Analysis of Variability in the Semiconductor Supply Chain

by

Joseph C. Levesque
Bachelor of Science in Chemical Engineering,
Massachusetts Institute of Technology (1995)

Submitted to the Department of Chemical Engineering and the Sloan School of Management in Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
and
Master of Science in Chemical Engineering

In Conjunction with the Leaders for Manufacturing Program at the Massachusetts Institute of Technology
June 2004

Signature of Author_____
Department of Chemical Engineering
Sloan School of Management
May 7, 2004

Certified by_____
Donald B. Rosenfield, Thesis Supervisor
Senior Lecturer, Sloan School of Management

Certified by_____
David Simchi-Levi, Thesis Supervisor
Professor of Civil and Environmental Engineering & Engineering Systems

Certified by_____
Gregory McRae, Thesis Reader
Professor of Chemical Engineering

Accepted by_____
Margaret Andrews, Executive Director of Masters Program
Sloan School of Management

Accepted by_____
Daniel Blankschtein, Chairman, Graduate Committee
Department of Chemical Engineering

Analysis of Variability in the Semiconductor Supply Chain
by
Joseph C. Levesque

Submitted to the Sloan School of Management and the Department of Chemical Engineering on May 7, 2004 in partial fulfillment of the Requirements for the Degrees of

Master of Business Administration and
Master of Science in Chemical Engineering

Abstract

While the pace of technical innovation in the semiconductor industry continues to accelerate, business processes and supply chain techniques have not kept up. Microprocessor performance improvement continues to follow Moore's Law, but increased variability has complicated efforts to accurately forecast demand and set inventory targets. Products are becoming more complex, often containing assemblies of multiple parts. Lifecycles are becoming shorter; made possible by technology breakthroughs and efficient manufacturing ramp-ups. Demand and supply are ever more stochastic and non-stationary. Inventory is one of the few ways that a firm can buffer themselves from the inherent and increasing variability, while still meeting required service levels.

We explore the sources of the variability in the semiconductor supply chain. On the supply side, we evaluate variability in throughput time, yield and other factors not explicitly considered in standard models. Here, we primarily focus on the natural stochasticity of the manufacturing process and disregard the variability arising from forecasting of these supply parameters. For demand, the natural stochastic process is not well understood, so we evaluate the forecast error and use it as a proxy for demand variability. We then apply these data to the base-stock model – constrained by its associated assumptions - to calculate inventory targets required to meet a certain level of service. Using a two-node base-stock model in conjunction with the actual variability data, we develop inventory estimates across the network and evaluate tradeoffs between different inventory strategies. We then determine what each variability parameter contributes to inventory. The combination of a simple yet representative model of the semiconductor supply chain with actual data from the variability characterization provides the tools to make powerful recommendations to reduce variability and decrease inventories throughout the supply network.

Thesis Supervisor: Donald Rosenfield
Senior Lecturer, Sloan School of Management

Thesis Supervisor: David Simchi-Levi
Professor of Civil and Environmental Engineering & Engineering Systems

This page is intentionally left blank.

Acknowledgements

No words can express my gratitude to my wife Kim, who has been my greatest supporter through the last two years. This work has taken us across the country and back again and having her with me through it all has made it a wonderful experience.

I would like to thank the Leaders for Manufacturing Program and Intel Corporation for providing me the opportunity to grow through working with some of the brightest and most capable people on the planet. It has been a pleasure to spend two years of my life learning from them.

This project could not have succeeded without the incredible support of Tony Newlin, Dennis Arnow and Karl Kempf. Tony's vision, unconditional support and ability to break down obstacles were invaluable. Dennis' coaching, mentorship and collaboration allowed me to grow intellectually and personally through the course of the project. Karl's technical leadership was motivating, challenging and always productive. Most of all, they each demonstrated their own unique brand of strong leadership. Without them, I would have not accomplished as much nor enjoyed myself as much as I did.

I am also grateful for the help of my thesis advisors, Don Rosenfield and David Simchi-Levi. Both provided valuable insight into the challenges of the project. I am most indebted to Don, who helped me work through numerous challenging issues and the result is much better for it. I would also like to thank my informal advisors, Steve Graves and Roy Welsch, who helped me to tackle important and difficult technical problems throughout the course of the project.

Finally, I would like to thank my parents, Madeleine Levesque and Ed Comeau for their love and support.

This page is intentionally left blank.

Table of Contents

Table of Figures

Table of Equations

1 INTRODUCTION AND OVERVIEW

The objective of this thesis is to present a framework for assessing the impact of variability in the semiconductor supply chain and implement it with actual data. We evaluate variability in the supply network related to supply and demand parameters and incorporate the data into a two-node, base-stock, inventory model which helps determine how much inventory should be kept to buffer from this variability. By identifying which variability sources are driving inventory, recommendations to reduce variability are made such that the company can reduce the inventory required to meet a particular service level. This research was conducted within Intel Corporation's eBusiness Group (eBG) from June through December 2003.

1.1 INDUSTRY AND COMPANY BACKGROUND

The semiconductor industry is unique among its high-technology peers in many ways. Intel, like may other high-tech firms, has extremely short lifecycles and large Research and Development budgets. However, unlike software companies, for example, semiconductor makers require large investments in manufacturing capacity. Semiconductor manufacturers spend billions of dollars on new factories every few years to stay on the top of the product development curve. Likewise, manufacturing of microprocessors is an exceedingly complex process, which requires not only superior technical skills, but also distinctive organizational capabilities such as strong leadership, collaboration and project management. This combination of cutting-edge technology, high capital costs, long manufacturing lead times, complex supply chains and short product lifecycles make for a challenging industry. These challenges provide the barriers to entry that make semiconductor manufacture such a lucrative business.

By understanding and mastering these complexities, Intel has been very successful since its founding in 1968. The company has grown to become the largest provider of computer microprocessors in the world, garnering over 80% market segment share[1] and over $25B in revenue in 2004[2]. However, slowing growth in the core microprocessor business (internally called Intel Architecture Group or IA) has prompted the company to diversify its product lines in recent years. Expansion into new markets has come partly from organic growth and partly from acquisitions. This has led to the formation of two additional business units within Intel. These businesses, the Intel Communications Group (ICG) and Wireless Communications and Computing Group (WCCG) are focused on high growth areas like wireless networking, flash memory, embedded communications devices and integrated processors for mobile devices. Recently, it was announced that these two groups are merging into one under the Intel Communications Group[3].

It should be noted that IA and ICG are dramatically different businesses. The x86 processor[4] is the most popular standard for desktop and mobile computing and competes with small players like AMD, Transmeta and Sun. On the other hand, ICG products are in much more competitive markets, with strong competitors like Qualcomm, Texas Instruments and Samsung. As a result, the demand forecasting and inventory management processes are quite different between the two businesses. In particular, the demand for ICG parts can be considered "perishable", since customers can readily get

---

[1] InfoWorld: http://www.infoworld.com/article/04/02/03/HNintelamd_1.html
[2] 2003 Intel Annual Report: http://www.intel.com/intel/annual03/
[3] Intel Press Release: http://www.intel.com/pressroom/archive/releases/20031210corp.htm
[4] The x86 processor is the general name for several generations of processors including the 386, 486 and Pentium®.

competitive products elsewhere in the marketplace. On the other hand, demand for Pentium® or Celeron® chips is less perishable, since switching suppliers often requires large investments of time and money. As a result, a customer is likely to wait for supply to become available, or take a substitute product with a similar specification rather than switching to a competitor. This market dynamic has dramatic implications for inventory management. Our analysis of supply chain variability focuses on the processor business, since this is where the majority of profit is made and where many of the supply chain challenges lie.

1.2 VARIABILITY IN THE SUPPLY NETWORK

By carrying inventory (or in some cases, extra capacity) manufacturing firms can insulate or buffer their performance from the effects of supply chain variability. The amount of inventory a company must hold to meet required or a desired customer service level (CSL) depends on the level of variability in supply and demand. If one could predict with 100% certainty[5] what future demand and supply will be, and there were sufficient flexibility in production rates, one would not need any inventory.

In many high-tech businesses, inventory is a relatively insignificant piece of the balance sheet. For example, Microsoft only holds $0.64B in inventories for a company with $32B in revenue[6]. Intel holds 3.5 times as much inventory to support slightly less revenue[7]. This is primarily because Microsoft has the benefit of quick manufacturing in response to new demand. Microsoft needs only to burn some more CDs if actual demand exceeds forecasted demand. This is in dramatic contrast to Intel, where manufacturing

---

[5] In this context, 100% certainty implies perfect forecasting and no variability.
[6] 2003 Microsoft Annual Report: http://www.microsoft.com/msft/ar.mspx
[7] 2003 Intel Annual Report: http://www.intel.com/intel/annual03/

lead times of several weeks require safety stocks to be held in order to service potential excess orders or shortages in production. In addition, the cost of a processor is far greater than the cost of a CD. Due to these factors, mistakes in supply chain planning can cost semiconductor companies billions of dollars in write-offs of obsolete inventory. Worse, in a constrained environment, production of the wrong product takes capacity from the right product, doubling the cost of mistakes.

An important corollary to the question of how much inventory to hold is where a firm should hold it. The microprocessor manufacturing operation includes two main parts, the Fab/Sort (F/S or Fab) process and the Assembly/Test (A/T) area. The first major inventory point lies between F/S and A/T and the second inventory point is between A/T and the customer. Material between F/S and A/T are held in what is called, Assembled Die Inventory (ADI) and finished goods are held in what is called Components Warehouse (CW). Our analysis uses a two-node model to evaluate the impact of variability on inventory levels. In particular, we evaluate throughput time (TPT) and yield variability of both F/S and A/T as well as demand variability to the end customer.

1.3 PROJECT APPROACH AND FINDINGS

The project was divided into two major components. First we quantified the variability from various sources in the current process. This was called the variability characterization part of the project. Then we used a two-node stochastic model of the semiconductor supply network to understand the impact of this variability on inventory. This was called the inventory analysis part of the project. The variability characterization provided a necessary foundation for the inventory analysis. Throughout both parts of the

project, we drew conclusions and developed recommendations to reduce variability and inventory in the network.

Quantification of variability can take on many forms, with diverse and often conflicting results. Details such as what metrics to use, which data sources to analyze and what level of granularity to evaluate require significant consideration. In addition, compromises must be made between what analysis is desired and what analysis is possible given the data available. We considered all available data sources, and developed a standard method of statistical analysis, which was consistently employed throughout the variability study. Another difficult aspect of the variability assessment is the question of what variability you want to evaluate. In particular, there are at least two types of variability that manifest themselves in a supply chain planning context. The first one is the natural stochastic nature of a given parameter. For example, the mean throughput time of the A/T process may be three days, but for a variety of reasons, some products emerge in two days and others are produced in four days. Expediting, machine breakdowns and re-testing are just some of the factors that cause TPT to behave stochastically. In addition to this "natural" source of variability relating the stochastic nature of sequential and re-entrant processing, there is the issue of forecasting, which is a separate supply chain variability problem. If we use forecasts of parameters like TPT to make production schedules and set inventory targets, then the difference between forecasted values and actual values of parameters like TPT introduce still more variability in the system. We refer to this source of variability as "forecast" variability. In general, the natural demand generation processes for most products are so poorly understood that we cannot evaluate the natural stochasticity of demand. The demand generation process

involves numerous macroeconomic factors and demand can change dramatically based on small changes in any one of these dependencies. As a result, we rely solely on forecast error as a measure of demand variability. On the other hand, supply variables change more slowly such that forecasts can be quite accurate. Therefore, we ignore forecast variability and we use natural variability to describe supply-side variability.

The major recommendation of the variability part of the project is to begin to measure the variability, reduce it where possible and manage the remainder. Specifically, we recommend monitoring the variability in the supply chain as you would a manufacturing process. This includes the collection of historical forecasts and actuals, implementation of the metrics and indicators described above, flagging of exceptions and subsequent investigation with the development of recommendations to prevent recurrence. Variability should then be reduced by using less detailed data where possible to simplify analyses. This can be accomplished by reducing the time horizon of planning and reducing the time or product granularity. Finally, we recommend managing the remainder by explicitly accounting for the variability inherent in the supply chain. Control limits or range forecasts should be developed. No changes should be made unless parameters are outside these limits.

In the second part of the project, we present a two-node supply chain model using supply and demand variability data[8] to calculate the cost of this variability, in terms of units of inventory. By identifying what factors are driving these costs, we make several recommendations for reducing variability in the system. The primary recommendation resulting from this work is that a paradigm shift – from judgment to data – is required.

---

[8] The supply and demand data shown in this thesis has been masked to protect confidentiality.

Specifically, Intel must change its inventory management strategy from "don't stock out" to "set a service level, and calculate an inventory target". Service level and variability should be analyzed - and inventory targets should be set - based on the attributes of the products. For example, stage of product lifecycle (ramp vs. end of life), relative volumes (high-volume vs. low-volume) and market attributes (desktop-value vs. mobile-performance) could be used to set targets since they are a major driver of variability. Finally, the company must quickly move toward global inventory optimization through data sharing, common tools and collaboration. This includes the development a multi-echelon stochastic optimization which should utilize a consistent and robust set of data sources. This will require better measurement of service levels and automated supply and demand data feeds.

We believe that if the recommendations described above are implemented, Intel will be well positioned to succeed in the coming era of intense global supply chain competition.

1.4 Overview of Thesis

In this chapter, we have discussed the supply chain challenges for Intel and the semiconductor industry as a whole. We have laid the foundation for our evaluation of the variability in the supply network and its impact on the corporation in terms of inventory necessary to meet required service levels.

In the next chapter, we review the background of the project as well as prior work in the field. Chapter 3 discusses our analysis of variability in supply and demand. Chapter 4 uses this characterization as a basis to present a two-node stochastic supply chain model

with which to analyze the impact of variability on inventory. In Chapter 5 we present conclusions and key insights for this thesis.

2 PROJECT DESCRIPTION AND LITERATURE REVIEW

In this section we discuss the background of the variability work that was done from June through December 2003 and the relevant literature that has formed the foundation for our research.

2.1 PROBLEM STATEMENT AND PROJECT DESCRIPTION

Like most forward-thinking manufacturing companies, Intel has dedicated significant resources toward improving supply chain efficiency. Many in the company believe that supply chain improvement is the next great challenge for the company to overcome. Like the implementation of "Theory of Constraints" and similar transformational ideas over the last 35 years, Intel's leaders expect supply chain optimization to help provide the stimulus for the next period of growth.

We strongly agree with the assessment that mastery of the supply chain is the next great frontier in Intel's corporate evolution. Intel has become successful by mastering technical product development in the 1980's and by mastering high-tech manufacturing in the 1990's. If successful in using the supply network to their advantage, we believe that Intel can experience another period of exponential growth. If not, then they may someday be overtaken by nimbler, more efficient, competitors. We believe that such expertise will not only provide the stimulus for the next period of growth, but may even provide the basis for another inflection point in the history of the company.

Intel has placed significant resources toward developing a set of capabilities and tools for achieving the goal of supply network excellence. In particular, the Edge-to-Edge (e2e) program was formed in 2000. Spanning the eBusiness Group (eBG) and Intel Supply Network Group (ISNG), the stated Edge-to-Edge Vision is to "Make Intel's

supply network planning capability a competitive advantage". The term Edge-to-Edge intimates the enormity of this task, with the transformation affecting everyone from one edge of the organization to the other. Together with cross-functional business unit teams, e2e has the responsibility to implement business process changes and new capabilities within the businesses. This is a challenging task, since it requires significant and simultaneous changes to data, processes and tools.

These challenges are significant in their own right, but the change agents who are tasked with making the supply chain a competitive advantage also have history to contend with. According to senior technical people in the company, such efforts have been tried numerous times in the past. In fact, one person counted seven different attempts at "supply chain management" or "reengineering" and noted that most of them failed because they were implemented under the guise of information technology projects, rather than business process changes requiring new capabilities, tools and training. Significant effort has been spent to avoid these problems in the current effort within e2e.

Within the last two years, several important initiatives have been launched within e2e. Some have been completed successfully, while others have struggled to meet their objectives. Managers within the organization often cite a lack of basic data, or failure to recognize significant resultant paradigm shifts as reasons for the lack of success within certain projects. When a project to implement multi-echelon inventory optimization was being readied for launch, management sponsored our work to identify, evaluate, quantify, reduce and manage the sources of variability which affect inventory strategy. Many of the observations and recommendations are applicable to other projects as well, and

significant collaboration with interested parties has occurred throughout the course of the effort.

2.2 CURRENT AND FUTURE SUPPLY NETWORK OPERATION

Intel has undergone significant transformation over the past few years. Some have been proactive changes designed to gain efficiency improvements. Others have come about as a result of the constant competitive pressures in the marketplace. One particularly striking example of a change that has come about as a result of the dynamic nature of the marketplace is the number and diversity of products that Intel provides. In decades past, Intel sold relatively few products. But today, Intel sells a dizzying array of products including desktop/mobile/server processors, 16/32/64 bit cpus, multi-chip products, chipsets, flash memory and wireless devices. Along with the increase in the number of products, the complexity of manufacturing and managing them has exploded as well.

In many ways, Intel's supply chain has not evolved with the escalating complexity of the business. One example of this is the fact that the majority of the tools that are used to manage the supply chain are spreadsheet-based. There have been improvements in specific areas, like the installation of an ERP system and the implementation of demand management software. However, these have been primarily localized improvements and have not been part of a holistic supply network information strategy. In addition, when these tools are not implemented in a cross-functional way, their usefulness to others in the organization is limited. For example, an ERP system implemented by finance is often virtually unusable to the supply chain organization because it has not been implemented to present data at the right level of granularity for

supply chain managers. Once architected, it often requires significant technical skills to acquire appropriate supply chain data from such a system. In the absence of sophisticated tools to manage the complex data, perform what-if scenarios and make decisions, a library of spreadsheets has been created to fill the void.

Nevertheless, the spreadsheets that are used to run the supply chain are quite robust. Many have embedded macros that get data from different data sources and perform complex transformations and calculations. Such analysis may be sophisticated, but it is often used in isolation. Though the results are shared across organizations in order to make planning decisions, the only person who understands the transformations that were done and the assumptions that underlie the calculations is the person who created it. And in some cases the spreadsheet has been handed down from the creator without documentation of these assumptions.

Such towers of "information isolation" create numerous problems in the implementation of advanced supply chain techniques. It has been shown that supply chain efficiency increases with additional information sharing. But sharing of results is not the same as sharing information. In particular, when a spreadsheet is sent from one person to another for review and action, if the person does not know what data was used in the analysis or what assumptions were made, they are likely to need clarification or question the results. One of the major goals of Edge-to-Edge is to reduce the amount of work that is required by creating and using spreadsheets in this way. The vision is that a group of supply chain experts will sit at the same table, with the same data and a set of numerical optimization tools with which to run scenarios, until all can agree to a supply chain strategy for a given planning horizon.

An initial step in this direction has been the implementation of linear optimization programs to optimize wafer starts over give time horizons. While the technical implementation of these tools has been outstanding, the business success has been questioned. Specifically, the data required to run the models takes hundreds of hours to collect and the optimality of the results has been debated. The reason, some believe, is that the data used to run the optimizations is highly variable. And since linear programs treat data as point values, this variability may be obscuring the true behavior of the independent variables. In other words, the optimizers may be optimizing noisy signals leading to non-optimal solutions. Worse yet, the optimizers could actually amplify the noise. The supply chain work completed from June to December 2003 has helped Intel understand this interaction between variability and important supply network variables.

2.3 LITERATURE REVIEW

In order to capitalize on the latest research and thinking in the fields of variability analysis, forecasting and inventory management, we based our work on several masters' theses from recent Leaders for Manufacturing and MIT Ph.D. graduates as well as industry experts.

Margeson [2003] develops a forecasting and inventory model for short lifecycle products with seasonal demand patterns. This paper was the precursor to our project. Margeson develops a new method of exponential smoothing designed to contend with the highly stochastic, non-stationary demand and lead times present in Intel's supply chain.

Graban [1999] presents a methodology for planning inventories under significant variability in his work at a semiconductor manufacturer. Extending prior work by Black

[1998], Graban develops a two-node stochastic supply chain model which allows for the evaluation of the impact of different sources of variability on the supply chain.

Black [1998] describes an extension to the base-stock inventory model in his work. He uses the principles and implications of the base-stock model and extends its reach by showing how yield variability can be incorporated.

Coughlin [1998] tackles the difficult questions of non-normality and stochasticity in his assessment of the inventory policies at an assembly facility. Gilpin [1995], Miller [1997], Hetzel [1993] also provide insights into variability analysis, forecasting and inventory management in their theses.

Willems [1998] provides numerical formulations of the stochastic optimization problem involved with the placement of safety stocks in a multi-echelon inventory context. Together with Graves, this work forms the foundation of one approach to multi-echelon inventory optimization. This approach formulates the stochastic nature of the problem as a deterministic optimization with the stochastic issues handled by appropriate safety stock levels.

The other vein of multi-echelon inventory optimization research has focused on the dynamic programming (DP) and simulation approach. Significant work in this field is described by Tayur [1999]. Although this approach is not utilized in our work, it is mentioned here for completeness.

In the area of forecasting, the foundational work is provided in Mullick, Chambers and Smith [1971]. More recently, Armstrong and Collopy [1992] and Armstrong and Fildes [1995] provide a comprehensive review of different forecasting methodologies in their papers from *The International Journal of Forecasting*. They

evaluate the performance of numerous forecasting error measurements in their relationship to decision-making, robustness to outliers, reliability, and sensitivity. In a subsequent article [2000], they provide an alternative error measure used in our work.

There is a rather large body of work concerned with forecasting of short lifecycle products in general and semiconductors in particular. Murty [2000] presents a simplified version using discretization techniques that overcome some of the difficulties associated with the DP approach.

To deal with the complexities of forecasting semiconductors, several combination methods have been developed. Cakanyildirim and Roundy [2002] present the SeDFAM demand forecast accuracy model which addresses autocorrelations and non-stationarity. Zhang, Et. Al. [2003] describes a combination technique which uses time series and marketing forecasts to develop demand signals.

In his brief article, A Process Control Approach to Forecast Measurement, Johnson [2003] brings numerous issues related to forecast measurement to the forefront. In particular, the response of people who are measured (often incorrectly) in the accuracy of their forecasts is evaluated.

Discussion of the psychological foundations of forecasting bias is presented in Maudlin's [2003] publication, "Thoughts from the Frontline".

In her analysis of current trends in manufacturing, Kilgore [2002] identifies several production issues that are causing problems in inventory management. These trends - including mass customization, outsourcing and shorter lifecycles - are challenging existing inventory tools. She discusses numerous inventory planning tools which purport to combat these challenges.

3 ANALYSIS OF VARIABILITY IN THE SUPPLY NETWORK

Business decisions depend greatly on whether we believe that the variation we see is indicative of a fundamental change in a system or whether it is "normal" variation. However, many aspects of business are not structured in a way that makes such variability differences explicit. For example, when a forecast is made for sales of a particular product in the 3rd Quarter, the value is usually presented as a single number. This type of report misleads us into thinking that there is no variability associated with the forecast. An equally troubling misuse of variability is to assign cause to natural random variations. For example, when new financial results are reported, pundits often attempt to link recent company actions with the increase or decrease, even though these fluctuations may just be the result of natural variability in the business climate.

Without a characterization of the background variability in a given process, there is no way to know whether the changes you observe are caused by special events (and therefore require action) or are simply natural variability in the process you are observing. The difference is important; because taking action on something that is exhibiting normal variability can amplify the noise and send the system out of control. This chapter describes such a variability characterization for Intel's supply network and the resulting recommendation: to measure the variability, reduce it where possible and manage the remainder. Specifically, we recommend monitoring the variability in the supply chain as you would a manufacturing process. Such monitoring should include the collection of historical forecasts and actuals, usage of metrics and indicators described above, flagging of exceptions and investigation with development of recommendations to prevent recurrence. Variability should then be reduced by using less detailed data where

possible to simplify analysis. This could be accomplished by reducing the time horizon of planning and reducing the time or product granularity. Finally, we recommend managing the remainder by explicitly accounting for the variability inherent in the supply chain. Control limits or range forecasts should be developed. No changes should be made unless parameters are outside these limits.

3.1 MEASURING VARIABILITY

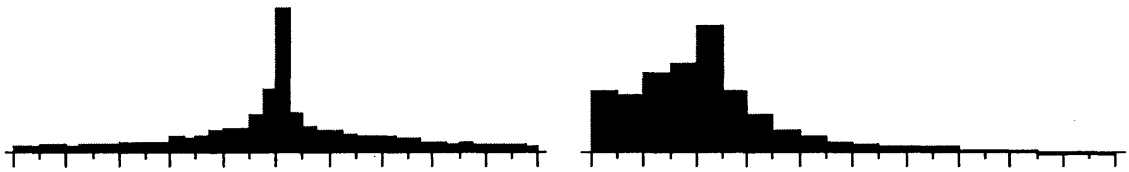
Traditionally, manufacturing organizations have completed variability analyses in order to understand and control process parameters which impact quality. The foundations of statistical process control lie with Deming [1986], Juran [1988] and Ishikawa [1985] but there has been significant recent work in applying these concepts to supply chain management.

There are several reasons that such a characterization of the semiconductor supply chain is needed. First, there is a desire for the company to use variability data to perform comparisons and benchmarking. Specifically, a firm may want to use information on variability of supply and demand to allocate capacity to certain products, to compare projects or simply to educate the organization. The second potential use for variability data is as an input into models, optimizations and simulations. Existing linear optimizations use precise data to develop specific recommendations concerning how many wafers to start without any sense of how noisy these signals are. A proper variability characterization provides the models with critical data concerning the quality of the inputs. In addition, future multi-echelon inventory optimization efforts at Intel will require data on the variability of yield, throughput time and demand.

3.1.1 Variability Data – Absolute vs. Relative Measurement

These two separate and distinct needs for variability data are both quite important, but require different types of variability data. For the purposes of comparison, relative variability (in percent) is desired, to account for differences in volume between products or businesses. On the other hand, numerical methods (optimizations and the like) require absolute (in units) variability, since these distributions are often closer to normal than relative distributions.

Figure 3-1. Absolute and Relative Variability Distributions



Representative Absolute Variability Distribution *Representative Relative Variability Distribution*

As one can see in Figure 3-1, the distribution of absolute variability is more normal, with a high, symmetrical peak. Through the tails don't die rapidly like other well-behaved distributions; it is clearly closer to a normal distribution than the highly skewed, relative variability distribution. This makes distributions of units more appropriate for numerical work even though relative measures are needed for comparison purposes.

3.1.2 Variability Types – Natural vs. Forecast

In the case of both absolute and relative data, a key question to answer in advance of performing a statistical variability analysis is which type of variability you wish to measure: natural variability or forecast variability. The distinction is subtle, but important. We define natural variability as the backward-looking, historical variability of

a parameter in a process. This type of variability is the variability of actual parameter results. For example, if the past three results for US GDP growth are 3%, 4% and 6%, then one measure of the natural variability is the range of those actual data points or 3%. Such a measure might be appropriate if the distribution of GDP results over time is deemed to be stationary, or unchanging. On the other hand, forecast variability is the variability that is introduced to the system by inaccurate forecasting. This type of variability is extremely important for non-stationary processes. A non-stationary process is one where the mean or variation is changing over time. In this case, the historical variability is not an accurate predictor of future variability. As a result, the primary indicator of process variability in such a process is the error in forecasts. Of course, this result depends on the assumption that the forecast is being used to make decisions within a business. If this is true, and the underlying process is non-stationary, then forecast variability is a significant contributor to overall system variability. A graphical description of natural and forecast variability is presented in Appendix A – Graphical Representation of Aggregation and Variability Type.

Many difficulties arise when one tries to quantify the contributions to variability from natural causes and from forecast error, since these terms occur naturally together. Adding to the difficulty is the fact that, within an organization, forecasts are often confused with targets, goals or some hybrid. This difference results from the level of confidence that the forecaster attributes to their number. In general, a target can be defined as a forecast which is attained over approximately half of the measurements. Or in statistical terms, a target is a forecast reached 50% of the time. A goal, on the other hand, is often considered to be more challenging than a target. In contrast to a target, a

goal may be set so that the value is only achieved 10% of the time. The assumptions underlying the forecasts are quite important and defined differently among different people, organizations and companies. As a result, one must be careful when using forecast error as the primary signal in a variability study.

In addition to the canonical differences between the two types of variability, there are implementation differences as well. The way an organization measures natural variability differs from the way that forecast variability is measured and monitored. The primary metric used to describe natural variability or variability of a stationary process typically is the standard deviation of the distribution of observations. While most models incorporate the assumption of normality, real data is rarely normal. In extreme cases of non-normality, standard deviation ceases to be an accurate depiction of variability. Needless to say, blindly entering mean and standard deviation values into models without checking the underlying distribution can lead to misleading results.

In the case of forecast variability, where historical variance is not an accurate predictor of future variance, we use differences between forecasts and actuals as a proxy for variability. In particular, forecast bias shows whether forecasts are, on average, higher or lower than actuals and is simply calculated as forecasted value minus actual value. Since a systematic bias should be easily corrected by making changes to the forecasting process, the error of forecasts can be more insightful than bias. The error of a given forecast point is the absolute value of the bias for that forecast point.[9] Lastly, the variability of the errors can be calculated as a measure of the spread of errors and hence the standard deviation of forecast errors. In both the case of natural and forecast

---

[9] While true for individual forecast points, this is not correct for summary statistics, like average error. This is due to the effects of averaging positive and negative errors.

variability, the relative variability can be calculated by dividing variability by the mean error. Since the variability of a parameter usually increases in proportion to its mean values, the coefficient of variation (CV) is a way to normalize the variability to compare across products. However, in most supply chain problems, the variability relative to mean demand is more insightful than variability to mean error. By scaling the variability to mean demand (typically actual demand, not forecasted demand) we can use these values to model inventory of products with different volumes. We call this relative variability the pseudo-CV.

Another difference between natural variability and forecast variability is in the data available and the nature of the distributions of the data. In the case of natural variability, there is a rather robust understanding of the underlying stochastic processes that can cause variability. For example, throughput time variability is caused by random events like machine breakdowns, re-testing due to quality failures and human error. There is non-random human intervention, in the form of expedited lots and opportunistic preventative maintenance, but the contribution of these to overall variation, we judge to be small. The judgment is supported by the fact that distributions of TPT are more normally distributed than other parameters studied.

In the case of forecast variability, there are weaker arguments describing the underlying stochastic process. In forecasting demand, for example, there has been copious research dedicated to a causal model of demand and the resulting variability of demand. However, most models fail to accurately predict the future because their model describing the stochastic nature of demand is not robust to new information. In addition, demand forecasts are subject to much more manipulation than the throughput time of a

process. Employees are diligently doing their best to meet the forecasts, and thus there is a high peak around zero error.

In addition, forecasters suffer from well known self-deception biases which have been characterized by psychologists. These include over-optimism, over-confidence and conservatism. Over-optimism manifests itself in the fact that there will typically be a higher number of over-forecasts (F-A>>0) than under-forecasts (F-A<<0) and they will be of greater magnitude. Over-confidence means that people are surprised more often than they expect to be. According to Maudlin [2003], when you ask people to make a forecast with 98% confidence, the correctness of their predictions is only about 60-70%! So if a forecaster is making a judgment with 50% confidence, we might expect far lower accuracy. Lastly conservatism bias leads people to cling to their forecasts until indisputable proof is presented to the contrary. This leads to large errors propagating longer than they might otherwise. Statistically speaking, this means that the right tail of the Forecast Error distribution will be longer and larger than the left tail, which we empirically see in the data.

3.1.3 Variability Metrics – MPE vs. APE

Since there are a significant number of different variability measurements, selection of metrics which minimize potential undesirable impacts is an important foundation to any variability characterization. There are several metrics discussed in the literature which could be used to measure forecast bias, forecast error and variability. Some result in misleading summary statistics. For example, Mean Percent Error is a common way to measure relative forecast bias.

Equation 1. Mean Percent Error

$$\text{MPE} = \frac{Forecasted Value - Actual Value}{Actual Value}$$

However, since our data has high numbers of positive outliers and since the MPE measurement is bounded by -100 and positive infinity (see Appendix B – Bounded Nature of MPE and APE), the average MPE over hundreds of observations can be skewed high. Though technically accurate, such high average error results often provide a barrier to discussion and an unfair representation of a forecaster's ability or track record.

A measurement which mitigates the effects of these outliers is desirable. Collopy and Armstrong [2000] discuss an alternative to MPE, called Average Percent Error (APE), in which the forecast bias is divided by the average of forecasts and actuals.

Equation 2. Average Percent Error

$$\text{APE} = \frac{Forecasted Value - Actual Value}{Avg(Forecast, Actual)}$$

This has the effect of minimizing the impact of outliers and bounds the results from -200 to 200. Though its usefulness in optimization is quite limited, it is an effective way to compare variability across businesses or products. We use both an error (absolute value) and a bias (signed) version of this APE measurement in our work, since the standard error measurement, Mean Absolute Percent Error (MAPE), suffers from the same skewness problem as MPE.

As discussed previously, forecast variability statistics suffer from significant data issues, like large positive outliers. It is tempting to simply exclude these outliers in order to create a well-behaved, truncated distribution. Such cropping not only underestimates the actual variability, but also excludes the points which often contain the most important

information. For example, a data point which represents forecast error of 1,000 percent probably presents opportunities for learning and continuous improvement. On the other hand, these points may represent extraordinary demand or extraordinary supply. From an inventory modeling perspective, such extraordinary demands can be handled quite differently within models. In early work in modeling safety stocks, Simpson [1958] describes the concept of maximum reasonable demand. He assumes that safety stock should be adequate to cover a certain maximum demand level. Above this level, he implicitly assumes, the firm would perform the extraordinary actions (like expediting) required to meet the higher demand. Other models, like those of Hanssmann [1959] do not require such an assumption. It is important to understand the interaction between this assumption of maximum reasonable demand and the large positive outliers we see in the demand forecast error distribution. Large differences in conclusions can result from using raw demand data versus applying the Simpson assumption to demand data.

The conservation equation for inventory is shown in Equation 3 below.

Equation 3. Conservation Equation for Inventory

$$BeginningOnHand + Supply - Demand = EndingOnHand$$

Assuming that the actual inventories themselves have zero variation[10], the safety stock required to meet demand with some specified service level is determined by the variability in supply and the variability in demand. Assuming that inventories can be determined accurately, supply and demand parameters account for all of the variability in

[10] As an aside, inventories themselves do not necessarily have zero variation. Often, policies like adjusting the inventory number in a computer system based on the physical count can actually amplify noise in the supply chain.

the supply chain. The specific parameters within these two groups are described in detail below.

3.2 VARIABILITY IN SUPPLY PARAMETERS

3.2.1 Sources of Variability in Supply

In most manufacturing supply networks, the primary sources of variation on the supply side of the conservation equation are throughput time and yield. However, if we specifically consider the semiconductor supply chain, we can postulate the existence of several secondary sources. For example, there is variability in how products get separated into speed bins after testing. A more obscure source of variability is the mapping between manufactured items and salable products. In the semiconductor manufacturing process, different wafer types made through different processes can produce the same product. A finished product is the result of a set of probabilistic events throughout the manufacturing process so there is a many-to-many relationship between wafers started and products produced. This complex relationship requires numerous mappings between production names and finished product names. These mappings are done manually and are prone to error; creating yet another source of variability in the process. However, like the speed bin parameter, it is difficult to determine what the relationship between the mapping variances and inventory might be. As a result, these sources of variability are not considered in our model.

On the other hand, the relationship of throughput time or yield variability to inventory is well known. The higher the variability of throughput time or yield, the more safety stock must be held to meet a given service level. This is because with higher

variability, there is less certainty that you will produce enough product - on time - to meet demand. Therefore, additional inventory is needed to prevent stock-outs.

3.2.2 Supply Variability – Data, Types and Metrics

As discussed before, the primary source of variability for supply parameters is natural variability. As opposed to demand parameters, forecast variability of supply parameters is not considered in our treatment because, although forecasts of these parameters are generated as part of the business process, the forecasts are usually quite accurate. This is a result of their relatively stationary behavior and well-behaved distributions[11] of TPT and yield data. Whereas a particular sku[12] sold to a customer may only exist for a few months before becoming obsolete, a wafer type[13] may exist for months or years. This gives engineers the ability to forecast supply-side values with greater accuracy.

The issue of data aggregation becomes important in the analysis of both supply and demand variability. In the case of supply parameters, data could be analyzed at the lot level or it could be aggregated to the weekly level[14]. For that matter, the data could be aggregated by family instead of lot and yearly instead of weekly. The key decision is which level of aggregation is most appropriate for the requirements of the supply chain work. The variability of TPT measured in weekly level of aggregation will be much higher than the level of variability as measured by yearly aggregation. Since variability

---

[11] In this case well-behaved implies a distribution with a high peak and rapidly dying tails.

[12] For the purposes of this analysis, a sku is defined as the lowest level in the product hierarchy. Several different skus make up a product family. It is primarily a demand-side term in that a customer orders a particular sku.

[13] A wafer type is a supply-side product description used to describe products in the factory. A single wafer type may produce several different skus and perhaps even skus within different product families. The rough equivalent to a sku in the supply side product hierarchy is called a die.

[14] The weekly level might consist of an average of all lots produced in a particular week.

drives inventory, dramatically different results and recommendations will result. The relationship between levels of aggregation is described by the sum of variances. Under the assumption of independent and identically distributed (i.i.d.) and normal data, the variance[15] of monthly data should be 2 times[16] that of weekly data. So, in theory, we could easily convert between different levels of aggregation. In practice, however, the data is auto-correlated[17] and non-normal which makes such conversions difficult. Since converted results for different levels of aggregation can be quite different, it becomes important to identify which level is appropriate to the decision that must be made, and analyze the raw data at that level of aggregation. The importance of aggregation is that it signifies a level of substitutability. If a particular sku cannot be substituted for another sku in the eyes of the customer, then the right level of aggregation should be the sku level. On the other hand, if one wafer can be substituted for another wafer in the Fab, then wafer level of aggregation may be appropriate. For supply parameters, we view the relevant aggregation as wafer or lot level for F/S and die level in A/T. A graphical representation of the concept of aggregation is shown in Appendix A – Graphical Representation of Aggregation and Variability Type.

3.2.3 Throughput Time Example of Supply Variability

To demonstrate the salient aspects of supply variability, we use TPT as an example. The TPT of a given lot is calculated by taking the time that the last processing step was completed and subtracting the time that the first processing step started. By

---

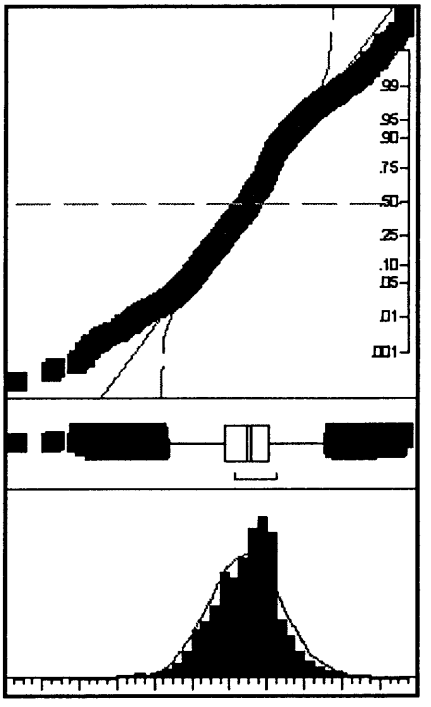
[15] This is true for absolute variance (measured in units, die, wafers, etc), not relative variance (measured in percent). The special properties of i.i.d variances do not apply to percentages.

[16] Two is derived from the square root of 4 weeks.

[17] Auto-correlation implies that variability in one week is linked to variability in another week and is more common for products with non-perishable demand. For example, if you overforecast the demand for this week, the forecast error for next week will tend to be overforecasted as well.

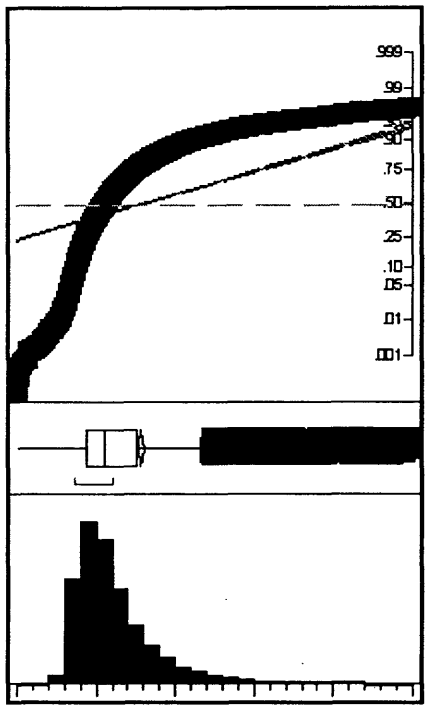
taking all of these throughout times for given lots, we plot the distribution of throughput times for the Fab section of the process. As shown below, it is well behaved and roughly normally distributed.

Figure 3-2. Distribution of F/S Throughput Time



However, A/T throughput time looks quite different. Rather than a normal curve, the distribution looks exponential in nature. There is a high concentration of data around the mean, but extremely long right tail.

Figure 3-3. Distribution of A/T Throughput Time



The long tail results from the advanced testing and diagnostics that are done in A/T such that certain units spend much more time to pass through. A/T is quite flexible and high priority lots can be expedited. Thus, the long tail may be a result of non-critical lots sitting idle while so-called "hot-lots" are processed first.

Our contention is that lots that sit idle due to expediting decisions should be excluded from the data set used for determining variability for supply chain planning purposes. In other words, the relevant A/T TPT should not include outliers that result from human intervention. In the words of Deming [1986], these are "special causes" and should be investigated, but should not be included in the assessment of common cause variability. However, the data to isolate such events is not readily available to supply chain professionals and even if it was, if would be too detailed. Such difficulties make it hard to differentiate the variability that safety stock is meant to buffer against from the

"extra" variability that should be ignored from a supply chain perspective. In some cases, distributions can be truncated to better describe the relevant level of variability and we note where this is done.

The F/S factories are usually not in the same locations as the A/T facilities, so shipping time must be considered in the measurement of TPT. Since transit time data was not available during this period, an assumption of 1 week transit[18] was made. In addition, we assumed that this time was deterministic, with no variability associated with it. The throughput time of shipping was added to A/T lead time to get overall A/T lead time for use in calculating finished goods warehouse (CW) safety stock. A/T mean lead time added to Fab TPT gives the expected overall lead time from raw materials to finished goods.

3.3 VARIABILITY IN DEMAND PARAMETERS

Given the lack of a fundamental understanding of the stochastic nature of demand generation, variability in demand parameters is much more difficult to ascertain than supply parameters. At Intel, several demand signals are used to provide information on which products are required over different time horizons. Though significant work is ongoing to rationalize these inputs into a single demand signal, such a "single-voice" forecast is currently unavailable.

3.3.1 Demand Variability Data, Types and Metrics

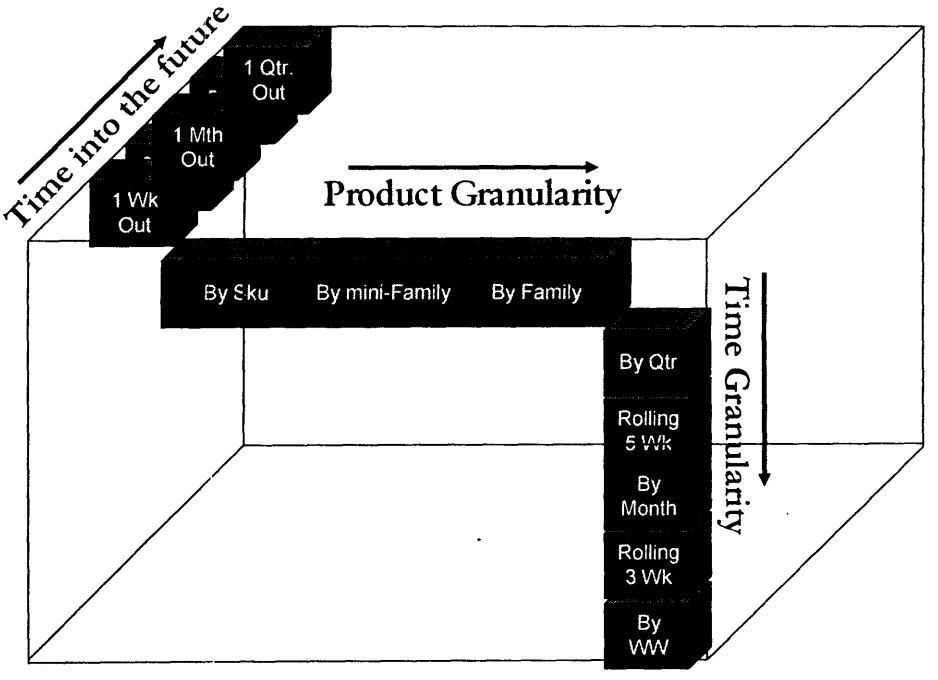
The calculations of demand variability are more complex than supply variability for several reasons. First, since we are dealing with forecast variability, rather than

---

[18] The transit time includes packing, shipping via air freight, customs, and unpacking. Customs is required because the F/S factories are usually in different countries than the A/T factories.

natural variability, we require two parameters (forecasts and actuals) rather than just one (actuals). In this case, we use the forecasted, projected or expected value of the parameter and compare it to the actual or observed value once the time has passed. This is unlike the supply variability analysis where we simply focus on the distribution of actuals around the mean. In addition, the non-stationary nature of demand requires that we add the "time into future" dimension to the analysis. Forecasts usually get better as time passes and ship-dates near, and there is a need to understand how the average variability changes through this "forecast horizon". Since different business decisions (i.e. material purchases) are made at different times, the variability at each forecast horizon is important to differentiate. Thus, the level of product aggregation, time aggregation and forecast horizon can be thought of as a 3 dimensional matrix.

Figure 3-4. Matrix of Aggregation and Forecast Horizon



Each cube in the matrix contains the appropriately aggregated forecasts and actuals necessary to determine bias, error and variability at a given level of granularity and planning horizon.

The appropriate level of aggregation should be chosen to accomplish two objectives. First, the aggregation must align with the data used to make forecast-based decisions. For example, if sku-level demand data is used to plan A/T production, then sku-level demand variability is the likely to be the aggregation of interest. Second, the data must be relevant to the supply chain problems at hand. For example, if the supply

chain is reset[19] weekly, then weekly time granularity in demand is required. In this example, daily data may be available, but it will be too detailed and noisy for supply chain analysis. Likewise, monthly data may also be available, but it will be at too high a level to be directly linked to the weekly decisions upon which the business runs. That being said, the data for the lowest level of aggregation is often unavailable.

A brief discussion of the demand identification and fulfillment process is required to understand the nature of demand variability. Intel currently uses a commits process to allocate product to customers. The process begins with customers describing their needs to Intel on a quarterly basis for the upcoming quarter. Intel rationalizes this demand against its projected supply for each product to determine what amount of each product will be "committed" to each customer. Prior to the start of a quarter, these commits are granted, and the customer can begin "booking" units into backlog. The deadline for booking backlog is several weeks into the current quarter, so a full picture of weekly demand is not available until part way through each quarter. In addition, the commercial terms of the commits are such that booked backlog does not have to be consumed in the week that it is booked. The units can be pushed further out into the quarter or cancelled with very little notice. In addition, a customer can request additional units or request to have future committed product pulled forward in time. Such requests are evaluated through a standardized weekly process.

This effective "zero-cancellation window" causes some unique and undesirable effects on the backlog of orders. Since push-outs are easier to accomplish than pull-ins or requests for extra product, customers often request commits for a higher amount than is

---

[19] Supply chain reset is a term to describe the periodic alignment of supply with demand, and the reallocation of supply and demand resources that results.

truly required early in the quarter. The theory is that, if the customer needs the material early, it will be there. Then, as time progresses, and perhaps certain pieces of projected demand do not get realized, these extra units are pushed out week by week throughout the quarter. Finally, at the end of the quarter, the customer either buys out their balance of backlog or they cancel it. This phenomenon is the considered to be primary cause of variability in the backlog demand signal.

3.3.2 Variability in Backlog

The following analyses show the bias and error of backlog over the planning horizon. Note that we occasionally use the word "forecast" in the context of backlog. To the extent that backlog is a predictor of actual demand; the backlog signal is a forecast. However, this use of the term forecast should not be confused with the marketing forecast demand signal discussed later in Section 3.3.3. This data represents backlog for the third quarter 2002 through the second quarter of 2003.

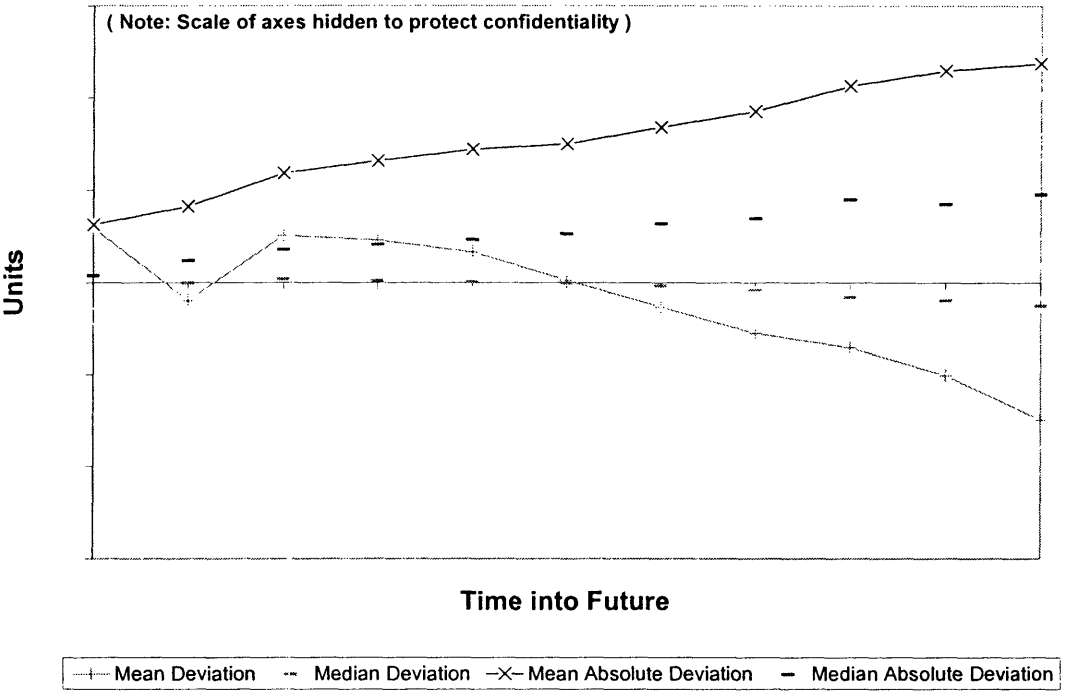
In analyses of this type, the x-axis represents time into the future, measured in weeks or months. Thus, a value of 2 represents the average of all forecast errors of 2-week-out[20] backlog. Therefore, the data may include projected backlog for week 3 as of week 1 and projected backlog for week 37 as of week 35. It should be noted that there are distinct differences between the two examples provided. First, there are differences across time that impact backlog error. For example, a week 1 projection for any forecast horizon is usually worse than a backlog projection made mid-quarter, due to the timing requirements for booking backlog. That is, the week 1 projection will have higher error than the mid-quarter projection, on average. This is what we term a "week-of-quarter"

[20] The term "2-week-out" (or any x-week-out) is used to denote a 2 (or any x) week forecast horizon.

effect. Also, there is a different level of uncertainty in forecasting for each quarter. The yearly seasonality means that the forecast errors of the 3rd and 4th quarters are generally higher than the forecast errors of the 1st and 2nd quarters. This "quarter-of-year" effect results from the start of school and Christmas seasons. Second, there are differences in variability between products with different attributes and characteristics. For example, demand for mobile products are more variable than desktop or server products. Likewise, boxed processors sold to retailers are more variable than processors sold in bulk to OEMs. Lastly, products with different production characteristics, like volume (high-volume/low-volume) or stage of product lifecycle (ramping/stable/end-of-life) have different variability profiles. Though this analysis generalizes the bias of all products over all periods, the data provides a reasonable basis for discussion and is indicative of overall trends in backlog variability.

The following graph shows both bias and error of backlog over the planning horizon where backlog is considered a forecast of projected demand. Bias is an indication of whether - on average - the forecasts are higher or lower than actuals. The error shows just how far off – on average - the forecasts are. The backlog and actuals are aggregated by sku and by week, which is the lowest level of data available for this analysis.

Figure 3-5. Absolute Bias and Error of Backlog

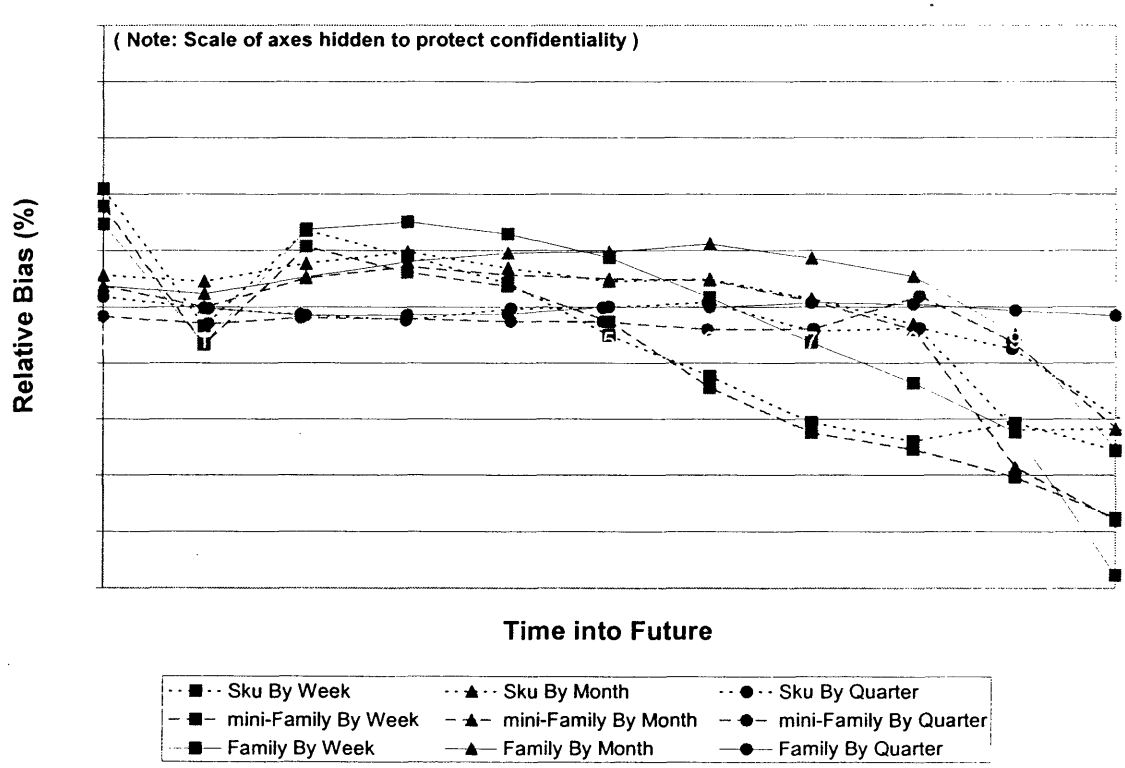


As you can see from Figure 3-5 the projected backlog bias (x points) is biased high within a 4 week horizon, except for 1-week-out. This is an indication of the late push-out behavior that is common among customers and described earlier. The bias in out weeks becomes negative due to incomplete bookings early in the quarter. It is interesting to note that the median deviations are much closer to zero than the mean deviations. This is an indication that the higher mean deviations (MD) are being driven by large outliers – probably high-volume products with moderate errors or moderate volume products with high errors. The median bias near zero indicates that most of the products in most weeks have near zero bias – on average. The (+) points shows the mean absolute deviation (MAD) increasing slowly and steadily over the forecast horizon. This shows that while

the odd nature of the bias is driven by effects from cancellation policies, the error is a relatively linear function of the time into the future of the forecast.

The average bias or MD of backlog should not be dramatically affected by the level of aggregation of the forecasts. We can imagine product A and B, both with projected backlog of 10 units. When actual sales of A are 5 and sales of B are 15, the bias is zero, just like the family bias[21]. There are some differences in how these values get averaged through summary statistics; and these differences are reflected in Figure 3-6 below. To confirm our intuition, we plot average percent error of different levels of aggregation across the forecast horizon.

Figure 3-6. Relative Bias of Backlog

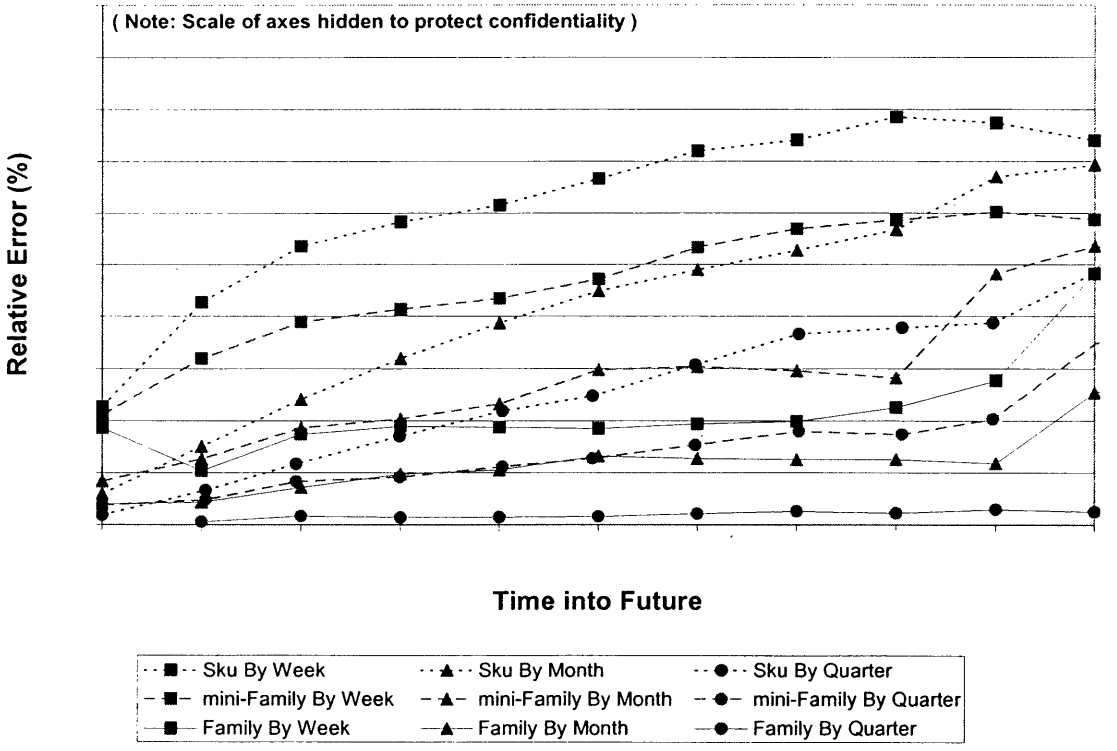


[21] The average sku-level bias (fcst-act) is [5+-5] = 0, while the family-level bias is [10+10] – [5+15] = 0.

The different levels of product aggregation (sku= blue dotted line, mini-family=green dashed line and family=red solid line) and different levels of time aggregation (week=square, month=triangle, quarter=circle) display very similar average bias behavior. Any differences can be explained by the subtle differences arising from averaging of different numbers of values and division by the average of backlog and actuals.

We often care more about how far off target the backlog is, rather than whether it is off high or off low. So in addition to calculating the average bias or Mean Deviation, we evaluate the average error or Mean Absolute Deviation. Whereas, the MD is relatively insensitive to aggregation, error is very dependant on the level of aggregation used. The notion that error is reduced with greater levels of aggregation is intuitive but difficult to quantify. The various levels of time and product aggregation describe different levels of interchangeability. If there is a forecast for 10 units of sku A and 10 units of sku B and the actual demand is 5 units of A and 15 units of B, then the average error on a sku level is 5 units. However, if A and B are in the same product family, and we aggregate to this family level, then the forecast is 20 and the actual is 20, so the error is zero. This is the benefit of pooling. However, the benefits gained from pooling of products vs. pooling of time are not straightforward. Shown below is a complete empirical analysis of the variability differences among different levels of aggregation for a single, high-volume microprocessor family in order to help quantify these differences.

Figure 3-7. Relative Error of Backlog



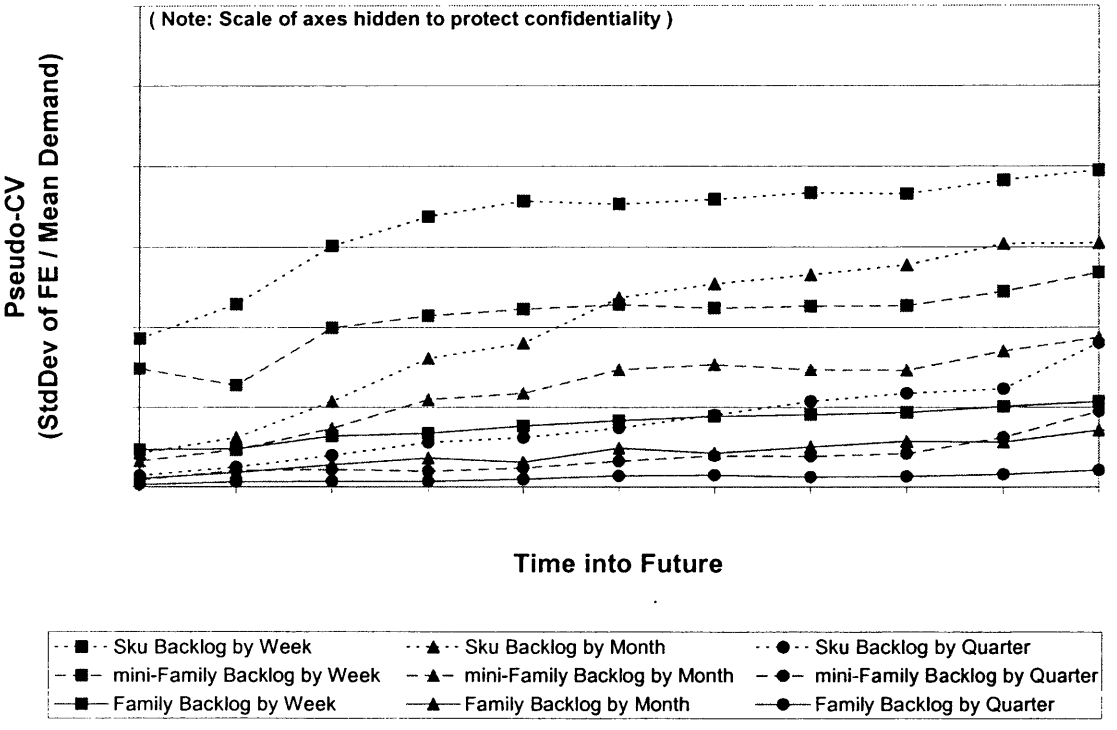
As you can see in Figure 3-7, the relative error of sku-level backlog by week is highest, while the relative error of family level aggregation by quarter is lowest. This makes sense, since family level, quarterly data allows for all errors in skus to cancel each other out, and errors between weeks to cancel each other out. We also see that the sku-weekly error is roughly twice the magnitude of sku-monthly error, which makes sense from a pooling perspective. What is interesting in this figure is the interaction between time and product aggregation. What we see is that, over the majority of the forecast horizon, the sku, monthly error is roughly equivalent to the mini-family, weekly error. So if one wanted to lower error to the level of monthly, one might consider continuing to plan weekly, while changing the level of product aggregation to mini-family. A similar relationship exists between mini-family, monthly and sku, quarterly.

When we evaluate forecast variability of a demand signal like backlog, we are interested in not only the bias and error of that signal, but also the variability of the signal. Whereas bias is a measure of how far off a forecast is, variability is a measure of the spread of the distribution of those errors. Errors that are often close to zero and tightly distributed around the mean indicate consistently good forecasting. On the contrary, if the errors are sometime far off, and spread widely around the mean, this indicates high variability in forecasting. There are several ways that this "spread of errors" can be calculated. One way is to calculate the standard deviation of the distribution of forecast bias or the standard deviation of Forecast Errors (what we call StdDevOfFE). This is often the best way to characterize variability for use in stochastic models and it bears close resemblance to the method of measuring natural variability by calculating the standard deviation of actuals around their mean. The StdDevOfFE is identical to the normal concept of standard deviation if, instead of calculating the spread around the mean, you look at the spread around the forecasts. This association is clarified further when we describe both mean values and forecasts as "expected values" in a mathematical sense. Another way to characterize variability is to calculate the Mean Square Error, by averaging the squared forecast errors. It can be shown that these two techniques yield the same results, when the average bias is zero (see Appendix C – Equivalence of MSE and StdDevOfFE) and either method can be used for input into models, optimizations and the like.

As discussed before, another use of variability data is to compare businesses, products, projects and programs. To accomplish this, we need a relative measure of variability which can account for differences in production volumes. Such a measure will

also allow us to compare the variability of different levels of aggregation. The measure we use is the pseudo-CV described in Section 3.1.2. This parameter is calculated by taking the standard deviation of Forecast Errors (StdDevOfFE) and dividing it by the mean actual demand. A comparison of the pseudo-CV for different levels of aggregation is shown in Figure 3-8 below.

Figure 3-8. Relative Variability of Backlog

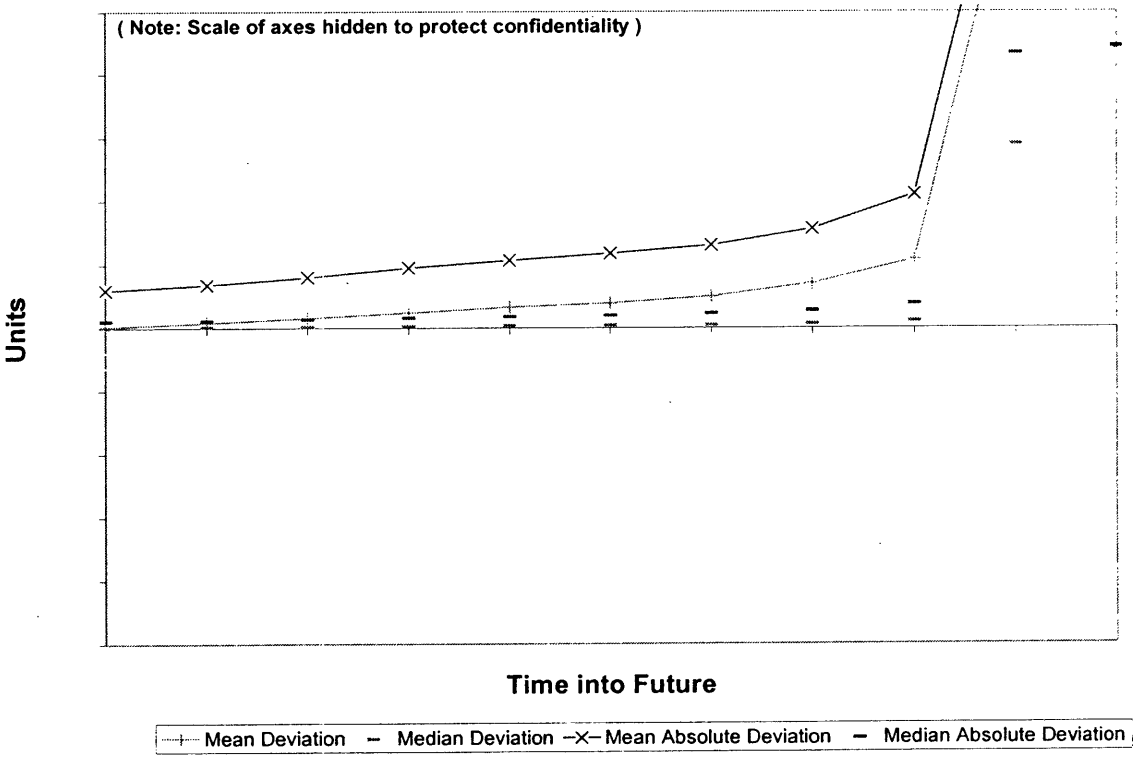


Like the forecast error, the spread of the errors or "relative variability" is significantly reduced with higher levels of aggregation as well. It is interesting to note that, in the out-weeks, the pseudo-CV is quite constant over the forecast horizon, for all levels of aggregation. This means that the variability of a 5-week-out forecast is about as variable as a 3-week-out or 7-week-out forecast. Only within the 3-week-out window does the variability improve significantly.

3.3.3 Variability of Forecasts

The second signal of demand used by Intel is the marketing forecast. This forecast is generated by industry and product experts in the marketing group who use a combination of economic models, customer forecasts, production capacity data, competitive information and intuition to generate quarterly forecasts for different products. The forecast is generated for several quarters into the future and revised monthly. In contrast to backlog, which is a snapshot of projected demand, the marketing forecast is a true forecast, subject to all of the biases that exist in forecasting. You can see these biases in Figure 3-9 below.

Figure 3-9. Absolute Bias of Forecasts

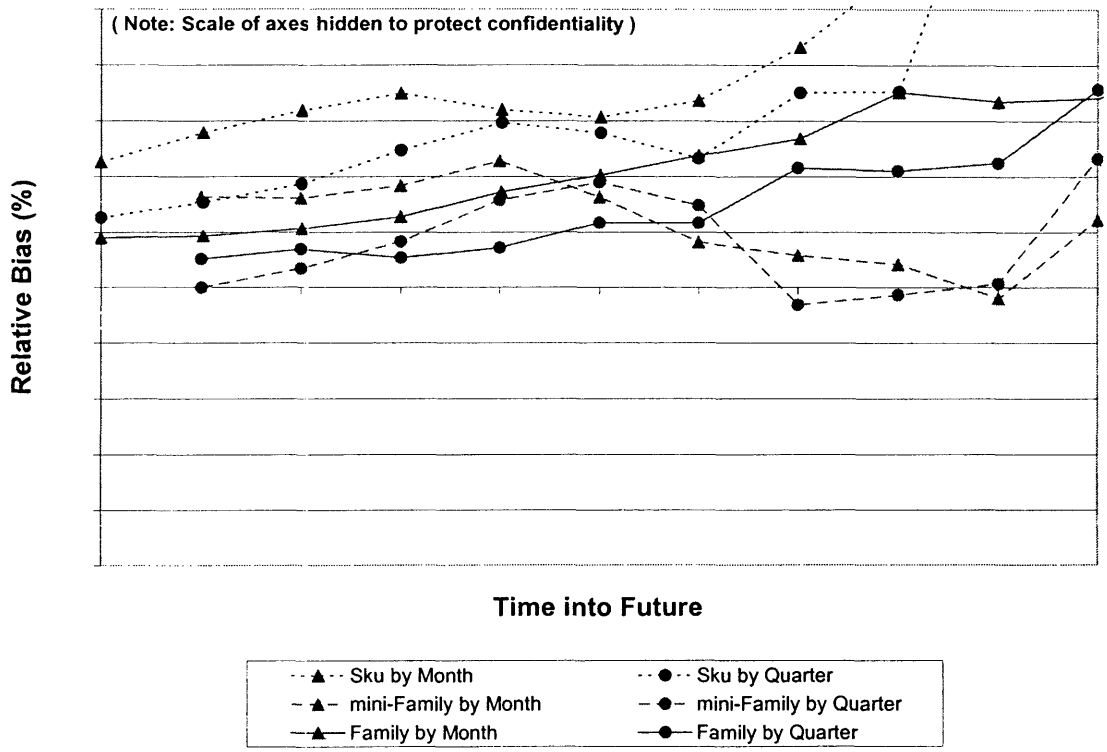


Whereas Backlog is a relatively unbiased predictor of actuals, the average bias or mean deviation (MD) of the marketing forecast is always positive. Again, the median

bias is close to zero, indicating that large errors in a few products are driving the mean deviation up. This data represents marketing forecasts for the third quarter 2002 through the second quarter of 2003. In the case of marketing demand, the x axis is in units of months into the future, rather than weeks. In order to obtain monthly level forecasts, the quarterly forecasts are rationalized with actual shipments within the quarter and divided out among months using historical seasonality factors (see Appendix D – Disaggregation of Quarterly Forecasts). Historically, Intel has used a 30-30-40 breakout, but we have found that 25% of the quarterly demand is realized in the 1st month, 30% in the 2nd month and 45% of demand is in the 3rd month of the quarter, we use this 25-30-45 breakdown in our analysis. The following charts do not show weekly error, since we don't have any additional information with which to disaggregate data into weekly buckets.

The relative bias among different levels of timing and product aggregation are shown below.

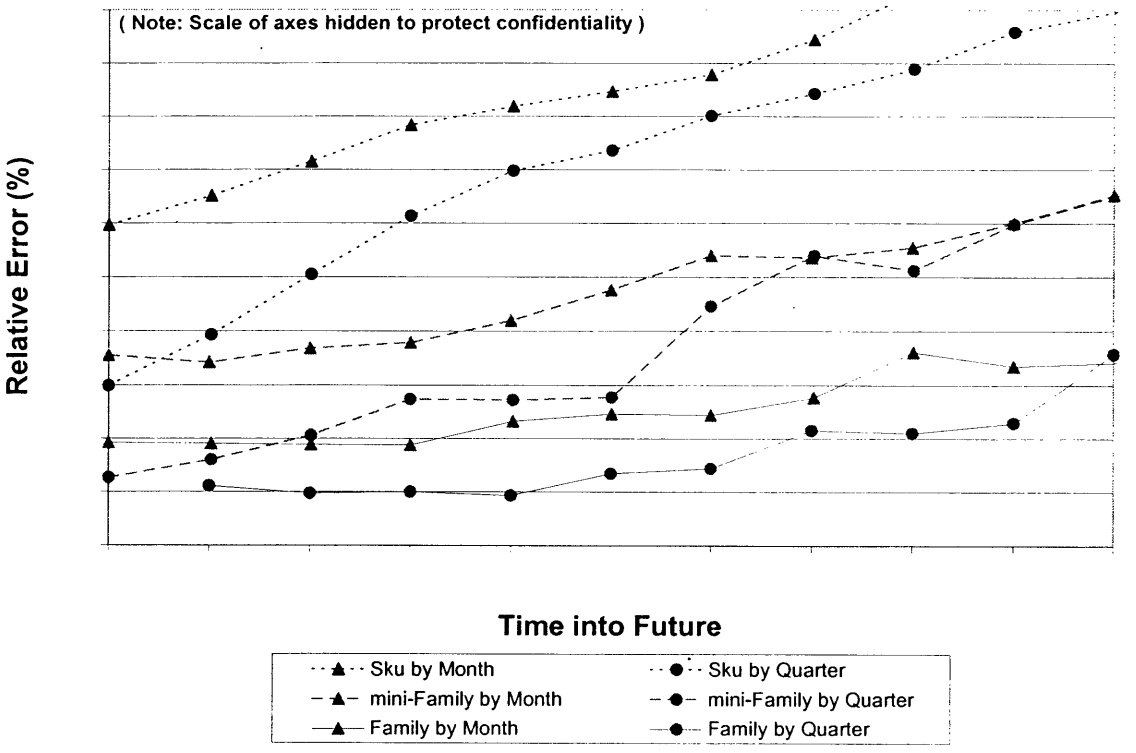
Figure 3-10. Relative Bias of Forecasts



It is interesting to note that marketing forecasts for this period are positively biased in all time horizons at all levels of product and time aggregation. There are several possible explanations for this behavior, but we will focus on one. The period of time within which this data was collected was notorious for being near the end of one of the worst downturns in the history of the semiconductor industry. The prediction for the market to regain momentum was probably a basis for some of the forecasts. Since the downturn continued far longer than originally thought, forecasters may be forgiven for their optimistic predictions.

The relative error among different levels of timing and product aggregation are shown below.

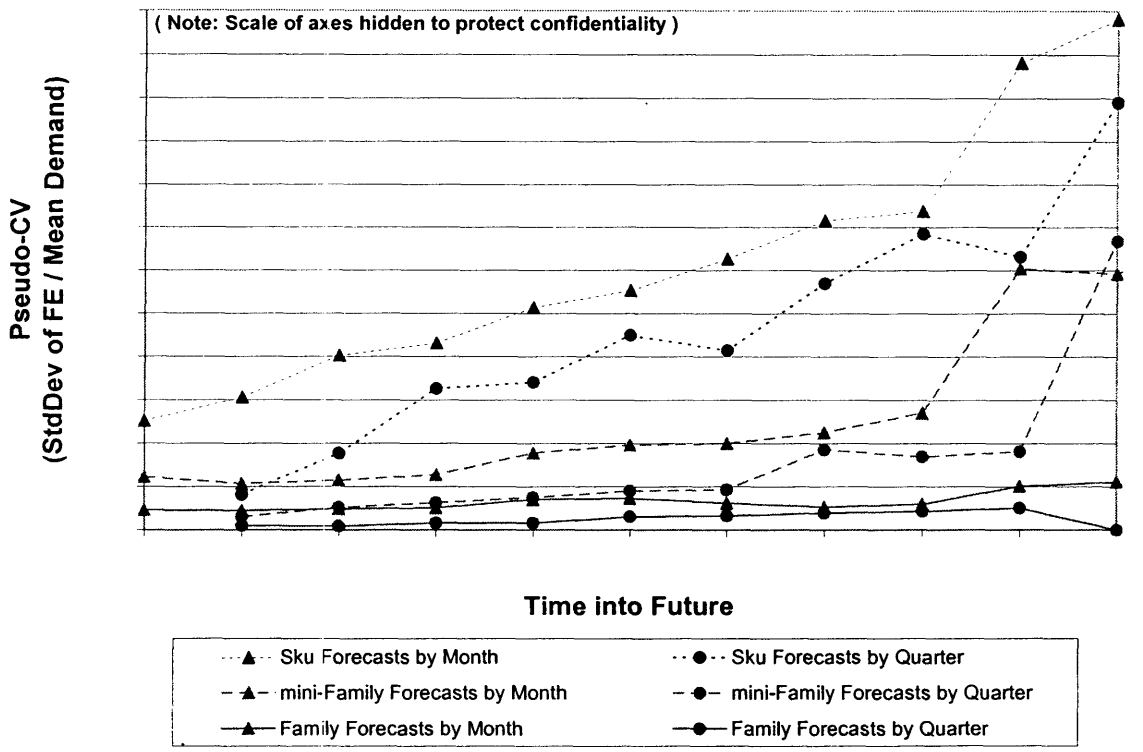
Figure 3-11. Relative Error of Forecasts



Of interest in Figure 3-11 is the fact that, while the sku-level forecast gets better (less error) as one gets closer to ship date, the error of family-level forecasts do not improve dramatically over the forecast horizon. We believe this is a result of two phenomena. First, the commits process sets a fairly firm limitation on the demand of processor-types (akin to a family) that can be met. Demand above supply capacity that is not committed-to is not counted as forecast error, because it never materializes as backlog. Second, because wafer starts are made 3 months into the future, this caps the total product family demand to the level of wafers which were started. Thus, the family level forecast made several months out is as good as it will get, while the mix of skus changes constantly and results in lower forecast error closer to delivery.

Though we may conclude that significant forecast error is a part of doing business in the semiconductor industry, it does not follow that the variability of forecast errors should also be high. That is, we may be able to deal with high average forecast error, but high variability is more difficult, since is results in our having low confidence that a particular forecast will be within a certain range. This is exactly the reason why we may choose to hold more inventory. The relative variability of the marketing forecast demand signal is shown below.

Figure 3-12. Relative Variability of Forecasts



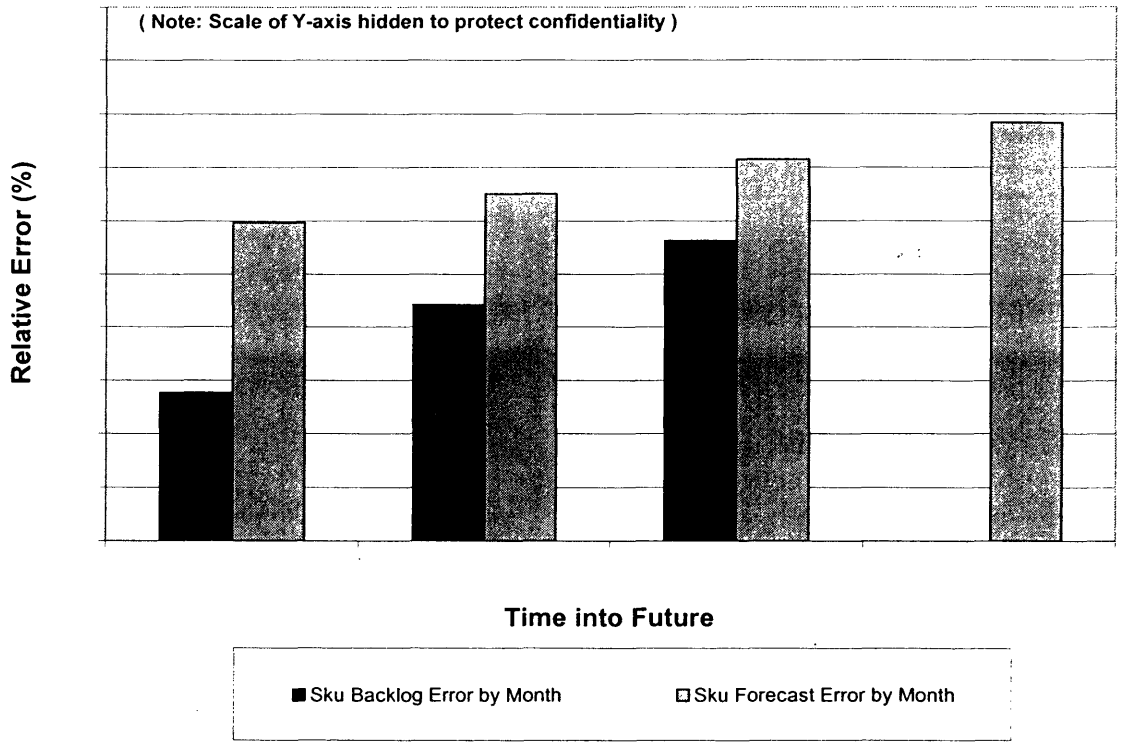
Again, we see the same phenomenon, where there is significant improvement in sku-level variability as we get closer to ship data but the same improvements are not seen in mini-family or family-level variability. In practice, if we could use mini-family data rather

than sku-level data, we would see a dramatic reduction in inventory required to meet a particular service level.

3.3.4 Comparison of Two Demand Signals

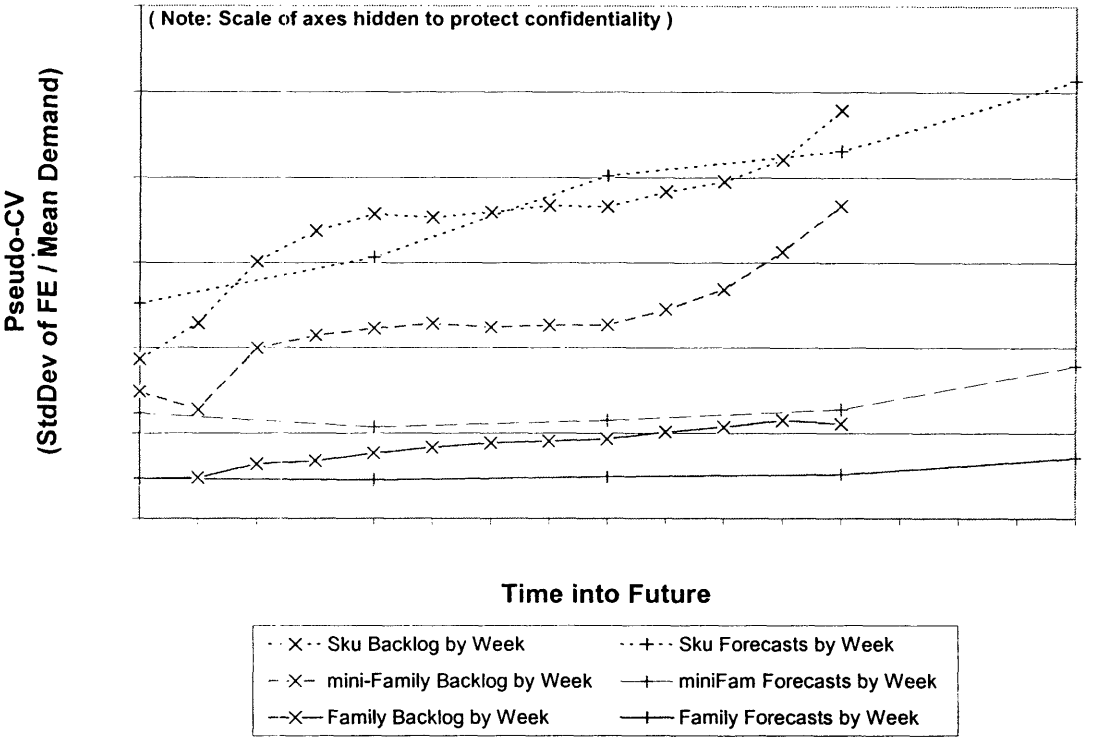
After analyzing both the backlog and forecasts in terms of bias error and variability, we compared them in order to see which signal is a better predictor of actual results. First, we compared the level of forecast error in each signal, and then we evaluated the variability of each signal as a measure of "the ability to forecast well". In order to do the error comparison, we aggregated the backlog data into monthly buckets in order to be in the same units as marketing forecasts. The result is shown in Figure 3-13 below.

Figure 3-13. Comparison of Backlog to Forecasts



The fact that the backlog demand signal has less error than the forecast demand signal is not surprising since it is more tactical in nature. Furthermore, backlog data is not available more than three months out, so comparison in the strategic forecast horizon is not possible. A comparison of backlog and forecasts variability is shown in Figure 3-14 below.

Figure 3-14. Variability of Backlog and Forecasts



If variability is a measurement of the firm's ability to forecast well, we see that forecasts are a better demand signal for mini-Family and family level of product aggregation. However, sku-level backlog is less variable in the immediate forecast horizon. Around the 3-week-out point, the forecast demand signal has lower variability until around the 7-week-out point, where backlog is again less variable. In the 2 to10-week-out horizon, backlog is roughly equivalent to forecasts in terms of variability.

As a point of note, many planners currently use backlog in the 0-4 week horizon, a combination of backlog and forecasts in the 5-8 week horizon and pure forecasts in the 8+ week horizon. Based on the variability results, we would recommend using backlog in the 0-2 week horizon, a combination of backlog and forecasts in the 2-10 week horizon and pure forecasts in the 10+ week horizon.

3.3.5 Implications of Bias, Error and Variability for the Supply Chain

After considering the bias, error and variability of demand, it is quite common to conclude that the Intel's forecasting is "bad" and should be improved. However, we do not necessarily draw this conclusion. First, good forecast error benchmarking data is exceedingly difficult to find. Second, the unique nature of Intel's supply chain makes comparisons with any such data specious. And third, for numerous reasons, forecasting the future is hard. The first reason that forecasting is hard is that there is high demand variability at business planning level (family, quarter) and it is dramatically higher at factory planning level (sku, .week). Second, there is gaming and judgment in the demand signal. That is, each customer believes that they will capture additional market share in sequential quarters. The sum of these values is much bigger than the size of the overall market. As a result, judgments must be made regarding quantity allocation. This gaming problem gets worse in a constrained environment. If customers believe that they may not get their requested quantity, then they will increase their order further. Third, macroeconomic factors cause massive shifts in supply/demand. This shifting is characterized by the bullwhip effect, where variability gets amplified as the demand signal propagates up the supply chain. Fourth, short microprocessor product lifecycles have little or no demand history. As a result, forecasters have to make judgments about

the uptake in the market rather than using historical data. Fifth, Intel produces thousands of differentiated products. This may seem like it would reduce variability, but the fact that a customer could receive several different products makes it difficult to forecast each product independently. Sixth, there are complex, interrelated product mappings in the production process, which make supply forecasting required to meet demand quite difficult. Seventh, manufacturing cannot easily react to forecast changes due to long wafer start times. With production lead time on the order of one quarter and product lifecycles not much longer than that, we see the classic newsboy problem where most of the supply must be produced before demand is known.

In light of these difficulties, one question worth asking is – 'if Intel can sell all that it makes, is forecast variability a problem?' While some people argue that it is not a problem, we strongly believe that such variability causes important secondary issues which can manifest themselves in two ways. First, fire-sale cannibalization of up-market products can occur if forecasted mix is wrong. If the firm makes too many low-speed products because its forecasts were poor, it will need to discount these processors to sell them. This will shift demand away from more expensive products and destroy margins and profits. Second, making the wrong products can cause production limitations in a capacity constrained market. If Intel is over-producing certain products, they are likely to be under producing others. The result, in a limited-capacity manufacturing environment, is high inventories of the over-produced product and missed sales of the under-produced products. So the next question is - 'if their demand signals were better, would Intel save money?' Again, we believe the answer is an emphatic yes, since the company would

surely require fewer inventories to meet the same level of service. Also, significant productivity improvements relating to less churn would result from less reaction to noise.

However, it is not the conclusion of this variability analysis that improvements in the forecasts are necessarily required. First, it is not obvious that the magnitude of the errors is dramatically worse than comparable companies in the industry. Second, it may not be possible to reduce the errors because of the difficulties in forecasting mentioned above. In other words, Intel should not say, 'when our forecasts get better, THEN our planning systems will work'. Instead, Intel should work to improve the robustness of the planning system to the variability inherent in the process.

As a result of the variability characterization, we conclude that higher levels of aggregation lead to lower error. Though higher level data has limited usefulness, this pooling benefit is clear. We also conclude that errors don't get much better as ship date approaches. Likewise, we find that the data is sparse, non-normal, dispersed and inadequate for supply chain analysis. Our recommendation is to modify the planning systems to account for the inherent variability in existing planning processes, rather than working from point forecasts and responding to all changes whether signal or noise. For example, Intel may not need to plan wafer starts for 9 months into the future, using sku level, weekly detail. In this case, we are just optimizing noise and we may get better results with less work by planning to the mini-family level of product aggregation. Specific recommendations which build from the conclusions of the variability characterization are presented in Section 5.

4 STOCHASTIC MODEL OF INTEL SUPPLY NETWORK

In this section, we discuss a model which helps to describe Intel's supply chain. The model is composed of two production nodes and two inventory nodes, which resembles the high-level structure of Intel's supply network. A graphical representation of the model is shown in Figure 4-1 below.

Figure 4-1. Two-Node Stochastic Model of Semiconductor Supply Network

As with all models, this one oversimplifies the supply chain and makes assumptions about the system which may not be perfectly accurate. However, this model is more useful than many other models because it is implemented using actual data. Instead of making gross assumptions about the average lead time or variability of demand, this model was created after the extensive variability characterization described in Section 3. As a result, we can make strong conclusions based on results provided by the model.

The primary recommendation resulting from this work is that a paradigm shift – from judgment to data – is required. Specifically, Intel must change its inventory management strategy from "don't stock out" to "set a service level, and calculate an inventory target". This will require better measurement of service levels and automated supply and demand data feeds. These service level and variability data should be analyzed and inventory targets should be set based on the attributes of the products. Finally, the company must quickly move toward global inventory optimization through data sharing, common tools and collaboration.

4.1 THEORETICAL BASIS FOR MODEL

4.1.1 The Base-stock Model

The supply chain model presented here is based on a version of the base-stock model described in Zimmerman et. al. [1974] among numerous others. The primary assumptions of the model are periodic review, no set-up costs, no lot sizing, normal variable distributions and infinite production capacity. While not all of our data cleanly meets the assumptions of the model, we find the base-stock model to be better than others in its usability and instructiveness in making tradeoffs. In the event that inventory decisions were to be based on such a model, advanced techniques and future work would be required to refine the model to deal with the edge cases (i.e. near full production capacity) and non-normality of certain variables. The basis for making these upgrades is the variability characterization provided in Section 3 and recommendations for future work discussed in Section 5.

Under this system, the safety stock, as well as total inventory, can be calculated to meet a particular service level, while taking into account certain sources of variability. If only considering demand variability, the equation is shown below.

Equation 4. Base-stock Equation with Demand Variability

$$\text{Base Stock} = \mu_d(r + \mu_{LT}) + z\sqrt{\sigma_D^2(r + \mu_{LT})}$$

Where: μ_d = average demand rate over lead time
r = review period
μ_{LT} = average lead time or throughput time
z = safety factor calculated from service level
σ_d = variability in demand

The pipeline stock, otherwise know as work-in-process or WIP, is represented by the average demand times lead time. The average demand times review period in the first

term represents the cycle stock. Mainly relevant for warehousing systems and batch production processes, the cycle stock is the quantity of goods required to meet demand until the next time that an order is placed. In the case of continuous production, like Intel's, we assume that the average demand over review period is equal to production over review period. Therefore, cycle stock (μ_d*r) is not a relevant concept under our set of assumptions and we remove it from the subsequent analysis. Furthermore, since safety stock buffers the system from variability, we will primarily focus on safety stock.

Safety stock is required to buffer from variability over replenishment period. The replenishment period is the throughput time required to produce a product plus the review period. In the equation described above, the only relevant variability is variability in demand. This can be calculated by determining the variance (or standard deviation squared) of forecast errors. Alternatively, assuming zero average bias in forecasts, we can calculate the Mean Squared Error as variability in demand. The z in the safety stock term represents the safety factor (z = 1.64 for 95% service level) implied by the specified service level. It is calculated as the inverse of the normal distribution. Combined with values for average demand, average lead time and review period, the overall inventory and safety stock can be calculated.

The safety stock portion of the base-stock model can be extended to account for variability in lead time as described by Equation 5.

Equation 5. Base-stock Equation with Demand and Lead Time Variability

$$\text{Safety Stock} = z\sqrt{\sigma_D^2(r+\mu_{LT}) + \mu_d^2\sigma_{LT}^2}$$

Where: μ_d = average demand over lead time
σ_{LT} = variability in lead time

The additional term in the equation represents the demand variability resulting from given variability in lead time. In a similar fashion, the model can be extended to account for variability in production yield. Black [1998] developed and used Equation 6 to optimize WIP in a CONWIP[22] process.

Equation 6. Base-stock Equation with Demand, Lead Time and Yield Variability

$$\text{Safety Stock} = z\sqrt{\sigma_D^2(r+\mu_{LT})+\mu_d^2\sigma_{LT}^2+\frac{\sigma_S^2\mu_d\mu_{LT}}{\mu_s}}$$

Where: μ_s = average yield
σ_s = variability in yield

The third safety stock term represents the variability in yield-adjusted demand caused by yield variance.

4.1.2 Modeling Multiple Nodes

The base-stock equation provides a mechanism to compute the safety stock of a single node, given variability information on demand, lead time and yield. There are many ways to connect these nodes to compute overall multi-echelon safety stocks. Graves and Willems [1988] assume fixed and guaranteed service times, bounded demand and no capacity constraints to formulate an optimization problem. Graban [1999] develops a serial model, where inventory is held to buffer variability in the closest upstream node. We develop our model using Graban's as a basis, though we compute inventory levels required to buffer from known variability in supply and demand, rather than to set WIP targets. We accept the fact that the neither the placement nor the quantity

[22] A CONWIP process is simply one which uses a policy of constant work-in-process as a way to manage in-process inventories.

of our calculated safety stock will be optimal. Rather, we use the values as a basis for evaluation of our current performance as well as a platform for performing tradeoffs.

In our two-node model, we assume that the purpose of die held in Assembled Die Inventory (ADI) is to buffer variability in the Fab/Sort (F/S) part of the production process. Likewise, we assume that the finished goods in Components Warehouse (CW) are held to buffer variability in Assembly/Test (A/T). In this way, we create a loosely coupled system in which extreme variability in Fab is assumed not to affect final service level. In fact, this is exactly what the purpose of strategic safety stock is: to decouple different parts of the system from each other.

Despite the decoupling of the two production areas, there are several links between the two nodes. First, the average demand observed by CW as final customer demand is transmitted up the supply chain, adjusted for yield. In other words, the F/S part of the process has to produce more than the amount demanded by customers, so that yield losses in A/T will be compensated for. Likewise, the demand variability witnessed by CW is transmitted up the supply chain, albeit in a more complex mathematical form (see Appendix E – Derivation of F/S Demand Variability). The upstream demand variability is affected not only by downstream demand variability, but also downstream yield variability. That is, if yield fluctuates wildly, the demand variability of the upstream node will be negatively impacted as well.

The service levels of the two stages are intimately linked to each other and the formulation of this model requires that a slightly modified definition of service level be adopted. The base-stock model assumes that service level is the amount of periods in which 100% of customer demand is met. We use this definition, except we replace the

words customer demand with downstream node. That is, the customer of ADI is A/T and therefore, the customer service level (CSL) represents the percentage of periods in which 100% of A/T demand is met.

A further discussion of service levels is required at this point. Despite the fact that the only customer service level that really matters is the CSL at the customer facing node, some notion of CSL at each node is required in order to determine how much stock to hold at in-process inventory points to meet certain levels of volatility in downstream nodes. However, the combined service level of two nodes whose service levels are less than 100% is less than either service level on its own. This can be seen by imagining that the downstream node has a service level of 95%, by itself. In order to provide that level of service (95%) to the customer, the downstream node would need to be provided 100% service from the upstream node. The notion of 100% service level is not germane to inventory analysis because, as a result of the properties of the normal distribution, such service would require infinite amounts of safety stock. When each node specifies a sub-100% service level, the final service level as seen by the customer is - at least – the product of the two upstream service levels. Thus, the minimum value of the CSL is the product of the two-node service levels and the maximum would be the value of the CSL of the higher node.

To see how these maxima and minima apply, we evaluate an example. If a die is requested by A/T from ADI but not received, it may not eventually be required at the time requested. Whether it is required to meet an order as requested is a function of demand variability, throughput time variability and yield variability. For example, if ADI misses a shipment to A/T, but provides the requested die in the following week AND the

TPT of that die is faster than average; then A/T still has a chance to provide the finished good to CW in time to meet that customer demand. Thus we can see that the product of the two service levels is the minimum CSL and we use this to find the worst case safety stock levels. The difference between this minimum value and the actual value can be dramatic. For example, if we imagine a situation where, in 50% of cases, the delinquent die is delivered to A/T in the week following the request, the overall service level jumps from 86% to nearly 90%.

4.1.3 Base Case Model Formulation

The base case model was formulated using data collected and analyzed through the variability characterization described in Section 3. In the case of supply variability, we calculated the mean and standard deviation for throughput time and yield for one product family[23] within both F/S and A/T. For demand, we used the relative (i.e. CV), family-level, variability in forecast errors as a proxy for the demand variability at CW.[24] We then combined the demand variability at CW with the yield variability in A/T to get an estimate of ADI demand variability as if the final customer demand variability were merely propagated upstream to A/T. Furthermore, we used the actual review periods for each node based on current business processes. The final customer service level was kept at a constant 86% for this entire analysis, because this was the best estimate of the actual

---

[23] The selected product represented a majority of Intel microprocessor volume over the time period studied.

[24] The mechanics of the calculation were completed as follows. First, calculate the forecast errors for each family over every time horizon. Second, determine the standard deviation of forecast errors for each family, for each time horizon. Third, calculate the average demand for each family over the time period studied. Fourth, divide the standard deviation of forecast errors by the mean demand to get the CV for each family over each time horizon. Fifth, average all of the family CVs in each time horizon. Lastly, multiply projected demand for a family by the horizon-specific CV (or weighted average CVs) of interest.

average service level under real conditions. The actual service levels at each node were assumed to be equal, for simplicity.[25]

As noted in Section 3, demand variability tends to be far more complex, interesting and impactful than supply variability. As a result, we spend much more time looking at sources and implications of demand variability. The way we calculate demand variability is a modified version of the technique used to get the data shown in Figure 3-8 and Figure 3-12. That is, for each individual sku, we calculate the standard deviation of forecast errors and the mean demand for each forecast horizon. By dividing the StdDevOfFE by Mean Demand, we get a pseudo-CV measurement for every sku in each time into the future horizon. We then average these pseudo-CVs to get the sku-level, horizon specific variability multiplier. If we take the projected demand (backlog or forecast) and multiply it by this calculated value, we get an estimate of variability (standard deviation of forecast bias) that we can use in the base-stock model. We complete the same analysis for mini-family and family levels of product aggregation and both backlog and forecasts for use in our inventory analysis scenarios.

An alternative method of developing variability estimates for given levels of demand is to do a variability regression, where error (MSE) is regressed against demand for all skus in each time horizon. Once the equation is determined, variability can be calculated by plugging the projected demand into the equation with the appropriate time horizon. This approach could be extended to include different equations for skus of different attributes to further refine the variability estimates. The problem with this

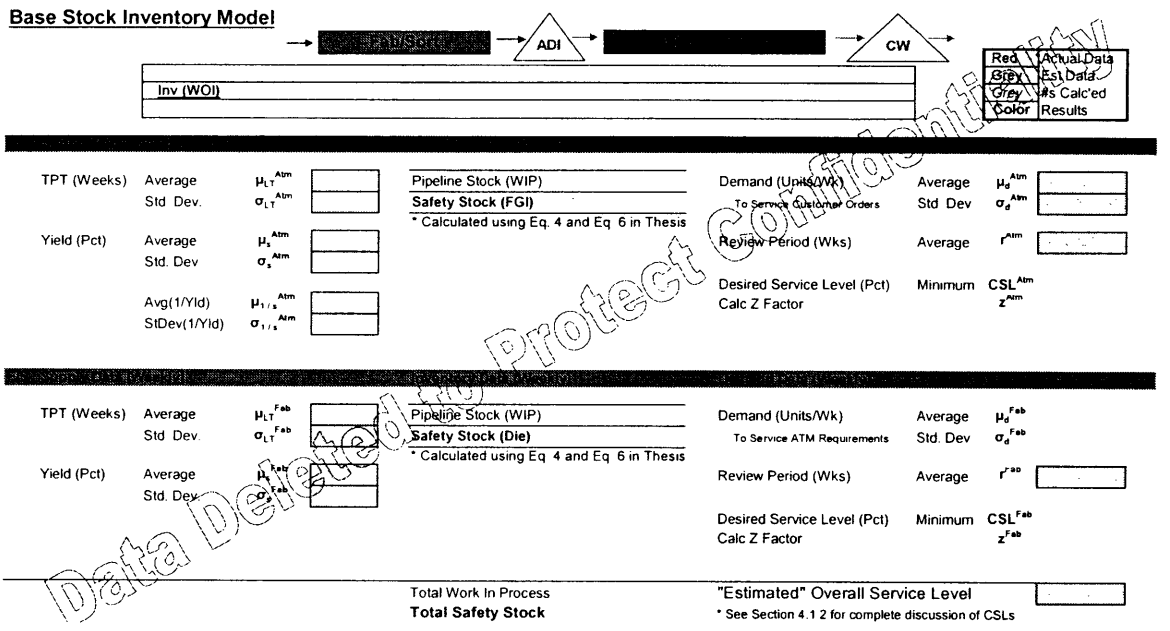
---

[25] With the final customer service level defined, but the two upstream service levels undefined, there are three variables and two unknowns. Requiring them to be equal removes one degree of freedom and results in upstream service levels of 92.74%. Again, the 86% represents a minimum CSL.

approach is that the relationship is quite weak, with R^2 values well below 0.5 for all regressions. The result is an equation which tends to underestimate variability when compared with the standard deviation of Forecast Errors. This is particularly troublesome because the MSE should actually overestimate the variability compared with StdDevOfFE, when there is a chronic bias in the data (see Appendix C – Equivalence of MSE and StdDevOfFE). As a result of these problems, this technique is not utilized in this analysis, but a more detailed description of the technique and its results is shown in Appendix F – Description of Demand Variability Regression.

The driver behind this analysis was to use real data, together with a realistic model, in order to quantify the inventory requirements and analyze tradeoffs. The actual data is proprietary and therefore is disguised as presented in this thesis. However, we discuss the relative results of the model runs as they relate to the actual data so that we can understand how changes in variability affect inventory requirements in the supply network. A representation of the model's user interface, with disguised data, is shown in Figure 4-2 below.

Figure 4-2. User Interface for Two-Node Stochastic Supply Network Model



4.2 USING MODEL TO IDENTIFY IMPACT OF VARIABILITY

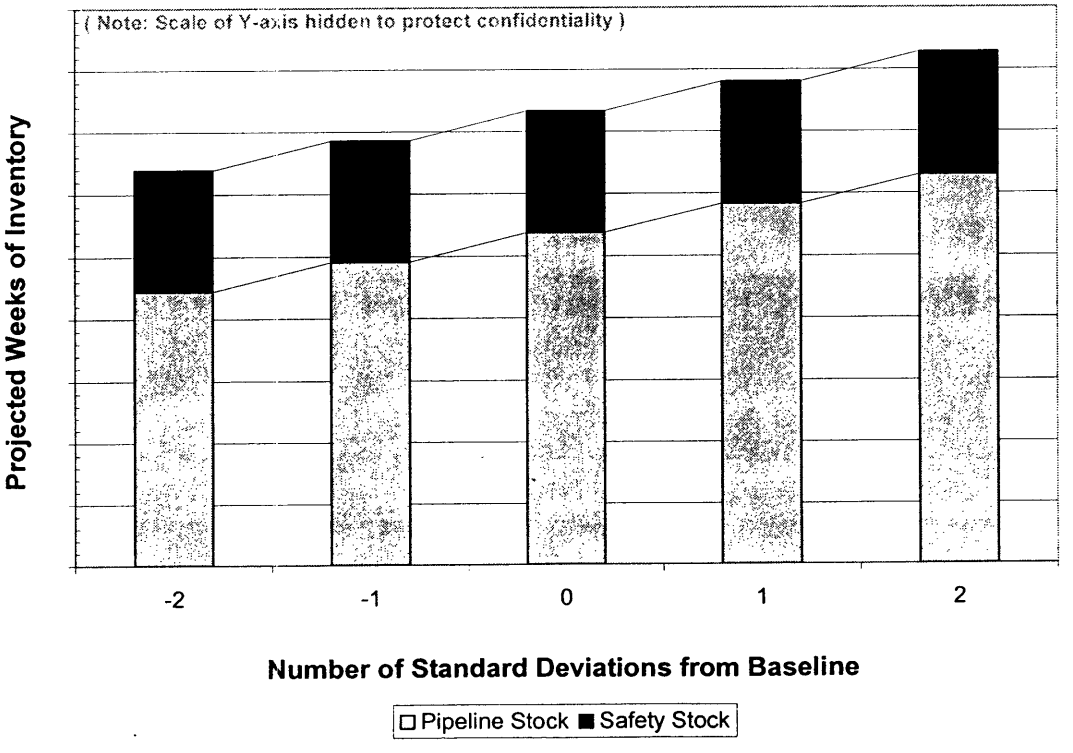
4.2.1 Impact of Changes in Parameters on Safety Stock

The relevant supply and demand variability data were entered into the two-node model and inventory requirements were calculated based on the base-stock equations described in Section 4.1.1. The results of the base case model indicate a requirement of 2.3 weeks of inventory in CW and 1.6 weeks in ADI for a total of 3.9 WOI. See Appendix G – Table of Inventory Analysis Results for complete model results.

In order to quantify the impact of each parameter's variability on the calculated inventory, a sensitivity analysis was completed. In this analysis, new inventory targets were calculated via the base-stock model, where the variability value of one parameter is increased, while all other parameter values are held constant. Further, this was done with both mean values and standard deviation values. Though inventory is explicitly held to

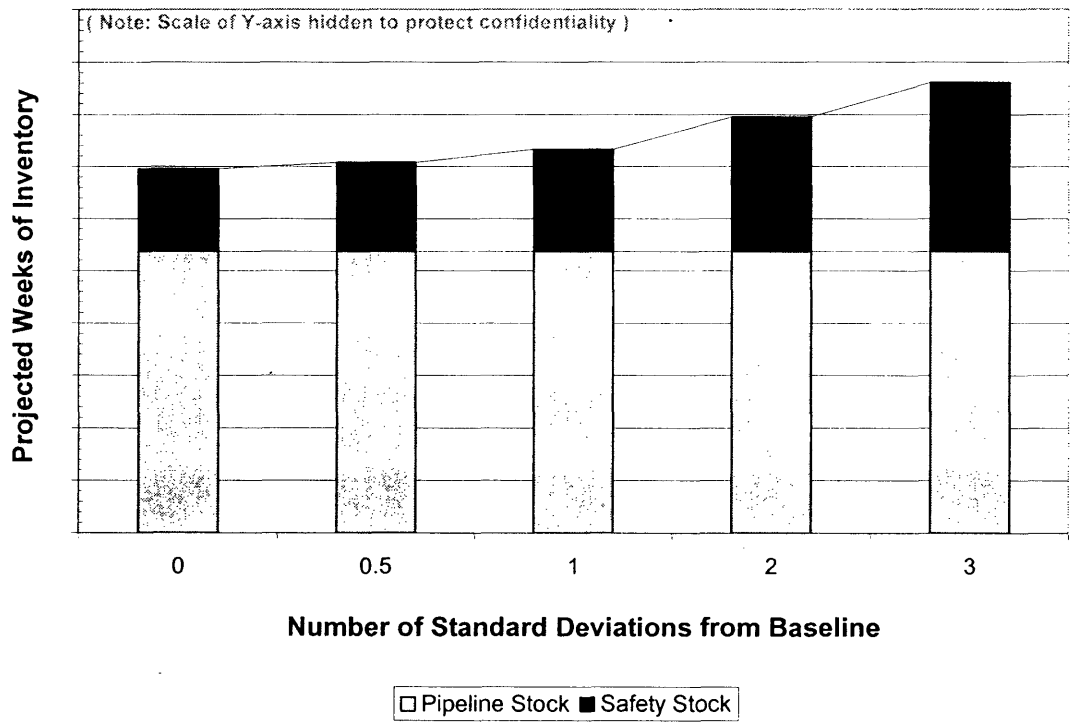
buffer from variability in lead time (for example) rather than average lead time, the average lead time is a part of the equation and its multiplicative effects cannot be ignored. Thus, to evaluate the effect of changes in mean value, inventory targets were calculated at the original mean value as well as mean value +/- 1 StdDev and the mean value +/-2 StdDev. Each new inventory target was subtracted from the original value and the difference was divided by the number of standard deviations. The result is an increase (or decrease) in safety stock per standard deviation increase (or decrease) in mean parameter value. As expected, the results show that changes in mean values have a large impact on pipeline stock (WIP) but virtually no impact on safety stock. The results of this analysis for one parameter are shown in Figure 4-3 below. Similar analyses were done for the remaining F/S and A/T parameters.

Figure 4-3. Impact of Change in Means on Safety Stock and Pipeline Stock



The impacts of changes in variability were measured by taking the original standard deviation and multiplying it by 0, 0.5, 1, 2 and 3. The new variability values were plugged into the base-stock equation and new inventory requirements were calculated. Again, the difference in safety stock requirements per standard deviation in variability was calculated for changes in each parameter. As expected, changes in variability had a significant impact on safety stock but not pipeline stock. The results of this analysis for one parameter are shown in Figure 4-4 below. Similar analyses were done for the remaining F/S and A/T parameters.

Figure 4-4. Impact on Change in Standard Deviation on Safety Stock and Pipeline Stock



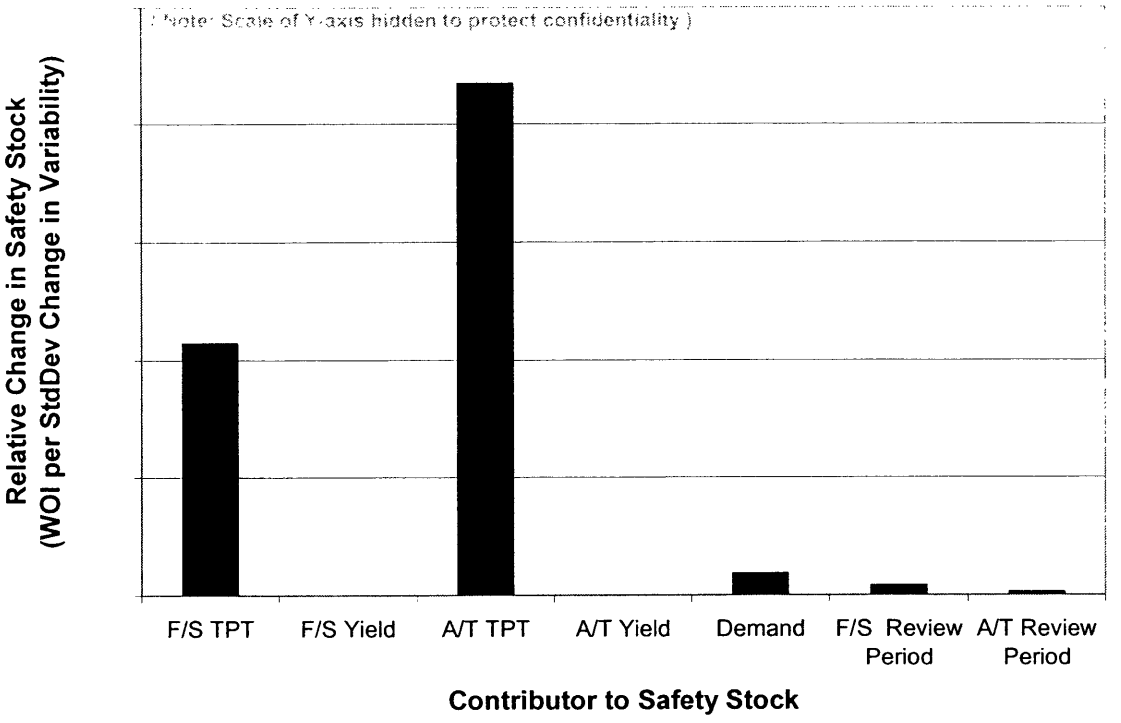
It is interesting to evaluate the base level of safety stock[26] which represents a situation where there is zero variability in the parameter. This is the baseline inventory required in the case where the evaluated parameter was deterministic, with all other parameters at their historic level of variability.

4.2.2 Relative Contributions of Variability to Safety Stock

For each parameter, the differences in safety stock (calculated from baseline variances) are divided by the number of standard deviations to get a "relative impact of variability" estimate measured as inventory per variability. Its interpretation is that, if variability increased or decreased by one standard deviation, then the calculated inventory increase would result. A more likely scenario is that the increase would be on the order of 0.1 StdDev, rather than an increase or decrease of a whole StdDev, but the results can easily be scaled as such. The result of these analyses for our five sources of variability is shown in Figure 4-5 below.

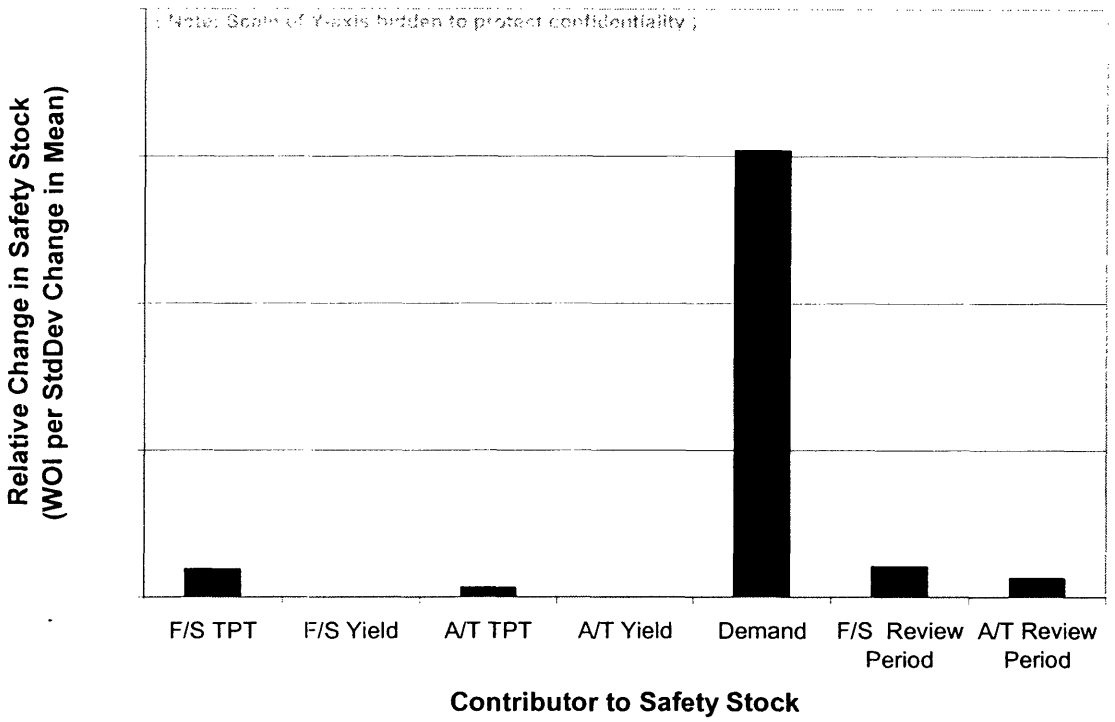
[26] The safety stock result associated with 0 variability multiplication factor is the leftmost bar in the graph.

Figure 4-5. Sensitivity Analysis on Standard Deviation of Sources of Variability



As you can see, changes to the variability in throughput time have the greatest impact on inventory requirements. This is because the calculated relative standard deviations of these values are far greater than that of either yield or demand. Furthermore, A/T TPT variability is higher than F/S TPT variability, even when scaled for the much longer F/S throughput time. For comparison purposes, the impact of review period is added to this graph even though this is in units of WOI/week of review period. A similar analysis was done for changes in mean values and is shown in Figure 4-6 below.

Figure 4-6. Sensitivity Analysis on Mean Values for Sources of Variability



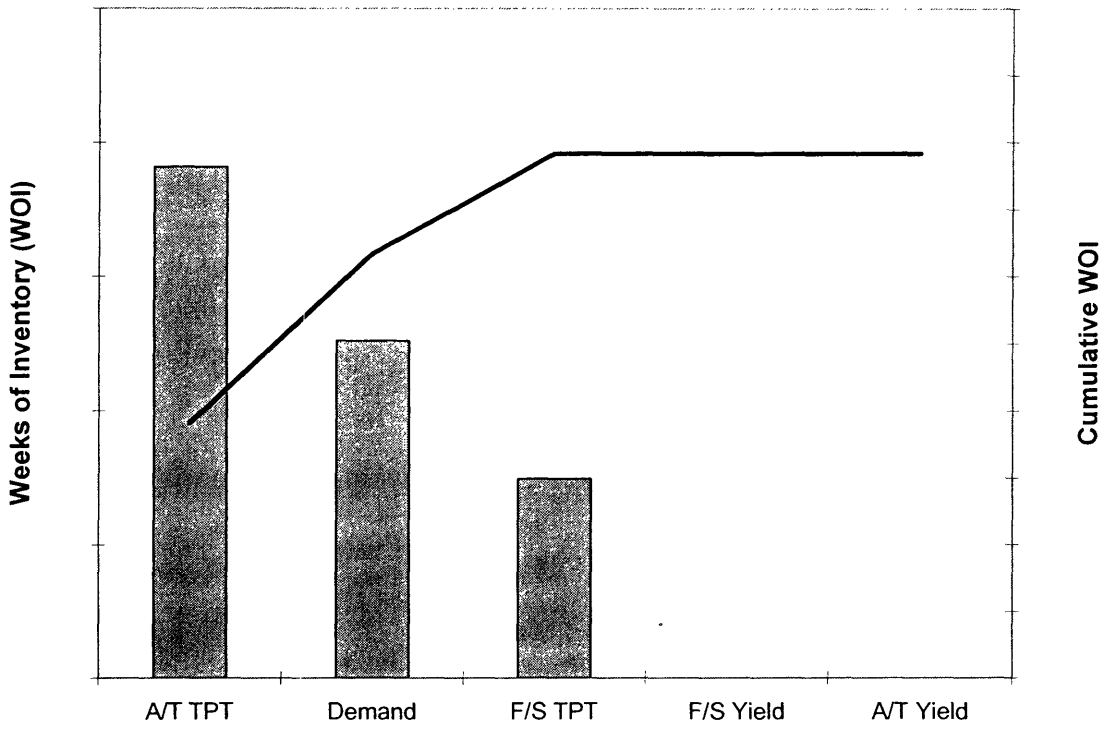
This "relative impact of means" chart clearly shows what is evident from looking at the base-stock equations - that the average demand has the greatest impact on the amount of inventory that must be held to meet a particular service level. Here, however we see the impact that review period can have, as the review period, measured in weeks has a larger impact on inventory than mean TPT values.

The sensitivity analysis described above is effective in describing the changes in inventory that result from differences in mean and variability of supply chain parameters. However, the results are highly dependant on the original calculated means and variances, since these are the numbers that are used to calculate differences. Such an analysis can help to focus improvement efforts on the sources of variability that are highest and thus causing more inventories to be held. While this is a valid analysis, it

does not help supply chain practitioners understand what level of inventory is being held to account for each different source of variability. To make this determination a slightly different approach is required. We take the safety stock portions of the base-stock Equation 4, Equation 5, Equation 6 above, representing inventory requirements adjusted for 1) demand, 2) demand and TPT and 3) demand, TPT and yield respectively. First we calculate the safety stock requirements under Equation 4, which accounts for demand variability. Then we calculate the safety stock with Equation 5. Subtracting the result in Equation 4 from the result in Equation 5 gives the amount of safety stock that is held to buffer from TPT variability. Similarly, subtracting Equation 5 from Equation 6 gives us the amount of stock needed to buffer from yield variability.[27] The result of this analysis for both A/T and F/S in the base case two-node model can be displayed as a Pareto chart with the highest contributors to total safety stock on the left side of the diagram and the smallest contributors on the right.

[27] Note that this methodology is required, since the square root function is not distributive. That is, $\text{sqrt}(a+b) \neq \text{sqrt}(a) + \text{sqrt}(b)$.

Figure 4-7. Pareto Chart of Base Case Results



As you can see in Figure 4-7 above, in the base case scenario, A/T TPT variability is the largest contributor to safety stock, with demand variability and F/S TPT as the next largest. Yield variability hardly plays a role in safety stock requirements. This is because the product analyzed is a mature product with high yields and small yield variance.

4.2.3 Effect of Data Sources on Safety Stock Requirements

We utilize the methodology described above to expand the base case model – and associated data inputs - to provide a more realistic representation of the actual supply chain. By using other sources of data to model the demand variability[28], we can eliminate certain deficiencies in the base case model. For example, the base case model uses family-level pseudo-CV data, along with volume forecasts to estimate demand

[28] Note that the supply variability data feeds are held constant in this analysis.

variability. The usage of family level product aggregation data implies that each sku within a given family is interchangeable with any other sku.[29] To the extent that skus are not interchangeable, the variability will be higher and the inventory requirements will be higher as well. Using sku level data may provide more realistic inventory requirements, particularly for A/T; where skus are not necessarily interchangeable. If one substitutes sku-level demand variability in the place of family-level variability in the base case model, the total safety stock required more than doubles from 3.9 to 8.2 weeks of inventory. See Appendix G – Table of Inventory Analysis Results for complete model results.

Furthermore, there are some cases where neither skus nor families are the appropriate level of aggregation from the customer's perspective. Consider a situation where the customer wants a desktop, Pentium® IV processor, but they may not care whether the unit is 2.2 GHz or 2.4 GHz. In this case, there is a level of aggregation somewhere between sku and family that represents the customers indifference toward a particular subset within a given product grouping. We call this level of aggregation a "mini-family". It is described as a unique combination of family[30], cache, vertical[31], market[32], brand[33], media[34] and package[35]. When the base case model is run using this level of aggregation, the calculated safety stock requirements are roughly 5.6 WOI,

---

[29] In later sections, different levels of product aggregation are used in the model, including certain cases where the two nodes use different levels of aggregation.
[30] Family refers to the internal product generation, and often corresponds with the brand name.
[31] Vertical refers to the type of computer a processor will be used in, such as desktop, mobile or server.
[32] Market refers to the market segmentation of he product, such as performance or value
[33] Brand is a description of the brand name of the product, like Pentium® or Xeon.
[34] Media describes how the unit is sold – either in trays to OEMs or in boxes to consumers.
[35] The package type represents how the microprocessor is mounted for installation into computers.

which lies between the results obtained using sku and family level data. See Appendix G – Table of Inventory Analysis Results for complete model results.

Another shortcoming of the base case model is the fact that the demand variability is assumed to be transmitted up the supply chain. As described in Section 4.1.2, the base case model simply adjusts the demand (backlog) variability for the yield variability in A/T to develop an estimate of F/S demand variability. Since the lead time in A/T is approximately 3 weeks, we use the 3-week-out backlog variability. So, in effect, the error of the 3-week-out backlog variability is transmitted up the supply chain – adjusted only for A/T yield - even though other demand signals are used to make F/S decisions in the real business processes. To address this issue, we can use a variety of different backlog signals to provide more representative variability input to the Fab/Sort part of the process. If we assume that the average lead time from the start of F/S though the end of the process is roughly 11 weeks, we can use the 11-week-out backlog variability data in the model. Because the variability of backlog 11-week-out is greater than the 3-week-out variability, the required safety stock is calculated to be 4.3 WOI rather than 3.9 WOI for the base case model. See Appendix G – Table of Inventory Analysis Results for complete model results.

Though the 11-week-out backlog data may be a more realistic representation of variability in the semiconductor supply chain, there are still some sub-optimal features of this approach. First, as discussed in Section 3.3.2, customers often don't book their backlog until a few weeks into the quarter, so the 11-week-out backlog signal may overestimate variability somewhat. Second, there is more flexibility in the manufacturing process than the 11-week-out forecast implies. It is not as if product is started 11 weeks

in the future - to meet projected backlog - and then the planners helplessly watch as orders change until there are nothing but unfilled orders and incorrect inventory at the end of the quarter. In fact, the product mix can be changed somewhat within the process, in order to adapt to changing product demands. So the 11-week-out forecast is not the only one which has any relevance for planning purpose. It is also true that the 1-week-out forecast is not appropriate, since many of the decisions made earlier will have caused products to be frozen into one configuration or another by then. Instead, some combination of the pseudo-CVs through the forecast horizon should be used. The simplest combination would be a simple arithmetic average of CV values across the forecast horizon. However, this does not account for the fact that the further-out horizons are relied-on more heavily than the near-term horizons. For example, the backlog variability of the 8-week-out horizon has more impact on inventory requirements than the 3-week-out horizon for the simple reason that by the time the product is 3 weeks away from delivery, many sources of flexibility that one has at 11-weeks-out have been depleted. Thus, if the lead time is 11 weeks, the 11-week-out value should be weighted more than the 3-week-out. We make the assertion that, in this case, the ratio should be weighted by time into the future. An example of this calculation is shown in Equation 7 below.

Equation 7. Weighted Average CV Calculation

$$\text{Average CV}_{\text{3WeeksOut}} = \frac{(CV_{1WeekOut} * 1) + (CV_{2WeekOut} * 2) + (CV_{3WeekOut} * 3)}{1 + 2 + 3}$$

When we calculate the "average" variability and use the 3-week-out weighted average for A/T and 11-week-out weighted average for F/S, we get results that are

slightly lower than previously calculated. The sku-level safety stock requirements are calculated to be 8.3 WOI rather than 8.6 WOI. See Appendix G – Table of Inventory Analysis Results for complete model results.

The business process for planning semiconductor wafer starts is a little more complex than the one implied by the model, even with the above upgrades. In particular, multiple demand signals are used for a variety of different planning purposes. Since the 11-week-out backlog often underestimates demand and overestimates variability, forecasts are typically used to plan long lead time events like wafer starts. As discussed in Section 3.3.3, forecasts have an entirely different set of characteristics than backlog, including chronic positive bias and significant variability. Nevertheless, since these forecasts are the demand signal used in the current business planning context, they should be incorporated into the inventory models. This is particularly true in sizing the ADI buffer since it shields the customer from variability resulting from long lead time processes like wafer starts. When one uses the average 13-week-out[36] forecast variability to buffer F/S from demand fluctuations rather than backlog variability, the safety stock requirements (based on family level aggregation) are calculated to be 3.8, which is slightly less than the base case requirement of 3.9 weeks. See Appendix G – Table of Inventory Analysis Results for complete model results.

So far we have evaluated safety stock requirements with different sources of data (backlog or forecasts), different analytical approaches to the time horizon (point-values or weighted-averages) and different levels of product aggregation (sku, mini-family or family). Though we have looked at the differences between sku-level data and family-

[36] The weighted-average 13-week-out forecast variability is used because only quarterly forecast data is available and it must be disaggregated before use in the inventory model.

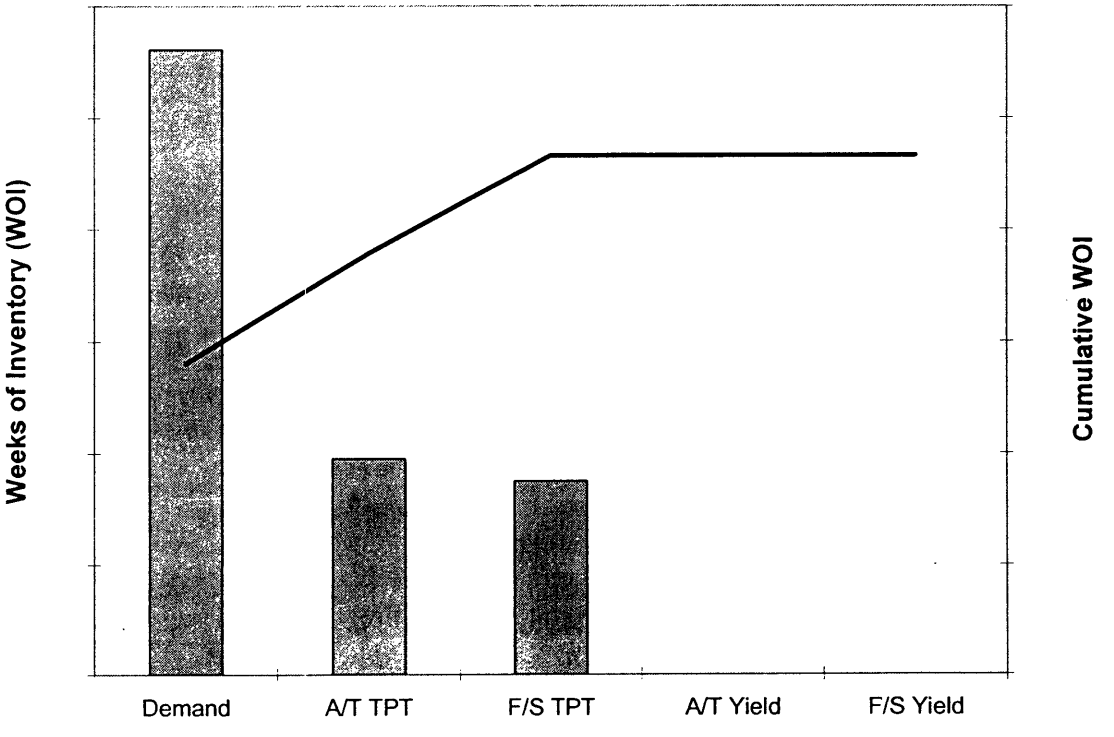
level data, we have always used the same level of product aggregation for each node. However, the same level of aggregation may not be appropriate for each of the nodes in the supply network. The level of product aggregation is a way of describing interchangeability and substitutability. If demand for one sku can not be met by another sku, then sku-level aggregation is appropriate. On the other hand, if a customer is indifferent to one sku within a mini-family and another sku within the same mini-family, then mini-family-level of aggregation is probably appropriate. The fact that the two nodes of our network have different requirements, demand patterns and levels of flexibility indicates differences in the required level of aggregation.

The Fab/Sort process is driven by wafer starts which are roughly equivalent to a product family level of aggregation. After a wafer is started, it can be routed any number of different ways through the factory, such that demand for different skus can be fulfilled. Furthermore, many of the final customizations which remove this flexibility occur in Assembly/Test. So for the purposes of inventory analysis, the ADI inventory buffer, which protects A/T from variability in F/S, the appropriate level of aggregation is likely to be family. Since package-type and other attributes are finally determined in A/T, the appropriate level of aggregation may either sku or mini-family, depending on the end-customer. Some customers can accept a variety of skus within a given mini-family, while others have qualified their equipment on a particular sku and do not have such flexibility. Using the sku level of aggregation for A/T and family level for F/S, the required safety stock is 4.7 weeks. Using the mini-family level of aggregation for A/T and family for F/S the required safety stock is 4.1. These are compared with the base case safety stock

requirement of 3.9 WOI. See Appendix G – Table of Inventory Analysis Results for complete model results.

By refining the base case model to include more realistic assumptions, actual demand signals and current business practices, we have a better understanding of how safety stock changes with different data sources and techniques. It is interesting to note that the sources of the largest components of variability actually changed between the base case and the final analysis. In the base case model, A/T TPT was the variability component which caused the most amount of safety stock to be held, while in the final scenario, demand variability was almost three times larger a contributor to safety stock than A/T TPT.

Figure 4-8. Pareto Chart of Final Model Results



As you can see in Figure 4-8, Demand variability (through both F/S and A/T) contributes nearly 3 of the 4.7 weeks of total inventory. A/T and F/S variability contribute nearly equivalent amounts – 1 week each. Again, this result meshes rather well with our intuitive understanding of the supply chain in that the majority or inventory is held to buffer from variability in demand. In an industry with such volatile and cyclical demand patterns, this is hardly a surprising result.

5 CONCLUSIONS, RECOMMENDATIONS AND FUTURE WORK

The analysis of variability in the semiconductor supply chain, completed from June through December 2003, accomplishes two objectives not often seen together in the supply chain literature. First, it presents a variability framework within which to understand the impact of aggregation and forecast horizon on supply chain variability of short lifecycle products. This enhanced insight can help to identify particular sources of variability to target for reduction or evaluate where inventory should be held to buffer from particular sources of variability. Second, it utilizes real supply and demand data to establish relationships between variability and inventory requirements. The use of real data helps validate the model against experience and allows for its use in running what-if scenarios to determine the inventory impact – and thus the financial implications – of undertaking projects to lower variability in the supply network. Based on the evaluation of the data and scenarios, we present a series of recommendations, broken into two major categories.

5.1 MEASURE, REDUCE AND MANAGE SUPPLY NETWORK VARIABILITY

The analysis of variability discussed in Chapter 3 forms the foundation for the inventory analysis and makes clear the need for a supply network variability strategy. Thus, we recommend that the supply chain group embark on an ambitious program to measure variability, reduce it where possible and manage the amount that remains. The reason for the recommendation is that such action is required to provide the data requirements needed to implement the kind of quantitative inventory program described

in the second recommendation. Again, the key pillars of this variability strategy are Measure, Reduce and Manage.

5.1.1 Measure

The first part of the proposed variability program is to monitor the variability in the supply chain as you would in a manufacturing process. This is a prerequisite for all quantitative inventory efforts. First, the company must collect data - forecasts and actuals - for yield, throughput time, demand and any other parameters of interest. The data should include specific information on which type of forecast – targets, goals or true forecasts - are being stored. Next, this data should be analyzed using the framework contained in this thesis and then evaluated via the metrics described above. Such analysis will provide the data needed for models and it will allow for comparison among projects, products and programs. Tracking the metrics described in this thesis, flagging exceptions, investigating excursions and developing recommendations to prevent recurrence will spur a continuous improvement loop which will help reduce the variability of the supply and demand processes.

5.1.2 Reduce

The second step of the variability strategy is to reduce the supply and demand variability in the network. This does not necessarily imply changes to the production process or sales and marketing efforts. Rather, with an understanding of the variability inherent in the system, one can begin to simplify analyses by using less detailed data where possible. Whether this is a reduction in the time horizon of planning or a change in the time or product granularity, lower variability and less non-value added work will be the result. Intel has already seen dramatic improvements by reducing the time horizon on

planning wafer starts from 9 months to 3 months without adverse effect. There are many other opportunities for such a reduction. For example, in most cases, the customer is indifferent to a variety of skus within a given mini-family. Thus if supply and demand were managed at the mini-family level through A/T, a significant (~25%[37]) reduction in inventory could be immediately realized over targeting inventory at the sku level.

5.1.3 Manage

Once variability in the supply network has been measured and reduced where possible, some amount will remain. In order to manage the remaining levels of variability, the company must begin to account for the variability inherent in the supply chain explicitly. Rather than one group padding demand forecasts and another group underestimating yield predictions, groups should do their best to accurately predict the future state of the business and rely on the variability data described in Section 5.1.1 to appropriately buffer the organization from the variability in the process. As part of this exercise, the planning groups should develop control limits or range forecasts so that they do not react to noise.

5.2 PARADIGM SHIFT FROM JUDGMENT TO DATA FOR INVENTORY MANAGEMENT

The inventory analysis part of this work - made possible by the variability characterization - showed that the Intel's heuristic and judgment-based inventory policies are reasonably close to those calculated through use of the two-node base-stock model

[37] The average reduction in inventory is 25% and ranges from 20% to 30%.

with actual data. However, it is the implementation of these targets - rather than the amounts themselves - that must be re-evaluated.

5.2.1 Calculate and Utilize Inventory Targets

Intel's historical inventory management strategy of "don't stock out" has historically worked, but only with large amounts of management, oversight and manipulation. In the new environment of lower profit margins and more complex products, such management is not only undesirable, but in some cases it is impossible. A more quantitative approach is required and the variability characterization together with the inventory analysis will allow the company to "set a service level, and calculate an inventory target". Again, this will allow the supply network to explicitly account for variability rather than padding it into the management of the business processes. In order to implement this recommendation, better measurement of service levels are required[38] and automated supply and demand data feeds are needed.

5.2.2 Attribute-based Inventory Targets

As part of the paradigm shift from heuristics to statistics, inventory targets should be calculated based on service level and variability data specific to a given products attributes rather than specific to the item itself. In contrast to keeping a given number of weeks of inventory for each sku in the product family, the variability of individual products should be analyzed and inventory targets set based on the salient attributes which are known to affect variability and service level. For example, variability analysis could be set based on stage of product lifecycle (ramp-up vs. end-of-life), relative

---

[38] See Jim Chows 2004 LFM Masters Thesis for detail and recommendation for improving the measurement and implementation of service levels in Intel's supply network.

volumes (high-volume vs. low-volume) and market attributes (desktop-value vs. mobile-performance). By doing this, the company can explicitly link variability to the major variability differentiators.

5.2.3 Work toward Global Inventory Optimization

The goal of this work and the supply network as a whole should be to work toward global inventory optimization through data sharing, common tools and collaboration. This thesis is one small step in the journey. Next steps include the development and utilization of a consistent and detailed set of data sources. The company must also develop effective and flexible reports and tools to allow management of this additional complexity. Finally, the team must develop and implement a robust multi-echelon stochastic optimization which can take all of the different types of variability into account and develop customized inventory targets based on attributes of the products. Only then will Intel realize the vision of the supply chain planning team sitting around the table with the data, tools and processes to make globally optimal decisions.

5.3 FUTURE WORK

While we believe that this work forms a strong foundation for future multi-echelon optimization work, we recognize the shortcomings of the model and assumptions under which we were required to develop our conclusions. First and foremost, Intel's supply chain is neither uncapacitated nor are the supply/demand data normal and stationary. Furthermore, there are many more than two nodes in the semiconductor supply chain. However, to the extent that we develop an easily understandable, multi-node model, using actual supply chain data, the model presented allows us to make assessments of the effects of variability and tradeoffs between sources of variability in

the supply network. Follow-on research should include the development of a multi-node network optimization using a commercial supply chain optimization algorithm. However, even commercial applications have shortcomings in the area of non-normal, non stationary parameters and this should be a subject of intense academic research in the future.

6 REFERENCES

Armstrong, Scott, J. and Fred Collopy. "Another Error Measure for Selection of the Best Forecasting Method: The Unbiased Absolute Percentage Error". http://www-marketing.wharton.upenn.edu/forecast/paperpdf/armstrong-unbiasedAPE.pdf. October 2000.

Armstrong, Scott, J. and Fred Collopy. "Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons". International Journal of Forecasting, 8 (1992), 69-80.

Armstrong, Scott, J. and Robert Fildes. "On the Selection of Error Measures For Comparisons Among Forecasting Methods". Journal of Forecasting, 14 (1995), 67-71.

Axsäter, S., Rosling, K., "Notes: Installation vs. Echelon Stock Policies for Multilevel Inventory Control," Management Science 39-10, 1993.

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.

Brown, R.G., "Smoothing, Forecasting and Prediction of Discrete Time Series," Prentice-Hall: Englewood Cliffs, NJ, 1963.

Çakanyildirim, M., Roundy, R., "SeDFAM: Semiconductor Demand Forecast Accuracy Model," IEE Transactions 34, 2002.

Chambers, John C., Mullick, Satinder K., and Smith, Donald D., "How to Choose the Right Forecasting Technique," Harvard Business Review, July-August 1971.

Coughlin, R. Lawrence, III. LFM Master's Thesis. Optimization and Measurement of a World-Wide Supply Chain. MIT Department of Mechanical Engineering and Computer Science, MIT Sloan School of Management 1998

Deming, W. Edwards. Out of the Crisis. Cambridge, MA: MIT Center for Advanced Engineering Studies, 1986.

Deming, W. Edwards. The New Economics for Industry, Government, Education. Cambridge, MA: MIT Center for Advanced Engineering Studies, 1993.

Gilpin, Brian C. LFM Master's Thesis: Management of the Supply Chain in a Rapid Product Development Environment. MIT Department of Electrical Engineering and Computer Science, MIT Sloan School of Management 1995

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 C. "Safety Stocks in Manufacturing Systems" Journal of Manufacturing and Operations Management, 1988, Vol. 1, No. 1, pp. 67-101.

Graves, Stephen C. "Safety Stocks in Manufacturing Systems" Sloan Working Paper WP 1894-87, June 1987

Graves, Stephen C. and Sean P. Willems "Optimizing Strategic Safety Stock Placement in Supply Chains" Working Paper, January 1998

Hanssmann, F. Optimal Inventory Location and Control in Production and Distribution Networks". Operations Research. Vol. 7, No 9, pp 483-498.

Hetzel, William B. LFM Master's Thesis: Cycle Time Reduction and Strategic Inventory Placement Across a Multistage Process. MIT Department of Chemical Engineering and Computer Science, MIT Sloan School of Management 1993.

Ishikawa, Kaoru. What is Total Quality Control? Englewood Cliffs, NJ: Prentice-Hall, Inc., 1985. Jones, Patricia, and Larry Kahaner. Say It and Live It. New York: Doubleday, 1995.

Johnson, George, A. "A Process Control Approach to Forecast Measurement". http://www.estepsoftware.com/ErrorMeasure.htm

Juran, J.M. and Frank M. Gryna. Juran's Quality Control Handbook. 4th edition, Milwaukee, WI: ASQC Quality Press, 1988.

Juran, J.M. Juran on Quality by Design: The New Steps for Planning Quality into Goods and Services. Milwaukee, WI: ASQC Quality Press, 1992.

Kapuscinski, Roman and Sridhar Tayur, "Optimal Policies and Simulation-Based Optimization for Capacitated Production Inventory Systems," in S. Tayur, R. Ganeshan and M. Magazine (eds.), Quantitative Models for Supply Chain Management, Kluwer Academic Publishers, Boston, 1999, chapter 2.

Kilgore, Stacy, "Balancing Supply and Demand" Forrester TechStrategy Report, March 2002.

Margeson, Wesley, D. LFM Master's Thesis: A Forecasting and Inventory Model for Short Lifecycle Products with Seasonal Demand Patterns. MIT Sloan School of Management and MIT Department of Mechanical Engineering and Computer Science, 2003.

Masters, James. Lecture Notes from ESD.260J, Logistics Systems. Demand Forecasting I, Four Fundamental Approaches.

Maudlin, John. "Your Inner Spock" from Thoughts from the Frontline. www.2000wave.com. November 14, 2003.

Miller, Michael P. LFM Masters Thesis: Business System Improvements Through Recognition of Process Variation. MIT Sloan School of Management, MIT Department of Chemical Engineering 1997

Murty, Katta G., "Supply Chain Management in the Computer Industry," Working Paper, University of Michigan Department of Industrial and Operations Engineering, 2000.

Nahmias, Steven. Productions and Operations Analysis. 3rd. Ed. Chicago: Richard D.Irwin, 1997

Sall, John, Ann Lehman and Lee Creighton. JMP Start Statistics. A Guide to Statistics and Data Analysis Using JMP and JMP IN Software. Second Edition. SAS Institute. Duxbury Press. 2001.

Simpson, K. F. "In Process Inventories". Operations Research. Vol. 6. pp.863-873. 1958.

Synder, Ralph, "Forecasting Sales of Slow and Fast Moving Inventories," European Journal of Operational Research 140, 2002.

Vining, Geoffrey, G. Statistical Methods for Engineers. Brooks/Cole Publishing Company. Duxbury Press. 1998.

Willems, Sean Peter MIT PhD Thesis: "Two Papers in Supply Chain Design: Supply Chain Configuration and Part Selection in Multigeneration Products" MIT Sloan School of Management, 1999

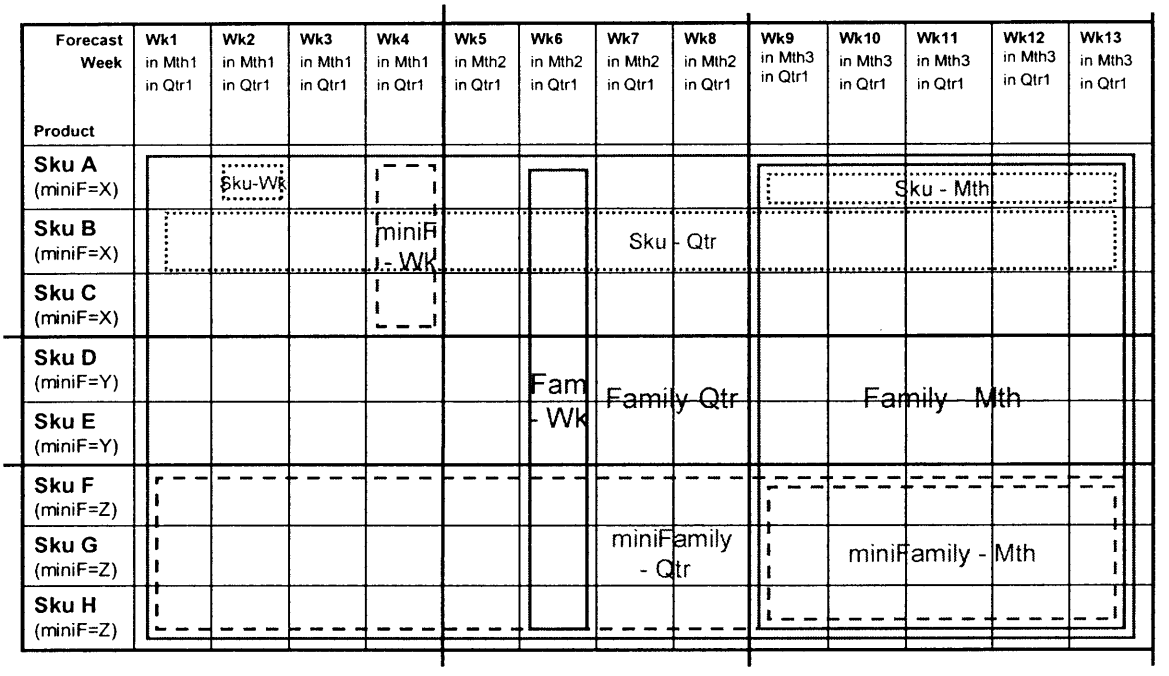
Winters, P.R., "Forecasting Sales by Exponentially Weighted Moving Averages," Management Science 6, 1960.

Zhang, Feng, Aliza Heching, Sarah Hood, Jonathan Hosking, Ying-tat Leung, Robin Roundy and Johnathan Wong. Industrial Data Experiments on Demand Forecasts Combination. Working Paper, School of Operations Research and Industrial Engineering, Cornell University. 2003.

Zimmerman, Hans-Jurgen and Michael G. Sovereign. Quantitative Models for Production Management, Prentice-Hall, Inc., Chapter 7, 1974.

APPENDIX A – GRAPHICAL REPRESENTATION OF AGGREGATION AND VARIABILITY TYPE

Aggregation



Legend:

Forecast Week – The week that is being forecasted
Mth1, Mth2, Mth3 – The month-of-quarter
Qtr1 – The quarter-of-year
Product – The specific product that is being forecasted for
Sku (A, B, C, D, E, F, G, H) – The lowest level of product
miniF, miniFamily (X, Y, Z) – Several skus with the similar attributes make up a mini-Family
Fam, Family – Several miniFamilies built from the same wafer type make up a Family

Procedure:

Imagine that each small square in the above matrix is filled with two numbers, a forecast and an actual. If you are using the level of aggregation of sku-weekly, then you calculate the forecast error of each individual square and average them to get the average error.

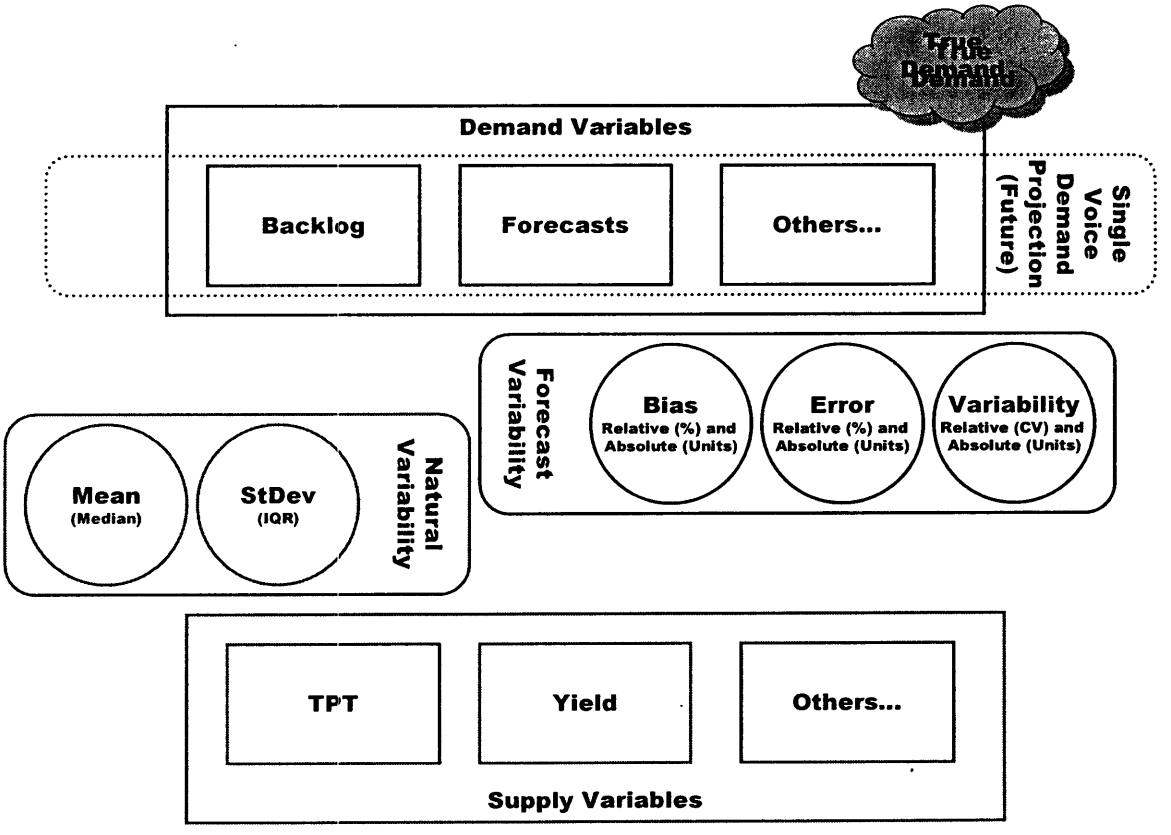
On the other hand, if you use the miniFamily-monthly level of aggregation, then you would first sum up all of the values of forecasts and actuals for a particular miniFamily

and month. Then you take the average of that smaller set of numbers. The average error will be lower than sku-weekly, because certain errors among weeks or between skus tend to cancel each other out when they are aggregated together. This is called the pooling effect.

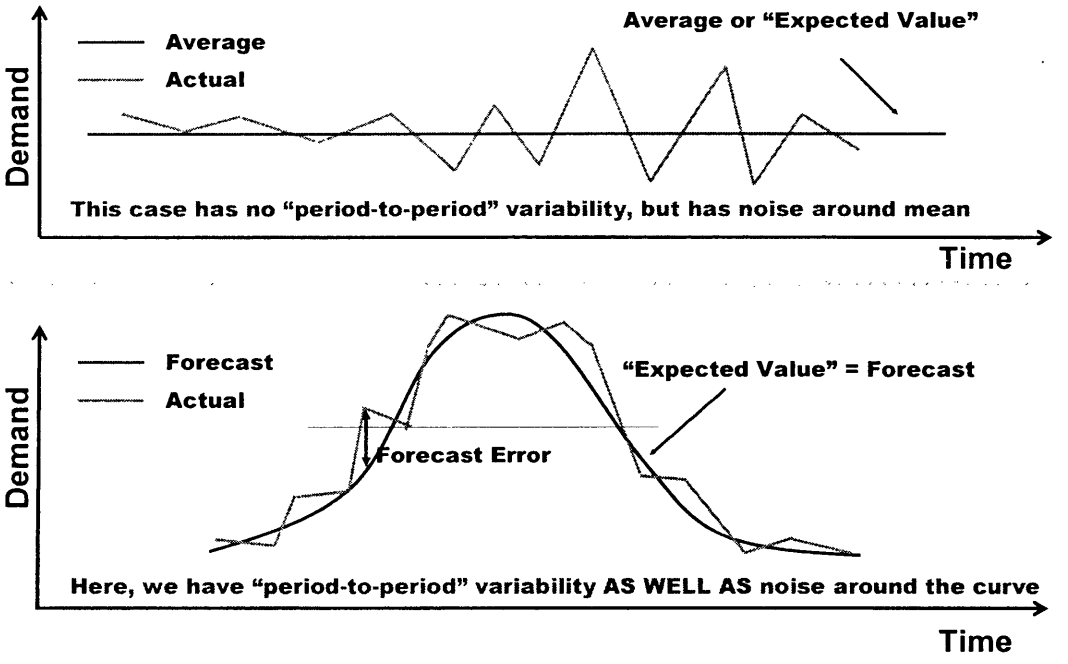
This is the impact of pooling on variability. The different dash and dotted lines indicate, pictorially, which values are summed to achieve different levels of aggregation.

Variability Type

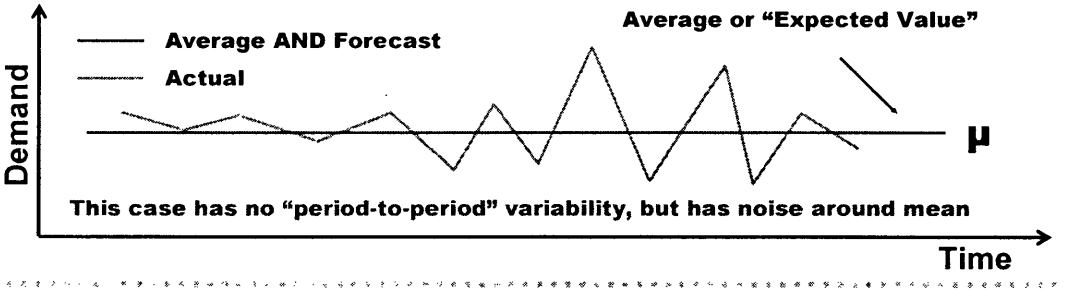
The diagram below helps draw the distinction between natural variability and forecast variability. The boxes contain the supply and demand parameters that are often measured, while the circles show which measurements of variability are primarily used for each type. True demand remains as a cloud hovering over the demand variables, because it is often difficult to identify true demand.

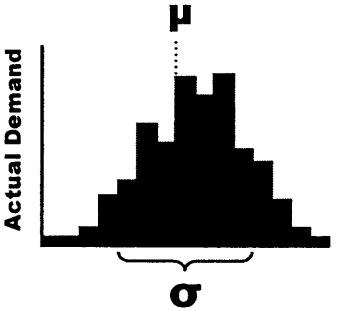


As seen in the picture below, variability can be broken into two parts; period-to-period variability and noise.



For stable demand, mean and standard deviation are key inputs



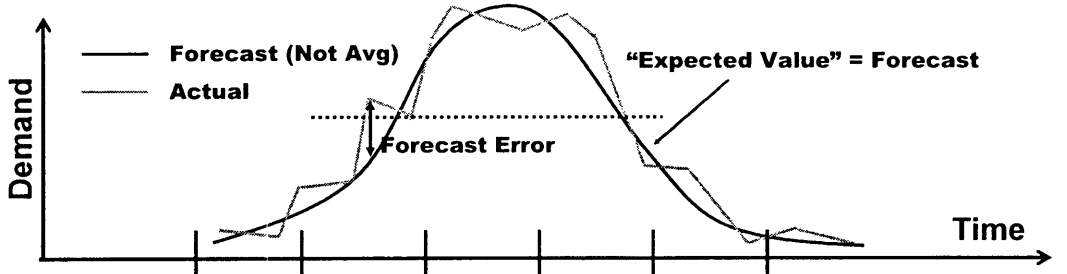


In this case you would enter into a model:

μ = Mean Demand over Time Horizon

σ = StDev of Demand over Time Horizon

For non-stationary demand, more complex inputs are required. Please note that the input data is fictional.



	Wk +0	Wk +4	Wk+8	Wk+12	Wk+16	
Expected Demand *(Similar to μ)*	300	500	1500	1500	800	i.e. Backlog
Average Forecast Error *(Similar to σ)*	150	200	600	700	350	i.e. Avg FE or MD
StDev of Forecast Error *(New Parameter)*	200	200	400	900	600	i.e. StDev of FE

In this case enter just mean and stdev, because the mean (dotted line) is not an accurate representation of expected demand.

Instead, enter expected demand for each week in the forecast horizon and average forecast error for each time horizon. But the forecast error is not deterministic...there is a range of errors, so you need to use the StDev of Forecast Errors.

APPENDIX B – BOUNDED NATURE OF MPE AND APE

Assuming only positive (no returns allowed) forecasts and actuals, take the two extremes:

MPE Case A Forecast = 1 unit, Actual = 1,000,000 units

Abs Error = -999,999 units

MPE = [-999,999 / 1,000,000] *100

MPE = = -99.99% ≈ -100%

MPE Case B Forecast = 1,000,000 units, Actual = 1 unit

Abs Error = 999,999 units

MPE = [999,999 / 1] *100

MPE = 99,999,999%

The reason why you can have unbounded positive errors and negative errors are bounded by -100% (except in the case of negative forecasts and actuals) is simply that you are dividing by actuals. If you divided by forecasts, you would have the opposite scenario. When plotted in a distribution, such data is usually skewed. For a parameter such as MPE, there are many values in the -100 to 100 range, but the existence of large positive outliers (1,000,000% or more) and the absence of large negative outliers to balance them, leaders to excessively high average errors. This is dramatically different from the symmetrical nature of the Average Percent Error measurement shown below. Again, assuming only positive (no returns allowed) forecasts and actuals, take the two extremes:

APE Case A Forecast = 1 unit, Actual = 1,000,000 units

Abs Error = -999,999 units

APE = [-999,999 / 500,000] *100

APE = -199.99% ≈ -200%

APE Case B Forecast = 1,000,000 units, Actual = 1 unit

Abs Error = 999,999 units

MPE = [999,999 / 500,000] *100

MPE = 200 %

APPENDIX C – EQUIVALENCE OF MSE AND STDDEVOFFE

Standard Deviation of Forecast Errors

Equation 8: Forecast Error

Forecast Error (FE) = (Forecast – Actual)

Equation 9: Standard Deviation

$$\text{Std Dev} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Where y is any variable:

Equation 10: Standard Deviation of Forecast Errors

We apply y=Average(FE):

$$\text{Std Dev of FE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(FE_i - \overline{FE}\right)^2}$$

If average forecast bias is 0:

Equation 11: Standard Deviation of FE with Zero Average Bias

$$\text{Std Dev of FE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(FE_i\right)^2}$$

Equation 12: Final Equation for Standard Deviation of Forecast Errors with Zero Average Bias

Set FE = F-A = Forecast-Actual:

$$\text{Std Dev of FE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(F_i - A_i\right)^2}$$

Mean Square Error

Equation 13: Mean Squared Error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(F_t - A_t)^2$$

Equation 14: Root Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_t - A_t)^2}$$

Note that Equation 12 is identical to Equation 14 for large n.

APPENDIX D – DISAGGREGATION OF QUARTERLY FORECASTS

Fcst Month / Plan Month	July (M1 in Q3)	August (M2 in Q3)	September (M3 in Q3)	October (M1 in Q4)	November (M2 in Q4)	December (M3 in Q4)
Jul (M1 in Q3)	MIF=0 $F_{calc}=0.3 * F_q$ **299,400**	MIF=1 $F_{calc}=0.3 * F_q$ **299,400**	MIF=2 $F_{calc}=0.4 * F_q$ **399,200**	MIF=3 $F_{calc}=0.3 * F_q$ **450,000**	MIF=4 $F_{calc}=0.3 * F_q$ **450,000**	MIF=5 $F_{calc}=0.4 * F_q$ **600,000**
	F_{q13} = 998,000 (QIF = 2/3)			**F_{q14} = 1,500,000** (QIF = 5/3)		
Aug (M2 in Q3)	MIF=-1 ***A1 = 265,824***	MIF=0 $F_{calc}=(F_q - A_1)*3/7$ **324,933**	MIF=1 $F_{calc}=(F_q - A_1)*4/7$ **433,243**	MIF=3 $F_{calc}=0.3 * F_q$ **486,000**	MIF=4 $F_{calc}=0.3 * F_q$ **486,000**	MIF=5 $F_{calc}=0.4 * F_q$ **648,000**
	F_{q13} = 1,024,000 (QIF = 1/3)			**F_{q14} = 1,620,000** (QIF = 4/3)		
Sep (M3 in Q3)	MIF=-2 ***A1 = 265,824***	MIF=-1 ***A2 = 269,954***	MIF=0 $F_{calc}=(F_q - A_1 - A_2)$ **369,222**	MIF=3 $F_{calc}=0.3 * F_q$ **438,900**	MIF=4 $F_{calc}=0.3 * F_q$ **438,900**	MIF=5 $F_{calc}=0.4 * F_q$ **585,200**
	F_{q13} = 905,000 (QIF = 0/3)			**F_{q14} = 1,463,000** (QIF = 3/3)		

Legend:

Fcst Month – The month that is being forecasted
Plan Month – The month in which the plan was made
M1, M2, M3 – The month-of-quarter
Q3 – The quarter-of-year
MIF – Months into Future
QIF – Quarters into Future
F_{calc} – Calculated forecast for a given month
F_q, F_{q13} – The quarterly forecast which is being disaggregated
A1, A2, A3 – The backward-looking actual demand for the given month

Procedure:

First Month-of-Quarter

For every month in the forecast horizon, the quarterly forecast is disaggregated by multiplying the quarterly forecast by the seasonality factors for each month-of-quarter. In this example, 30% for 1st month, 30% for 2nd month and 40% for 3rd month. Any forecast more than 1-quarter-out remains disaggregated by the 30/30/40 seasonality factors.

Second Month-of-Quarter

In the second month of the quarter, there are actual demand results available for the first month. We realign the forecasts by subtracting the first month result from the quarterly forecast and then multiplying the results by the adjusted seasonality factors. In this example, 0.3 / [1.0 - 0.3] (=0.43) for the second month and 0.4 / [1.0 – 0.3] (=0.57) for the third month. Any forecast more than 1-quarter-out remains disaggregated by the 30/30/40 seasonality factors.

Third Month-of-Quarter

In the third, and last, month of the quarter, the actual results for the first and second months are available, so the quarterly forecast is realigned to these actual results. The actuals are subtracted from the quarterly forecast (which may be revised from the prior two months) and the remainder is the new forecast for the third month. Any forecast more than 1-quarter-out remains disaggregated by the 30/30/40 seasonality factors.

APPENDIX E – DERIVATION OF F/S DEMAND VARIABILITY

Let $Y = X / Z$

Where: Y = F/S Demand
X = A/T Demand
Z = A/T Yield

$E(Y) = E_z [E(X \mid Z)]$
$= E_z [\mu_x / Z]$
$= \mu_x E_z [1 / Z]$

Equation 15: Yield-Adjusted F/S Demand

$$\mathbf{E(Y) = \mu_x E_z [1 / Z]}$$

$Var(Y) = E_z[Var\ Y \mid Z] + Var_z[E(Y \mid Z)]$
$= E_z[\sigma_x^2 / z] + Var_z[\mu_x / z]$
$= \sigma_x^2 * E[1 / Z] + \mu_x^2\ Var_z[1 / Z]$

Equation 16: Variability of Yield-Adjusted F/S Demand

$$\mathbf{Var(Y) = \sigma_x^2 * E[1 / Z] + \mu_x^2\ Var_z[1 / Z]}$$

APPENDIX F – DESCRIPTION OF DEMAND VARIABILITY REGRESSION

Though we primarily use the pseudo-CV method of demand variability estimation in our work, we explored many other techniques. One of the techniques which we spent much time and thought on is the regression methodology. The concept is that the level of demand variability should be related to the level of actual demand. Thus one can regress variability, measured as Mean Square Error (MSE), against demand on a log scale and obtain an equation which could be used to estimate variability given a new level of demand. In this case, sku level, weekly demand was used to develop such an equation.

However, as we have seen in this work, the variability of demand is highly dependant on other factors as well. In particular, the level of time and product aggregation has a dramatic impact, as well as the forecast horizon. For the case of time aggregation, we judged the usefulness of monthly or quarterly demand variability to be low, so we neglected to account for this effect. However, the level of product aggregation and forecast horizon were judged to have significant theoretical and practical impact and were included in subsequent analyses. We included the log of the number of weeks into the future as a dependant variable in the regression, and a separate regression was done for each level of product aggregation. The regression details are detailed at the end of this Appendix and the resulting equation for sku level variability follows.

Equation 17: Regression Equation for Sku Level Demand Variability

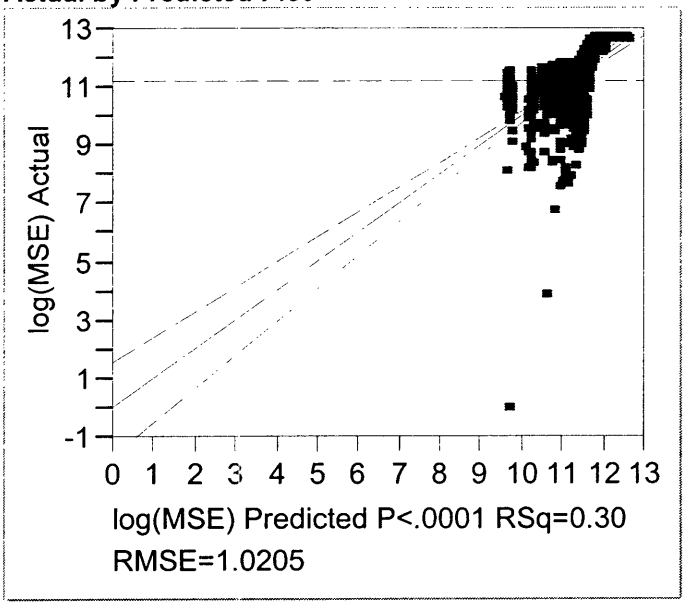
$$\text{Estimated MSE} = 10^{c} \times WeeksIntoFuture^{1.80} \times WklyDemand^{1.64}$$

Where: c = a constant[39]

The results given by this equation consistently under-represent the variability we see in the empirical data and the results generated by the pseudo-CV method. We believe that this is due to the large scatter and poor R^2 of the sku-level regression and the family and mini-family regressions are worse. This is why we did not use the regression results to complete the inventory analysis, even though we believe that the technique holds much promise in helping estimate variability for such analyses.

[39] The constant is not given in this treatment in order to protect confidentiality of the data.

Level_Of_Prod_Aggr=Family, Level_Of_Time_Aggr=Week
Response log(MSE)
Actual by Predicted Plot



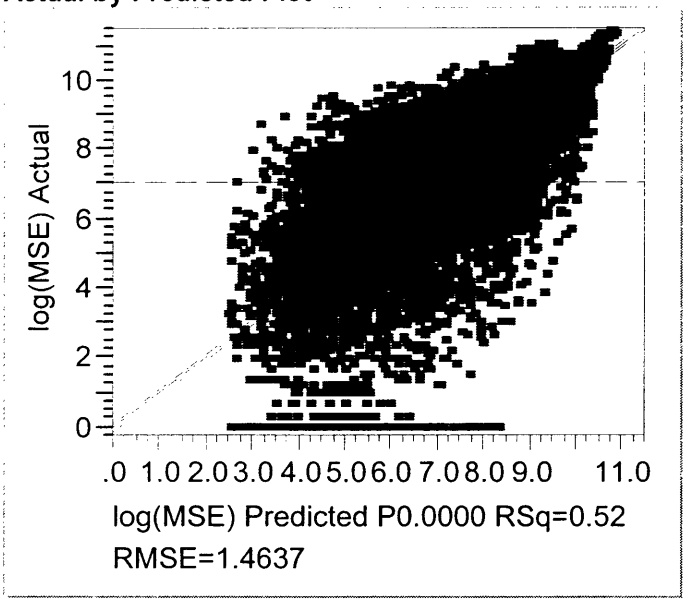
Summary of Fit

RSquare	0.295907
RSquare Adj	0.293869
Root Mean Square Error	1.020528
Mean of Response	11.19823
Observations (or Sum Wgts)	694

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	302.4496	151.225	145.2022
Error	691	719.6608	1.041	Prob > F
C. Total	693	1022.1103		<.0001

Level_Of_Prod_Aggr=Sku, Level_Of_Time_Aggr=Week
Response log(MSE)
Actual by Predicted Plot



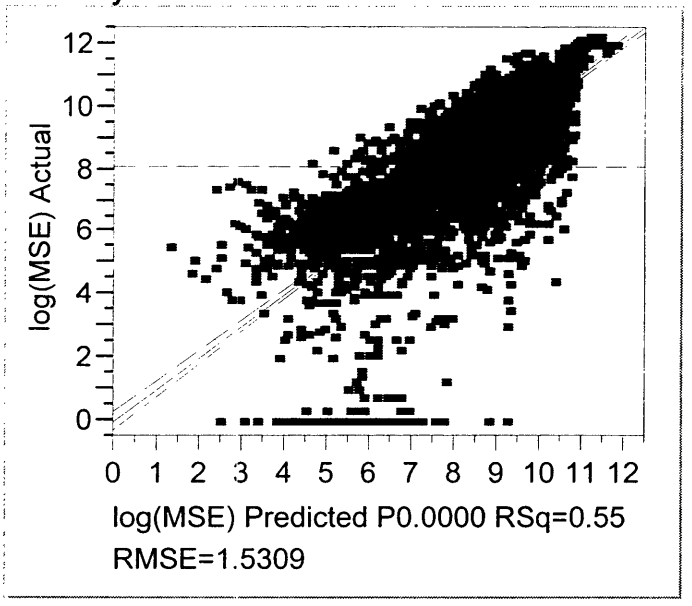
Summary of Fit

RSquare	0.520025
RSquare Adj	0.519982
Root Mean Square Error	1.463696
Mean of Response	7.092724
Observations (or Sum Wgts)	22284

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	51718.036	25859.0	12070.08
Error	22281	47734.948	2.1	Prob > F
C. Total	22283	99452.984		0.0000

Level_Of_Prod_Aggr=miniFamily, Level_Of_Time_Aggr=Week
Response log(MSE)
Actual by Predicted Plot



Summary of Fit

RSquare	0.554485
RSquare Adj	0.554305
Root Mean Square Error	1.530926
Mean of Response	8.088953
Observations (or Sum Wgts)	4960

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	14459.519	7229.76	3084.719
Error	4957	11617.887	2.34	Prob > F
C. Total	4959	26077.406		0.0000

APPENDIX G – TABLE OF INVENTORY ANALYSIS RESULTS

Total Safety Stock Requirements

F/S Data Backlog	A/T Data Backlog	3 Wk Out for F/S 3 Wk Out for A/T	11 Wk Out for F/S 3 Wk Out for A/T	0-11 Wk Out (Avg) for F/S 0-3 Wk Out (Avg) for A/T
Sku	Sku	8.2	8.6	8.3
mini-Family	mini-Family	5.6	5.3	5.6
Family	Family	3.9	4.3	4.1

F/S Data Forecasts	A/T Data Backlog	9 Wk Out for F/S 0-3 Wk Out (Avg) for A/T	13 Wk Out (Avg) for F/S 0-3 Wk Out (Avg) for A/T
Sku	Sku	9.1	8.4
mini-Family	mini-Family	5.0	5.0
Family	Family	3.8	3.8

F/S Data Forecasts	A/T Data Backlog	13 Wk Out (Avg) for F/S 0-3 Wk Out (Avg) for A/T
mini-Family	Sku	5.5
Family	Sku	4.7
Family	mini-Family	4.1

Base Case Model

Final Version Model

Safety Stock Requirements for F/S and A/T

F/S Data Backlog	A/T Data Backlog	3 Wk Out for F/S 3 Wk Out for A/T	11 Wk Out for F/S 3 Wk Out for A/T	0-11 Wk Out (Avg) for F/S 0-3 Wk Out (Avg) for A/T
Sku	Sku	A/T=4.8 | F/S=3.4	A/T=5.2 | F/S=3.4	A/T=5.1 | F/S=3.2
mini-Family	mini-Family	A/T=2.9 | F/S=2.7	A/T=2.6 | F/S=2.7	A/T=3 | F/S=2.6
Family	Family	A/T=1.6 | F/S=2.3	A/T=1.9 | F/S=2.3	A/T=1.7 | F/S=2.3

F/S Data Forecasts	A/T Data Backlog	9 Wk Out for F/S 0-3 Wk Out (Avg) for A/T	13 Wk Out (Avg) for F/S 0-3 Wk Out (Avg) for A/T
Sku	Sku	A/T=5.9 | F/S=3.2	A/T=5.2 | F/S=3.2
mini-Family	mini-Family	A/T=2.4 | F/S=2.6	A/T=2.3 | F/S=2.6
Family	Family	A/T=1.5 | F/S=2.3	A/T=1.5 | F/S=2.3

F/S Data Forecasts	A/T Data Backlog	13 Wk Out (Avg) for F/S 0-3 Wk Out (Avg) for A/T
mini-Family	Sku	A/T=2.3 | F/S=3.2
Family	Sku	A/T=1.5 | F/S=3.2
Family	mini-Family	A/T=1.5 | F/S=2.6

Base Case Model

Final Version Model