Strategies for Manufacturing Low Volume Semiconductor Products in a High Volume Manufacturing Environment

by

Robert L. Scholtz III
B.S. Mechanical Engineering, Purdue University, 1995

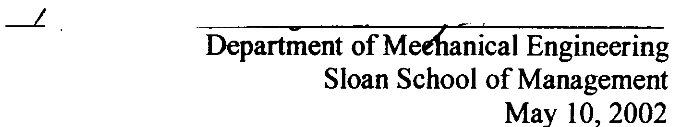
Submitted to the MIT Department of Mechanical Engineering and to the Sloan School of Management in partial fulfillment of the requirements for the Degrees of

Master of Science in Mechanical Engineering
and
Master of Business Administration

at the
Massachusetts Institute of Technology
June 2002

© Massachusetts Institute of Technology, 2002, All rights reserved

Signature of Author



Department of Mechanical Engineering
Sloan School of Management
May 10, 2002

Certified by

Jesús A. del Alamo
Professor of Electrical Engineering
Thesis Supervisor

Certified by

Stephen C. Graves
Leaders for Manufacturing Professor of Management and Engineering
Thesis Supervisor

Certified by

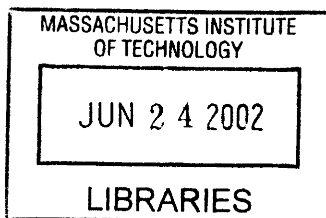
Stanley B. Gershwin
Senior Research Scientist, Department of Mechanical Engineering
Thesis Reader

Accepted by

Ain A. Sonin
Chairman, Department Committee on Graduate Students
Department of Mechanical Engineering

Accepted by

Margaret Andrews
Executive Director of Master's Program
Sloan School of Management



BARKER

Strategies for Manufacturing Low Volume Semiconductor Products in a High Volume Manufacturing Environment

by
Robert L. Scholtz III

Submitted to the MIT Department of Mechanical Engineering and to the Sloan School of Management on May 10, 2002 in partial fulfillment of the requirements for the Degrees of Master of Science in Mechanical Engineering
and
Master of Business Administration

ABSTRACT

The rapid growth of the digital communications market has prompted several large semiconductor manufacturers, including Intel Corporation, to begin the design and manufacture of communication ICs. The communications ICs are currently produced in much lower volumes than products such as microprocessors and memory. These low-volume products have been reported to cause operational problems, such as excessive cost, slow throughput time, and low yield when manufactured in semiconductor fabs designed for high volume manufacturing.

This thesis examines the operational problems caused by the manufacture of low-volume semiconductor products and explores potential improvements. A financial model was developed to compare the cost of manufacturing low-volume products using several different strategies in existing high-volume fabs. The model results demonstrated that mask set cost, a fixed cost, becomes a very large component of total production cost as the product volume is reduced. Further, this model identified multi-product wafers, a scheme of fabricating several products on a single wafer, as a strategy with potential for savings up to approximately 75% of the manufacturing cost of low-volume products.

A second financial model was developed to consider more detailed aspects of fabricating products on multi-product wafers. This model considered the sensitivity of the potential cost savings to changes in demand and changes to the design of multi-product wafers. This model also demonstrated that significant savings are possible with the multi-product wafer strategy, especially if the products are carefully matched (by die size and demand) with other products on the multi-product wafer.

Finally, a brief organizational study was conducted to analyze the implementation of a multi-product wafer manufacturing process for the production of low-volume CMOS ICs at Intel Corporation.

Thesis Supervisors: Jesus del Alamo
Professor of Electrical Engineering

Stephen Graves
Professor of Management

Thesis Reader: Stanley Gershwin
Department of Mechanical Engineering

Acknowledgements

Dedicated to my wife,
Jennifer

I would like to thank my internship supervisor, Brian Kelly, for his advice and support throughout my study of low-volume manufacturing strategies. I would also like to thank Todd Kinkade for joining me in developing the models presented here, Dave Fanger for his technical advice, Chris Richard for his manufacturing strategy advice, and Intel's MPS working group for welcoming me and allowing me to contribute to its charter.

I would like to thank my thesis supervisors Steve Graves and Jesus del Alamo, and my thesis reader Stan Gershwin, for their patient guidance and support as this project has evolved.

I would also like to thank my LFM colleague, Bob Kelly, for many helpful discussions about Intel and its manufacturing processes.

Finally, I would like to thank the Leaders for Manufacturing Program and Intel Corporation for making this work possible.

Table of Contents

ABSTRACT 3

Acknowledgements 4
Table of Contents 5

Chapter 1: A transition to low-volume IC manufacturing 7
1.1 The Problem 8
1.2 Sources of potential solutions 12
1.3 Thesis Structure 12

Chapter 2: Low-volume IC manufacturing 14
2.1 Technical challenges of low-volume IC manufacturing 14
2.1.1 Photolithography 14
2.1.2 Chemical Mechanical Polishing 16
2.1.3 Thins Films 17
2.1.4 Sorting 18
2.1.5 Other processes 18
2.1.6 New Product Introductions 18
2.2 Existing examples of low-volume IC manufacturing 19
2.2.1 "Shuttles" for prototypes and low-volume products 19
2.2.2 Intel's low-volume fabs 21
2.2.3 Mini-fab – Ericsson 23
2.2.4 Work cell layout 24
2.3 Alternatives for low-volume IC manufacturing 24
2.3.1 Multi-product wafers 25
2.3.2 Multi-product lots 27
2.3.3 Reduced NPI effort 28
2.3.4 Small lot size 29
2.4 Conclusions 29

Chapter 3: Financial analysis of potential low-volume manufacturing strategies 30
3.1 Development of a financial model 30
3.1.1 Calculation of the wafer quantity requirement 31
3.1.2 Tailoring the wafer requirement to each LVM strategy 34
3.1.3 Wafer cost components 37
3.1.4 Mask set cost 43
3.1.5 Die cost calculation 45
3.2 Low-Volume Manufacturing Cost Model results 45
3.2.1 Die cost comparisons with increasing die size 47
3.2.2 Impact of die yield on manufacturing strategy 51
3.2.3 Impact of New Product Introductions (NPIs) on wafer cost 53
3.3 Mapping the lowest cost LVM strategies 54
3.4 Conclusions 56

Chapter 4: Development and results of the Multi-product Wafer Cost Model 57
4.1 MPW Cost Model calculations 58
4.1.1 Modeling an MPW with multiple die sizes 58
4.1.2 Optimization of die ratio 59
4.1.3 Total cost calculation 61
4.2 Results of the MPW Cost Model 62
4.2.1 Sample MPW Cost Model results 63
4.2.2 MPW sensitivity to demand changes 66
4.3 MPW Cost Model conclusions 68

Chapter 5: Organizational Challenges of Implementing MPWs 69
5.1 Background of the MPS Working Group 69
5.2 Strategic design lens 69
5.3 Political lens 72
5.4 Cultural lens 73
5.5 Conclusions 74

Chapter 6: Conclusions and recommendations for further research 75

Appendix A: List of Acronyms 78

Appendix B: Low-Volume Cost Model results 79
Comparison of low-volume manufacturing strategies' die costs 79
Yield impact on MPW die cost 81

References 82

Chapter 1: A transition to low-volume IC manufacturing

As the market for personal computers has matured in the past decade, the growth rate of the microprocessor business has slowed considerably. Microprocessor manufacturers, who have become accustomed to strong "double-digit" growth, are being forced to look beyond the processor to find attractive growth opportunities.

Over the past few years, Intel Corporation has been entering the communications markets such as Ethernet and wireless in an attempt to find markets for ICs with high-expected growth. Table 1.1 compares the expected growth and size of the PC market and a few communications equipment markets. While the communications markets are currently relatively miniscule, the growth potential is remarkable.

Table 1.1: Forecast annual growth of selected worldwide technology markets, 2000 – 2006

Market	Annual Growth Rate	2002 Projected Market Size ($-million)
Personal Computers	-10 – 10%	190,000
10 Gigabit Ethernet	221%	714
Metro Ethernet	48%	155
Network Processors	120%	194

Source: Daiwoo Securities[1], InfoTech Trends[2]

BusinessWeek Magazine reported that Intel has pumped more than $10 billion into 34 acquisitions over the past three years. These deals were expected by CEO Craig Barrett to enable business units such as the Communications Group and the Wireless Communications and Computing Group to grow 50% annually [3].

One of the results of these acquisitions has been the creation of a diverse product line that includes both high-value, low-volume components and lower-value, high-volume components. The products are currently produced by a combination of internal and externally subcontracted

manufacturing, with much of the external manufacturing at subcontracted semiconductor foundries being a legacy from the manufacturing strategies of the original acquisitions.

A key strategy with the new businesses is to bring the manufacturing of many of the new products into Intel's fabs. Intel management believes it can achieve several benefits by bringing the manufacturing in house. The new products can benefit from Intel's logic technology, which is considered to be superior to the logic performance offered by the foundries. Additionally, the products are produced with greater confidentiality, thus protecting the intellectual property from competitor's eyes for a longer period. Further, internal manufacturing would simplify the supply chain by reducing the number of suppliers that must be managed.

1.1 The Problem

The shift of the product portfolio to lower volume, high mix products without a change to the manufacturing strategy will potentially cause significant delays and unnecessarily high costs in the production of these products. The product-process matrix as proposed by Hayes and Wheelwright demonstrates that the manufacturing process should match the product being produced. Ideally, the products map into the diagonal of the proposed product-process matrix [4]. This product-process matrix is shown in Figure 1.1.

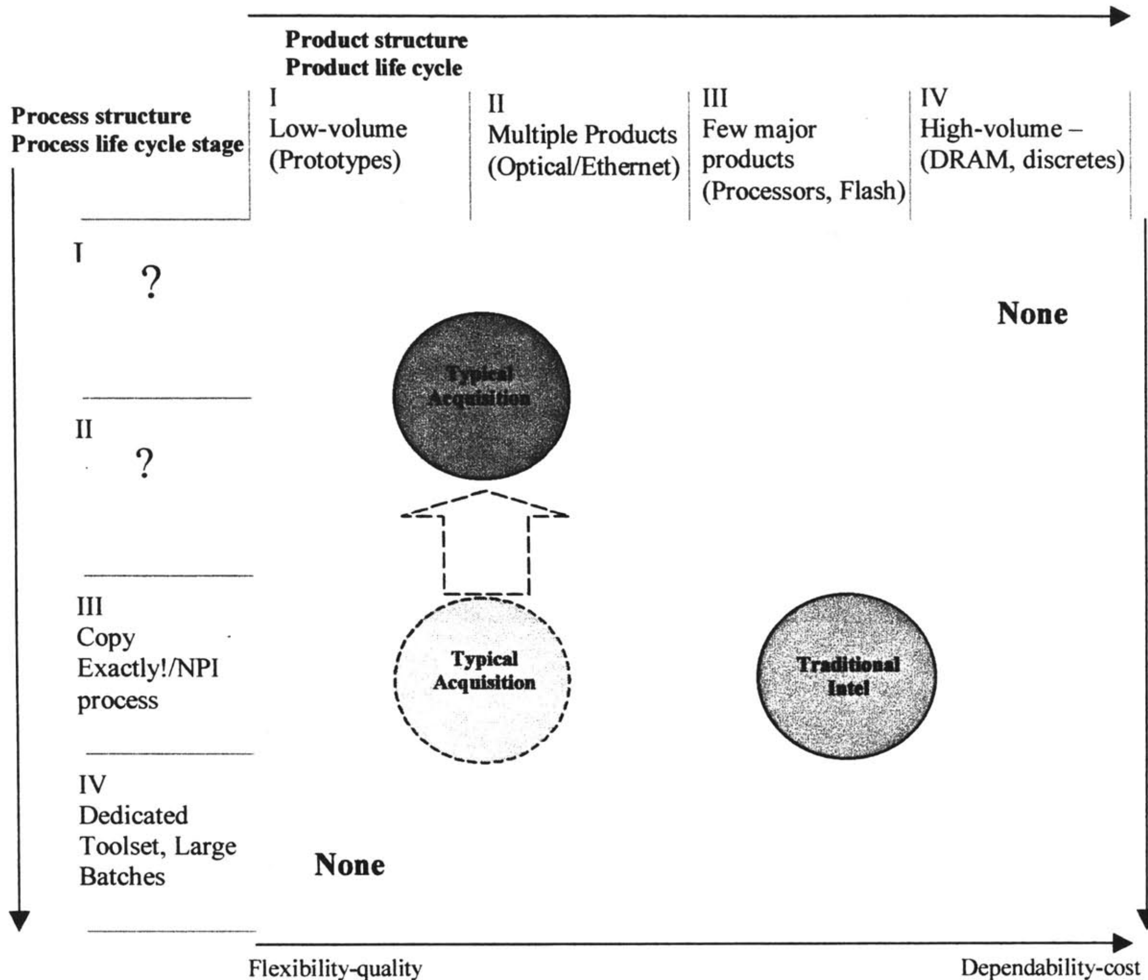


Figure 1.1: The Product-Process matrix adapted to semiconductor manufacturing [4]

Intel has created a manufacturing system that is tailored to the needs of the products it has traditionally manufactured. Those products, primarily microprocessors and flash memory chips, are produced in volumes of millions of units per year and are represented by the "Traditional Intel" circle in the above figure. The production of such high-volume components requires that the manufacturing process be focused on repeatability and high yield. The combination of Intel's Copy Exactly! (CE!) Methodology and New Product Introduction (NPI) process has done just that. Intel's CE! Methodology is the process by which new manufacturing processes are transferred from a development fab to a multiple high-volume manufacturing (HVM) fabs [5]. This rigorous process ensures that identical products can be produced at multiple fabs and that

high yields are achieved quickly during the ramp to high-volume production. The NPI process optimizes the parameters of the manufacturing process to each product. This process enables Intel to achieve high yields, which we will see is a strong lever in reducing manufacturing costs of low-volume products.

Intel's recent entry into communications markets has added new product categories to the product portfolio. The "Typical Acquisition" circle on the product-process matrix of Figure 1 represents these product categories. One can see that these lower volume products, if manufactured with the current process, fall off of the diagonal in which the process is well matched to the process. As suggested by the product-process matrix, it is not ideal to blindly push low-volume products into the same process used by Intel's high-volume products. Ideally, these low-volume products should have a manufacturing process tailored to their specific needs, thus bringing the products into the ideal diagonal of the product-process matrix. But what process should be used? And can Intel's fabs accommodate a fabrication methodology specific to low-volume products? These are some of the questions that this thesis will attempt to address. To begin to explore these questions, the problems of running low-volume components in the existing fabs are described below.

Problems with low-volume products produced in high-volume fabs are caused both by the physical limits of the manufacturing process and the incentives created by management to govern the process. The primary problems are as follows:

> **High setup costs per wafer** – The high mix of products inherent with low-volume production means that more setups will have to be conducted in the fab. Since the manufacturing process is optimized for each product, the tools must be set to different parameters for each product change. Making matters worse, the low-volume wafers cannot spread this setup across a large number of wafers.
>
> **More New Product Introductions (NPIs)** – An NPI is the time-consuming and labor-intensive procedure of optimizing the manufacturing process for each new product. The procedure is designed to maximize the yield of the new product. Low-volume products

suffer two negative effects from NPIs. First, the cost of the additional labor and the lost fab tool time is not spread over millions of units, so the cost per die is high. Second, the low-volumes prevent yield enhancements from significantly improving the manufacturing costs. Yield improvements do not significantly reduce the manufacturing capacity required because so few wafers are manufactured. Thus, a large improvement in yield would only create a relatively small reduction in wafers required.

Slow tool run rates – Low-volume products cause some of the tools to run slower because of the increased inspection required to assure good quality in the early stages of a product's life. High-volume products are able to rely on statistical samples to reduce the number of inspections required by each lot.

Slow throughput time – The combination of the above problems increases the time required for a product to be manufactured in the fab. But this is not all. The incentives used to motivate the workforce and management include metrics such as tool utilization, which are made worse by the aforementioned effects. So, the low-volume products receive a lower priority, which causes them to sit idle in the fab while the higher volume products move through rapidly.

Poor yield – Product yield is often improved over time as learning occurs. With low-volume products, the production run is too short to implement improvements to the process that may improve yield. While it may not make sense to expend the effort to improve the yield of some low-volume products, any low yield numbers tend to raise red flags in the Intel culture where profitability has been tied to yield for many years.

Opportunity Cost – Because many of the problems listed here effectively reduce the capacity of the fab, the opportunity cost of other products that could have been run may become an issue. The fab capacities are typically constrained during the ramp of a new process technology, so any "wasted" tool time is a lost opportunity to produce another product. Due to its complexity, the opportunity cost was considered beyond the scope of this thesis and will not be explored further.

1.2 Sources of potential solutions

This thesis explores and compares the potential of several strategies for manufacturing low-volume microelectronic products in the context of a high-volume IC manufacturer. Potential solutions to the problem of manufacturing low-volume products were collected from a variety of sources. Intel's process and industrial engineers were asked for ideas to improve the current process to accommodate low-volume products. A comparison of the major wafer foundries' processes for manufacturing prototypes was conducted. Other literature suggesting alternative methods of manufacturing semiconductors were also considered by the author.

Some constraints were placed on the range of potential solutions considered to enable Intel to continue its manufacturing process development cycle while manufacturing low-volume products. Thus, the scope of this thesis was limited to the products that can be manufactured in Intel's existing CMOS logic fabs. It was assumed here that the process technologies used by these low-volume communication products would be identical to or relatively minor revisions of the process technologies used to produce microprocessors. It was also assumed that the total combined volume of Intel's low-volume products is too low to justify the development and construction of a fab designed solely for the purpose of manufacturing low-volume wafers.

1.3 Thesis Structure

This thesis is divided into six chapters plus an Appendix. Chapter 2 delves into the technical challenges of producing low-volume products and considers several solutions to these problems. Chapter 3 discusses the development and results of a model used to evaluate various manufacturing strategies for low-volume products. The model, known as the Low-Volume Manufacturing (LVM) Cost Model, enables the user to determine the best manufacturing strategy (of those considered) for a given product. This chapter identified a process called "Multi-Product Wafers" to have potential for significant cost savings. Chapter 4 takes the cost modeling a step further with a more detailed focus on multi-product wafers. Chapter 5 steps

away from the technical issues and considers the organizational challenges that will accompany the implementation of the multi-product wafer strategy for producing low-volume products. Finally, Chapter 6 revisits the Hayes-Wheelwright product-process matrix, summarizes this thesis, and considers topics for further research.

Chapter 2: Low-volume IC manufacturing

As mentioned in Chapter 1 there are many problems that occur in an IC fab when low-volume products are introduced. The introduction of low-volume products into a high-volume wafer fab has different impacts on the various functional areas. The effect is most pronounced in the photolithography area, but other areas are also impacted. This chapter will explore the most dramatic effects low-volume products have on fabs, then will investigate what has been done in the past with low-volume semiconductor product manufacturing. This will allow us to synthesize low-volume manufacturing strategy options for further investigation in this thesis.

2.1 Technical challenges of low-volume IC manufacturing

This section explores the challenges that low-volume products present to the operation of a high-volume semiconductor fabrication facility, or fab. The low-volume products not only affect the performance of the different tools used in IC production, but also some processes such as the New Product Introduction process.

2.1.1 Photolithography

Photolithography is the process of transferring circuit patterns to the silicon wafers. These patterns act as masks to allow ion implantation or etching to be performed on narrowly designated areas of the wafer. The photolithography tools, known as steppers, are the most expensive tools in a fab. Because of this, the steppers are typically the bottlenecks in the fabrication process. As the bottleneck, any wasted time on the stepper is lost fab capacity. Thus fab incentive systems are designed to maximize the stepper utilization. The steppers are designed to rapidly and accurately project patterns on individual wafers. A simple diagram of the photolithography stepping process is shown in Figure 2.1.

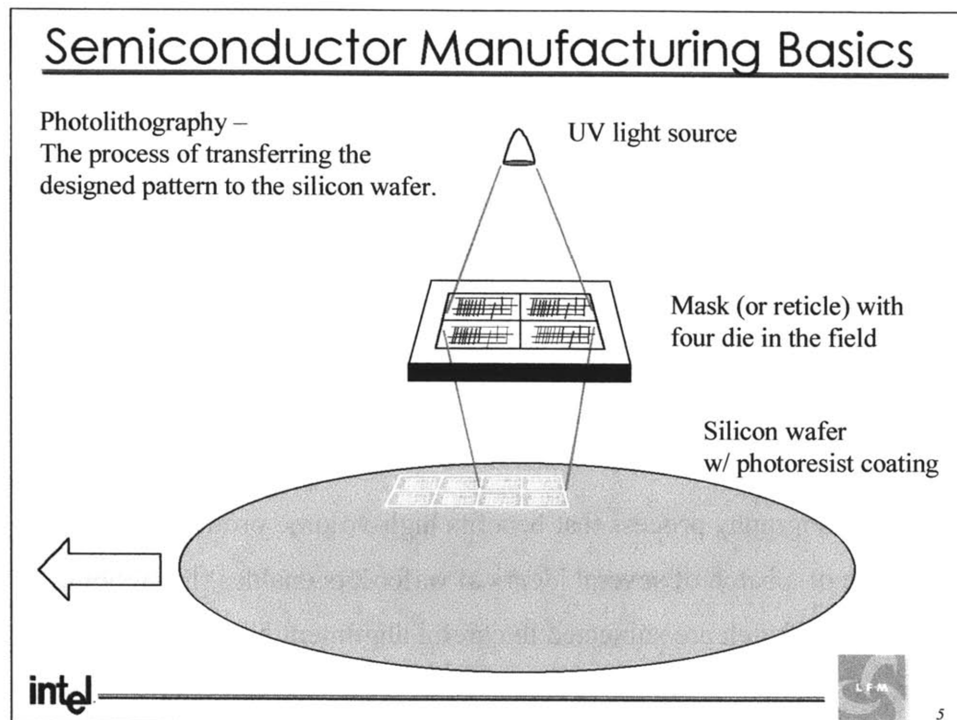


Figure 2.1 The photolithography process

Low-volume production runs have three primary effects on the photolithography area. These effects are: 1) more setups, 2) slower run rates, and 3) more mask sets[1] are required. The increased setup time and slower run rates reduce the production capacity of the fab by reducing the number of wafers per hour that can be patterned. Because a new mask set is required for each product, a portfolio of many low-volume products can significantly increase the capital expense and fixed costs.

The setup time is a function of two processes that typically occur during a changeover to a different product. First, the reticle[2] must be changed and aligned. The steppers typically have an automated magazine that holds several reticles to speed up the process. Once the reticle is changed, it must be carefully aligned to ensure that the pattern that will be printed on the wafer

[1] A mask set is a set of patterns, typically chromium patterns on a quartz glass window, that are projected onto the silicon wafer surface with UV light. Each mask, also known as a reticle, is analogous to a "negative" of a photograph. Since each layer of the semiconductor manufacturing process requires a unique reticle, a mask set is typically comprised of 20 to 30 reticles.
[2] See footnote 1 (above)

will be in focus and correctly positioned relative to the previous and subsequent layers. The reticle change and alignment process causes a down time on the order of five minutes.

One part of the photolithography process is the curing of the photoresist[3], which is conducted in an oven. Different products may require a different photoresist recipe, which can require different oven temperatures. An oven temperature change requires a period of roughly 30 minutes to achieve stability. Thus, product changes can create significant delays in the photolithography area.

An interesting feature of the lithography process that benefits high-volume products is that learning during the processing of a batch of several identical wafer lots enables the stepping rate to increase. The early wafers in a batch are subjected to careful alignment and mid- and post-process checks. However, as a statistical sample is collected, fewer checks are conducted and the wafer lots progress significantly faster (on the order of 20 – 30 % faster run rate).

The mask sets used to transfer circuit designs to the wafer, become a proportionately greater expense for low-volume products. Because a new mask set is required for every product, low-volume products cannot spread the associated mask costs across the sale of millions of units. Further, the cost of producing mask sets has been increasing as process technologies use finer and finer line widths. These smaller feature sizes in the semiconductor design require more expensive electron-beam equipment and phase shift coatings to produce the mask set. For example, a recent article published in the Semiconductor Business News reported the cost of a mask set for a 0.13-micron design runs about $650,000. The cost of a mask set for 0.10 micron design is expected to rise to $1.5 to $2 million. [6]

2.1.2 Chemical Mechanical Polishing

The chemical mechanical polishing (CMP) process is used to planarize the surface of a wafer after layers of material have been added during the fabrication process. This process is

[3] Photoresist is the light sensitive coating that enables the circuit pattern to be transferred to the surface of the silicon wafer.

performed by pressing the wafer against a rotating pad in the presence of chemically activated, abrasive slurry. The combination of chemical and mechanical material removal enables a nearly perfect flat surface to be created. The resulting smooth surface allows the photolithography steps to achieve a more precise focus, resulting in better yield and finer line width capability[6].

The CMP tools see a loss in productivity with low-volume products primarily due to the additional setups that are required. However, the loss of productivity depends on the tool type and product mix. Some CMP tools that have multiple spindles (and can run several wafers at a time) are able to process wafers to different specifications simultaneously. In other cases, the tool must be completely cleared of one product before another can begin its processing.

The effectiveness of CMP process is also dependent on learning from wafers fabricated during and after the NPI process. The material removal rate is partially a function of the pattern density of the product being polished. Because the pattern density varies from product to product, the optimal CMP tool settings differ for various products. Low-volume products, with fewer wafers required, cannot spread the cost of this learning across as many products.

Later in this thesis, we will see that CMP may cause problems with wafers with multiple products fabricated on them. The steep gradients of pattern density caused by placing different products in close proximity make the process difficult to optimize for maximum yield.

2.1.3 Thins Films

The thin films tools deposit very thin layers of metal or dielectric material on the wafers. These tools are chemical reactors that typically use a process known as chemical vapor deposition (CVD). Low-volume products have a relatively small but measurable impact on the productivity of the thin films tools. This impact is simply due to the additional setups required with low-volume products.

2.1.4 Sorting

The sort area of the fab tests the individual dies on a wafer for functionality and speed. This knowledge enables several decisions to be made. It enables the fab manager to understand the process capability independent of the packaging process. It also allows defective dies to be discarded prior to the packaging step.

Sort requires expensive probing tools to measure test structures on the wafer. Every product requires a corresponding probe card that has tiny pins that contact the die in the appropriate contact pads. These probe cards are a capital expense requiring tens of thousands of dollars per product. Very low-volume products typically skip the sort step, at the expense of packaging nonfunctional die.

2.1.5 Other processes

The other processes in the wafer fab, such as etch and ion implant are not as affected by low-volume products. These tools either have small setup times or run products in small batches regardless of the total number of wafers coming through the area. Additionally, these tools comprise a smaller percentage of the total fab capital spending than the tools described in sections 2.1.1 through 2.1.4. Thus a small increase in setup time will have an even smaller impact on overall spending.

2.1.6 New Product Introductions

New Product Introductions (NPIs) consume vast amounts of tool, operator, and engineering time. Every time a new product is launched in a fab, the process is optimized for that product. There are baseline tool settings for each process, but Intel has found that product yields can be significantly improved by choosing the best settings for each product. For example, the CMP results are largely affected by the density of features on a product. Several wafers of a new product may run through the CMP process with different polish times for each. The wafers are then inspected to determine which setting provides the best planarization without excess material

removal. This process of selecting the best settings for an NPI can take several hours for each step in the semiconductor manufacturing process. Adding the setup times for every layer results in a delay on the order of weeks.

The high cost of setups for NPIs is not a problem for a fab that only runs a few stable products. The NPI cost is spread across many wafers, and the yield improvement justifies the cost. As some fabs start manufacturing low-volume "custom" products, the cost and lost productivity due to the NPI process will become a more significant impact to wafer cost. For low-volume products, the cost of the NPI may not be worth the yield improvement that it drives.

2.2 Existing examples of low-volume IC manufacturing

Several companies, including Intel have manufactured low-volume products in their fabs. A few low-volume manufacturing strategies have been proposed, but most fabs seem to treat low-volume products as "second class" and give priority to high-volume products. This section is intended to look at a cross-section of some of the strategies that have been used to manufacture low-volume chips that may apply to the low-volume communication chips on Intel's product roadmap. Each strategy has various costs and benefits depending on the manufacturing situation.

2.2.1 "Shuttles" for prototypes and low-volume products

The so-called "shuttle" methodology has been effectively implemented by the semiconductor foundries to support the prototyping needs of fabless chip design houses. Shuttles are multi-product wafers (MPWs) on which the products of several customers are jointly fabricated, thus allowing the customers to share the mask set and other fixed costs. The shuttle programs are run in an analogous fashion to a shuttle bus. The foundries publish a schedule of the "departure dates" for a wafer lot to begin the fabrication process for a specific manufacturing process technology. Each process technology typically has shuttle departure schedules that range from weekly to quarterly departures, depending on the demand. For each shuttle, the mask set is divided into a number of seats of fixed size and price. Customers are offered the purchase one or more seats prior to the departure date. Then, the customers must submit their chip designs for

the shuttle by the departure date. The shuttles do not wait for late customers to ensure the satisfaction of the customers who submitted their designs on schedule. Thus, a late passenger simply waits for the next shuttle. Once the components are fabricated, the wafers are diced and packaged, and then the appropriate units are sent to each customer.

An example of a well-implemented shuttle program is Taiwan Semiconductor Manufacturing Company's (TSMC's) CyberShuttle program. The CyberShuttle program has a schedule including 16 different process technologies with an aggregate of over 100 shuttles per year. A typical shuttle run from TSMC will provide the customer with forty packaged units, with additional units available at extra cost. The TSMC program is focused on providing prototypes at relatively low cost and fast turnaround time. However, the program does not provide support for production parts. Thus once an acceptable prototype has been obtained, the product goes into production using standard high-volume manufacturing methods. [8]

The shuttle process provides several benefits to both the supplier and customer. The customers are able to share the high fixed cost of the mask set, resulting in significant savings over a dedicated product run. The lower cost enables the designers to try several iterations of a product before finalizing a design. The frequent departures of shuttles support the ability to try multiple iterations and allow flexibility in scheduling product launches. The foundry benefits by allowing potential customers to "sample" its service for relatively low cost and by selling the shuttle seats at a higher price than a corresponding seat price on a regular production wafer. In the fab, many of the potential low-volume runs are consolidated into fewer standard runs. This helps bring the average batch size up, which results in higher tool utilization and lower wafer costs.

Shuttles also have some challenges associated with them. Because the shuttles typically have many different products "on-board", the process cannot be optimized for any single product, and the yields may be lower than production parts. Also, since different customers own the products on a single prototyping wafer and intellectual property must be protected, TSMC cannot provide whole wafers to any customers. While the designers can get functional information from the packaged units, there are often tests that the designers would like to perform while the dies are still part of a whole wafer. By committing that the shuttles will leave on time, TSMC is taking

some risk that the shuttle will not be filled. Since the cost of a shuttle wafer is the same whether there are ten passengers or a single passenger, TSMC could end up subsidizing under-subscribed shuttles. Of course, once the demand for prototypes is well known, the pricing and schedules can be set to accommodate demand variability risk.

The success of the shuttle programs at the foundries suggests that the MPW concept is worth exploring for low-volume manufacturing.

2.2.2 Intel's low-volume fabs

Intel does have some fabs that produce low-volume products. However, these fabs tend to produce products that are at the start or the tail of the product lifecycle as depicted in Figure 2.2. The location of the products on the product lifecycle has an important impact on the manufacturing strategy used. The primary method of dealing with low volumes is to make low-volume products wait in queue until there is enough demand requiring several production lots or until the deadline is fast approaching.

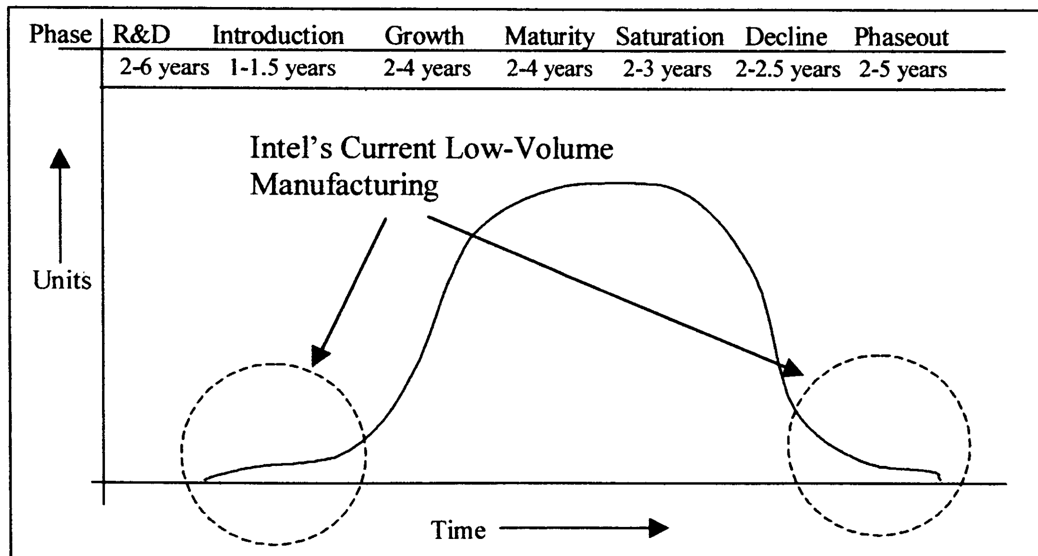


Figure 2.2: The Intel product lifecycle [9]

The first type of Intel fab that frequently runs low-volume products is the Technology Development fab, or TD fab. During the introduction phase of a new logic CMOS product or process, the manufacturing process is being refined in the TD fabs for transfer to Intel's high-volume fabs. The anticipation of transferring the process to other fabs combined with Intel's Copy Exactly! methodology forces the low-volume TD fabs to operate like high-volume fabs. Thus, there is an attempt to maintain large batch sizes when possible, and single-lot products have to wait. [4]

The second type of Intel fab that produces low-volume products is one that produces products that are being phased out. These products are typically produced on an old process technology and are concentrated to a single fab. This concentration creates a wide diversity of products and demand requirements in that fab. However, one benefit of this situation is that the fab can be more flexible in its operation and scheduling of production since the products and processes are guaranteed not to be transferred to other fabs. The Copy Exactly! methodology does not necessarily apply at the tail of the lifecycle.

The low-volume manufacturing process at the end of a process's life cycle is somewhat constrained by the tools used in the fab. The manufacturing processes at the tail are still virtually the same as they were during high-volume manufacturing, so the tools are reused as the process demand shrinks. Because the tools and mask sets are already bought and paid for, it does not make sense to purchase and qualify new low-volume tools for the tail of the product. Excess tools are disposed of or upgraded for use with new process technologies.

The primary process for improving the efficiency of running low-volume products at the end of the life-cycle is to delay production of small batches until there is justification to run a larger batch. Unfortunately, this results in many small lots waiting in queue until the delivery dates approach, and then the lots must be expedited through the fab. Once the lots are completed, significant finished goods inventory must be held which is an additional expense.

Overall, the current methods of manufacturing low-volume products at Intel are not compatible with the requirements for new low-volume communications products. The tradeoff with these

methods is to sacrifice throughput time in order to reduce wafer cost. In the communications sector, this is unacceptable.

2.2.3 Mini-fab – Ericsson

One option for the production of low-volume products is to separate them completely from the high-volume production fabs. However, a drawback of this approach is that a minimum toolset is required to manufacture the wafers. That is, at least one of each type of tool (stepper, implant, polish, etc.) is required. This is only a problem if there are not enough low-volume products to use a significant portion of the fab capacity. Once established, the low-volume fab can focus primarily on reducing setup times due to frequent change-overs, rather than yield or tool utilization.

Ericsson, the telecom and cell phone company based in Sweden, opened what it calls a "mini-fab" in 1994. This fab had an output of only 10,000 wafers per year compared to approximately 300,000 wafers per year at an Intel plant. The Ericsson mini-fab was designed to enable fast design cycle times for custom ASICs (Application Specific Integrated Circuits). While Ericsson did not have the competency or desire to manage the mass production of high-volume chips, the mini-fab enabled Ericsson to develop prototypes, and produce small production runs without revealing its circuit designs to an outside contractor [10].

Ericsson considered the mini-fab to be a strategic asset. Prior to the construction of the mini-fab, Ericsson was dependent on external foundries for both manufacturing knowledge and manufacturing capacity. The mini-fab, which was designed through a partnership with Texas Instruments, enabled Ericsson to build knowledge in the semiconductor processing area. The fab may have made Ericsson more competitive in the marketplace by improving its design capability and manufacturing flexibility, but was not implemented to reduce manufacturing costs, a key goal of this thesis. Intel already has an abundance of opportunities to learn about CMOS wafer processing, thus building a separate fab for new low-volume products does not appear to be an economical option.

2.2.4 Work cell layout

The use of a cellular arrangement of tools for semiconductor manufacturing has been suggested in the manufacturing system design literature as an alternative to the traditional job shop layout of wafer fabs [11]. This cellular tool arrangement has proven to be successful in automotive and aerospace machining applications. The manufacturing cell brings the various tools needed to manufacture a wafer into close proximity in the order of the process steps. The primary benefits of a cellular manufacturing arrangement are that transport times would be nearly eliminated, and single wafer processing could be enabled. Single wafer processing would have the effects of reducing the quantity of WIP[4] in the system and the total throughput time. Such a cellular fab would employ small tools that have only small capacity. However, it is not clear how the traditional re-entrant flow, in which a wafer visits a particular tool multiple times during the semiconductor fabrication process, would be accommodated. Presumably, an entire tool set, including low-capacity and highly accurate photolithography steppers, would be purchased for every layer in a process. The cellular arrangement of the semiconductor manufacturing tools is not compatible with Intel's high-volume fab designs, so it was not considered further in this thesis.

2.3 Alternatives for low-volume IC manufacturing

Intel's plans to produce low-volume products hinge on the desire to use standard Intel process technologies (logic CMOS) with minor modifications to those processes. This desire is created by two factors. First, the capacity required to produce the low-volume products is small compared to Intel's processor and flash memory products. Thus, there is a general consensus that a new fab is not needed to accommodate the low-volume products. Rather, the extra capacity required can be "built in" to the requirements for new process technologies. Second, the development of a process technology is a time consuming and expensive endeavor that is specific to the tools used for production. Redeveloping a process for low-volume tools would require a significant investment that has yet to be justified.

---

[4] Work in Process

Based on the initial desire to produce Intel's low-volume products in Intel's high-volume manufacturing fabs, the following options for low-volume manufacturing were generated through brainstorming and discussions with industrial engineers. It should be noted that many of the strategies could be used simultaneously, but for simplicity they will be discussed separately.

2.3.1 Multi-product wafers

Multi-product wafers, as the name indicates, are wafers that have more that one product fabricated on them. The foundries introduced the use of multi-product wafers with their shuttle programs for prototypes, such as the CyberShuttle program at TSMC. Multi-product wafers have some limitations, but also some very significant cost benefits.

The benefits of using a multi-product wafer for product manufacturing include sharing mask set costs, sharing tool setup times, and potentially sharing wafer costs. The mask sets are continually increasing in cost as the process technologies require finer and finer line widths. By placing several products on the mask set, the cost could be shared by those products. In the fab, the multi-product wafer would only drive a single NPI effort. Thus the setup time per product could be driven down substantially. Finally, it may be possible to share wafer costs between products that have less than a single lot (a lot is a box of 25 wafers) of demand.

Multi-product wafers also come with several challenges. Some of the challenges include dealing with product mix that is fixed once the masks are created, potentially lower yields, ensuring pattern density uniformity for the polish step, and increased coordination between designers. The fixed product mix is caused by the fact that once a mask set is created, it cannot be changed. For example, consider a mask set that creates patterns for two of hypothetical product "A" and four of product "B" as illustrated in Figure 2.2. The mask set will only produce the products in the 2:4 ratio regardless of what the market desires. If for some unforeseen reason, the "A" products are a hit, but the "B" products falter, up to two thirds of the silicon may be wasted producing the "A" products.

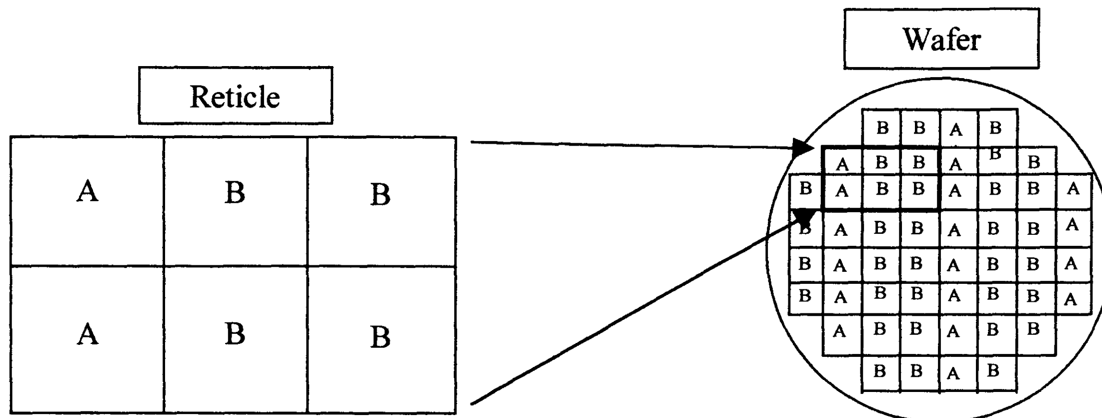


Figure 2.2: Sample MPW reticle (mask) and wafer layouts

There is also some concern that the multi-product wafers will cause an unacceptable decline in die yield. Because the fabrication processes will not be optimized on a product level, as is currently done with dedicated wafers, the yield may drop to levels that outweigh the savings. The primary cause of the yield concerns is the CMP step of the wafer fabrication process. Because various products have different features, the resulting designs have different pattern densities. The issue with having various pattern densities on a wafer is that the polish step is not as effective and cannot be controlled as well. The result of polishing a wafer with die of different densities is like sanding a knotty board. The board (wafer) is dished out in the soft areas, and bulged where there are knots (highly dense components). The density problem can be somewhat alleviated by introducing dummy structures into the designs of less dense chips. The dummy structures ideally have no impact on the circuit performance while raising the pattern density to a common level. However, depending on the design, the dummy structures may consume valuable silicon area.

The physical die size of the products slated for MPW manufacturing has impact on the design of MPWs. The dies must be able to fit within the available area of the reticle field. Thus, the size of the reticle field places a constraint on the number of different products that can be manufactured on an MPW. We will see the effects of this constraint in Chapters 3 and 4.

The sharing of wafers with MPWs would require the use of packaging equipment that can cut, pick, and sort different sized chips from a single wafer. Intel's wafer saws cannot currently cut different sized products from a single wafer. This limitation would cause a significant amount of wasted silicon, since only a single product would be cut from each MPW. Laser saws that have the capability to cut out different-sized dies are currently available on the market, but have not been integrated into Intel's packaging process. The decision of whether to purchase and integrate a new packaging technology will depend on the relative value of the potentially wasted silicon and the laser saw technology. The financial implications of this will be explored further in Chapter 3.

Finally, MPWs require coordination and synchronization of product launches. The products to be produced on an MPW must go to market at nearly the same time to prevent one product from delaying the launch of another. This may create significant challenges since one design group may not have the resources to launch multiple products simultaneously, and separate design groups would have to coordinate their product launches. However, if a "critical mass" of new product launches is obtained, the coordination become much easier. If there is a constant stream of new products to be launched, there will be enough products that can be relatively easily grouped together without attempting to change the product launch schedule.

2.3.2 Multi-product lots

The concept of multi-product lots (MPLs) was created to find a way to reduce the setup times inherent in running many low-volume products in a high-volume fab. The theory is that mask sets are not shared but the process setups and NPIs are shared. Essentially the process is made generic for a group of products. Operationally, this strategy would be identical to the multi-product wafer strategy except at the photolithography stage. At the lithography bay in the fab, a standard 25 wafer lot would begin its processing of the first product until, say, eight wafers were completed. At that point, the reticle (or mask) would be changed and the exposure of the second product would continue.

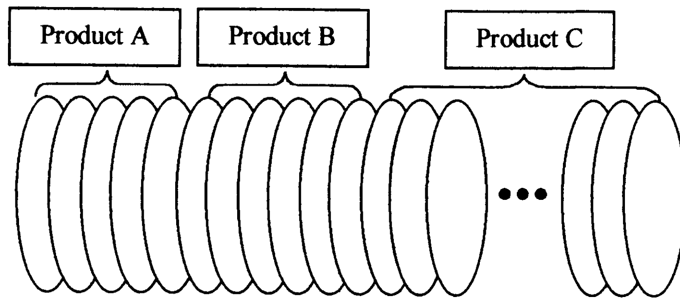


Figure 2.3: A multi-product lot with three products

The primary cost increase is lower utilization of the photolithography tools due to the reticle change time. There would also be potential CMP and yield problems, because the tool settings would not be optimized for each product. This strategy would also require wafer level tracking (as opposed to the current lot level tracking) to know which wafers in a lot belong to which group. The multi-product lot idea would only be useful in the case of products that have a volume requirement that is less than a full lot. If a full lot or more is required, it simply does not make sense to add a reticle change to the photolithography process.

2.3.3 Reduced NPI effort

When a new product is introduced, the engineers go through a standard procedure of tailoring the manufacturing process parameters to that product. The procedure (known as an NPI) enables the fab to run the product at the highest yield possible. An NPI involves running up to three wafers through a process step, then sending them to metrology for analysis. The remainder of the lot waits for the results to return and the tools to be adjusted before further processing. The waiting and setup times can amount to several hours of lost tool productivity and increased cycle time at each process step. Presumably, the cost savings from higher yield outweighs the overhead of performing the NPI. However, in the context of low-volume products with fewer wafers being processed there may not be enough savings due to yield to justify a full NPI yield improvement effort. This strategy would have to be developed through engineering and experimentation to determine the yield improvement per hour of tool and engineering time spent. Then the appropriate yield target would be estimated by balancing the NPI cost with the cost of wasted silicon, based on the forecast demand for the product.

2.3.4 Small lot size

One option for processing low-volume products that does not dramatically change the way a fab operates is to run smaller wafer lots. Typically, products are run in lots of 25 wafers. Some low-volume products are not expected to require a full wafer lot to satisfy the entire lifetime demand! Thus, running a whole lot wastes wafers. Additionally, running smaller lots can have the advantage of lower WIP and faster cycle times. The main drawback of running smaller lots is that the tool setup times are spread over fewer wafers so the tool utilization is much worse than that observed with high-volume manufacturing.

2.4 Conclusions

In this chapter, we have explored various options for manufacturing low-volume products. Most of the current methods of manufacturing low-volume products were eliminated due to incompatibility with Intel's high-volume fabs. The remaining strategies, multi-product wafers, multi-product lots, reduced NPI effort, and small lot sizes, will be compared in more detail in the following chapter.

Chapter 3: Financial analysis of potential low-volume manufacturing strategies

In Chapter 2, several of the options for manufacturing low-volume products were discussed. However, the discussion alone does not help select an appropriate strategy. As a reminder, the goals of choosing the best strategy are to manufacture low-volume products at low cost, with fast throughput times, and high flexibility. In this chapter, the costs of the potential low-volume manufacturing strategies will be compared using a financial model, referred to as the Low-Volume Manufacturing Cost Model (LVM Cost Model). First, the development of the model will be discussed in general. Next, the tailoring of the model to the specific manufacturing strategies will be shown. Finally, the results of the model are displayed and compared.

3.1 Development of a financial model

The model developed here was intended to capture the total cost of producing a common product with different manufacturing strategies. The LVM Cost Model is organized as shown in Figure 3.1. First, the quantity of wafers required is calculated from the forecast demand and the combination of expected factory parameters (e.g. die size, wafer size, and die yield) and the specific manufacturing strategy to be used (e.g. Standard, Multi-product Wafer, Multi-product Lot, etc.). Once determined, the wafer quantity is converted to the number of required lots.

Next the wafer cost is determined. The final wafer cost can vary based on the number of lots run together in a chain, or cascade. Also, the first two lots or a product's lifecycle are burdened with the cost of the NPI.

The final wafer cost is converted to "cost per die" by dividing by the number of dies per wafer. The mask set cost is also divided among the total lifecycle die produced with that mask. The sum of these two costs gives us a final Die Cost that can be used to compare the different manufacturing strategies for products with various die sizes and demand forecasts.

Sections 3.1.1 through 3.1.3 describe the equations used to model the die cost. These sections are organized to follow the model roadmap depicted in Figure 3.1. Section 3.1.1 describes the calculation of the wafer requirement parameters used by all of the manufacturing strategies, including the calculation of "lot requirement". Section 3.1.2 describes the calculations specific to the individual manufacturing strategies considered in the model. Finally, Section 3.1.3 describes the equations used to calculate the final wafer cost and die cost.

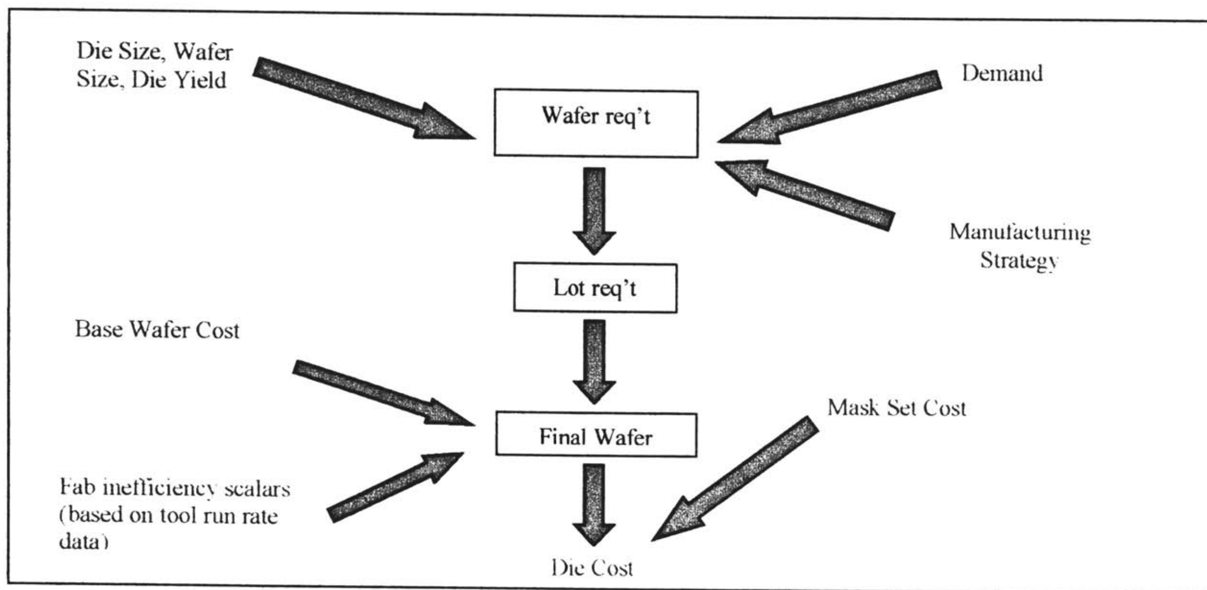


Figure 3.1: Roadmap of the Low-Volume Manufacturing Cost Model

3.1.1 Calculation of the wafer quantity requirement

The number of wafers required to satisfy demand is a function of product die size and factory parameters such as yield and wafer size. For simplicity, the demand forecasts for the products to be modeled were lumped into a single time period. This simplification eliminated the need to calculate inventory held between periods.

Die size

Die size is the size of a single chip. Typical die sizes vary from 100 mils x 100 mils to about 500 mils x 500 mils. The units of measurement of the dies are "mils", or thousandths of an inch. The die size determines how many die will fit on a wafer. Clearly, a product with large dies will require more wafers than a product with small dies to satisfy a particular unit demand. For this

initial, Low-Volume Manufacturing Cost model, the dies were assumed to be square. This constraint will be relaxed in a more refined model presented in Chapter 4.

Gross die per wafer

The size of the wafer has an obvious impact on the number of die that can be produced on the wafer. The number of whole die printed on a wafer, called Gross Die per Wafer is calculated by,

Eq. 3.1 $$GrossD/W = \frac{\pi \times (WaferDiameter / 2)^2}{A} - EdgeEffectCorrection$$

where, *WaferDiameter* is the diameter of the raw silicon wafer, and A is the die area. The examples presented in this thesis assume a 300mm wafer diameter. The value, *EdgeEffectCorrection*, is empirically obtained based on die size and wafer size that accounts for the loss of partially fabricated dies that overlap the edge of the wafer because the dies are square in shape and the wafers are round in shape. The edge effect is most significant for smaller wafers and large die.

Line yield

Line yield, or wafer yield, is the ratio of good wafers after completing all of the process steps to the total wafers started. A small percentage of wafers are broken during handling or are processed incorrectly and must be scrapped. The line yield is typically quite high in modern factories, so it is assumed to be 100% in the LVM Cost Model.

Die yield

Die yield is the ratio of good dies to total dies fabricated on a wafer. Defects caused by particles or scratches on the wafer surface cause some dies to be non-functional upon completion of the fabrication process. Since the defects on a wafer are assumed to be randomly distributed across a wafer, products with large die are more likely to encounter a killer defect. Thus, large dies suffer poorer die yields than small dies.

The model uses a proprietary die yield calculation, so a similar calculation derived from the famous Murphy yield model [12] is presented here. The die yield can be characterized by the

following equation in which the defect density is approximated by the gamma distribution function:

Eq. 3.2 $$Y_{die} = \frac{Y_0}{(1 + D_0 A / \alpha)^{\alpha}}$$

where, Y_0 is the linear yield function constant which reflects that portions of the wafer area result in zero yield despite the absence of point defects. D_0 is the mean defect density, which represents the number of killer defects per unit area. The clustering parameter, α, accounts for the tendency of defects to cluster on the wafer. The parameters Y_0, D_0 and α, must be determined experimentally or empirically, and they will depend on the wafer size, the process technology being used, and the cleanliness of the fab.

Table 3.2 demonstrates the effect of die size and defect density. In this example, Y_0 and α are held constant at values of 0.90 and 15, respectively. The defect density was varied for three different die sizes. It is seen that small die generally have a higher die yield than large die and are much less affected by increases in defect density.

Table 3.2: Sample Calculation of die yield with varying die size and defect density

		Y_{die}		
Die Dimensions	Die Area (cm^2)	D_0=0.10	D_0=0.25	D_0=0.40
100 mil x 100 mil	0.06	89%	89%	88%
300 mil x 300 mil	0.58	85%	78%	71%
500 mil x 500 mil	1.61	77%	60%	48%

Note: Y_0 = 0.90 and α = 15 in this example.

Die per wafer

The die yield and gross die per wafer are used to calculate the number of good die per wafer. This is given by:

Eq. 3.3 $$D/W = GrossD/W \times Y_{die}$$

Now that we have the number of dies that a single wafer yields, we are nearly ready to calculate the total number of wafers required to satisfy a given quantity of demand. However, the next section will show that the number of wafers required may also depend on the strategy used to manufacture those wafers.

3.1.2 Tailoring the wafer requirement to each LVM strategy

The various manufacturing strategies have an effect on the number of wafers required. The different methods of combining products into common lots or on multi-product wafers cause unique changes in the calculations. The calculations that are unique to the individual manufacturing strategies are described in this section.

Standard

The standard manufacturing strategy assumes that there is only one product produced per lot and there are 25 wafers in each lot. This is the standard method of running high-volume products at Intel and is used as a baseline for comparison with the low-volume manufacturing strategies. The whole number of wafers required to satisfy the total lifecycle demand of the product is calculated by:

Eq. 3.4 $$WaferRequirement = \frac{Demand}{D/W}$$

It follows that the number of lots required to satisfy demand is represented by:

Eq. 3.5 $$LotRequirement = \frac{WaferRequirement}{Wafers/Lot}$$

where, *Wafers/Lot* is the number of wafers in a box, typically 25. Both the wafer requirement and lot requirement values were forced to be integers, so the values are rounded up to the next whole number.

Small lot size

The first low-volume manufacturing strategy considered was small lot sizes. The lot size is the number of wafers that move together through the fab in a single box. The option of using smaller lot sizes can potentially reduce waste by producing fewer wafers in the case of extremely low demand where fewer than one lot of wafers will satisfy a product's lifetime demand. Small lot sizes were modeled by simply reducing the constant *Wafers/Lot* in Equation 3.5.

Multi-product wafers

A more complicated low-volume manufacturing strategy to model is the multi-product wafer. For this initial model, two simplifying assumptions were made. First, the model assumes all of the products of an MPW have the same die size. Second, the model assumes that the products have the same total demand. These assumptions, while not very realistic, enabled the model to quickly assess the potential of using MPW as an LVM strategy. A more detailed model, described in Chapter 4 will eliminate these assumptions.

Multi-product wafers were considered in two types. The first type assumes that only one product is harvested from each wafer. For example, in the case of a four-product MPW only one of the products is assembled into packages and tested while the other three are simply discarded. This may occur due to difficulties in dicing wafers with nonstandard scribe lines or difficulties in logistics. There is a clear tradeoff in this model since much of the silicon is wasted. The more products that are fabricated on an MPW, the more silicon is wasted, but the mask set cost and NPI and setup cost are spread over more products as well. Thus, the number of products placed on the MPW must be balanced with the relative cost between wafers and mask sets. The number of wafers required with the "MPW" strategy is given by:

Eq. 3.6 $$WaferRequirement_{MPW} = \frac{Demand}{D/W_{MPW}} \times (\#products/wafer)^2$$

where D/W_{MPW} is the number of good die per wafer for a MPW and *#products/wafer* is the number of different products fabricated on an MPW. *Demand* is the quantity of dies required for each individual product.

The term, *#products/wafer*, appears squared in Equation 3.6 because:

1. Each additional product reduces the number of dies per product on each wafer
2. Only one product is packaged from each wafer

The second type of MPW assumes that all products on the MPW are packaged and tested. This eliminates the problem of wasting silicon when many products are placed on an MPW. Because this is considered a special case of the MPW it was designated the "MPW-AT" strategy. As discussed in Section 2.3.1, this MPW-AT strategy requires more advanced wafer sawing techniques than are currently used at Intel and, if implemented, would require additional tool purchases and process development.

The number of wafers required to satisfy demand when the MPW-AT strategy is used and all products are packaged for each wafer is given by:

Eq. 3.7 $$WaferRequirement_{MPW-AT} = \frac{Demand}{D/W_{MPW}} \times (\#products/wafer)$$

where, MPW-AT denotes that the MPW strategy is used with all products *Assembled and Tested*, or packaged. Here we see that *#products/wafer* is not squared, thus fewer wafers are needed if all products can be packaged from each wafer as is shown below.

Both of the MPW strategies were considered with 2, 4 and 8 products per wafer. Additionally, the MPWs were modeled with a worse die yield, because the manufacturing processes cannot be optimized for each specific product on an MPW. This was accomplished by incrementing the defect density when calculating wafer requirements for the MPW and MPW-AT strategies.

Multi-product lots

Multi-product lots (MPLs) were modeled in a similar method to the MPWs. To review, an MPL is a lot of wafers with different products fabricated on different wafers of that lot. The process

technology is the same for each product so setup times are shared. However, unique mask sets are required for each product so mask set cost cannot be shared.

Again, a simplifying assumption was made that the products on an MPL are the same die size and have the same unit demand. With MPLs there is no problem separating the different products because each product is on its own wafer. However, the MPLs may have a lower yield than wafers run with standard production techniques since the setup procedures are shared between the various products. So, the manufacturing processes cannot be optimized for a specific product on an MPL. The number of wafers required to satisfy demand when the MPL strategy is given by:

Eq. 3.8 $$WaferRequirement_{MPL} = \frac{Demand}{D/W_{MPL}} \times (\#products/lot)$$

where *#products/lot* is the number of different products fabricated in an MPL. Again, *Demand* is the quantity of dies required for each individual product.

3.1.3 Wafer cost components

As discussed in Chapter 2, many tools become less efficient when small batches are run due to setup times and test requirements. The financial model will try to capture the sensitivity of capital cost to the number of lots required for each product.

Base wafer cost

The wafer costs used in the model were based on what Intel Finance calls the "Yielded Wafer Cost", which includes labor, capital, overhead, and raw materials. This measure of cost is calculated on a routine basis to estimate the wafer costs of each process technology over time. The estimates are based on the total expected expenses divided by the total expected wafers. Thus, the wafers produced during the ramp phase of production are extremely expensive because there are relatively few wafers being produced while an entire fab is depreciated. Later in the process lifecycle, the wafer cost declines as the depreciation is spread across many more wafers. A processed silicon wafer can cost anywhere from several hundred dollars to tens of thousands

of dollars depending on the process technology and position on the lifecycle. Typical wafer costs are broken down as shown in Table 3.1.

Table 3.1: Typical Wafer Cost Breakdown, 2001 [13]

Depreciation	48%
Raw Materials	32%
Direct Labor	10%
Overhead and other	10%

The largest expense, depreciation, includes the expenses of both the fabs and the tools. Because depreciation constitutes such a large portion of wafer cost, it is easy to see why there is an emphasis on tool utilization. The second highest expense, raw materials, is the cost of the raw silicon wafers, chemicals, and gasses required to manufacture ICs. Direct labor and overhead account for the remainder of the base wafer cost.

To give a representative example of base wafer costs, the third quarter 2001 average prices of wafers produced by foundries are shown below in Figure 3.2. The higher prices for process technologies with smaller feature sizes reflect the higher cost of new processing tools.

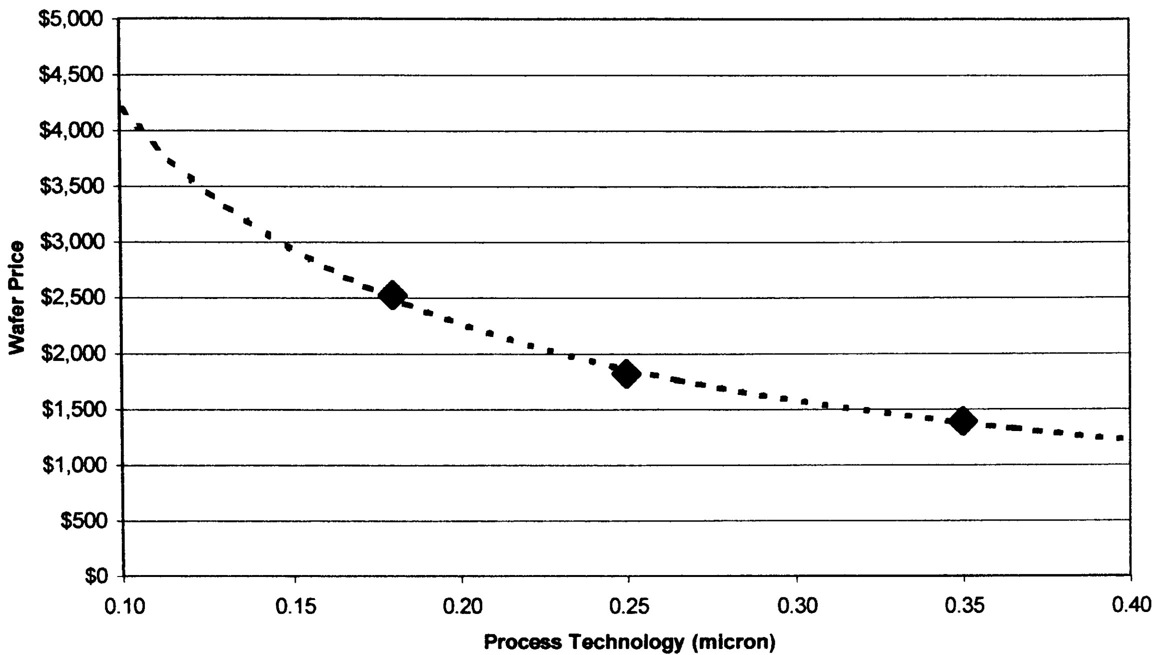


Figure 3.2: Q3 2001 average foundry wafer prices by CMOS process technology [14].

Fab tool utilization

In order to maximize the fab tool utilization, fab managers attempt to reduce the number of changeovers by running several consecutive lots of each product. These consecutive lots are known as a "cascade". The typical cascade size for high-volume products is four lots, or 100 wafers. However, some low-volume products may require only a single lot of wafers (or less) to meet the forecast lifetime demand. These low-volume products cause a proportionally higher amount of tool down time. Because the tool depreciation is the largest portion of the wafer cost, lost tool run time has the potential to dramatically increase wafer cost. This potential increase in wafer cost is not currently accounted for at a product level by Intel's finance department. The current accounting method spreads the cost of the extra capacity requirement across all of the wafers processed per time period. However, some Intel employees fear the addition of many low-volume products could cause enough of a productivity drop that the high-volume product managers would want a wafer cost based on productivity measured at the product level.

To model the fab efficiency loss caused by low-volume products, real data for representative tools was collected with the help of various industrial engineers. A typical tool efficiency curve is represented in Figure 3.4. The efficiency improves rapidly as the cascade size increases, but eventually diminishing returns slow the potential improvement.

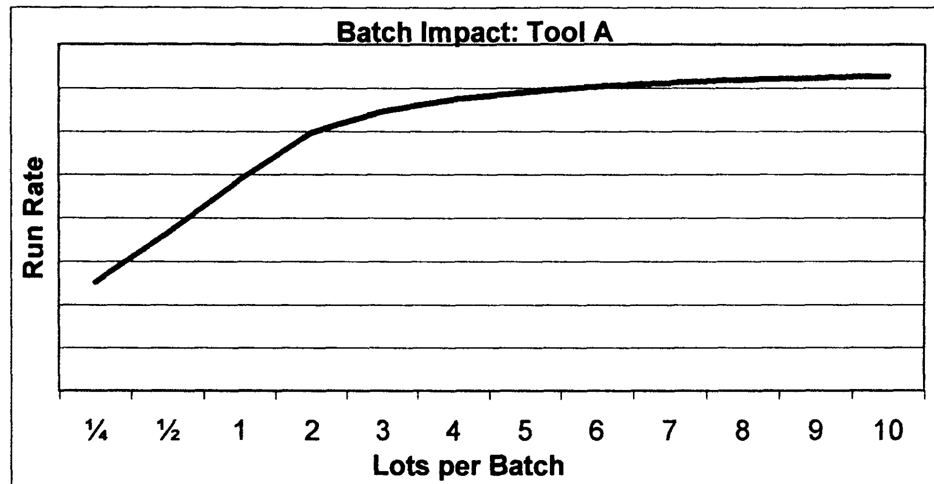


Figure 3.4: Sample fab tool efficiency curve.

To account for this in the model, tool efficiency curves were collected for those tools that suffer the most impact from low-volume manufacturing as described in Chapter 2, then converted into "tool inefficiency" factors as follows:

Eq. 3.9 $e_{tnc} = \frac{rr_{nc}}{rr_{n4}}$

where, rr_{nc} is the run rate (in wafers per hour) for tool n with cascade size c. A baseline cascade size of four lots was used in this equation, and the inefficiency factors were calculated for 1/4, 1/2, 1, 2, 3, and 4 lot cascades. Thus, a typical product with cascade size of four lots will have an inefficiency factor of "1". A low-volume product with fewer lots in a cascade will an inefficiency factor greater than "1".

These tool inefficiency factors were used in conjunction with the weighted cost of each tool type to create a composite fab inefficiency factor for each cascade size. The weighted cost of each tool was determined by:

Eq. 3.10 $w_{tn} = \frac{cost_n}{\sum_{i=1}^{n} cost_i}$

Where, $cost_n$ is the cost of tool number n.

It follows that the overall fab inefficiency factor for cascade size c, is calculated as follows:

Eq. 3.11 $ef_c = w_{t1} \times e_{t1c} + w_{t2} \times e_{t2c} + ... + w_{tn} \times e_{tnc}$

To demonstrate the calculation of the tool inefficiency factors, suppose we want to find the fab inefficiency factor for a cascade size of two lots. First we must collect the run rates for the tools at the cascade sizes of both the proposed two lots and the standard four lots. In this example, we will consider four hypothetical tool types designated by letters A, B, C, and D.

For Tool A, suppose the run rates for two-lot and four-lot cascades are rr_{A2}=20 wafers/hour and rr_{A4}=30 wafers/hour, respectively. Thus, the tool inefficiency factor, e_{tA2}=30/20=1.5 because tool A requires 1.5 times the processing time per wafer when processing two-lot batches compared to the baseline of four-lot batches of wafers. If cost of tool A capacity is $5 million and the entire toolset cost is $10 million, then the weighted tool cost, w_{tA}=$5million/$10 million = 0.5. Using these calculations for the rest of the tools we can construct Table 3.2.

Table 3.2 Sample tool cost weighting calculation

Tool	rr_{n2} (wafers/hour)	rr_{n4} (wafers/hour)	Tool Inefficiency Factor	Tool Cost ($-millions)	Tool Cost Weighting
A	20	30	1.5	5.0	0.5
B	30	33	1.1	2.0	0.2
C	15	30	2.0	1.5	0.15
D	10	12	1.2	1.5	0.15

Using Equation 3.11 we find that the fab inefficiency factor for a cascade size of two lots is then:

$$ef_2 = (1.5 \times 0.5) + (1.1 \times 0.2) + (2.0 \times 0.15) + (1.2 \times 0.15) = 1.45$$

The fab inefficiency factor can then be used to charge low-volume products for the extra fab and tool time they consume. Based on conversations with industrial engineers, an estimate was made that direct labor costs would only increase by half the amount depreciation increases. Thus, the depreciation and direct labor components of base wafer cost are multiplied by the fab inefficiency factors to obtain a final wafer cost as follows:

Eq. 3.12 $wc_f = ef_c \times wc_d + wc_{rm} + \{[(ef_c - 1)/2] + 1\} \times wc_{dl} + wc_o$

where, ef_c is the fab inefficiency factor for cascade size c, wc_d is depreciation portion of base wafer cost, wc_{rm} is the raw material portion of base wafer cost, wc_{dl} is the direct labor portion of base wafer cost, wc_o is the overhead portion of base wafer cost.

An illustration of the final wafer cost calculation using a representative wafer cost of $3000 (for a four-lot cascade) is shown here. First, the wafer cost must be broken into its components using the values from Table 3.1. The resulting wafer cost components are as follows:

$$wc_d = \$3000 \times 0.48 = \$1440$$
$$wc_{rm} = \$3000 \times 0.32 = \$960$$
$$wc_{dl} = \$3000 \times 0.10 = \$300$$
$$wc_o = \$3000 \times 0.10 = \$300$$

Then using Equation 3.12, we find the final wafer cost of producing wafers in cascades of two lots.

$$wc_f = 1.45 \times \$1440 + \$960 + ((1.45-1)/2 + 1) \times \$300 + \$300 = \$3716$$

To further enhance the model, a switch was created to enable the user to neglect the effect of tool efficiencies by setting them to the value of "1". For simplicity, an implicit assumption is made that if the lifetime demand is less than four lots, then those lots are run together in a cascade.

NPI costs

New Product Introductions affect the fab efficiency in the same fashion as cascade size. NPIs similarly require more tool run time and operator time than existing products. The NPI typically only affects the first two lots of a production run, so the effect is only modeled for the first two lots of a production run here. The final wafer cost, wc_{npi}, of those first two lots is calculated using the following equation:

Eq. 3.13 $$wc_{npi} = ef_c \times e_{npi} \times wc_d + wc_{rm} + \{[(ef_c - 1)/2] + 1\} \times e_{npi} / 2 \times wc_{dl} + wc_o$$

where, e_{npi} is the fab inefficiency factor caused by the NPI process. This equation is identical to Eq. 3.12 with the exception of the added NPI inefficiency factor to the depreciation and direct labor components of the wafer cost.

The NPI fab inefficiency factors, introduced in Eq. 3.13, were based on models of capacity impact due to NPIs as created by one of Intel's industrial engineering teams. The NPI impact on labor cost was estimated to be half the impact of the tool capacity impact.

Using the values from the previous example and an NPI inefficiency factor of 4, we can estimate the cost of a wafer during the NPI process (using a cascade size of two) using Equation 3.17 as follows:

$$wc_{npi} = 1.45 \times 4 \times \$1440 + \$960 + ((1.45-1)/2 + 1) \times 4/2 \times \$300 + \$300 = \$10{,}347$$

A switch was also created to enable the modeler to neglect the effect of NPIs by setting the NPI inefficiency factors to the value of "1". This allows us to determine the potential savings created by reducing or eliminating the NPI effort.

3.1.4 Mask set cost

The mask set is a fixed cost that is generally not included in the wafer costs. So the mask set cost can be divided across all of the dies manufactured with a particular mask. Mask sets have been increasing in cost as the process technologies demand smaller line widths and require more

layers (because each layer gets a unique reticle). The mask set cost for a given process technology may vary based on product complexity, number of layers, and e-beam write time. In this model, typical mask set costs were used for each process technology. The mask set costs from Intel are confidential, but some representative costs reported in press releases from TSMC are shown in Figure 3.3.

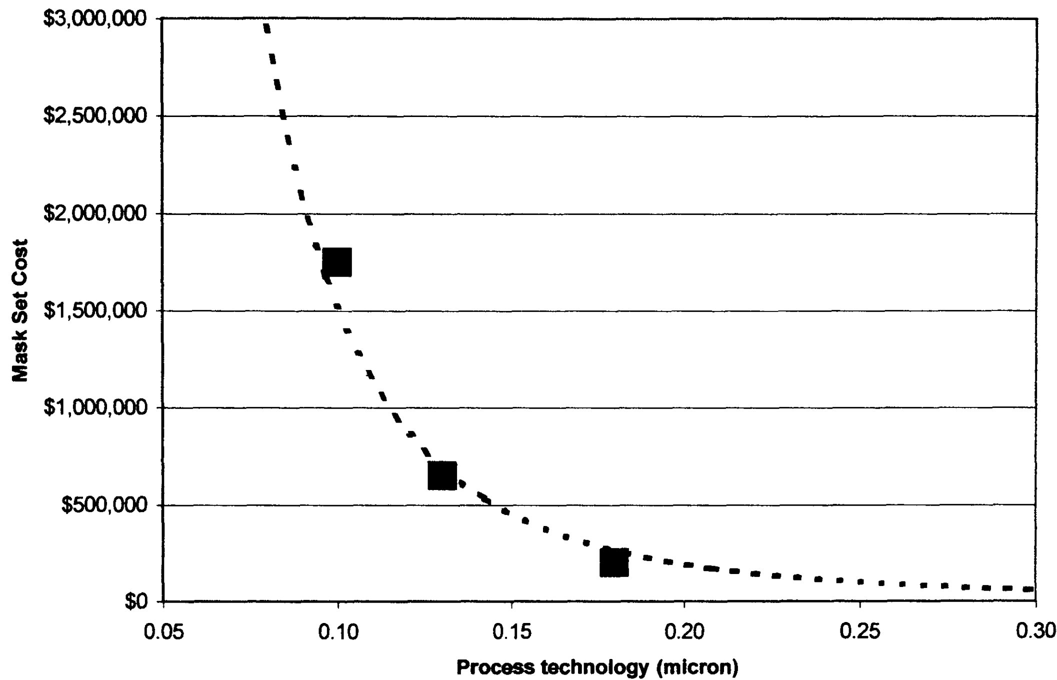


Figure 3.3 Mask set costs are increasing with each new process technology [6][15]

The LVM Cost Model allocates the mask set cost among the entire quantity of good die produced using that mask as follows:

Eq. 3.14 $MaskCharge/Die = MaskSetCost \, / \sum_{i=1}^{n} GoodDieQuantity_i$

Where, *MaskCharge/Die* is the value of the mask set that is allocated to each die, *MaskSetCost* is the cost of a mask set for a given project, *GoodDieQuantity* is the number of good die produced of product *i*, and *n* is the number of products designed into the mask set.

Clearly there are economies in maximizing the use of a mask set. There are limits to the number of exposures a mask set can make, but this quantity is much higher than the low-volume runs being considered here.

3.1.5 Die cost calculation

As mentioned at the beginning of this chapter, the objective of the low-volume manufacturing model is to determine the total manufacturing cost per die for different manufacturing strategies. This cost per die is calculated by summing the cost components described in the preceding sections. The die cost is given by:

Eq. 3.15

$$DieCost = \frac{wc_{npi} \times \#wafers_{npi} + wc_f \times (WaferRequirement - \#wafers_{npi})}{\sum_{i=1}^{n} GoodDieQuantity_i} + MaskCharge/Die$$

where, $\#wafers_{npi}$ is the number of wafer that undergo the NPI process, typically 50 wafers. The die cost can now be used to compare various manufacturing strategies.

3.2 Low-Volume Manufacturing Cost Model results

Initial results using the model just described show that there are significant differences between the costs of low-volume products and high-volume products using the standard manufacturing method. A hypothetical pair of products, one low-volume product and one high-volume product, illustrates the effects of product quantity and mask cost on total die cost. Figure 3.4 shows the resulting die cost comparison. For added clarity, the additional cost caused by the NPI and fab inefficiencies are shown separate from the base wafer cost in this figure. As defined previously, there is no additional fab inefficiency for the HVM case.

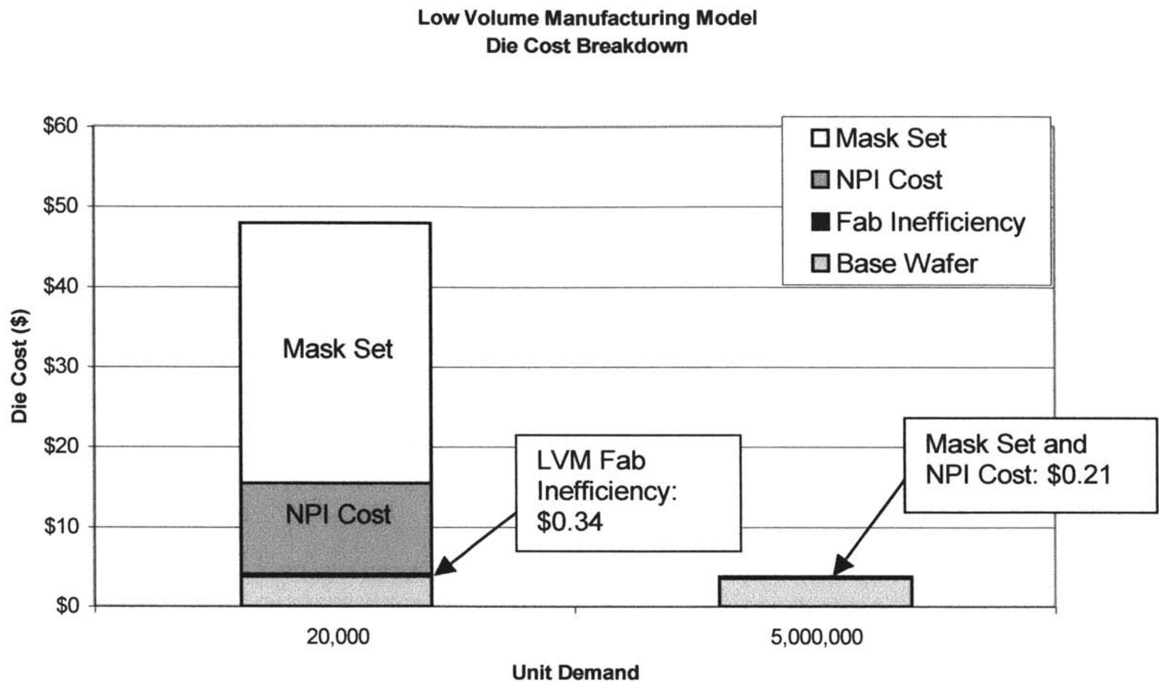


Figure 3.4: Comparison of the die costs of two products with identical die size.

It is readily apparent that the mask set is the majority of the cost of the low-volume product. Interestingly, the fab inefficiency is small for the low-volume product, but the NPI cost is the second highest cost component.

For the product with higher demand, the percent of die cost attributed to the mask set diminishes to a nearly negligible level. The high-volume product cost is dominated by the base wafer cost, which is comprised of depreciation, raw material, direct labor and overhead. Hence, yield and tool utilization are the largest levers in reducing the cost of this product. This example suggests that sharing the mask set costs and NPI costs, with MPWs, might make sense for low-volume products, but not for high-volume products. In the following sections, we will explore how die cost varies using each of the potential low-volume manufacturing strategies discussed in Section 2.3.

3.2.1 Die cost comparisons with increasing die size

Since the model allows many variables to be tested at once, the general results will be displayed with only one variable changing at a time. The strongest influences on the appropriate manufacturing strategy were die size and product demand.

The model was run with hypothetical, but reasonable inputs for a 0.13 micron logic CMOS process technology. A mask set cost of $650,000 and a 300mm wafer cost of $3000 were chosen as inputs. The cost impacts of NPIs and fab inefficiency are included in these results. Tabular versions of the results can also be found in Appendix B.

The most striking result of the model output is the magnitude of savings that can be obtained by using a low-volume manufacturing strategy such as MPW. Figure 3.5 shows the die cost of a relatively small sized product (100 mils x 100 mils) as a function of total demand using each of the different strategies modeled. In this case, the standard manufacturing method is the highest cost over the modeled range of demand. The small lot size and multi-product lot strategies show some improvement over the standard method. These strategies can save approximately 10% to 30% of the total manufacturing cost. However, to achieve savings of greater than 50%, the MPW strategies must be used.

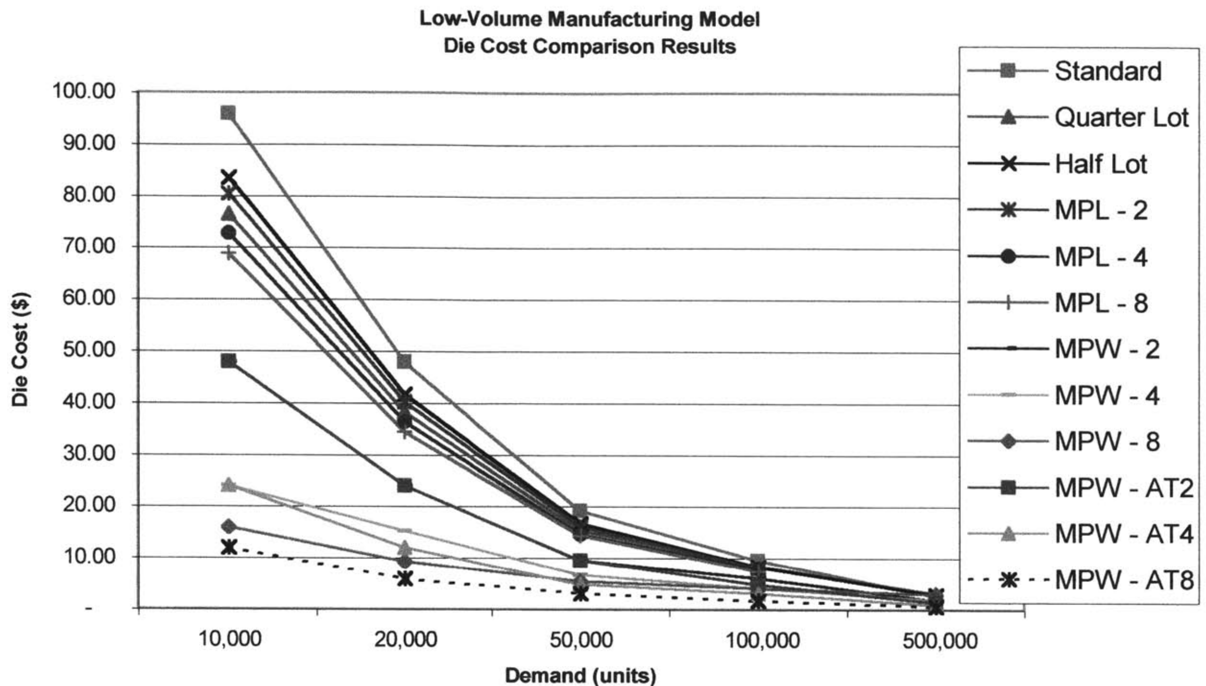


Figure 3.5: Model results for products with small dies (100 mils x 100 mils).

The MPW strategy with eight products and shared wafers (MPW-AT8) distributes the fixed costs across the most products, thus it is the most cost effective strategy. The eight product MPW with no wafer sharing (MPW-8) starts out as an effective strategy, but it is eclipsed by the MPW-AT4 strategy near the 50,000 unit mark. Even though the MPW-8 strategy shares the fixed cost between more products, the wasted silicon costs eventually erode those savings.

The next model run was completed with identical parameters, except the die size was increased to 300 mils x 300 mils, a size more typical of communication chips and microprocessors. The results, as shown in Figure 3.6, are much less cluttered because some of the strategies became either infeasible or simply would not make sense as explained in Chapter 2. For example, the two product multi-product lot (MPL-2) was not a viable strategy in this case because more than two lots were required to satisfy demand. It does not make sense to change reticles mid-lot, when the reticle change can occur between lots and tracking complexity can be significantly reduced. The larger die size reduces the effectiveness of the MPW strategies somewhat in this

case. The MPW-AT8 strategy is again the most effective, but the savings are not as high as with small die sizes.

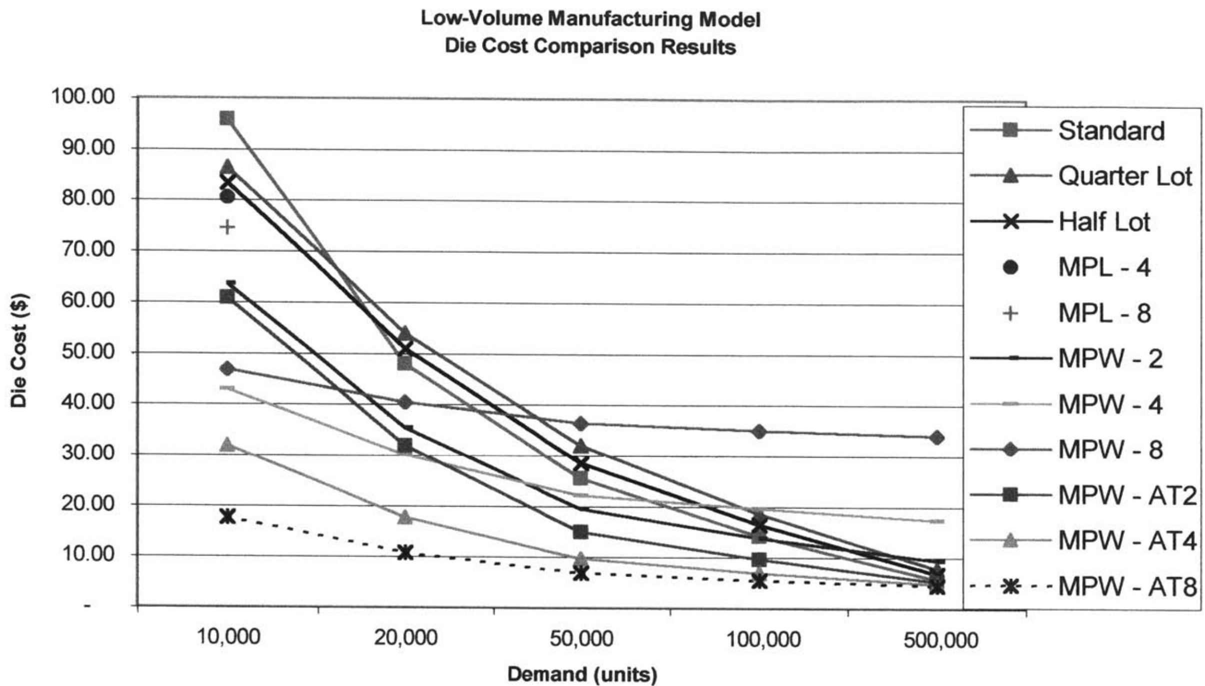


Figure 3.6: Model results for products with medium sized dies (300 mils x 300 mils).

The high silicon waste of MPW-8 pushes that strategy into last place (highest cost) at volumes greater than 50,000 units. Also, as multiple lots become required to meet demand for some of the manufacturing options, the small lot size strategies (Half Lot and Quarter Lot) are more expensive than the standard manufacturing method. The small lot sizes cause additional unnecessary setup costs when a full lot could be run with a single setup.

The model results for a product with large dies (500 mils x 500 mils) become even more sparsely populated. The MPL options are discarded here because, as seen before, multiple lots of each product are required. The options of four and eight product MPWs are now infeasible because the dies are too large to fit into the reticle field in those multiples. Only the two-product MPW is feasible in this test. As expected, the small lot size options are worse than the standard manufacturing strategy. The MPW-2 strategy is only better than standard at extremely low volumes, and even the MPW AT-2 strategy loses out to the standard method at 100,000 units and

above. The dual effects of a large die size, more wafers required to meet demand and lower die yield, have diminished the effectiveness of the MPW strategy when demand is greater than 100,000 units.

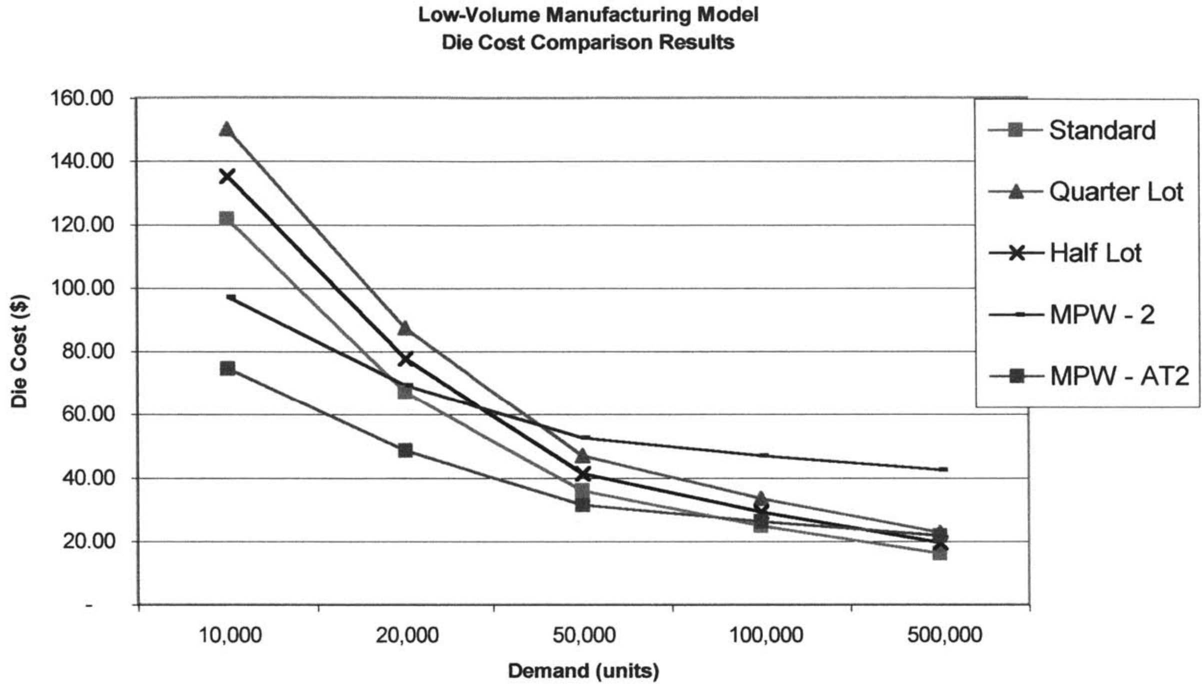


Figure 3.7: Model results for products with large dies (500 mils x 500 mils).

The above comparisons of the potential manufacturing strategies have a few common themes. First the MPW strategies with shared wafers are the best strategies for low-volume products and can result in significant savings. Second, the MPW strategies that do not share wafers are beneficial for low-volume products, but the benefits are quickly eroded by silicon waste as the demand rises. The MPL and small lot size strategies never show much promise as effective cost savers because they must assume the entire cost of a mask set. Finally, the standard manufacturing method narrows the gap between its costs and the other strategies' costs as the demand rises.

3.2.2 Impact of die yield on manufacturing strategy

The previous analysis assumed that the MPW strategies would suffer from a higher defect density than the standard manufacturing method because the manufacturing process cannot be optimized for each individual product on the wafer. The higher defect density results in a lower die yield for the MPW strategies, but the impact is different depending on the product's die size. The following sensitivity analysis demonstrates the effect of yield on the chosen manufacturing strategy.

A comparison of die cost using the standard strategy and the MPW-AT strategy was conducted with three different values of defect density entered into the die yield equation. These values of defect density are listed in Table 3.3. The high yield case assumes that the standard and MPW strategies have approximately the same defect density, D_0. The expected yield case is just that, the expected defect density for MPWs based on the author's discussions with industrial engineers. The low yield case assumes a defect density that is higher than expected for the MPW, but within the realm of possibility.

Table 3.3: Sample MPW defect density values

Test Run	MPW Defect Density (defects/cm^2)
High yield	0.3
Expected yield	0.5
Low yield	0.7

The plots in Figure 3.8 show that the yield has a relatively small effect on the die cost when using MPW-AT strategy for medium sized die. Tabular data are presented in Appendix B. Only when the demand volumes reach 500,000, does the lower MPW yield start to have a significant negative impact relative to the standard manufacturing strategy. This result means that efforts to improve the yield of low-volume products with small die sizes may be wasted effort. It makes logical sense that the yield has a small impact in low-volume products. Recalling the die cost breakdown in Figure 3.4, the wafer cost component (the portion affected by die yield) was a very small percentage of the total die cost.

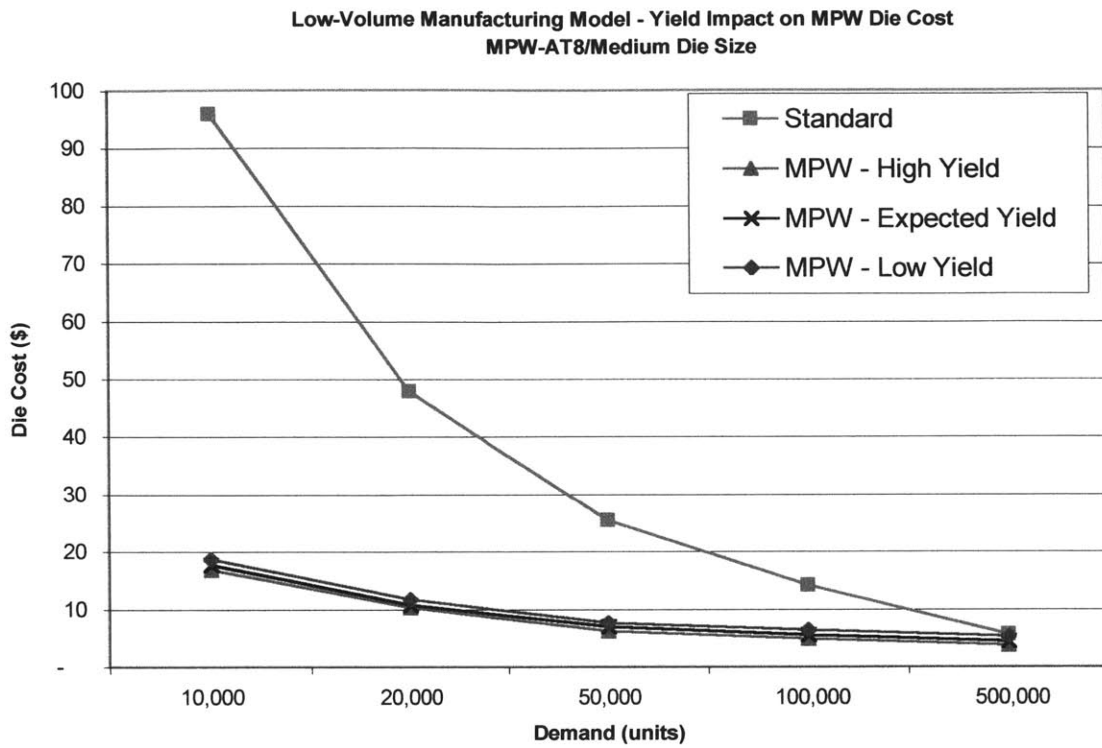


Figure 3.8: Effect of yield on die cost for medium size dies (300 mils x 300 mils).

As seen in Eq. 3.2, a change in defect density should have a greater impact on a product with a larger die size. Figure 3.9 shows this is the case. With a large die, the range of demand volume that makes the MPW-AT strategy look attractive can change significantly. For example, if the yield turns out to be the expected value, the MPW strategy will create a cost savings for products with demand as high as about 100,000 units. However, if the yield were worse than expected, the MPW strategy is only effective for products with demand well below 50,000 units. Conversely, higher than expected die yield extends the applicability of the MPW strategy out to 500,000 units. This example shows that large die create more risk in the application of low-volume manufacturing strategies because any change in yield has a large impact on the cost savings.

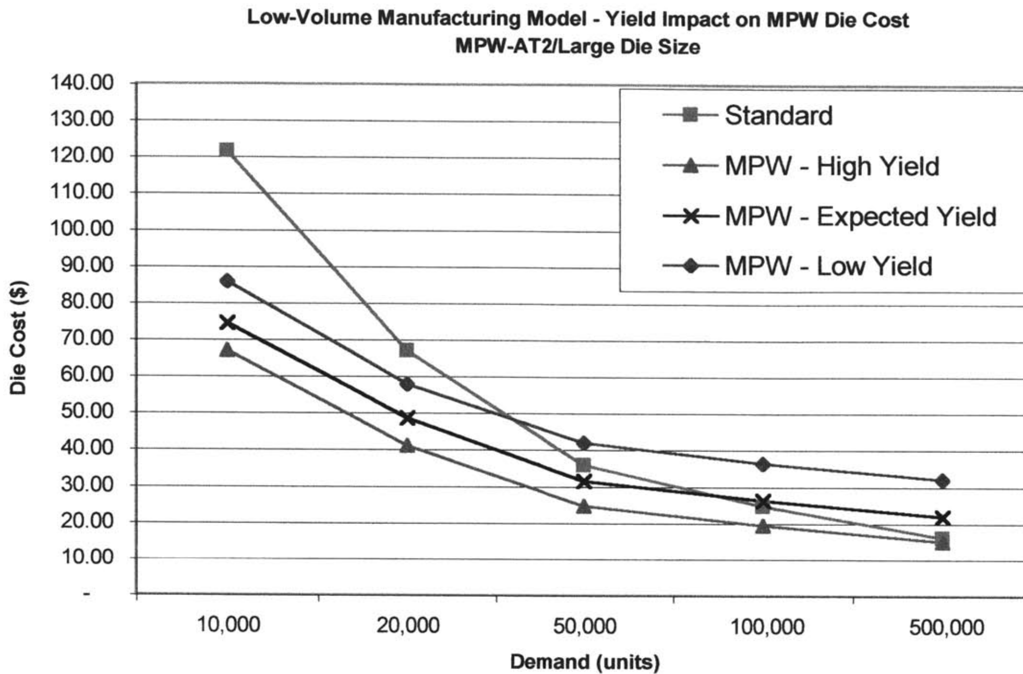


Figure 3.9: Effect of yield on die cost for large dies (500 mils x 500 mils).

3.2.3 Impact of New Product Introductions (NPIs) on wafer cost

In the previous section, it was shown that die yield has only a small effect on the effectiveness of the MPW strategies' cost savings when the products have small or medium sized die. So, a logical, but perhaps counterintuitive, next step is to explore potential savings by allowing the yield to worsen somewhat. Because most of the yield improvement occurs during the NPI process in which the fabrication process is optimized for each product, perhaps some savings could occur by reducing this effort. The MPW strategy automatically splits the NPI cost between several products, but even more savings may be attainable.

Table 3.4 shows the cost of the NPI effort as a percent of total die cost, across the matrix of product demand and die size. In each cell the lowest cost manufacturing strategy was chosen. For example, the MPW-AT8 strategy is used in the upper left corner and the standard strategy is used in the lower right corner.

Table 3.4: Percent of total die cost due to NPI effort

Unit Demand	Small Die →					Large Die
10,000	24%	34%	28%	26%	26%	20%
20,000	24%	29%	23%	21%	20%	17%
50,000	34%	23%	14%	13%	12%	17%
100,000	29%	17%	9%	8%	16%	11%
500,000	16%	5%	2%	8%	5%	3%
1,000,000	10%	3%	1%	5%	3%	1%

The NPI cost is nonlinear because of the discrete nature of manufacturing in lots of 25 wafers and the fact that only the first two lots of a product are affected by the NPI effort. Thus, the percent savings are maximized when only two lots are run to meet demand. The table shows that the NPI tends to cause the highest percent of the total die cost when small die are manufactured in low volumes. This is the same region that is least affected by yield increases! Thus, there may be potential to further reduce wafer costs by reducing the NPI effort, when producing products than fall in the upper left corner of Table 3.4.

3.3 Mapping the lowest cost LVM strategies

Now that the MPWs have been shown to create potential savings over standard production methods, one may wonder: Which, if any, MPW strategy should be used for my product, and how much will I save? Tables 3.5 and 3.6 present a sample of how the model can be used to create a map to match product types with a manufacturing strategy. For each combination of die size and demand, the minimum cost manufacturing strategy was selected and the associated savings is displayed.

Table 3.5: Potential Cost Savings using LVM strategies

Unit Demand	Small Die →					Large Die
10,000	88%	84%	82%	61%	39%	28%
20,000	88%	83%	78%	62%	28%	9%
50,000	84%	78%	73%	47%	12%	0%
100,000	83%	76%	61%	31%	0%	0%
500,000	76%	50%	21%	0%	0%	0%
1,000,000	67%	33%	5%	0%	0%	0%

Table 3.6: LVM Strategies used to achieve maximum cost savings

Unit Demand	Small Die →					Large Die
10000						
20000				MPW-AT4	MPW-AT2	
50000		MPW-AT8				
100000						
500000					Standard	
1000000						

The above savings saving are impressive, but some of the assumptions may be unrealistic. For example, it may not be possible to synchronize the launch of eight products on the same process. Further, the process for sharing wafers, so that multiple products can be assembled and tested from a common wafer, has not yet been developed. Tables 3.7 and 3.8 explore the potential cost savings if MPWs are limited to a maximum of four products and wafer sharing is not allowed (i.e. MPW-AT is not available). The area and magnitude of savings are significantly reduced, but 20% -75% cost savings over standard production for some products can still save millions of dollars.

Table 3.7: Potential Cost Savings with restricted LVM strategies

Unit Demand	Small Die →					Large Die
10000	75%	65%	55%	32%	20%	0%
20000	68%	59%	38%	23%	0%	0%
50000	65%	40%	24%	0%	0%	0%
100000	57%	29%	2%	0%	0%	0%
500000	33%	0%	0%	0%	0%	0%
1000000	15%	0%	0%	0%	0%	0%

Table 3.8: Restricted LVM strategies used

Unit Demand	Small Die →					Large Die
10000					MPW-2	
20000		MPW-4		MPW-2		
50000			MPW-2			
100000		MPW-2	MPW-2		Standard	
500000	MPW-2					
1000000	MPW-2					

The model user must remember that a different set of maps must be created for each process technology that has different mask set costs, wafer costs, or wafer sizes. While the trends will be the same, the borders will move significantly when the factory parameters are modified.

3.4 Conclusions

In this chapter, the Low-Volume Manufacturing Cost Model was developed to explore the costs and benefits of various concepts for manufacturing low-volume semiconductor products. The model identified multi-product wafers as a solution that may create dramatic cost savings. The model's results allowed the comparison of multiple yield or demand scenarios. It also helped define the limited applicable range of some strategies such as small lot size and multi-product lots.

However, the assumptions used in the LVM Cost Model limit its applicability with real-world products that have different die sizes and demands. These limitations inspired the creation of a more detailed model, focusing on the MPW and MPW-AT strategies, that will be described in Chapter 4.

Chapter 4: Development and results of the Multi-product Wafer Cost Model

The previous chapter established the MPW and MPW-AT strategies as best options for manufacturing low-volume products, but there were significant assumptions made in modeling those strategies. This chapter will investigate the impact on cost savings caused by relaxing some of the assumptions made in the Low-Volume Manufacturing Cost Model. Those assumptions were:

- All products on an MPW have identical die size.
- All die have square dimensions.
- All products have identical demand.

These assumptions limited the Low-Volume Manufacturing Cost Model to calculating cost savings for hypothetical products because most products have different die sizes, and most products are not square. Allowing products with multiple die sizes to be placed on an MPW will enable the calculation of savings for real products on the product roadmap. The limitation of identical demand for each product obviously limits the capability of the tool to make estimates of real costs. Real products almost certainly have different demand requirements (unless they are complementary products.) Further, the demand may change. One may wonder what will happen to the manufacturing costs if, for example, demand for one product on an MPW doubles, while the demand for another product on the same wafer declines.

In order to address these issues, the Multi-Product Wafer Cost Model was developed to model real products and the application of MPWs for those products. It also enables the user to assess the financial impact of demand changes once an MPW is designed and the mask set is manufactured. The MPW Cost Model is based on the calculations and equations used in the LVM Cost Model. The major changes include the elimination of the assumptions listed above and the elimination of the other potential low-volume manufacturing strategies (MPL, Small lot size) due to their limited application.

4.1 MPW Cost Model calculations

This section will discuss the calculations and assumptions used to model the design of MPWs. This more advanced model is termed the MPW Cost Model, and only its differences from the LVM Cost Model will be discussed in this section.

4.1.1 Modeling an MPW with multiple die sizes

The first step in modeling an MPW with multiple products is to determine the reticle field dimensions and divide it into smaller areas for the dies to fit into. The field dimensions are typically fixed depending on the photolithography stepper being used, so the model user must enter the reticle field dimensions.

The reticle field is then divided into many smaller areas, termed "seats", which the die will be placed onto. The model user is asked to determine the seat size. The seat size is usually the size of the smallest die to be placed on the MPW, or a size that can evenly divide each of the die intended for the MPW. Because a large die is allowed to take multiple seats, there is no problem with making the seat size very small. The seat size will have an impact on how efficiently the field area is utilized. Figure 4.1 illustrates the division of a reticle field into seats.

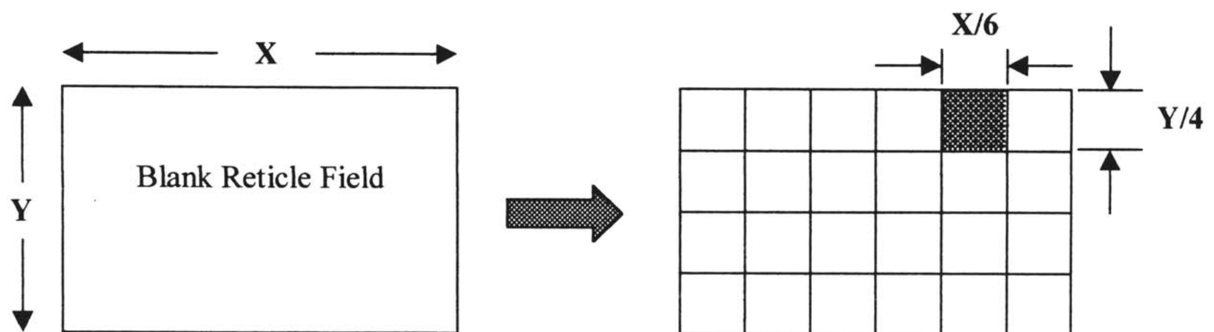


Figure 4.1: Sample division of the blank reticle field into "seats".

Once the seat size is set, the products can start to fill the reticle. First, it must be determined how many seats each product will consume for a single die placed on the reticle. As seen in the example in Figure 4.2, Product A does not completely fill the space it requires on the reticle, and the excess space is wasted. A better selection of the seat size can help reduce this waste.

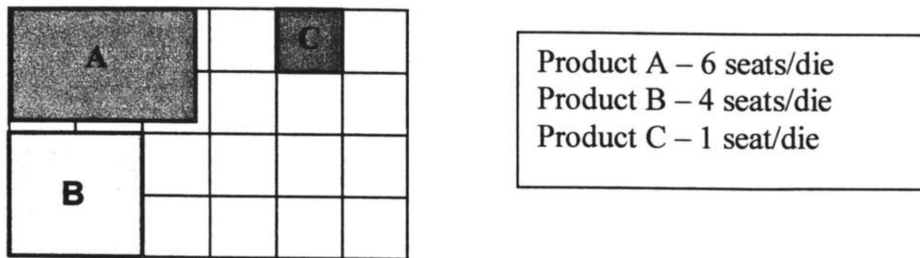


Figure 4.2: Example of determining the number of seats used by each die.

Then the products must be placed on the field in the appropriate ratios to meet demand. The above example can have many different die ratios to accommodate different demands. The model keeps track of the number of seats used as the modeler determines the appropriate ratio. It is acceptable to have extra seats open after the ratios have been determined, but obviously one cannot use more seats than there are available on the reticle. For simplicity, this model does not attempt to create a layout of the dies to be placed on the MPW. A consequence of this simplification is that the model may allow certain combinations of dies to be placed on an MPW that in reality would not fit due to the geometries of the dies and the reticle field. Thus, the modeler must manually check that any desirable combination of dies is physically possible.

4.1.2 Optimization of die ratio

The optimal ratio of product die is the ratio that minimizes the number of lots that must be produced. Determining the optimal ratio and the number of lots that must be produced is somewhat complicated by the fact that the different sized die will have different yields. The process begins by determining the number of whole fields that can be printed on the wafer using an equation similar to Eq. 3.6 as follows:

Eq. 4.1 $$GrossFields/W = \frac{\pi \times (WaferDiameter/2)^2}{FieldArea} - EdgeEffectCorrection$$

where, *FieldArea* is the area of the reticle field.

The number of good die per wafer for each product on the MPW is then calculated by:

Eq. 4.2 $D_A/W = GrossFields/W \times D_A/Field \times Y_{dieA}$

where, $D_A/Field$ is the number of product A dies per field, which is selected manually or by the algorithm below, and Y_{dieA} is the die yield of product "A" as calculated with Equation 3.2.

Because the demand may differ from product to product on an MPW, there may be large differences between the numbers of wafers and lots required by the individual products. The model assumes that the products to be combined on an MPW are launched in the same period of time, and that excess dies are either held in inventory or discarded. The wafer requirement is calculated for each product using

Eq. 4.3 $$WaferRequirement_A = \frac{Demand_A}{D_A/W} \times \frac{1}{LineYield}$$

where, $Demand_A$ is the total market demand for product "A" over its lifecycle. *LineYield* is the percent of whole wafers that survive the manufacturing process (assumed to be 100% here).

It follows that the lot requirement for product "A" is

Eq. 4.4 $$LotRequirement_A = \frac{WaferRequirement_A}{Wafers/Lot}$$

The highest number of lots required for an individual product is chosen so that demand for all of the products can be covered. However, if there is a large difference between the highest and lowest number of lots required, a shift in the ratio of products per reticle may need to be changed. A Visual Basic routine was written to automatically optimize the ratios of different dies on an MPW. The algorithm assures that the demand for all of the products is met while

minimizing the number of lots that must be run, which in turn minimizes the cost of manufacturing the MPW. The optimizer works as follows:

1. Initialize the number of dies per reticle field to "1" for all products.
2. Determine which product drives the most lots
3. Increment the number of die per reticle field for the product identified in Step 2.
4. Check that the number of seats used does not exceed the number of seats available
5. If the number of seats is exceeded, decrement the number of dies per field and stop. Otherwise, repeat the loop starting at Step 2.

One current weakness of the die ratio optimizer is that it does not automatically check the geometry of its results to be sure that the die can physically fit into the field. For example, three 200 mil x 200 mil die would not fit in a field with dimensions of 500mil x 300mil even though there is enough area. The results of the optimizer must be checked by hand before committing to a particular die layout. A long-term improvement to the model would be the creation of an algorithm to automatically check the suggested layout for issues such feasible geometries, area usage, scribe lines for cutting out the dies, and other Design for Manufacturing rules.

4.1.3 Total cost calculation

Once the number of die per field for each product is set, the total MPW cost can be calculated and compared to the cost of manufacturing each product separately. The MPW cost model was kept simple and does not include the NPI effects or batch efficiency factors that were used in the LVM Cost Model. The costs were estimated by simply adding the total mask set costs to the total wafer costs for each case. Thus, the total cost of manufacturing the lifecycle demand of an MPW is

Eq. 4.5 $TotalCost_{MPW} = MaskSetCost + (MaxLotsRequired \times Wafers/Lot \times wc_f)$

where, *MaskSetCost* is the cost of the mask set shared by the products on the MPW, *MaxLotsRequired* is the number of lots that must be run to satisfy all products' demand, and wc_f is the final wafer cost calculated in Equation 3.4.

Similarly, the total cost of the same products manufactured using the standard production strategy is calculated by,

Eq. 4.6 $TotalCost_{STD} = MaskSetCost \times n + wc_f \times Wafers / Lot \times \sum_{i=1}^{n} LotRequirement_i$

where, n is the number of products on the competing MPW, and $LotRequirement_i$ is the number of lots that must be run to satisfy product i's lifecycle demand.

Once the total costs for each option are calculated, the savings created (or lost) by using the MPW strategy is observed. It is then possible to manually adjust the combinations of products placed together on an MPW to look for better solutions. It is possible that certain combinations of products complement each other from demand and die size perspectives and produce better solutions than simply grouping as many products as possible on an MPW.

4.2 Results of the MPW Cost Model

The MPW Cost Model is intended to estimate the potential savings by using the MPW strategy for specific products. Ideally, a manufacturing manager would scan the roadmap of products to be produced in a short time period. The manager would identify products that may be candidates for an MPW using the general tables created with the LVM Cost Model such as Table 3.3. Finally, the manager could use the MPW cost model to determine the savings created by different versions of MPWs and decide how to proceed.

By exercising the MPW Cost Model, the following rules of thumb for designing MPWs have been identified:

- Small dies and low volume create the most significant savings when designing MPWs.
- Products with similar characteristics (demand and die size) reduce the wasted silicon, thus help reduce cost.

- It is often better to create an MPW with a small number of similar products (as measured by die size and demand), than an MPW with many dissimilar products.
- Demand changes do affect the savings, but it typically takes a large increase in demand to overcome the initial MPW savings.

4.2.1 Sample MPW Cost Model results

To illustrate the use of the MPW Cost Model a sample case has been prepared using the generic parameters for a 0.13 micron logic CMOS process as discussed in Chapter 3. Four hypothetical products, listed in Table 4.1, were created for this exercise. The products, while all considered low-volume products, span a fairly wide range of die sizes and demand requirements. Product D with a large die size and relatively high volume does not appear to be a good match for the other products, but we should check the model results.

These are the hypothetical process characteristics used in this example:

- 0.13 micron
- Mask set cost: $650,000
- Wafer cost: $3,000
- Wafer diameter: 300mm
- Reticle field: 1000 mils x 800 mils

Table 4.1: Sample MPW Product Parameters

	Die x-dimension (mils)	Die y-dimension (mils)	Unit Demand (units)
Product A	100	100	50,000
Product B	200	200	50,000
Product C	300	300	100,000
Product D	400	400	200,000

These products were all entered into the model and the die ratios were optimized using the algorithm described above in section 4.1.2. The die ratio was then checked to be sure that the solution was feasible. The resulting layout is illustrated in Figure 4.1.

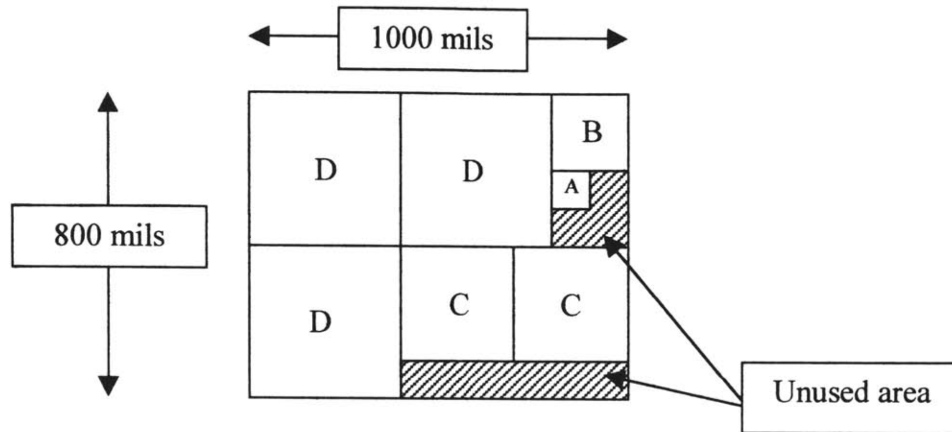


Figure 4.1: Initial reticle layout of a four-product MPW.

The projected cost savings by creating an MPW of these four products are shown below in Table 4.2. A savings of only 8% is not very compelling, so we should examine the results. Product D, with its high demand volume, drives the production quantity of 48 lots. Product C only requires 29 lots with the layout shown. Thus, Product D is driving an additional 19 lots in which 40% of the field is wasted. While the mask set cost savings were a significant $1.95 million, the silicon wafer costs increased by over $1.5 million. This combination of products generates too much wasted silicon to achieve significant savings.

Table 4.2: Cost Comparison of MPW-AT vs. Standard manufacturing of products A, B, C, and D

MPW Wafer Cost	MPW Mask Cost	Total MPW Cost	Standard Wafer Cost	Standard Mask Costs	Total Standard Cost	% savings w/MPW	MPW $ savings
$3,600,000	$650,000	$4,250,000	$2,025,000	$2,600,000	$4,625,000	8%	$375,000

The apparent imbalance caused by Product D leads us to try the design of a three product MPW, leaving out Product D. The resulting layout is shown below in Figure 4.2 with Product C now driving the bulk of the wafer requirements. In this case, only ten lots are required to satisfy the demand for Product C, nine lots are required to satisfy the demand for Product B, and eight lots

for Product A. Interestingly, there is room for additional copies of Product A and Product B, but adding those copies will not reduce the total number of lots required to be manufactured to meet demand. We will see in the next section that adding those additional copies can help protect against demand variation and mix changes. If there was certainty in the product mix, the products could be packed more efficiently, and then the field size could be reduced to reduce the amount of wasted silicon.

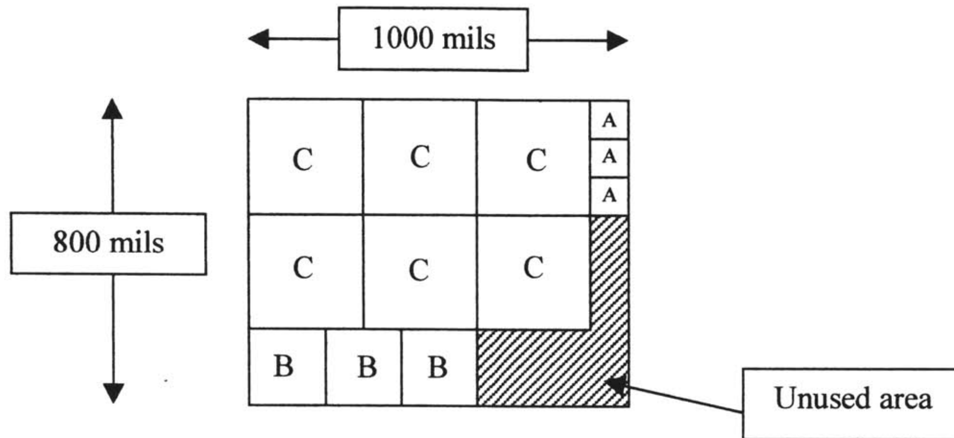


Figure 4.2: Reticle layout of a three-product MPW.

The following results in Table 4.3 show what happens if we drop Product D off of the MPW design. As discussed above, Product D was not a very close match to the other products and was causing the bulk of the waste. The revised MPW, with only Products A, B, and C, provides a significant 43% savings over the standard individual production method. A closer look at the results reveals that the mask set savings were an impressive $1.3 million and the silicon wafer costs, while higher with the MPW, were not nearly enough to offset the mask cost savings.

Table 4.3 : Cost Comparison of MPW-AT vs. Standard manufacturing of products A, B, and C

MPW Wafer Cost	MPW Mask Cost	Total MPW Cost	Individual Wafer Cost	Individual Mask Costs	Total Individual Cost	% savings w/MPW	MPW $ savings
$750,000	$650,000	$1,400,000	$525,000	$1,950,000	$2,475,000	43%	$1,075,000

Other combinations were considered, such as the two-product MPW with Product A and Product B, but the savings were less than 43% in all other cases. Thus, those results are not shown here.

4.2.2 MPW sensitivity to demand changes

One potential drawback of the MPW strategy is that once the mask set is created, the product mix is fairly fixed. It is not possible to change the ratio of die on a mask[5]. To get a feel for the potential sensitivity to product mix and demand changes, four additional variations were conducted with the results of the MPW described above with Products A, B, and C. Since demand forecasts are always wrong and actual demand may fluctuate up or down considerably, the case of twice the predicted demand was considered here. Table 4.4 below shows the resulting cost savings of the MPW, for four different scenarios. In this case, the change to any of the individual demands reduces the total savings. In fact, doubling the demand of all of the products reduces the savings. It should be noted, however, that even though the doubling of demand for one product can reduce the savings, it would take more than a tripling of forecast demand to completely wipe out the profits.

Table 4.4: Demand Sensitivity of Sample MPW

Change in Demand	Cost Savings with MPW Strategy
Baseline	$1,150,000
Demand for A doubles	$700,000
Demand for B doubles	$625,000
Demand for C doubles	$775,000
All demand doubles	$850,000

One method of protecting against demand changes is to fill the mask set with excess die if there is available space. In the example above, the entire reticle field was not filled, even though there was space available for some extra copies of A's and B's. Because Product C, the largest die

[5] As a side note, it may be possible to design the mask set such that sections of it can be "blinded" and only a portion of the mask set would have to be printed. In the worst case, however, the desired dies are located around the edges of the mask so the entire mask must be printed to fabricate these dies.

with the highest demand, was the cost driver, it took the bulk of the space, then A's and B's were fit into the extra space until their demand would just be met. A better solution is to fill the extra field space with copies of the lower volume A and B dies as shown if Figure 4.3. This buffer protects against an upswing in demand for these products.

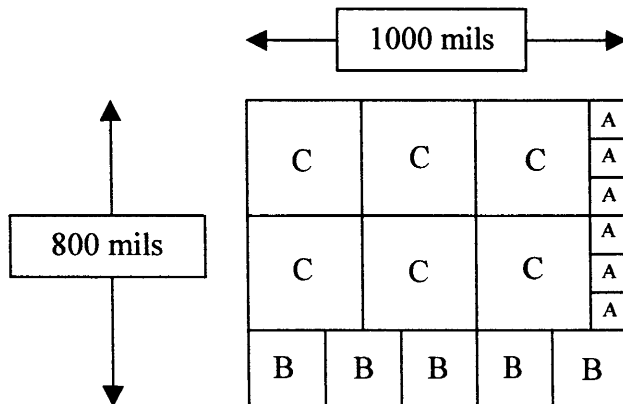


Figure 4.3: Reticle layout of the three-product MPW with excess area filled.

As shown below in Table 4.5, the excess die printing maintains the total savings if demand for Product A doubles, and $300,000 of the savings is protected if the demand for Product B doubles. Because the number of Product C dies was not changed, there is no impact on the effect of the doubling of demand for Product C.

Table 4.5: Demand Sensitivity with excess A and B dies printed

	Cost Savings with MPW Strategy
Baseline	$1,075,000
Demand A doubles	$1,075,000
Demand B doubles	$1,150,000
Demand C doubles	$775,000
All demand doubles	$850,000

4.3 MPW Cost Model conclusions

The creation of the MPW Cost Model has allowed the exploration of some of the more "real world" concerns of using an MPW strategy for production. This model has shown that it is not an optimal solution to place as many products as possible on an MPW. The design of an MPW must consider the die size, the demand, and the potential for demand variability when choosing products for an MPW.

This model has also shown that the opportunity for savings is more realistically in the 40-50% range, rather than the 75-90% range for MPWs. Unless complementary products, in which demand for one drives the demand for the other, can be placed together, the losses due to demand differences and mix changes limit the total savings. While this savings is much lower than the ideal, it can still amount to millions of dollars over the lifetimes of several low-volume products.

Chapter 5: Organizational Challenges of Implementing MPWs

Chapters 3 and 4 modeled the multi-product wafer process and show that significant savings are possible with the MPW strategy. However, the MPW manufacturing strategy cannot be implemented simply by flicking a switch. There are a number of engineering and organizational challenges that must be solved to make the process a success.

To analyze the organizational challenges of implementing the MPW process at Intel, the three classic lenses of organizational study will be utilized here. These lenses, or perspectives, are known as the strategic design, political, and cultural lenses [16].

5.1 Background of the MPS Working Group

As I began investigating the cost-saving potential of MPWs, I became aware of a small group forming within Intel to create a "shuttle" program similar to TSMC's CyberShuttle program. Similar to the CyberShuttle program's purpose, the Intel team's purpose was to improve the speed and cost of the prototyping process. The Intel group was later named and will be referred to as the Multi-Product Shuttle (MPS) working group.

I was able to join the working group creating this program in hopes of sharing my analysis and to learn more about the challenges of manufacturing MPWs. It was fortunate that the team was just beginning to develop the goals and mission that it would follow. By joining the team at the beginning, I was able to expand the scope of the process being developed to include low-volume production products in addition to prototypes.

5.2 Strategic design lens

The strategic design lens focuses on how information flows and tasks occur in the organization. Ideally, the structure of the organization supports the strategy being pursued. Intel's organization is a matrix of functional and business groups. A partial representation of Intel's organization chart is shown below in Figure 5.1. This chart shows that the MPS Team is located in the Technology and Manufacturing Group, which also contains the fabs. Thus, there is an

organizational proximity to the manufacturing side of the process. The links to the finance group and the business group will be the more difficult connections to make and maintain.

The organization chart also points out the location of the author of this thesis, who was located in the Strategic Capacity Planning Group.

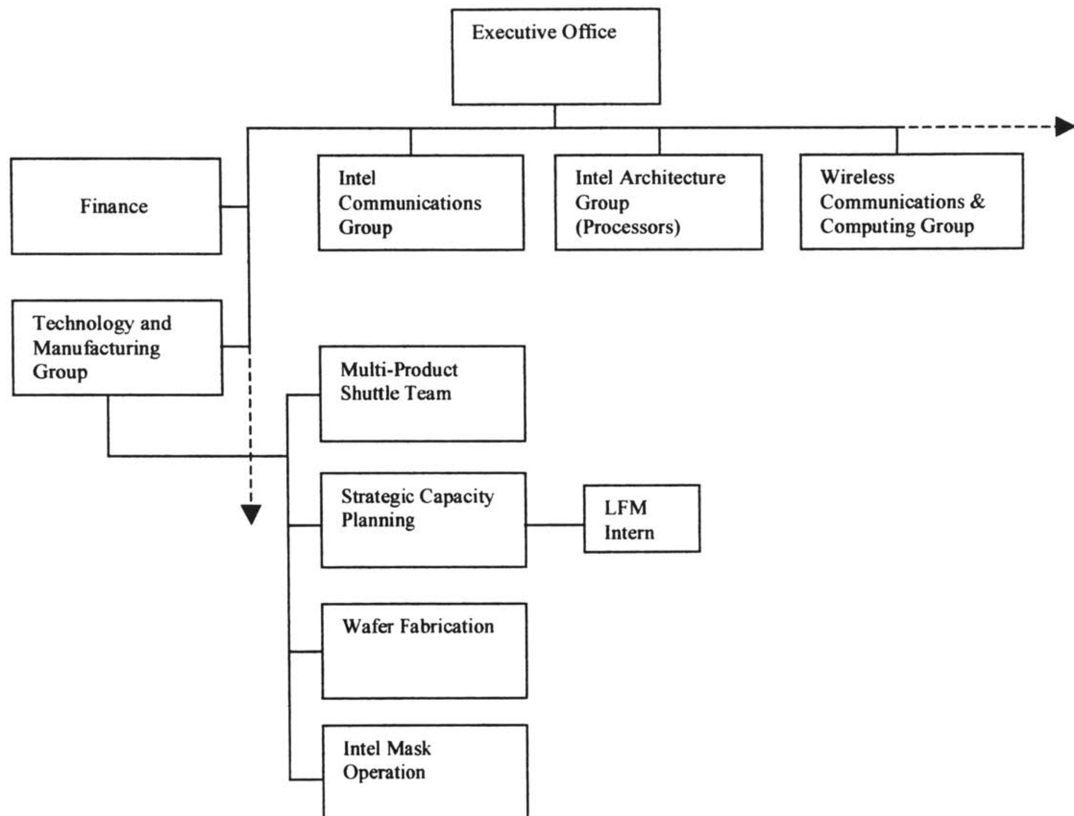


Figure 5.1: Partial Intel "Org Chart"

To analyze the process of creating an MPW, the process must be broken into sub-processes. There are three primary steps to creating an MPW. First, the designs of the chips to be manufactured must come together so a mask set can be produced. Second, the products must be manufactured in a fab. Finally, the products and financial charges must be divided and submitted to the appropriate customers. Each step requires either organizational structure or a process to ensure that the information gets to the right group at the right time.

The initial design of the MPW will require a standard set of information and a schedule. Further, the business groups will probably require some training to help determine which products are best suited to the MPW manufacturing process. Tables similar to the ones shown in Chapter 3 may be useful to the business groups by demonstrating that products with small die sizes and low volume requirements can benefit from significant cost savings made possible with the MPW strategy.

For budget planning, the business groups will also want to know how much the product will cost to manufacture on an MPW. This will be difficult to determine until an understanding of demand is achieved because the cost of each product depends on the number of other products produced on the MPW. A table similar to the one shown in Table 5.1 has been created with the output of the MPW Cost Model to give the business groups an estimate of the costs that may be expected. This table assumes that the other products have typical die sizes and demands.

Table 5.1: Sample MPW price schedule (values disguised)

Sample MPW Price Schedule

Four Passenger MPW ($/product per batch)

Die Quantity (units)	Die Size: SMALL →					LARGE
FEW	$ 50,000	$100,000	$110,000	$110,000	$115,000	$120,000
↓	$ 50,000	$100,000	$110,000	$115,000	$120,000	$140,000
↓	$ 80,000	$100,000	$110,000	$120,000	$135,000	$170,000
↓	$100,000	$100,000	$120,000	$130,000	$160,000	$200,000
MANY	$150,000	$112,000	$150,000	$160,000	$200,000	$300,000

Six Passenger MPW ($/product per batch)

Die Quantity (units)	Die Size: SMALL →					LARGE
FEW	$ 50,000	$ 70,000	$ 75,000	$ 75,000	$ 75,000	$ 80,000
↓	$ 50,000	$ 70,000	$ 75,000	$ 75,000	$ 80,000	$ 95,000
↓	$ 55,000	$ 70,000	$ 75,000	$ 80,000	$ 90,000	$115,000
↓	$ 70,000	$ 70,000	$ 80,000	$ 85,000	$105,000	$135,000
MANY	$100,000	$ 75,000	$100,000	$105,000	$135,000	$200,000

Coordinating across design and manufacturing groups to create a common method of running MPWs is a key to achieving the strategic objectives of this process. Because there is demand for the service being provided by the MPS team, if it cannot provide the service other groups will. The fabs and mask operation certainly will not like multiple versions of the MPW process, so it is important to make sure the processes created meet the needs of all of the stakeholders. The MPS team will have to continue to create contacts and publicize the benefits and results of the process as they occur.

5.3 Political lens

The political lens focuses on the locus of power and influence within an organization. The success of an organization can depend on the support of many stakeholders. At a high level, the MPS group is well positioned from the political perspective. The group is positioned within the Technology and Manufacturing Group, which is the most likely group to object to the process. The MPW process will most likely introduce more low-volume runs into the fabs as new communication chips are developed. These low-volume runs of prototypes and products will worsen the overall fab utilization and slow the cycle time. Thus, MPWs may damage the metrics that the fab managers' performance is measured with. The MPS team, with its organizational location, should be able to internally market itself as a cost saver, and change the metrics that the process affects. The MPW process is being requested by the Intel Communication Group (ICG) operations, so there should not be any resistance from that area. However, the MPS team needs to continue to solicit customer requirements from the ICG operations. The Mask Operation should also benefit from the MPW process because it should reduce the demand for unique mask sets. This lower demand should help it to reduce capital expenditures on the very expensive equipment required to make next generation masks. Even if the capital expenditure is not reduced, there is at least a possibility of improved cycle time in the mask shop. Again, the MPS team needs input, if not a working group member from the Mask Operation.

The business groups that drive the majority of the capacity in the fabs, such as the processor designers of the Intel Architecture Group, may have some objections to the introduction of many low-volume products into the fabs. Those groups may fear losing production capacity or a degradation of production quality due to the different procedures used in producing MPWs or other low-volume runs. The MPS team can best reduce the fears by recognizing the potential

problems and by working with the business groups to avoid them. Further, the MPS team can offer its fast and inexpensive prototyping service to the business groups with higher volume products to potentially gain their support.

The Finance department is the last area that will require extensive partnering during the establishment of the MPW process. The division of the cost of a few lots across several business groups will probably seem like an unnecessary and complicated procedure to the finance people. Of course, it is the division of the costs that makes the procedure attractive in the first place!

5.4 Cultural lens

The Cultural lens focuses on the organization's history and the assumptions and meanings driven by that history. Intel's success in manufacturing semiconductors can be partially attributed to its obsession for creating standard processes. Because the use of MPWs to reduce manufacturing costs potentially violates the Design for Manufacturing rules already in place at Intel, it may be difficult to make exceptions for MPWs in the fabs. It will have to be recognized that the MPWs, while taking more fab time than high-volume products, are actually saving time by consolidating several low-volume products.

Another aspect of the MPW process that may cause cultural problems is the lower yield of wafers produced using that process. The yield of products on an MPW is unlikely to compare favorably with the yield of high-volume products because the manufacturing process cannot be optimized for a specific process. As shown in Chapter 3, improving the yield of low-volume products is not nearly as important as saving mask set costs. These changes in cost saving priorities will potentially cause confusion if they are not carefully managed. Early discussions with several manufacturing engineers, revealed a general mistrust of ideas that violated established design rules. Based on Intel's data-based engineering culture, the best way to counter these objections will be with real data. Thus, it will be crucial to accumulate yield data of products produced with an MPW and show that the lower yields are not detrimental to the cost savings.

5.5 Conclusions

Overall, the keys to success in implementing the MPW process will be creating a process that is technically feasible and forging supporting relationships with the business groups, the manufacturing groups, and the finance groups. As was discussed with the MPS working group prior to the author's return to MIT, the MPS working group should recruit members from the mask operation and finance groups to support and promote the team's interests. The process will need advocates throughout the organization to both define the process and to help achieve a critical mass of customers to fill up the MPWs on a regular basis.

Chapter 6: Conclusions and recommendations for further research

This thesis has shown that low-volume products have a much different cost structure from high-volume products. This is primarily due to the distribution of fixed cost across fewer dies. The cost modeling has shown that there is potential for significant improvement in the cost of manufacturing low-volume products through the use of a multi-product wafer manufacturing strategy. Revisiting the product-process matrix discussed in Chapter 1, the two question marks in Figure 1.1 can potentially be replaced by the two MPW strategies. Figure 6.1 shows the proposed revision to the product-process matrix.

Product structure
Product life cycle

I Low-volume (Prototypes)
II Multiple Products (Optical/Ethernet)
III Few major products (Processors, Flash)
IV High-volume – (DRAM, discretes)

Process structure
Process life cycle stage

I **MPW**
II **MPW-AT**
III Copy Exactly!/NPI process
IV Dedicated Toolset, Large Batches

Typical Acquisition
Traditional Intel
None
None

Flexibility-quality
Dependability-cost

Figure 6.1: Revised Hayes-Wheelwright product-process matrix.

The MPW strategy in which wafers are not shared between groups will work best for extremely low-volume products or prototypes. The MPW-AT strategy, which requires that the MPWs be carefully cut apart to separate the various dies, is a strategy that will make sense for products with demand requirements of less than roughly 500,000 units, depending on die size.

The other low-volume manufacturing strategies considered in this thesis such as small lot sizes, have some merit for limited applications or in combination with the MPW strategies. The MPL and small lot size strategies, with their relative inability to share fixed costs, simply cannot compete with MPWs in reducing the manufacturing costs of low-volume products. The reduction of the NPI effort would benefit low-volume products in both standard and MPW manufacturing strategies.

The introduction of many low-volume products and processes will also potentially cause some organizational problems. Intel's organization and culture has evolved to support the design and manufacture of high-volume microprocessors and memory chips. If Intel plans to achieve excellence in the communications markets, it will have to adapt its organization to meet the specific needs of that market, which include low-volume products, low cost, and fast time-to-market. The processes required to meet those needs will undoubtedly increase the complexity of running a wafer fab. Fab managers, engineers and operators, will need a clear understanding of the differences in priorities of high-volume and low-volume products.

This thesis has shown that there are several options for manufacturing low-volume products. Each option has unique technical and financial costs and benefits, and of those options considered here, the MPW processes appear to be the best solutions for manufacturing low-volume products in high-volume semiconductor fabs.

Issues or unknowns in the implementation of MPWs that were not resolved by this thesis and warrant further investigation are listed below:

Actual yield of MPWs – the actual die yield of chips produced on an MPW has not yet been measured by the fabs. An understanding of the actual yield compared to the expected yield is needed to fully understand the tradeoff between mask set cost and wasted silicon.

NPI effort vs. yield improvement – the NPI process is a much larger part of the cost of producing low-volume products than high-volume products as was shown in Chapter 3. It would be valuable to understand how much of the NPI process can be eliminated while still achieving acceptable yield.

MPW die separation – to fully take advantage of the MPW-AT strategy for low-volume products, the capability to separate different dies from the MPW must be developed or acquired. Laser wafer saws, which have the ability to cut odd shapes from a wafer, are currently available on the market but are not integrated into Intel's assembly/test processes. Alternatively, low-volume products could be designed to have standard die sizes.

Pattern density as a MPW product selection criteria – all of the analysis in this thesis neglected product pattern density as a potential criteria for combining products on an MPW. An understanding of the range of pattern densities than can be produced on an MPW with acceptable results is needed. Also, how much difference can the use of dummy structures make in the overall pattern density of a product?

MPW product reliability – because the MPW strategy has only been used for prototypes in the past, an understanding of whether the MPW process has an impact on product reliability should be developed.

Development of a low-volume product "culture" – the manufacture of low-volume products with MPWs will require the established culture of "focusing on yield improvement" to become more flexible. This focus is part of what has made Intel so successful in the past, so implementing such a change may be extremely difficult. A detailed look at how Intel's organization can best support the emerging communication markets would be valuable.

Appendix A: List of Acronyms

ASIC – application specific integrated circuit

AT – assembly/test, the process of separating, packaging, and packaging individual dies

CE! – Copy Exactly!, Intel's process for transferring manufacturing process technologies from development to high-volume fabs

CMOS – complementary metal oxide semiconductor, a transistor design typically used for logic circuits

CMP – chemical mechanical polishing

CVD – chemical vapor deposition

HVM – high-volume manufacturing

IAG – Intel Architecture Group

IC – integrated circuit

ICG – Intel Communications Group

LVM – low-volume manufacturing

MPL – multi-product lot

MPS – Multi-Product Shuttle

MPW – multi-product wafer

NPI – new product introduction

PC – personal computer

TD – technology development

TMG – Technology and Manufacturing Group

TSMC – Taiwan Semiconductor Manufacturing Company

WIP – work in process

Appendix B: Low-Volume Cost Model results

These data tables were used to create the figures used in the analysis shown in Chapter 3.

Comparison of low-volume manufacturing strategies' die costs

Die Size: 100 mils

			Unit Demand		
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	95.90	47.95	19.18	9.59	2.55
Quarter Lot	76.43	38.21	15.29	8.42	3.15
Half Lot	83.48	41.74	16.70	8.35	2.80
MPL - 2	80.45	40.22	16.09	8.04	99,999.00
MPL - 4	72.72	36.36	14.54	7.92	99,999.00
MPL - 8	68.86	34.43	14.42	7.36	99,999.00
MPW - 2	47.95	23.97	9.59	6.08	1.72
MPW - 4	23.97	15.21	6.71	4.11	1.87
MPW - 8	15.92	9.32	5.42	4.03	2.91
MPW - AT2	47.95	23.97	9.59	4.79	1.42
MPW - AT4	23.97	11.99	4.79	3.04	0.86
MPW - AT8	11.99	5.99	3.04	1.68	0.60

Die size: 300 mils

			Unit Demand		
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	95.90	47.95	25.47	14.16	5.69
Quarter Lot	86.49	53.99	32.02	18.52	7.73
Half Lot	83.48	50.98	28.53	16.39	6.74
MPL - 2	99,999.00	99,999.00	99,999.00	99,999.00	99,999.00
MPL - 4	80.59	99,999.00	99,999.00	99,999.00	99,999.00
MPL - 8	74.58	99,999.00	99,999.00	99,999.00	99,999.00
MPW - 2	63.67	35.41	19.43	13.85	9.46
MPW - 4	42.93	29.93	22.13	19.52	17.37
MPW - 8	46.85	40.35	36.45	35.05	33.96
MPW - AT2	60.85	31.84	14.92	9.71	5.25
MPW - AT4	31.84	17.71	9.71	6.93	4.73
MPW - AT8	17.71	10.73	6.93	5.53	4.47

Appendix B: Low-Volume Cost Model results (cont.)

Die Size: 500 mils

			Unit Demand		
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	121.70	67.06	35.85	24.69	16.07
Quarter Lot	150.03	87.44	46.97	33.48	22.68
Half Lot	135.24	77.80	41.30	29.26	19.57
MPL - 2	99,999.00	99,999.00	99,999.00	99,999.00	99,999.00
MPL - 4	99,999.00	99,999.00	99,999.00	99,999.00	99,999.00
MPL - 8	99,999.00	99,999.00	99,999.00	99,999.00	99,999.00
MPW - 2	97.15	69.26	52.53	46.95	42.56
MPW - 4	99,999.00	99,999.00	99,999.00	99,999.00	99,999.00
MPW - 8	99,999.00	99,999.00	99,999.00	99,999.00	99,999.00
MPW - AT2	74.58	48.57	31.46	26.26	21.80
MPW - AT4	99,999.00	99,999.00	99,999.00	99,999.00	99,999.00
MPW - AT8	99,999.00	99,999.00	99,999.00	99,999.00	99,999.00

Appendix B: Low-Volume Cost Model results (cont.)

Yield impact on MPW die cost

Medium Die Size: 300 mils

High Yield -- Defect Density: 0.3 cm^{-1}

	Unit Demand				
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	95.90	47.95	25.47	14.16	5.69
MPW - AT8	16.76	10.26	6.17	4.87	3.78

Expected Yield -- Defect Density: 0.5 cm^{-1}

	Unit Demand				
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	95.90	47.95	25.47	14.16	5.69
MPW - AT8	17.71	10.73	6.93	5.53	4.47

Low Yield -- Defect Density: 0.7 cm^{-1}

	Unit Demand				
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	95.90	47.95	25.47	14.16	5.69
MPW - AT8	18.65	11.67	7.68	6.38	5.32

Large Die Size: 500 mils

High Yield -- Defect Density: 0.3 cm^{-1}

	Unit Demand				
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	121.70	67.06	35.85	24.69	16.07
MPW - AT2	67.06	41.05	24.69	19.49	15.03

Expected Yield -- Defect Density: 0.5 cm^{-1}

	Unit Demand				
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	121.70	67.06	35.85	24.69	16.07
MPW - AT2	74.58	48.57	31.46	26.26	21.80

Low Yield -- Defect Density: 0.7 cm^{-1}

	Unit Demand				
Mfg. Strategy	**10,000**	**20,000**	**50,000**	**100,000**	**500,000**
Standard	121.70	67.06	35.85	24.69	16.07
MPW - AT2	85.86	57.98	41.99	36.42	31.95

References

[1] Kim, Tae. "Still gloomy global PC market", Daewoo Securities – PC sector, September 11, 2001.

[2a] InfoTech Trends. "01Q3-Forecast market for network processors(NPUS)", www.infotechtrends.com, 2002.

[2b] InfoTech Trends. "02Q1-Forecast market for metro Ethernets", www.infotechtrends.com, 2002.

[2c] InfoTech Trends. "01Q3-Forecast market for 10 Gigabit Ethernet", www.infotechtrends.com, 2002.

[3] Edwards, Cliff. "Intel", Business Week, The McGraw-Hill companies, October 15, 2001, pp. 80-90.

[4] Hayes, Robert, Wheelwright, Steven. "Link manufacturing process and product life cycles", Harvard Business Review, January-February 1979.

[5] Mlynarczyk, Michele. "Improving Synergy in Multi-site Microprocessor Manufacturing: An Analysis of a Copy Exactly Approach", Thesis, Massachusetts Institute of Technology, 1995.

[6] LaPedus, Mark. "0.10-micron technology could drive up mask costs to $2 million per set, says TSMC", Semiconductor Business News, January 30, 2002.

[7] Xiao, Hong. Introduction to Semiconductor Manufacturing Technology, Prentice Hall, 2000.

[8] TSMC website. Cybershuttle page, www.tsmc.com, 2001.

[9] Froot, Kenneth. "Intel Corporation, 1992", Harvard Business School Case Study, 9-292-106, 1993

[10] Engde, Kurt-Ingvar. "Ericsson's VLSI Mini-fab Strategy: Low Volume VLSI Fab to Ensure Short Time to Market for Ericsson Telecom Systems and Products", IEEE/CPMT International Electronics Manufacturing Symposium, 0-7803-2037, 1995.

[11] Guo, Ruey-Shan, Slama, Mike, Griffin, Ralph, Holman, Ken. "A Work Cell Manufacturing System for VLSI Fabrication", IEEE/CHMT International Electronics Manufacturing Technology Symposium, 0-7803-1424, 1993.

[12] Berglund, Neil. "A Unified Yield Model Incorporating Both Defect and Parametric Effects", IEEE Transactions on Semiconductor Manufacturing, Vol. 9, No. 3, August 1996.

[13]Chan, Mansoon. "Introduction to CMOS Technology", presentation, University of California at Berkley, 2001.

[14] Blackwell, Jason. "Foundry wafer prices: Still hang tough, but for how long?", Semiconductor Business News, November 6, 2001.

[15] Hutcheson, G. "VLSI's Reticle Forecast", Yield Management Solutions, Autumn 2000.

[16] Ancona, Deborah, Kochan, Thomas, Scully, Maureen, Van Maanen, John, Westney, D. Eleanor. Organizational Behavior and Processes, South-Western College Publishing, 1999.

Other interesting resources not discussed in this thesis

[17] Dasgupta, Abhijit, Magrab, E. B., Anand, D. K., Eisinger, Karlheinz, McLeish, James, Torres, Myra, Lall, Pradeep, Dishongh, Terrance. "Perspectives to Understand Risks in the Electronic Industry", IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A. Vol. 20, No. 4, S1070-9886(97), 1999.

[18] Gedaly, Jerry. "Economic Justification of a 6" GaAs Wafer Fab", GaAs Mantech, 1999.

[19] Leachman, Robert, Plummer, John, Sato-Misawa, Nancy. "Understanding Fab Economics", Report CSM-47, Competitive Semiconductor Manufacturing Program, University of California at Berkley, 1999.

[20] Maly, W., Jacobs, H., Kersch, A. "Estimation of Wafer Cost for Technology Design", IEEE Transactions, 0-7803-1450-6, 1993.

[21] MOSIS website. cost page, www.mosis.org, 2001.

[22] Bonn, Jeffrey, Sisler, Sharon, Tivnan, Patricia. "Balancing Mask and Lithography Costs", IEEE Transactions, 07803-6555-0, 2001.