Policies in an Environment of Stochastic Yields and Substitutable Demands," Operations Research, Vol. 40, No. 5, Sep-Oct 1992, pp. 999-1017.

Black, Bryan E. LFM Master's Thesis: "Utilizing the Principles and Implications of the Base-stock Model to Improve Supply Chain Performance," MIT Sloan School of Management and MIT Department of Electrical Engineering and Computer Science, 1998.

Cakanyildirim, Metin and Roundy, Robin. "SeDRAM: Semiconductor Demand Forecast Accuracy Model," Technical Report, Cornell, Nov 1999.

Cakanyildirim, Metin and Roundy, Robin. "Evaluation of Capacity Planning Practices for the Semiconductor Industry," Technical Report, Cornell, Aug 2000.

Casella, George and Roger L. Berger. Statistical Inference. Belmont, CA: Duxbury Press, 1990.

Chopra, Sunil, Gilles Reinhardt and Maqbool Dada, "The Effect of Lead Time Uncertainty on Safety Stocks," Decision Sciences, Vol. 35, No. 1, Winter 2004.

Chow, Jim. LFM Master's Thesis: "Analysis of New Approaches to Improve Customer Responsiveness of Intel's Microprocessor Supply Chain," MIT Department of Civil and Environmental Engineering and Sloan School of Management 2004.

Clark, Andrew J. and Herbert Scarf, "Optimal Policies for a Multi-Echelon Inventory Problem" Management Science, Vol. 6, No. 4. (Jul 1960), pp. 475-490.

Cook, Miles and Russ Hagey. "Why companies flunk supply-chain 101: only 33 percent correctly measure supply-chain performance; few use the right incentives," Journal of Business Strategy, Vol. 24, No. 4 (2003), pp. 35-42.

Eilon, Samuel and Joseph Elmaleh. "Adaptive Limits in Inventory Control," Management Science, Vol. 16, No. 8, Application Series (Apr 1970), pp. B533-B548.

Eppen, Gary and R, Kipp Martin. "Determining Safety Stock in the Presence of Stochastic Lead Time and Demand," Management Science, Vol. 34, No. 11, Nov. 1988, pp. 1380-1390.

Ernst, Ricardo, Jose-Luis Guerrero, and Alan Roshwalb, "A Quality Control Approach for Monitoring Inventory Stock Levels," The Journal of the Operational Research Society, Vol. 44, No. 11 (Nov 1993), pp. 1115-1127.

ESD Symposium Committee, "ESD Symposium Committee Overview: Engineering Systems Research and Practice," Proceedings of the ESD Internal Symposium, May 2002, pp. 1-29.

Ettl, Markus, Gerald Feigin, Grace Lin and David Yao. "A Supply Network Model with Base-Stock Control and Service Requirements," IBM Research Report 1996.

Glasserman, Paul. "Allocating Production Capacity among Multiple Products," Operations Research, Vol. 44, No. 5, Sep-Oct 1996, pp. 724-734.

Graban, Mark. LFM Master's Thesis: "An Inventory Planning Methodology for a Semiconductor Manufacturer With Significant sources of Variability," MIT Department of Mechanical Engineering and Computer Science, MIT Sloan School of Management 1999.

Graves, Stephen. "A Tactical Planning Model for a Job Shop," Operations Research, Vol. 34, No. 4, Jul-Aug 1986, pp. 522-533.

Graves, Stephen, David Kletter, William Hetzel. "A Dynamic Model for Requirements Planning to Supply Chain Optimization," Operations Research, Vol. 46, Supp. No. 3, May-Jun 1998, pp. S35-S49.

Graves, Stephen and Sean Willems. "Optimizing Strategic Safety Stock Placement in Supply Chains," Manufacturing & Service Operations Management, Vol. 2, No. 1, Winter 2000, pp. 68-83.

Hausman, Warren and Rein Peterson. "Multiproduct Production Scheduling for Style Goods with Limited Capacity, Forecast Revisions and Terminal Delivery," Management Science, Vol. 18, No. 7, Theory Series, Mar 1972, pp. 370-383.

Hinkley, D. V. "On the Ratio of Two Correlated Normal Random Variables," Biometrika, Vol. 56, No. 3. (Dec., 1969), pp. 635-639.

Hsu, Arthur and Yehuda Bassok, "Random Yield and Random Demand in a Production System with Downward Substitution," Operations Research, Vol. 47, No. 2, Mar-Apr 1999, pp. 277-290.

Lee, Hau L. and Candace Yano. "Production Control in Multistage Systems with Variable Yield Losses," Operations Research, Vol. 36, No. 2, Mar-Apr 1988, pp. 269-278.

Levesque, Joseph. LFM Master's Thesis: "Analysis of Variability in the Semiconductor Supply Chain," MIT Department of Chemical Engineering and Sloan School of Management 2004.

Nahmias, Steven. Production and operations analysis. Boston, MA: McGraw-Hill, 2001.

Quirk, Michael and Julian Serda. Semiconductor Manufacturing Technology. Upper Saddle River, N.J.: Prentice Hall, 2001.

Veinott Jr., Arthur F. "Optimal Policy for a Multi-Product, Dynamic, Nonstationary Inventory Problem" Management Science, Vol. 12, No. 3, Series A, Sciences (Nov 1965), pp. 206-222.

Zipkin, Paul. Foundations of Inventory Management. Boston, MA: McGraw-Hill, 2000.

Appendix A – Literature Review

Managing business operations under uncertainty is a well-studied area with the field of operations research. We provide references to the work that most closely aligns with the ideas examined in this thesis.

Clark and Scarf (1960) and Veinott (1965) are among the first researchers to analyze inventory systems in a deep and meaningful way. Their work provides much of the basis for subsequent work in the field.

Inventory models dealing with different forms of uncertainty are well studied and several books have been written on the subject. Nahmias (2001) provides an introduction to these models while Zipkin (2000) provides a rigorous treatment.

There are several academic papers that provide the basis for the models in the books discussed above or extensions to the basic models. Several papers examining the impact of demand and yield uncertainty have been written. Bitran and Dasu (1992) discuss ordering policies when yields are random and demand is substitutable (as in the semiconductor industry). Hsu and Bassok (1999) further analyze this situation with several different solution methods. Lee and Yano (1988) discuss similar problems and provide results for an application to a light-emitting diode manufacturer. Ettl et al (1996) provide a supply network model with several realistic assumptions (e.g. non-stationary demands and stochastic lead times).

Eppen and Martin (1988) discuss possible flaws in the standard approximations for lead time uncertainty. Chopra et al (2004) extend this work and provide greater details. These works are of interest because they point out how important it is to pay attention to the assumptions you make when modeling.

Product allocation and scheduling models are well studied and are relevant to the issues examined in this thesis. Hausman and Peterson (1972) study scheduling under limited capacity and forecast revisions. Graves (1986) proposes a model for the operation of a job shop that allows for meaningful analysis for complex operations. Bitran et al (1986) discusses production planning for so-called style goods. These are products that have short selling seasons and stochastic demand. Microprocessors fall into this category due to the rapid pace of technological change. Glasserman (1996) looks at allocating production capacity among multiple items. Finally, we note the paper of Graves, Kletter, and Hetzel (1998) that examines requirements planning in multistage systems. Their Dynamic Requirements Planning (DRP) helps set inventory levels between stages in the supply chain.

Related to the product allocation papers are those of strategic inventory placement. Graves and Willems (2000) propose a dynamic programming algorithm to solve the problem of optimal safety stock placement in a supply chain. An application of this work was written by Billington et al (2004) and was a finalist in the INFORMS Edelman competition for the best application of operations research.

Specifically in the semiconductor industry, Cakanyildirim and Roundy's (1999) SeDRAM paper examines demand forecasting for the semiconductor industry in detail. They discuss methods to estimate variance and covariance of demand forecast errors and allows for correlations across time and products. In another paper, Cakanyildirim and Roundy (2000) discuss the evolution of capacity planning in the industry and make suggestions for improvements.

Several MIT LFM theses have examined issues in the semiconductor industry. Levesque (2004) provides a detailed analysis of variability for both supply and demand parameters. Chow (2004) discusses the idea of service level and setting inventory targets. Black (1998) proposes one method for dealing with yield variability in inventory planning while Graban (1999) extends this work to evaluate the impact of different sources of variability.

Introducing quality control ideas into inventory management is an idea discussed in this thesis as several companies have started looking into this as a way to reduce variability. Eilton and Elmaleh (1970) examine setting adaptive upper and lower inventory limits based on forecasting techniques. Two decades later, Ernst, Guerrero, and Roshwalb (1993) examine using quality control techniques to monitor inventory levels for accuracy.

Appendix B – Simulation Details

The code below was used to complete the simulations used in section 5.2.3 of this thesis.

```
Sub Simulation()
    Application.ScreenUpdating = True
    Range("B21:AA21").Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.ClearContents
    j = 0
    While (j < 100000)
        Worksheets("DYV-Const95").Calculate
        Dim prod(1 To 13) As Double
        Dim inv(1 To 13) As Double
        For i = 1 To 13
            prod(i) = Cells(13, 1 + i).Value
        Next i
        For i = 1 To 13
            inv(i) = Cells(17, 1 + i).Value
        Next i

        For i = 1 To 13
            Cells(j + 21, 1 + i).Value = prod(i)
        Next i
        For i = 1 To 13
            Cells(j + 21, 14 + i).Value = inv(i)
        Next i

        j = j + 1
    Wend
End Sub
```

The code below was used to complete the simulations used in section 5.3.1 of this thesis.

```
Sub ComboSimulation()
    Application.ScreenUpdating = True
    Range("B40:CC40").Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.ClearContents
    j = 0
    While (j < 100000)
        Worksheets("DYV-CombinedSheet95").Calculate
        Dim prod1(1 To 13) As Double
        Dim inv1(1 To 13) As Double
        Dim prod2(1 To 13) As Double
        Dim inv2(1 To 13) As Double
        Dim prod3(1 To 13) As Double
        Dim inv3(1 To 13) As Double
        For i = 1 To 13
            prod1(i) = Cells(18, 1 + i).Value
        Next i
        For i = 1 To 13
            inv1(i) = Cells(22, 1 + i).Value
        Next i
        For i = 1 To 13
            prod2(i) = Cells(25, 1 + i).Value
        Next i
        For i = 1 To 13
            inv2(i) = Cells(29, 1 + i).Value
        Next i
        For i = 1 To 13
            prod3(i) = Cells(32, 1 + i).Value
        Next i
        For i = 1 To 13
            inv3(i) = Cells(36, 1 + i).Value
        Next i

        For i = 1 To 13
            Cells(j + 40, 1 + i).Value = prod1(i)
        Next i
        For i = 1 To 13
            Cells(j + 40, 14 + i).Value = inv1(i)
        Next i
        For i = 1 To 13
            Cells(j + 40, 28 + i).Value = prod2(i)
        Next i
        For i = 1 To 13
            Cells(j + 40, 41 + i).Value = inv2(i)
```

```
        Next i
        For i = 1 To 13
            Cells(j + 40, 55 + i).Value = prod3(i)
        Next i
        For i = 1 To 13
            Cells(j + 40, 68 + i).Value = inv3(i)
        Next i

        j = j + 1
    Wend
    Application.ScreenUpdating = True
End Sub
```

The confidence intervals for the mean production level and mean inventory level for the section 5.3.1 simulation study are included below.

Mean(P)	StDev(P)	CI(P)	Mean(I)	StDev(I)	CI(I)
1152.38	352.62	2.19	494.60	300.39	1.86
1152.53	352.56	2.19	496.11	300.71	1.86
1153.20	313.05	1.94	503.85	305.51	1.89

Figure 44: Confidence intervals corresponding to Figure 21 results

Mean(P)	StDev(P)	CI(P)	Mean(I)	StDev(I)	CI(I)
1152.38	352.62	2.19	494.60	300.39	1.86
1152.53	352.56	2.19	496.11	300.71	1.86
1153.20	313.05	1.94	503.85	305.51	1.89
1154.35	351.45	2.18	492.68	299.49	1.86
1156.36	350.48	2.17	514.34	303.70	1.88
1160.81	281.63	1.75	560.60	319.23	1.98
1153.81	351.16	2.18	493.44	299.22	1.85
1153.61	351.16	2.18	491.32	300.01	1.86
1152.96	301.63	1.87	484.33	307.95	1.91
1153.12	351.56	2.18	493.60	299.51	1.86
1154.82	350.70	2.17	511.66	304.36	1.89
1158.16	275.02	1.70	543.09	325.22	2.02

Figure 45: Confidence intervals corresponding to Figure 22 results

Mean(P)	StDev(P)	CI(P)	Mean(I)	StDev(I)	CI(I)
10990.88	3893.11	24.13	5772.46	3500.25	21.69
10992.46	3893.01	24.13	5789.53	3504.59	21.72
11000.10	3453.61	21.41	5877.98	3558.74	22.06

Figure 46: Confidence intervals corresponding to Figure 23 results