13

Reducing Variability in a Semiconductor Manufacturing Environment

by

Minja Johanna Penttilä

M.S., Materials Science, Helsinki University of Technology (2000)

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

Master of Business Administration and
Master of Science in Electrical Engineering and Computer Science

In conjunction with the Leaders for Manufacturing Program at
the Massachusetts Institute of Technology
June 2005

Signature of Author_____
Sloan School of Management
Department of Electrical Engineering and Computer Science
May 6, 2005

Certified by_____
Thomas Roemer
Assistant Professor of Operations Management
Thesis Advisor

Certified by_____
Roy E. Welsch
Professor of Statistics and Management Science, Director CCREMS
Thesis Advisor

Accepted by_____
David Capodilupo
Executive Director, MBA Program
Sloan School of Management

Accepted by_____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses
Department of Electrical Engineering and Computer Science

Reducing Variability in a Semiconductor Manufacturing Environment

by
Minja Johanna Penttilä

Submitted to the Sloan School of Management and the Department of Electrical Engineering and Computer Science on May 6, 2005 in the Partial Fulfillment of the Requirements for the Degrees of
Master of Business Administration
and
Master of Science in Electrical Engineering and Computer Science

ABSTRACT

The main drivers in today's flash memory business are low cost and flexibility. Low cost requires high tool utilizations, whereas flexibility and ability to respond quickly to changing customer demands require short throughput times. There is, however, an inherent operational conflict with achieving both high utilization and short cycle time simultaneously. Intel's flash memory factory is striving for shorter manufacturing throughput times without reducing tool utilizations. One of the major components in throughput time today is queuing time caused partly by variability in the manufacturing environment. Being able to reduce this variability component could result in improvements in throughput time.

In this work, Factory Physics methods are used to analyze variability in the manufacturing flow. First, potential high variability areas in the flow are identified. Second, manufacturing data is analyzed to find the main sources of variability. Third, ways to reduce variability are investigated. Finally, means to align manufacturing metrics with variability reduction efforts and the effect of metrics on organizational culture and change implementation are discussed.

During the study it was found out that the lithography area reduces the overall manufacturing flow variability. It was also found out, that the area is highly utilized and is thus introducing non-value adding queuing time for the product throughput time. Arriving material flow was identified to be the main source of variability. Recommendations for improving the area performance include optimizing tool dedications, standardizing operator decision making, and changing preventive maintenance operations.

The key takeaway from this study is the importance of metrics alignment. Metrics are the most powerful incentives for operator behavior. Unless the daily floor level performance measurements are aligned to support the organizational goals, implementing new operations management methods to reduce variability will be challenging.

Thesis Advisors:

Thomas Roemer
Assistant Professor of Operations Management

Roy E. Welsch
Professor of Statistics and Management Science, Director CCREMS

ACKNOWLEDGEMENTS

I would like to thank the Leaders for Manufacturing program, Don Rosenfield and the program staff for this great two-year journey. I would also like to thank my thesis Advisors Thomas Roemer and Roy Welsch for their guidance during my internship project.

I would like to acknowledge Intel Corporation for providing me the opportunity to learn from one of the best semiconductor manufacturers in the world. I am especially grateful for Fab11 Plant Manager Dave Baglee for sponsoring my internship and for my company supervisor Jim Blumenschein for his time and guidance. In addition, I would like to thank Eric Stebbins, Zane Rakes, Randy Ortega, Tee Welton, Justin Sanchez, Kyle Lee, Tina Volzer, and George Castellano and the rest of the litho staff for their help and support.

These past two years would not have been the same without my awesome LFM and Sloan classmates; special thanks to my LFM summer team for countless laughs during those long summer evenings.

Most importantly, I would like to thank my family and loved ones for their infinite support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 PROBLEM STATEMENT

Intel Corporation is the world's largest semiconductor manufacturer. The company manufactures microprocessors (MPUs), flash memories, chipsets and related software for computers, servers, and consumer electronic products. This thesis work was conducted in Fab11 in Rio Rancho, New Mexico, where Intel manufactures mainly flash memory products. Flash is a non-volatile memory chip that holds its content even when the power in the mother device has been switched off. Flash memory is manufactured using standard semiconductor manufacturing technologies: implant and diffusion to introduce elements into silicon to form n- and p-type regions, photolithography to pattern various layer structures, and metallization to form conductive electrical circuitry.

The industry environment for flash products is very different than that for microprocessors. Intel has been the dominant player in the microprocessor industry since the company was founded. Conversely, in the highly segmented commodity-like flash business the company faces new challenges. Intense price competition and lower profit margins create strong pressures for manufacturers to manage their cost structures effectively. Since depreciation cost from capital equipment is one of the main components contributing to silicon wafer cost, tool utilization has become an important measure. In addition, timing of market entry and flexibility to respond to volatile customer demand create pressures for short manufacturing cycle times.

According to Factory Physics [1], there is an inherent conflict with high tool utilization and low cycle time; cycle time approaches infinity when utilization approaches 100%. Cycle time is comprised of process time and queue time. From the customer's point of view only process time is value adding time. Queuing time on the other hand is non-value adding and can be introduced to the system due to a variety of reasons, such as waiting for an operator or a tool to become available. Queuing time is dependent among other things on variability in the manufacturing environment.

---

[1] W.J. Hopp, M.L. Spearman: Factory Physics, 2nd Edition, McGraw-Hill 2001 New York

Fab11 needs to stay competitive in the intense flash industry environment. This can only be accomplished by achieving a low cost but highly flexible manufacturing environment. Currently, wafer cost is the dominating driver in the organization, which results in high forced tool utilizations. At the same time the factory suffers from high cycle times. In addition, current performance metrics in use in the facility might be promoting the wrong kind of behavior and hiding critical problems instead of supporting the factory mission.

1.2 MOTIVATION, OBJECTIVES AND METHODOLOGY

The purpose of this thesis work was to help Fab11 to achieve operational excellence by reducing barriers to improved cycle time. The work focused on analyzing manufacturing variability and related performance metrics. The primary goals were to understand and reduce the sources of manufacturing variability, as well as to help Fab11 understand how to improve their manufacturing performance measurement systems.

The variability analysis work was done as part of an existing operations improvement task force team. The analysis used concepts both from Factory Physics as well as from lean manufacturing. Factory Physics methods were chosen since they supported on-going research activities in the organization. The metrics alignment work was conducted alongside with a company internal metrics team. Material was found from company internal documents and literature, as well as through interviews and direct observations on the floor level.

1.3 ORGANIZATION OF THESIS

Chapter 1 presented the problem statement, objectives and methodologies. Chapter 2 will discuss flash memory technology and industry dynamics. Factory Physics concepts are introduced in more detail in Chapter 3. Chapter 4 focuses on variability analysis and presents the developed analysis method together with results and recommendations. Chapter 5 discusses the importance of metrics and presents weaknesses and improvement recommendations for the current metrics system. Organizational challenges for change implementation are covered in Chapter 6, and finally, thesis conclusions and key insights are given in Chapter 7.

2 MEMORY INDUSTRY ANALYSIS

This chapter discusses flash memory structure, technologies as well as the dynamics in flash memory industry.

2.1 WHAT IS A FLASH MEMORY?

Flash is a non-volatile memory chip that holds its content even when the power in the mother device has been switched off. Flash is a form of EEPROM (electrically erasable programmable read-only-memory), which means that memory units, called blocks, can be electrically erased and reprogrammed. Unlike in EEPROM, flash allows multiple memory cells to be erased or reprogrammed in one operation which gives it higher speed. Other benefits include reliability, low-cost, high-density storage, operation at low voltages and retaining data when power is off. Typical applications of flash memory include memory sticks, digital cameras, cell phones, printers, and handheld computers. [2,3]

Flash memory is manufactured using standard semiconductor manufacturing processes: wafer fabrication, electrical die sort, die packaging and device testing. Flash memory cell structure is presented in the following figure:

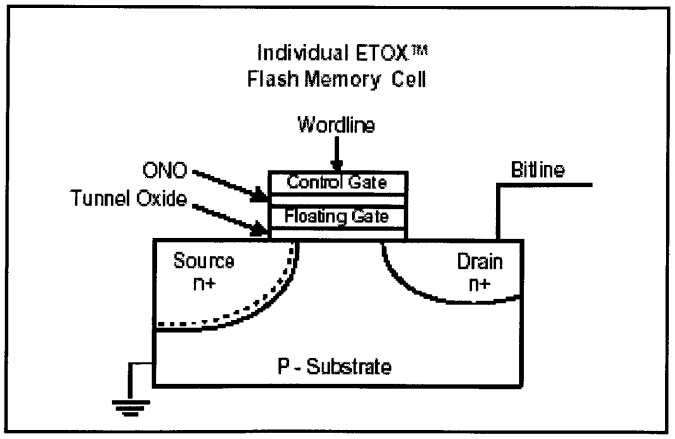

Figure 1. Structure of a flash memory cell.[4]

[2] http://en.wikipedia.org/wiki/Flash_memory
[3] "What is flash memory?" October 2002. http://www.intel.com/design/flash/articles/what.htm
[4] www.intel.com

During the wafer fabrication phase memory cell structures are built on "empty" silicon wafers using standard semiconductor manufacturing technologies, such as implant and diffusion to introduce elements into silicon to form n- and p-type regions, photolithography to pattern various layer structures, and metallization to form conductive electrical circuitry. Through these processes each memory cell receives its physical transistor structure with source, drain and gate regions, as was shown in Figure 1. Hundreds of individual memory dies are formed on one silicon wafer. Wafer fabrication operations are comprised of hundreds of highly sensitive process steps. Though all dies on a wafer go through the same manufacturing flow, exposure on different areas on the silicon wafer surface might differ. Therefore, functional performance level of individual dies on a same silicon wafer might be different. In die sort phase individual dies are classified into different product groups based on their performance. After sort dies are packaged to make them easier to test, to handle and to assemble by the end customer, for example a camera manufacturer.

2.1.1 Flash Memory Technologies

Flash memory comes in two different types: NOR (Not Or) which is traditionally used for code storage (e.g. operating systems), and NAND (Not And) which is suitable for data storage (e.g. digital pictures). The names refer to the type of Bolean algebra logic operations used in storage cells.

NOR flash was invented by Intel in 1988. NOR is a high read speed type with full memory interface that allows random access to any cell location. This makes it suitable for storage of program code that needs to be infrequently updated. NAND flash was developed by Samsung and Toshiba in 1989. It is fast erase and write type which makes it more suitable for mass-storage devices that experience consistent need to update information, like memory cards in digital cameras. NAND has a significant cost-performance advantage over NOR, since NAND has lower cost per bit resulting from smaller physical size of memory cells. NAND can also achieve considerably higher memory densities: up to several Giga bytes vs. NOR Mega byte range. [5]

[5] http://en.wikipedia.org/wiki/Flash_memory

2.2 FIVE-FORCES ANALYSIS

Semiconductor Industry Association (SIA) reported an 18% growth in semiconductor sales between 2002 and 2003. The estimate for 2004 showed a 28% growth from 2003, resulting in the overall semiconductor industry size of $214 billion. Predictions for future vary, but the SIA anticipated the sales to be $223 billion in 2005, $221 billion in 2006 and around $250 billion in 2007. Semiconductor industry is highly cyclical, and industry trends have been observed to follow closely the overall economic trends. This relationship is understandable, since the overall economic environment affects the demand of computers and consumer electronic products which are the main users of semiconductor products. While personal computers and large computer systems remain critical market segments, other segments like communications, handheld devices and automotive applications have experienced a rapid growth. [6]

Barriers To Entry

Intel's co-founder Gordon Moore forecasted in his famous law from 1965 that the number of transistors per chip will double every 2 years. Instead of a consumer pull, semiconductor industry has been driven by Moore's law resulting in technology push and continuous need to invest in manufacturing process technologies. For chip manufacturers Moore's law means exponentially growing transistor counts per chip, increased number of chips per wafer and thus increased revenue per wafer. However, shrinking chip feature sizes mean also increased investments and technical challenges. A good example is lithography, a key technology driver for the industry, which continuously needs to develop to enable the patterning of smaller and smaller device dimensions. Today top of the line lithography tools used in volume production cost around $15 million a piece and use 193nm light sources to pattern average feature sizes of 90nm. Further investments in new photolithography technologies, such as EUV (extreme ultraviolet), are needed to reach average feature sizes of around 30nm.[7]

---

[6] M. Singer: "Report Chills Chip Futures", http://www.internews.com/infra/article.php/3379791
[7] www.news.com: "Intel sheds light on future chip technology", August 1, 2004.

In 2003, the industry leader Intel spent $3.7 billion on capital investments and $4.4 billion on R&D. The majority of the capital investments were spent on factories and equipment. Continuous investments in advanced manufacturing process technology enable Intel to increase uniqueness of its products: shrink transistor sizes and increase the number of transistors per chip, decrease the size of a chip or offer new features such as increased processing speed, lower power consumption or lower manufacturing cost. As an example, new 300mm wafer process yields twice the number of dies per wafer than older 200mm process, having thus a significant impact on manufacturing economies.[8] Older tool generations lack technical capability and manufacturing economies. The steep learning curve requiring high continuous investments in manufacturing process technology, high level of required technical knowledge as well as the dominant position of Intel in the market place can be seen as being the main barriers of entry for the microprocessor industry.

However, it is important to notice that the industry environment for flash products is very different from microprocessors. Microprocessors, such as Pentium chips, are high end and hard to manufacture, whereas flash is lower technology and can be seen as being more of a commodity. Due to these lower production technology requirements learning curve and capital requirements are lower making it easier for companies to enter the market. Some companies even function "fabless" meaning that they do not have any manufacturing facilities. Due to lower barriers to entry and positive future demand predictions, several players have entered the markets making the flash industry much more segmented than the microprocessor industry.

Competition

In the microprocessor industry barriers to entry have significantly limited competition. Intel, who invented the microprocessor, has been the dominant player in the industry since the company was founded. Today, Intel still continues to dominate with its leading edge chip generations resulting in 82% market share in PC processors (see Figure 2).

[8] Intel Annual Report 2003

Intel's main direct competitor in the microprocessor segment is AMD (Advanced Micro Devices) with 17% market share. [8]

In flash rivalry is intense. The leading suppliers in non-volatile memory business today are Samsung, Intel, Spansion (AMD – Fujitsu joint venture), ST Microelectronics and Toshiba. Samsung and Toshiba-Sandisk dominate the NAND market, whereas the main NOR players are AMD, Intel, ST Microelectronics, and Sharp. The relative market shares of the players are shown in Figure 2.

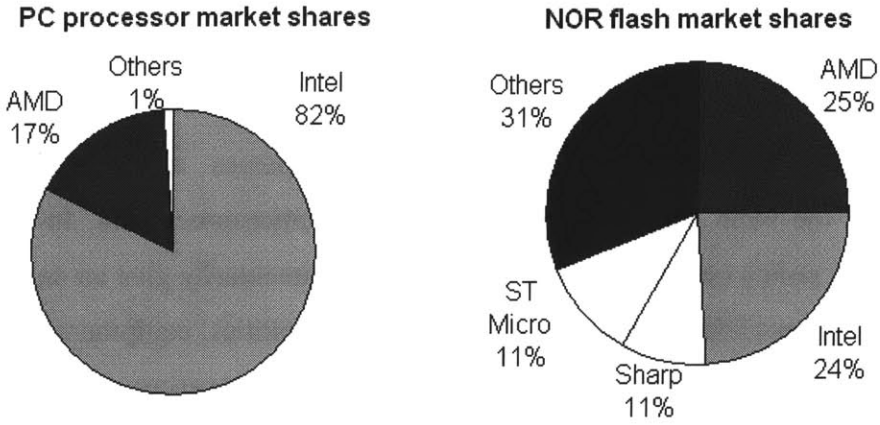

Figure 2. Industry players and market shares in PC processors and NOR flash. [9, 10]

Intel supplies only NOR flash. In addition to the large semiconductor companies the industry has few fabless players in both memory categories. Due to intense price competition and economies of scale, value capturing is hard and profit margins in flash are much lower than in microprocessors. To distribute risk the main flash players excluding Sandisk do not focus solely on flash, but have also other semiconductor products in their portfolios. It has been estimated that AMD gets 41% of its total revenues from flash products, STMicroelectronics 10%, and Samsung and Intel only 6%. [10, 11]

Price competition in flash creates strong pressures for the manufacturers to manage their cost structures effectively. Depreciation costs from capital equipment are the main

[9] http://news.com.com/AMD+gains+market+share,+but+so+does+Intel/2100-1006_3-5557740.html
[10] JPMorgan Asia Pacific Equity Research, Bhavin Shah, Hong Kong 12 Jul 2004. "Tech Hardware Supply Chain"
[11] WRHambrecht + CO, Semiconductor Devices Research Report, June 17, 2004

component contributing to wafer cost. To lower depreciation costs semiconductor manufacturers use the so called waterfall approach for their fabrication facilities: New facilities are used to manufacture high end microprocessor products which give the highest profit margins. As facilities and equipment grow older and it becomes too expensive to update them to accommodate to the demands of latest microprocessor generations, facilities start producing lower technology flash products. This way equipment can be used in-house as long as possible. The down side is that companies to which flash is only a minor part of total revenues will not be able to enjoy the economic and technical advances brought by state-of-the-art flash fabrication facilities.

Suppliers and Buyers

Semiconductor equipment manufacturers, chemical companies and silicon wafer manufacturers are the main suppliers for both flash and microprocessors. In spite of Intel's copy exactly policy (see Chapter 6), which might momentarily give an equipment supplier a benefit by providing an entry point to all Intel's facilities, equipment suppliers face intense competition and continuous need for technical innovation. Chemical and silicon wafer suppliers on the other hand are selling commodity products. Thus, only cutting-edge equipment suppliers are seen as having power in the value chain. Demand for semiconductor products is highly cyclical, and since equipment manufacturers are furthest away from the end consumer, they suffer the highest demand volatilities due to the bullwhip effect.

Traditionally microprocessor buyers have not had power over chip makers. However, the strengthening of AMD as a competitor has offered buyers a choice. As a result large OEM customers have increased their power to bargain over prices and delivery schedules. In the flash industry, buyers on the other hand tend to extract considerable value. Due to the cost pressures for their own products buyers select their flash supplier mainly based on price. Depending on the application, memory density, speed and physical chip size can also be important factors. The situation is especially tough during times when flash supply equals or exceeds demand, because existing flash manufacturers

are not able to establish very long term sales contracts with their customers. Flash inventories are expensive and face the risk of obsolescence.

Substitutes and Emerging Technologies

NOR flash memory products, that Intel also manufactures, have started to face a threat from NAND flash. Traditionally NOR has been the higher revenue generating segment in the flash markets; NOR 2004 revenues were around $10 billion versus NAND around $6 billion. However, NAND is predicted to take over NOR in 2007 with revenues of around $12 billion. The rise in NAND revenues and demand comes from the increased data storage requirements for consumer electronics and handset devices. To compete with NAND, NOR players like Intel and AMD have developed their own high density data storage NOR products (StrataFlash and Mirrorbit) that use so called Multi-level cell (MLC) technology. Typically NOR products can only store one bit of data in one memory cell. MLC technology allows memory cells to store 2 bits of data thus doubling memory density with the same price. [11]

Semiconductor flash memories do not currently have direct substitutes. However, research is being conducted for several alternative technologies that might pose a threat in the longer term, especially once flash CMOS based process technology starts to approach its size limitations: [12]

- Magnetoresistive Random Access Memory (MRAM) is a non-volatile memory technology with high enough speed to become a potential replacement for flash. MRAM uses magnetic charge instead of electrical charge to store information in a magnetically polarized thin film. Other advantages are reduced cost and smaller size. MRAM technology is backed up by for example IBM, Motorola and Infineon.
- Ovonyx Unified Memory (OUM) uses a phase-changing thin film alloy to store information. The manufacturing process follows the traditional CMOS logic flow used in flash but requires fewer process steps. The resulting advantages are

[12] WRHambrecht + CO, Semiconductor Devices Research Report, June 17, 2004

shorter cycle times and smaller die sizes. Intel is one of the investors in this technology.

- FeRAM (Ferroelectric RAM) has a ferroelectric capacitor and a MOS transistor. The structure is faster and has lower power consumption than existing flash technologies. Companies interested in this technology include for example Infineon and Samsung.
- Bistable polymer chains can be used to store information by switching their molecular structure from one stage to the other with electric field. Polymers will maintain their structure even when the electrical field is turned off. Such "smart" structures can be built molecule by molecule enabling application customized designs. In addition the technology has the potential for greater storage capacity with lower price. For example Intel, Samsung and IBM are investing in this technology.

2.3 INTEL AS AN INDUSTRY PLAYER

Intel is the world's largest semiconductor manufacturer with 2003 revenues of $30.1 billion. The company manufacturers microprocessors, flash memories, chipsets and related software for computers, servers, cellular phones, digital cameras and various other consumer electronics products. Intel was established in 1968 and has currently 78,000 employees. Intel has a global presence with 27% of 2003 revenues coming from the Americas, 24% from Europe, 9% from Japan and the remaining 40% from the rest of Asia Pacific. In the past 5 years the main shift in revenue source has been from the Americas to Asia Pacific, when countries like China have started investing heavily in their IT infrastructure. [13]

More than 75% of Intel's manufacturing was conducted in the US: California, Arizona, New Mexico, Massachusetts, Oregon and Colorado. The remaining 25% of manufacturing was done in Ireland and Israel. The majority of Intel's fabrication facilities use 200mm (8-inch) wafers. By the end of 2004 the company had three fabs using the new 300mm (12-inch) wafer process. In addition to the established 130nm line-width

[13] Intel Annual Report 2003

process technology some factories are capable of producing microprocessors with the more advanced 90nm process. Flash memory is manufactured using 130nm and 180nm process technologies. In addition of being high volume, some facilities produce also a high mix of products. Intel tends to manufacture products in more than one facility at a time to reduce the supply capability risk. [13]

At the end of 2004 Intel was divided into two main operational units: The Architecture Group and The Communications Group. The Architecture Group contributed 87% of company's 2003 net revenue. [14] Intel's mission in 2004 was to "do a great job for customers, employees and stockholders by being the preeminent building block supplier to the worldwide Internet economy". To support this mission and new customer requirements, the company announced a significant reorganization of its business units in January 2005, switching from product focus into technology platforms. Simultaneously two new organizations were created to address opportunities in new growth areas. The new business units are [15]:

- The Mobility Group developing platforms for notebook PCs and handheld computing and communications devices
- The Digital Enterprise Group developing computing and communications infrastructure platforms for business customers
- The Digital Home Group developing computing and communications platforms for the emerging digital home, with emphasis on living room entertainment applications and consumer electronics devices
- The Digital Health Group developing products for healthcare research, diagnostics and personal healthcare
- The Channel Products Group focused on developing and selling Intel products to meet the needs of local markets worldwide

Intel's suppliers include a large number of raw material, process equipment and service providers. Intel supplies its products both directly to large Original Equipment

[14] Intel Annual Report 2003
[15] www.intel.com

Manufacturers (OEMs), Original Design Manufacturers (ODMs) and resellers as well as to small and medium size businesses via distributor channels. The channel distributor segment is especially important for Intel in emerging markets. The two Intel internal units are Fab Sort Manufacturing (FSM) and Assembly Test Manufacturing (ATM). FSM includes wafer processing, testing and sort, whereas ATM includes die testing and packaging. The following figure gives a simplified illustration of Intel's value chain: [16]

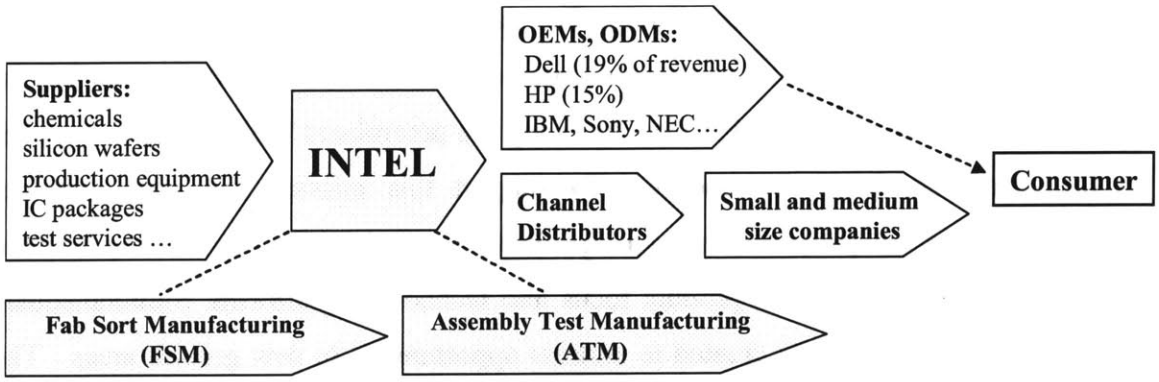

Figure 3. Intel's operational value chain.

Historically Intel has captured a lion's share of the microprocessor market value due to continuous technical innovations, lack of competitors, and smart positioning in the microprocessor value chain through "Intel Inside" supply strategy. As a result of its dominant position Intel has been able to capture high margins and control the speed of development by correctly timing the introduction of new chip versions. After the emergence of AMD Intel started facing competition in value capturing for the first time. In addition to large OEMs creating pricing pressures, the switch to lower profit margin products, like flash memory, has influenced company revenues.

To respond to emerging competition in microprocessors and to accommodate to the new competitive environment in flash products Intel has been investing heavily on value adding services and features, such as product architectures and platforms. A platform is a collection of silicon and software components that together add more value to the end

[16] Citigroup Smith Barney: The Global Supply Chain Handbook v3.1. July 16, 2004. pp. 92

user than they do individually. The recent reorganization of business units is a clear signal of a new strategic direction. Platform strategy in flash is intended to change a commodity product into something more valuable to ease the pressures for low cost manufacturing. Other complementary assets include strong brand, worldwide sales and marketing network, process technology know-how and Intel Capital securing future innovations. Despite the maturing industry and new competitive environments, Intel is strongly positioned with complementary assets that could become a critical advantage over competitors in the future. Accordingly, the long term financial metric forecasts are good, as shown in the following table:

Table 1. Historical and predicted future performance of Intel financial metrics.[17]

Financial metric	Historical performance (1992-2003)	Forecast performance (2004-2013)
Revenue growth	16.6%	7.0%
Gross margins	55.9%	58.8%
Operating margins	27.7%	32.3%
Average EPS	$0.66	$1.66
EPS growth	15.6%	11.3%

2.3.1 Intel Fab11 [18]

Intel's New Mexico site has 465,000 square feet or 9.7 football fields of clean room area, the largest continuous clean room in the world. The site incorporates two fabrication facilities, Fab11 and Fab11X. Fab11X is a leading edge facility producing 90nm line width microprocessors with 300mm wafer technology.

Fab11 is a 200mm wafer facility producing flash memory products. The facility has been operational since 1995. It has 300,000 square feet or 6 football fields of class 1 (1 particle per cubic feet of air) clean room area. The facility produces thousands of wafers per week, and shipped its 1 billionth die in November 2002. In addition to high volume Fab11 has also high mix production of flash, integrated flash and chip set memory

[17] Bernstein Research Call, US Semiconductors, August 9, 2004
[18] www.intel.com and company internal material

products. Product lines include flash memories ranging from single chip capacities from 8Mbits up to 256Mbits. The main product line devices are used in cellular phones.

The Fab11 mission today is to "be Intel's most flexible, highest volume, and lowest cost manufacturer of leading edge products". This mission is well aligned with current market conditions but is very challenging. As was discussed earlier, the characteristics of the flash memory industry today are multiple players, intense price competition and low profit margins. Flexibility is important to be able to quickly respond to changing market demand and customer order requirements. High volumes are needed for dilution of capital costs. Low cost structure again is critical for successful price competition. Leading edge products require continuous technology development but act as a competitive advantage and barriers to entry against competitors. The mission is even more challenging considering the high volume / high product mix manufacturing environment in Fab11. Today Fab11 is striving to meet all three goals simultaneously. However, as will be seen in the following chapters there is an inherent conflict between flexibility, high volume and low cost.

3 FACTORY PHYSICS CONCEPTS

Factory Physics, as described by Hopp and Spearman, provides a set of fundamentals to explain the relationships and behaviors that exist in a manufacturing operation. Factory Physics methodologies give tools to analyze existing operations, design improvement efforts and analyze tradeoffs. The following chapters shortly introduce the basic terminology and key concepts.

3.1 KEY TERMINOLOGY [19]

Nominal process time t_0 is the average time to complete one job without any distractions in the process (a.k.a. theoretical process time, raw process time).

Availability A is defined as the long term fraction of a time when a work station is not down for repair. In other words if tool availability is 80%, then for 80% of the time the tool is available for running production.

Effective process time t_e is the average time to complete one job including distractions such as setups, downtime, etc. Effective process time can be expressed using nominal process time and availability:

$$t_e = \frac{t_0}{A} \qquad \text{Eq. (1)}$$

If availability is 100%, in other words a work station has no unplanned failures and is constantly operational except for planned downtime, effective process time equals natural process time. In practice, however, most work stations experience some amount of random unplanned outages which reduces availability. Effective process time takes these events into account and is therefore longer than the theoretical process time.

Utilization u is defined as the probability a work station is busy. This includes the time when the work station is working on jobs or has parts waiting but is unable to process

[19] W.J. Hopp, M.L. Spearman: Factory Physics, 2nd Edition, McGraw-Hill 2001 New York

them due to setup or machine failure. Utilization can be described using arrival rate r_a and effective tool process time t_e:

$$u = r_a t_e \qquad \text{Eq. (2)}$$

If utilization is more than one, a work station is always busy. Jobs arrive at the station faster than they can be processed, $r_a > t_e$. If utilization is << 1, jobs are processed faster than they arrive and work station experiences idle time due to lack of material.

Utilization term U is expressed as

$$U = \frac{1}{1-u} \qquad \text{Eq. (3)}$$

Cycle time CT is the average time a job spends in the system, starting from the time it enters to when it exits. Cycle time is composed of mean effective process time and queuing time CT_q:

$$CT = t_e + CT_q \qquad \text{Eq. (4)}$$

From a customer point of view only effective process time is value adding. Queuing time on the other hand is non-value adding and can be introduced to the system due to variety of reasons, for example waiting for an operator or a tool to become available. Queuing time depends on effective process time, variability V and utilization term, as shown in the following equation. Variability is discussed more in Chapter 3.2.

$$CT_q = VUt_e \qquad \text{Eq. (5)}$$

Thus combining the previous two equations, cycle time can be expressed as:

$$CT = t_e + VUt_e = t_e(1+VU) \qquad \text{Eq. (6)}$$

Little's law is a fundamental relationship between cycle time, work in process WIP (inventory between start and end of the manufacturing line) and throughput TH (output of a production process):

$$WIP = CT * TH \qquad \text{Eq. (7)}$$

This useful but some times counterintuitive relationship states that WIP levels depend on cycle time and throughput. Thus, if you want to keep throughput constant in your manufacturing system but reduce cycle time, you should reduce inventory levels.

3.2 VARIABILITY [20]

From Equations 3 and 6 above it follows that

$$\frac{CT}{t_e} = V(\frac{1}{1-u}) + 1 \qquad \text{Eq. (8)}$$

The term CT/t_e is the so called cycle time X-factor, since it describes the relationship between the realized cycle time and the goal (effective process time). The higher the X-factor, the more there is queuing time in the process. From the equation it can be seen that when utilization approaches 100%, cycle time factor approaches infinity. In other words high tool utilization is going to cost you in longer cycle times.

In addition to utilization, as was shown in Equation 6, cycle time depends also on variability V. In practice, process times are never quite steady but they vary around some average value due to for example unexpected equipment breakages, delays in operator availabilities or deviations in material quality. In other words, there is variability in the system which creates queuing time. The relationship between utilization, cycle time and variability can be illustrated with the following factory characteristic curves:

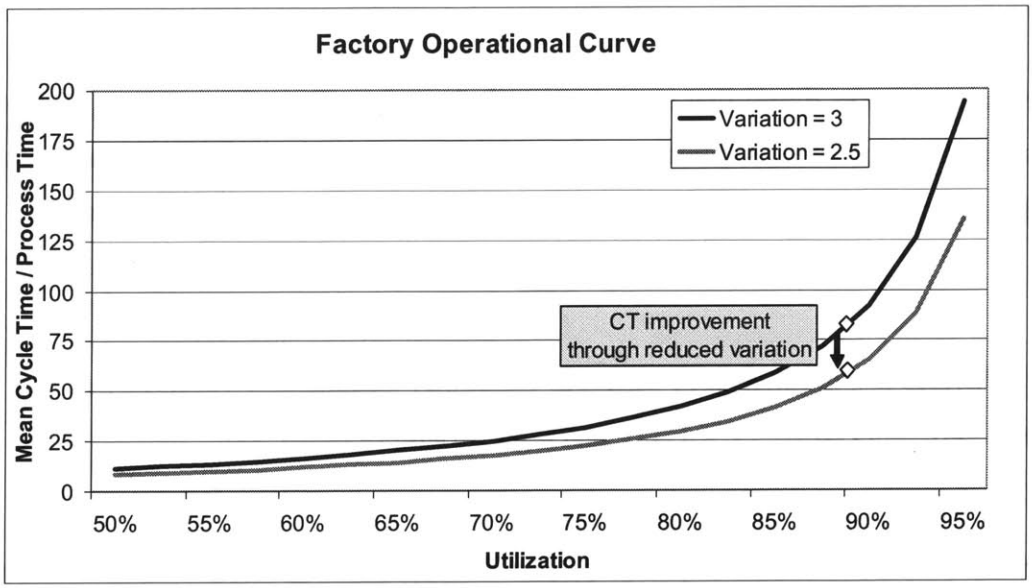

Figure 4. The effect of variability on system cycle time (www.fabtime.com).

[20] W.J. Hopp, M.L. Spearman: Factory Physics, 2nd Edition, McGraw-Hill 2001 New York

Figure 4 illustrates two identical manufacturing environments that experience different degrees of variability. At a fixed utilization level the environment with more variability experiences higher cycle time than the factory environment with less variability. Thus reductions in cycle time while keeping utilization constant can be achieved by reducing variability.

Three elements can be used to buffer manufacturing system against variability: inventory, capacity and cycle time. Increasing variability always degrades the performance of a production system, but the three buffers make it possible to choose how variability will affect manufacturing. Choosing the right buffer strategy depends on the production environment and the overall business strategy of the company. In a manufacturing environment where timely customer delivery is crucial, variability should not be buffered with cycle time but either with increased capacity and/or increased inventory levels (increases in invested capital). Similarly, if additional capacity is extremely expensive, the system can accommodate to variability with increased cycle time and/or increased inventory levels. However, accommodating to variability is always a worse option than trying to reduce it, since buffering will cost in high inventory levels, access capacity, or prolonged cycle times.

Variability represents things that are not normal; deviations from the average. Variability is thus a description of a distribution and can be quantified using the term coefficient of variation, c, which is defined as sample standard deviation divided by sample mean:

$$c_e = \frac{s_e}{x_e} \qquad \text{Eq. (9)}$$

Factory Physics has defined three different variability classes: low variability ($c < 0.75$); moderate variability ($0.75 \leq c < 1.33$); and high variability ($c \geq 1.33$).

3.2.1 Variability in Manufacturing Flow

Variability of a work station depends on two factors: arriving material flow and actual processing time at the work station. Variations in the arriving material flow are described using a distribution of times-between-arrivals. Variations in the work station's effective

process time are described using a distribution of process times. Thus variation can be expressed as:

$$V = \frac{(c_a)^2}{2} + \frac{(c_e)^2}{2} \qquad \text{Eq. (10)}$$

where c_a = coefficient of variation of time-between-arrivals; and c_e = coefficient of variation of effective process time.

Variability tends to propagate through the manufacturing flow. Departures from a work station are arrivals to the next work station downstream; thus ignoring stockers and transportation, variations in departures at tool n become arrival variation for tool n+1. As a result, variability early in the process flow can be seen being worse than variability later in the process.

Variability of departures (c_d) depends on that of the arrival flow (c_a) and effective process time (c_e). The relationship depends also on tool utilization and can be described with the following equation:

$$c_d^2 = u^2 c_e^2 + (1 - u^2) c_a^2 \qquad \text{Eq. (11)}$$

From this equation it can be seen that as utilization increases, effective process time variation becomes more significant for departure variation than variation in arrivals. Therefore, for a low utilization tool, regardless of the tool's own effective process time variability, departure variability from that tool tends to repeat the variability of the arrivals to that tool. In other words, low utilization tools do not change manufacturing flow variability, but simply just pass it on. On the other hand, high utilization tools with high effective process time variability have high departure variability regardless of the arrival variability. High utilization tools with low effective process time variability have low departure variability regardless of the arrival variability. In other words, flow variability through a high utilization tool reflects the tool's own effective process time variability.

It is important to notice that variability discussed in this study occurs in the manufacturing environment due to poor operations management. It increases non-value

adding queuing time and is therefore highly undesirable. This type of variability is different than variability that has been introduced into the system as part of a strategic choice by the company; for example to increase sales through more product variations or customization.

As was shown earlier, work station variability depends on variations in arrivals and process times. The following chapters discuss the main sources of variability in these two areas.

3.2.2 Sources of Arrival Variability

Typical sources of arrival variability are

- upstream variability
- batching
- starts policy

Upstream variability is created for example during tool outages in upstream operations. This can be a significant source of arrival variability for a work station if the upstream station happens to have a poor reliability.

Batching is another source of arrival variability in situations where downstream operations use different batch sizes than upstream operations. This is very typical in a semiconductor facility, since manufacturing equipment comes from multiple different vendors and has not been designed to handle equal batch sizes (same number of lots or wafers per operation). In such cases, the coefficient of variation for arrivals is not zero even when batches are delivered in regular intervals. This is due to the different interarrival times that batch parts see: for the first part in a batch interarrival time is the time since the previous batch was delivered, whereas for all the remaining parts in the batch it is zero (since they all arrived at the same time with the first part).

Starts policy is a third potential source of arrival variability. It refers to the schedule how new material, in the case of semiconductor facility empty wafers, is released to the line. Often start schedules are controlled by demand forecasts. In volatile industries demand volumes can vary dramatically causing variations in material release schedules. From

variability point of view, steady material release reflecting the capacity of the first work station would be the best policy.

3.2.3 Sources of Effective Process Time Variability

Typical sources of effective process time variability are

- variations in raw process time
- unscheduled outages, repair times
- operator availability and material handling
- scheduled set ups, preventive maintenance
- rework, priority lots

Variability in raw processing time, also called natural variation, is typical in manual operations. In highly automated environments raw process time variations tend to be low, but can be caused, for example, by differences in machine or material qualities.

Unscheduled outages can be caused for example by unexpected machine failures. The effect of unscheduled failures as well as scheduled maintenance work on variability is significant but often underestimated and poorly understood. According to Factory Physics long and infrequent repairs induce more variability on effective process time than short and frequent ones, since higher levels of work in process buffers are needed to prevent downstream operations from starving during long repairs. Higher WIP in line, as stated by Little's law, results in higher cycle times if throughput (bottle neck rate) remains constant.

Operator availability might be a problem in environments that have experienced head count cuts or do not have optimized labor resources. If operators need to run multiple tools simultaneously, they might not be able to program equipment or release finished material forward on time, thus increasing queuing time. Poorly managed break schedules might also cause unnecessary pauses in production.

In a high product mix environment setups could be frequently required due to product customization. Operator instructions or product-tool allocations might not be optimal

resulting in too many setups and inefficient use of capacity. In addition, rework and priority lots which are typical in a semiconductor manufacturing environment tend to distract the manufacturing flow and cause variability.

3.3 CHALLENGES FOR OPERATIONS MANAGEMENT IN A SEMICONDUCTOR ENVIRONMENT

As became apparent from chapter 2, the main competitive drivers in the flash industry are flexibility and price. Flexibility and ability to respond rapidly to changing demand drives short cycle times, whereas low price drives high tool utilizations. As was seen from Figure 4, it is impossible to achieve 100% utilization while having low cycle times. Finding the right balance between these two drivers seems to be a major challenge in many semiconductor fabrication facilities.

Cost is often the main driver also for manufacturing operations. Extremely high capital investments and the way wafer costs are calculated leads to high tool utilizations. High utilization together with variability causes long cycle times with a high proportion of queuing time. Long cycle times can result in several problems. According to Little's law increased cycle times cause the WIP levels to increase, assuming that bottle neck rate (throughput) stays constant. This excess inventory ties up capital and causes losses due to decreased yield and increased risk of material becoming obsolete. If cycle times become longer than customer lead times, material has to be released into the factory based on demand forecast instead of secured orders. These forecasts could be very inaccurate in the rapidly changing environments resulting in poor order fulfillment or excess inventory.

Semiconductor manufacturing environments suffer also from other characters that make efficient operations management challenging. The manufacturing flow is very complicated due to the high number of re-entry and rework operations as well as priority lots. In this complex flow there is typically no one bottleneck tool or operation, but the bottleneck tends to shift depending on tool excursions. Over time Fab11 has gone through capacity changes, tool purchases and technology improvements which further complicate bottle neck identification. In addition, according to company terms, Fab11 has a high mix - high volume manufacturing environment. Volumes for different product series also vary

greatly. Due to the waterfall principle (see Chapter 2) the facility has a collection of aging machines that have different technical capabilities and that suffer from unscheduled outages. Data collection systems are also not as sophisticated as in newer facilities and fail to capture some useful operational details.

Fab11 staff hopes that Factory Physics principles will help them to get a better understanding of their manufacturing system and to identify improvement opportunities in order to survive in the increasingly competitive environment. However, the major challenge will not be to understand Factory Physics concepts, but to change current metrics and ways of thinking, and to overcome change resistance in the organization. The following chapters describe the methods and results from analyzing variability in the manufacturing flow with Factory Physics concepts, as well as the challenges of implementing change for the current performance metrics system and organization.

4 VARIABILITY ANALYSIS

The concepts introduced in the previous chapter were used to develop a process for analyzing manufacturing flow variability in Fab11. The three main steps in the process were:

1. Find an area in the factory that is potentially suffering from high variability using Factory Physics analysis methods.
2. Analyze flow variability factors in the selected area (arrivals and process time) to find the major sources of variation.
3. Identify improvement opportunities.

The following chapters describe the detailed process steps, analysis methods, results and recommendations.

4.1 STEP 1 – SELECTING FOCUS AREA

The overall factory analysis was done using a method developed by Factory Physics. From Equation 6 it follows that

$$VU = \frac{CT}{t_e} - 1 \qquad \text{Eq. (12)}$$

Since true cycle times and effective process times (goals) were recorded in the company data base, VU values could easily be calculated. In this case, effective process time was a fixed value, whereas cycle times were true average lot process durations varying from day-to-day. Using a data collection period of 13 weeks and grouping individual operations under tool areas, the graph shown in Figure 5 was produced.

Data points represent tool areas and each data point includes several tools. For example WE1 is comprised of type X tools in the wet etch area; WE2 represents type Y tools in the same area. VU value is plotted on the y-axis and normalized true cycle times are on the x-axis. Since VU ~ (true cycle time / goal cycle time), the term can be seen as a cycle time factor: the higher the value is, the worse the true cycle time performance is when compared to goal. In other words, the more queuing there is in total cycle time. Tool

areas located on the right hand side of the graph have the highest total cycle time. Thus in order to reduce the overall product cycle time, focus should be placed on fixing first the tool areas in the upper right hand corner of the graph, with high VU factor and high cycle times. Reducing variability in these tool areas would result in biggest time savings in the overall product cycle time.

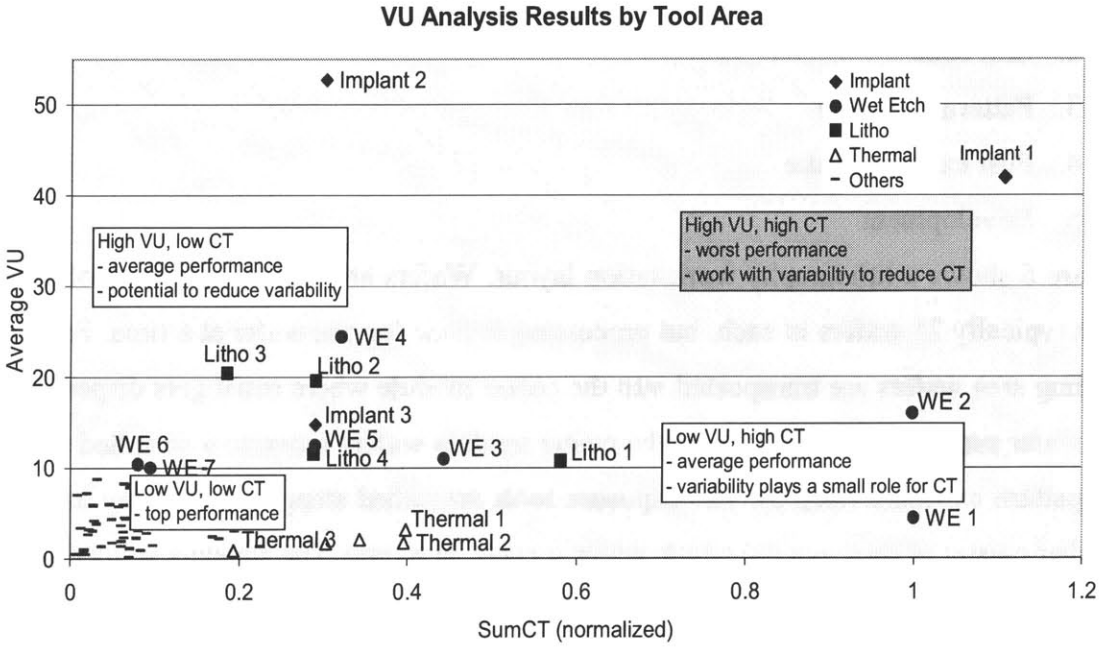

Figure 5. Factory Physics VU analysis results for the main tool areas in the factory.

Based on the analysis, the worst tool areas were found from implant and wet etch operations. However, since task forces to fix performance were already in place in these areas, the third worst area, lithography (litho), was chosen as the focus for this work. Lithography area was seen as an important target due to its importance for the overall manufacturing flow (high number of re-entry) and due to the upcoming capacity changes: changes occurring in the first half of 2005 will cause a 35% increase in the number of operations per tool, assuming current product mix. Litho area supervisors were familiar with lean manufacturing concepts and together with the operators were known to be supportive for the project. Additionally, the lithography area is often seen as the desired bottleneck for the factory due to most expensive equipment. To make the most of the investments tool utilizations are kept high which often creates queuing time for the process.

4.1.1 Lithography Area Introduction

Silicon dies are comprised of several layers of different patterning. When layers are built on top of each other, the result is a three dimensional structure of connected wiring. Lithography is the transfer of a layer pattern from a reticle (mask) onto silicon wafer coated with photosensitive resist. The simplified process flow is:

1. Resist coat
2. Soft bake
3. Pattern exposure
4. Post exposure bake
5. Development

Figure 6 shows a lithography workstation layout. Wafers are loaded into the tool in lots with typically 25 wafers in each, but processing is done for one wafer at a time. From the loading area wafers are transported into the coater module where resist gets dispensed on the wafer surface (steps 1-2). From the coater module wafers move to a so called stepper for pattern exposure (step 3). The exposure tools are called steppers since they expose a smaller cluster of dies, not the whole wafer surface, at a time. The steppers studied in this work use the "I-line" ultra violet wavelength of 365nm. After exposure wafers are moved to the developer unit where the patterned resist is developed (steps 4-5).

In most cases the tool operator has to load and unload wafer lots manually but wafer transfer between the modules is done with automated serial arms. An operator controls the work station from a close by computer terminal.

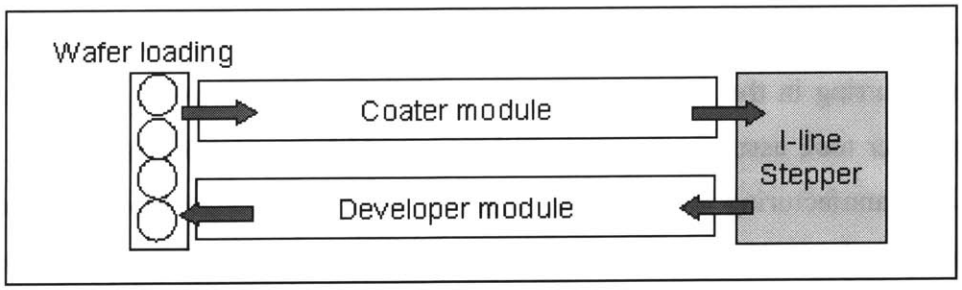

Figure 6. A simplified lithography workstation layout.

4.2 STEP 2 – ANALYZING VARIABILITY COMPONENTS

As was discussed in chapter 3.2, manufacturing flow variability is comprised of three components: variation in the arriving material and variation in process times resulting in variation in departing material. In the litho area these components could be analyzed either by tool or by operation (wafer layer), since the area is running several re-entry operations. Both approaches were used to provide feedback on area performance, to find the major variability contributors, and to compare individual tools. Finally, tool availabilities over time were analyzed.

For both by tool and by operation approaches the analysis method was the same: first the time-between-arrivals, time-between-departures and process time data was collected. After that data points in each three areas were combined into distributions and coefficient-of-variations were calculated. Data was plotted on graphs to analyze the results. In the case of operations all needed data was readily available in the plant data bases, whereas in the case of tools parts of the data had to be statistically pooled.

4.2.1 Assumptions

Data Collection Period

The data collection period should provide easy data handling but also serve the purpose of the analysis. Too short a time period, say few weeks, could give a false representation of the true situation due to a tool outage or a sick operator. On the other hand, the longer the time period the more the variations in the data get evened out and hidden in averages. Also, the amount of data increases rapidly with time. For this study a data collection period of thirteen weeks was chosen. For some variables this had to be reduced to ten weeks due to limitations in the data recording systems. All data was collected during August and October of 2004.

Product Mix

During the observed time period the factory was manufacturing several product types using four main process technologies. One technology was clearly dominating with 94% volume share. The next largest processes had 3% and 2% volume shares. Only these three technologies were included in variability calculations.

Unit Quantity

Wafers move in lots of 25 through production. Data were collected on a per lot basis instead of a per wafer basis. Occasionally, a lot could have fewer than 25 wafers due to testing, failed products or engineering requests. A sample study revealed, however, that 97% of the lots moving through lithography area had 24 or more wafers. Thus the chosen per lot approach was believed to yield adequate accuracy.

Arrival Variation

The time-between-arrivals data was obtained using real time-between-departures data from the previous, i.e. feeding, operation. This assumption ignores lot transportation between work stations. Lots are moved with an automated material handling system that picks up finished lots from work stations and distributes them to central stockers. Operators then call lots from central stockers to their work stations. As a result time-between-departures do not translate into time-between-arrivals with 100% accuracy. However, including the effects from transportation would have been very difficult. Lots are picked up and transported fairly rapidly, and thus believed not to cause too many inaccuracies for the lot interarrival times.

Tool Dedications

Lithography tools are capable of running multiple different operations (layers), but setups might be required when switching between operations. Required setup times depend for example on what the operations in question are and how much time has passed since the tool ran the operation last time. Setup might be a simple and short mask change, or require a longer "send-ahead" lot to test process quality. Area supervisors are responsible for deciding what operations to run with which tools and in what order. To ease the operator decision making supervisors have created a tool-operation dedication matrix that tells which operations (layers) to run with which tools. Main considerations when creating these dedications are feeder operations' stocker locations, product volumes, technical tool capabilities and total available tool capacity.

The dedication matrix was used in the analysis to determine what operations the tools were running. It should be noticed that the reality might differ somewhat from the matrix depending on daily work in process levels and maintenance schedules. However, a sample test showed that tools obeyed their primary dedications whenever WIP was available. Also, the matrix changes over time to respond to changes in product mix and capacity, for example due to a long repair or new qualification requirements. Thus the dedications used in this analysis would not necessarily be valid for longer time periods.

Pooling Arrival Distributions

The tool-operation dedication matrix (see above) was used to determine what operations were arriving to each tool. In most cases tools were dedicated to run more than one operation, in which case the following statistical pooling equations were used to combine several arrival distributions into one:

$$Stdev = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2 + n_3 s_3^2 + ... + n_k s_k^2}{n_1 + n_2 + n_3 + ... + n_k}} \qquad \text{Eq. (13)}$$

$$Average = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3 + ... + n_k \bar{x}_k}{n_1 + n_2 + n_3 + ... + n_k} \qquad \text{Eq. (14)}$$

where

n = sample size (# of lots)

s = standard deviation

x = average

Interarrival times can be assumed to be exponential and thus memoryless, since it is hard to predict when the next arrival will occur. Thus these equations are a suitable method for pooling. However, the equation for pooled standard deviation assumes that arrival operations are independent. This is not the case in reality, since for example a power outage could influence arrivals from all operations. Yet finding out these correlations would be extremely difficult.

Process Time Variation

Factory Physics lists equations for calculating process time variations in cases where tools have random failures and require setups and rework. However, using these equations in practice turned out to be very complicated, mainly due to lack of data and slightly different definitions for availability and utilization between Intel and Factory Physics. Thus only raw process times ("move out – move in") were chosen to represent process time data. This data does not take into account variability caused by disturbances in availability (such as unscheduled outages). Therefore availability variability was studied separately.

Departure Variation

Time-between-departures data was obtained from real movements recorded in the company data base, just like time-between-arrivals. Equation 11 could have been used to calculate departure variations. However, the equation gave different departure variation values than real data. The inaccuracies were believed to be caused by utilization and process time variation terms: Intel's definition for tool utilization is slightly different from Factory Physics definition. Also, as explained above, process time variations in this work were estimated slightly differently from Factory Physics approach. Thus real data for departures was seen as the most reliable method.

4.2.2 Flow Variability Results by Operation

Figure 7 summarizes lithography flow variability results by operation for the main process technology. Only part of the operations was included for confidentiality. The figure lists ten litho operations (layers) on the x-axis. It should be noted that these are not consecutive wafer operations, but wafers visit other tool areas before returning into litho. The columns represent coefficient-of-variations for arriving material, departing material and process times. The solid line represents the change in flow variability, in other words the difference between arrival and departure variations. The straight horizontal line shows the limit for high variability; 1.33 as defined by Factory Physics.

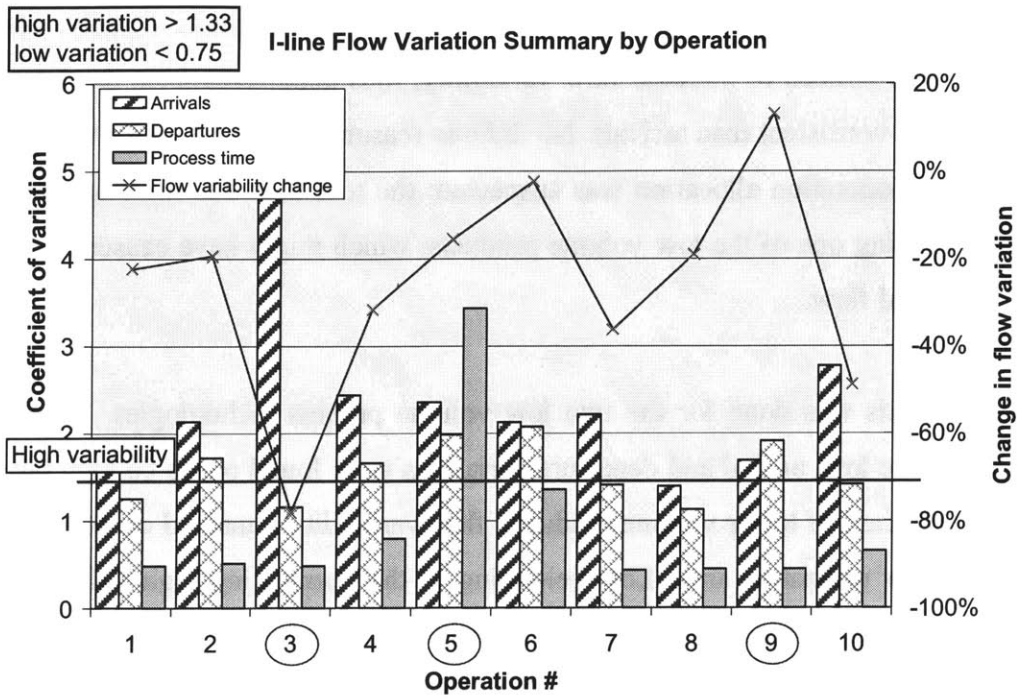

Figure 7. Summary of lithography flow variability analysis for selected operations.

The first observation from the graph was that arrival variations were high for all operations. Departure variations were lower than arrival variations for all other operations except for operation number 9. Thus the lithography area in general seemed to be reducing the overall manufacturing flow variability. In most cases process time variations were low, which was the expected performance in a highly automated environment. According to chapter 3.2.1 the combination of flow variability reduction and low process time variations indicates high tool utilizations. Hence the lithography area could be a local bottleneck in the process. Being a bottleneck might be desirable due to the extremely high capital investments However, highly utilized tools are not always available when material arrives and thus tends to build queues. Thus the proportion of queuing time in the total cycle time is increased in this area.

Three operations were circled due to unique performances: operation number 3 had very high arrival variability. This indicated potential problems with the feeder operation or a poor match in batch sizes in the two adjacent operations. Operation number 5 had very high process time variability. The reason was likely a technical conversion work that was

interfering with production at the two tools running this operation. Finally, operation number 9 was noticed to increase flow variability, thus departures from the lithography area were less consistent than arrivals. No definite reason was identified for this behavior, but poor tool-operation allocation was suspected: the tool that was running operation 9 was also running one of the low volume products, which might have caused distractions for the material flow.

Similar analysis was done for the two low volume process technologies. Process time variations were low, arrival and departure variations were found out to be high but equal. Hence, in the case of lower volume products, flow variability remained constant and litho acted as a low utilization area. Lots belonging to the dominating high volume process technology had on average lower arrival variability than the lower volume products. This made sense since high volume lots were often prioritized over the other two technologies. Lower volume technologies tended to have higher standard deviations for time-between-arrivals due to more inconsistent material flow through the factory.

4.2.3 Flow Variability Results by Tool

Figure 8 summarizes lithography flow variability results for individual tools including all the three process technologies. The graph is similar to the one shown in the previous chapter for operations.

Results were similar to the previous graph: overall the area seemed to be reducing manufacturing flow variability, and in general process time variations were small. Some problem tools could, however, be identified: Tool D was increasing overall flow variability since departing material flow had higher variations than arriving material flow. One of the reasons could have been operator behavior: the tool was running mainly one of the low volume products which tend to receive less attention from operators than the high volume main product. Tool F on the other hand was experiencing very high arrival variations. Though it might appear that this tool must have been running material from the worst feeder operation, surprisingly this was not the case. High arrival variability was caused by running a combination of high volume (low time-between-arrival average) and low volume (high standard deviation) products. Due to the way operations were pooled

into one distribution (Equations 13 and 14), the high standard deviations typical for the low volume arrivals caused the overall coefficient of variation to be very high. The three worst tools in terms of arrival variability were all running similar combinations. Tools running only low volume technology, even multiple layers, had lower arrival variations. This was an important finding and showed that high arrival variations could be avoided by better allocating operations from different product technologies to tools.

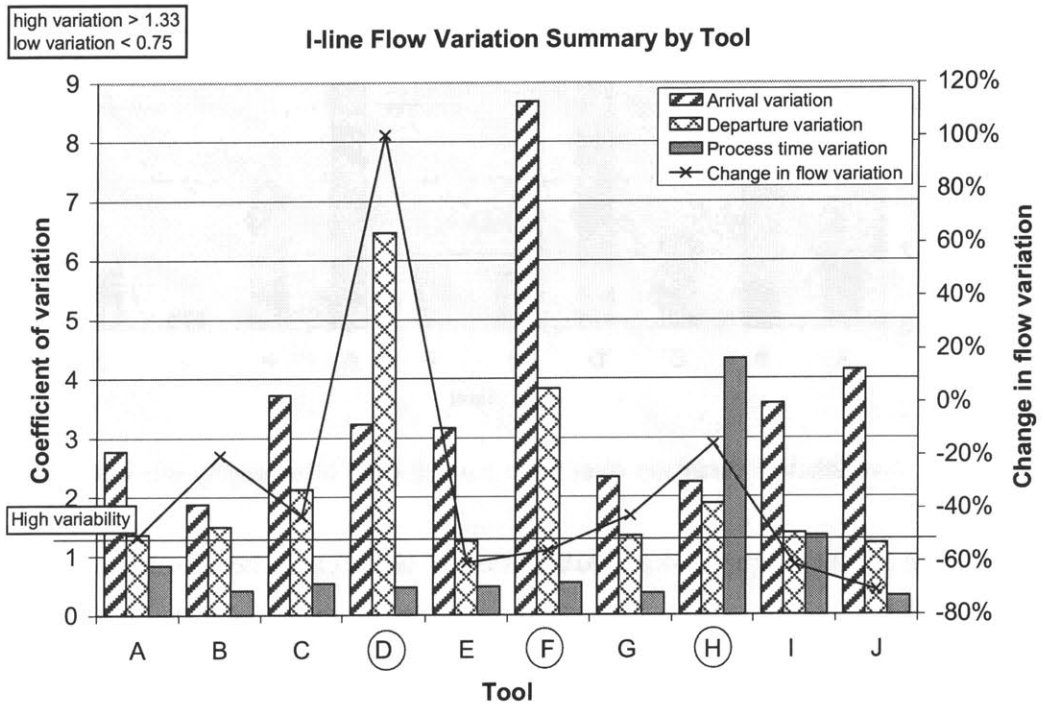

Figure 8. Summary of lithography flow variability analysis for selected tools.

Finally, tool H had very high process time variability. Average process times looked normal, so the variation was unlikely caused by continuous lack of operator resources. Standard deviation was very high, so there might have been few lots that experienced problematic qualifications or testing and were thus "trapped" at the tool for longer periods. Technical conversion at the tool might have caused the same issue, if lots were already logged in to the tool when conversion was started.

Due to company specific data recording systems a modified approach had to be used to analyze process time variations. The approach did not account for tool availabilities as

the Factory Physics method would have done. Therefore tool availability variations over time were studied separately. Figure 9 summarizes availability variations for the selected tools.

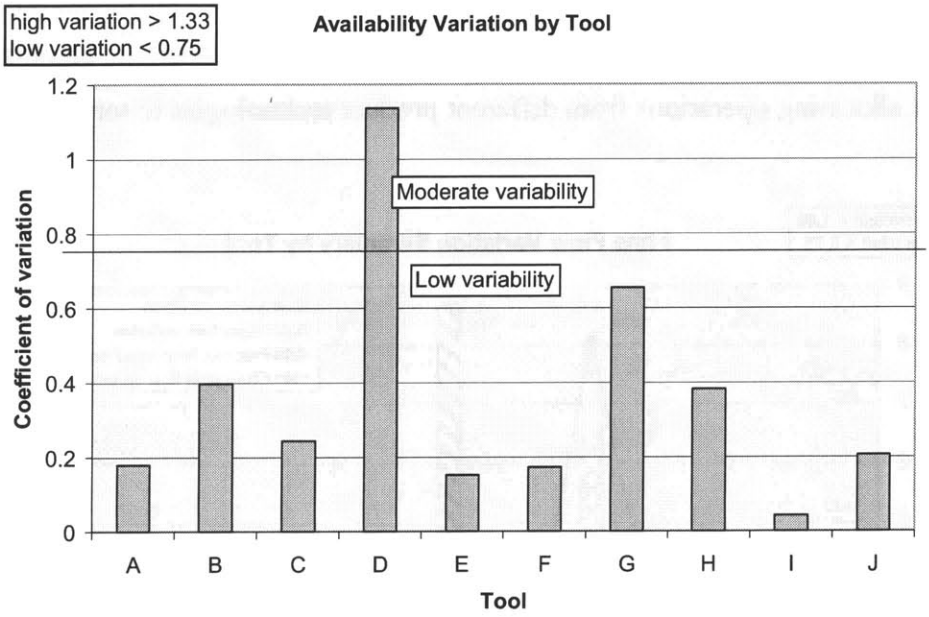

Figure 9. Availability variation over time for selected lithography work stations.

As can be seen from the graph, tool availabilities do not seem to be a problem in this area since most tools experience low availability variations. Only tool D shows moderate variability. This tool was out of production 6 out of 13 weeks during the data collection period. This severe outage caused availability variability and weakened the tool performance. As was seen from Figure 8, tool D suffered from high departure variations which could be partly explained by this severe availability problem.

4.3 STEP 3 – IDENTIFYING IMPROVEMENT OPPORTUNITIES

Chapter 3.2 listed some of the main causes of arrival and process time variations. Once variability analysis was done, these lists were used to help to find the major improvement opportunities. Feedback on major problems contributing to variability was also requested from lithography area supervisors, since they tend to have the best understanding of their area's day-to-day performance. Table 2 summarizes the main problems identified by supervisors. Their feedback was well in line with previous analysis results.

The first two items in the table, arrival flow control and tool-operation dedications, affect the time-between-arrivals variation. Human factors, non-routine work and tool health on the other hand are related to process time variations.

Table 2. Main topics contributing to variability in the lithography area according to area supervisors.

Subject	Concerns
Arrival flow control	Variable inventory profile; no control over upstream processes; poor visibility.
Tool-operation dedications	High product mix and tools' technical characteristics cause setups; complicated qualification rules; lack of optimal tool-operation allocation scheme.
Human factors	Area size and layout combined with tight headcount resources results in wait-to-tech times; lack of automated run decision making; staffing inconsistencies; lack of standardization.
Non-routine work	Engineering requests; hot lots; technical upgrades.
Tool health (availability variability)	Aging tools often under repair; automation issues; tool quality tests not up to date with used materials; preventive maintenance inconsistencies.

The main finding from the feedback and from the flow variability analysis was that arrivals seemed to be the main source of variability in the lithography area for the main process technology. In general, process time variations were low and the area was reducing flow variability, which according to Equation 11 indicates high tool utilizations. Tool availabilities were found to be good. Thus in order to improve area performance the emphasis should be placed first on fixing arrival variability. However, since highly utilized tools tend to accumulate queues, in which case arrival variability becomes insignificant, also items affecting process time variations and utilization were seen as important improvement targets. Moreover, the slightly customized way of estimating process time variations in this work did not fully bring out the variability in the tool operations.

4.3.1 Improving Arrival Variations

4.3.1.1 Feeder operations

In order to understand the major source of variability better, feeder operations were investigated in more details to identify those upstream operations that caused the biggest problems for the lithography area. The following figure shows arrival variations by lithography feeder operation for the main process technology (operation numbers do not necessarily match with the ones in Figure 7). Feeder operations looked similar also for the two other process technologies.

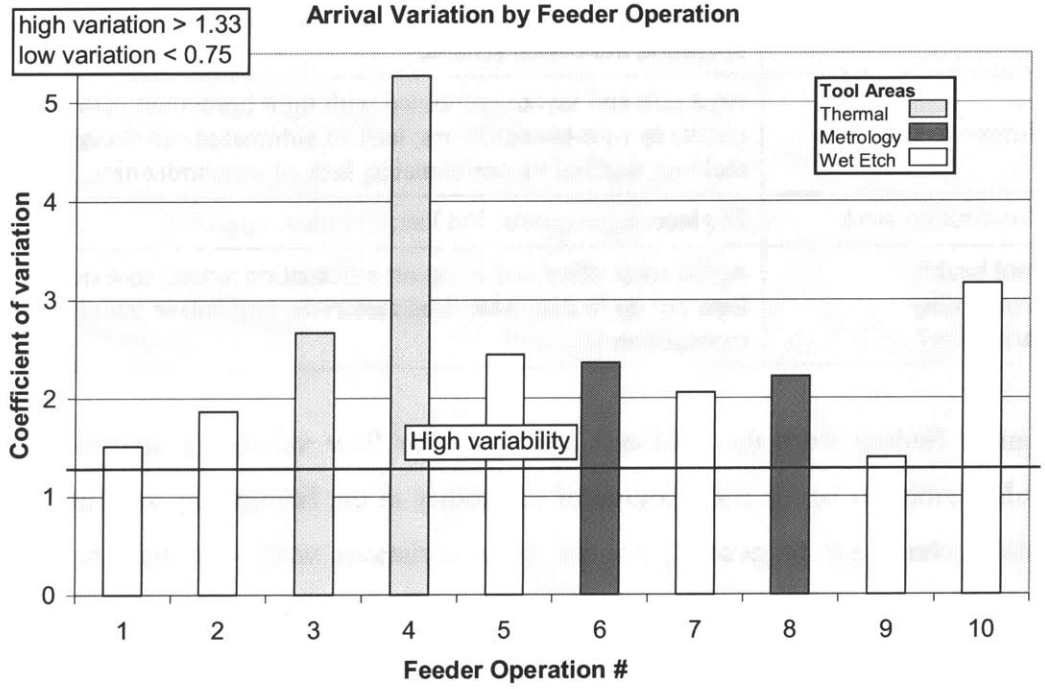

Figure 10. Variability of arriving material flows into the lithography area classified by feeder tool type.

As can be seen from the figure, feeder operations number 3, 4 and 10 caused the highest arrival variations to the lithography area. For operation 3 the reason was believed to be in the so called staging operation. Staging means that the thermal operation in question and its feeder operation were linked: for process technical reasons material has to proceed to the next operation within a certain time period. Thus processing of the material won't start unless a tool is available in the next operation. These limitations reduce flow flexibility and cause variations. Batching and operator availability were believed to be the

main causes for high arrival variation from feeder operations number 4 and 10. If a feeder operation uses different batch size than lithography tools, arrival variability increases. This is especially true for feeder operations with very long processing times. Some of the feeders were wafer handling operations and could be classified as sub-operations. Operators tend to focus on running the main operations and leave sub-operations with less attention, which is why these areas suffer from operator availability. Poor tool health causing availability problems could also be contributing to high arrival variations for all feeder operations.

Changing the staging operation might be very difficult due to technical quality requirements, but it could be possible to modify batch sizes for the thermal feeders. However, operator awareness on the effects of batching and tool health on downstream variability should be improved first.

4.3.1.2 Starts Policy and Tool-Operation Dedications

Other factors affecting arrival variations in addition to upstream variability are starts policy and tool-operation dedications. Starts policy was not seen as the major problem area: Fab11 starts were fairly consistent during the observation period and improvements in the schedule would likely be balanced out by operations upstream from lithography. In addition, certain established upper level management metrics would create serious resistance to any changes in starts policy (see Chapter 6 for details). Even though starts release schedule is fairly consistent, overall starts policy is still far from perfect. Start volumes are based on demand estimates, not the situation in the manufacturing line. Thus material is pushed in to the flow even though factory is congested. This excess WIP influences variability and increases cycle times.

Tool-operation dedications were seen as a major way to control arrival variations. Since tools tend to run multiple operations that each have different arrival patterns, the final arrival variability for a tool depends on how different arrival streams are combined (statistical pooling equations). A simple trial was made to illustrate the effect: A group of four identical tools that all have the same technical capabilities were chosen. The tools were responsible for running 3 layers from the main product technology and 6 layers

from one of the smaller volume technologies. Simply by reorganizing layer dedications between tools without changing capacities, the average arrival variation by tool could be reduced by more than 30%, as can be seen from the following figure:

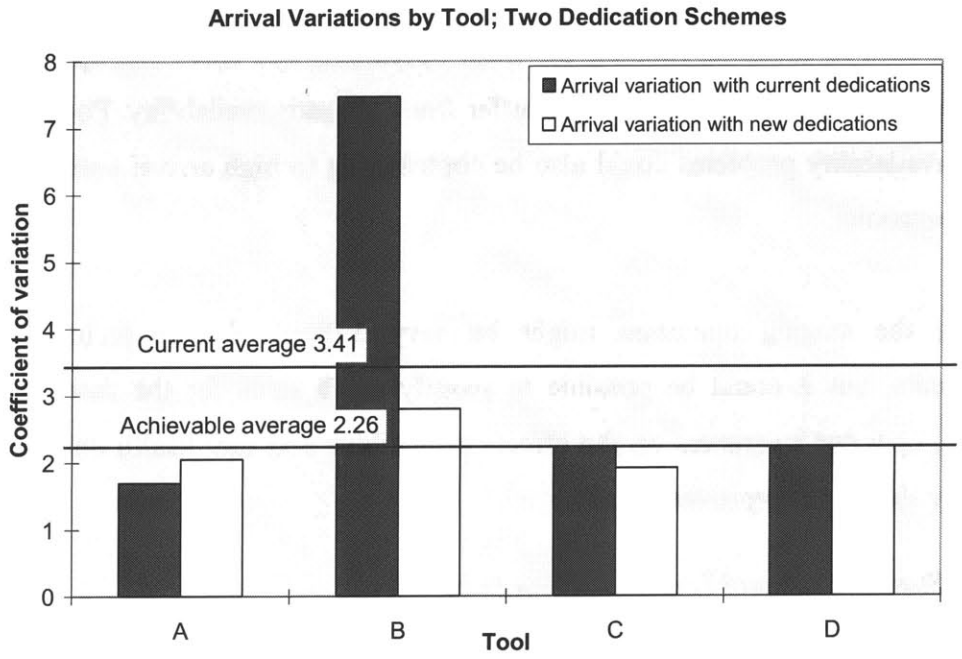

Figure 11. An example illustrating the effect of tool-operation dedications on arrival variations.

Due to the high number of tools, operations, changing capacity requirements, tools' different technical capabilities and restrictions tools' physical locations, optimizing the whole dedication matrix would require significant effort to build appropriate software tools. Though such an application was seen very useful for the organization, the task was not further pursued within this project due to approaching capacity changes in the area. A simple concluding rule is that to reach low arrival variability the best option is not to mix different process technologies, and to run as few layers as possible.

In the case of low volume technologies lithography tools were seen as having low utilizations which indicates that there were enough tools dedicated for the lower volume operations. Even though dedication schemes for lower volume technologies seemed optimal, the setups were likely eating away capacity from the main process technology.

4.3.2 Process Time

Though process time variations were found out to be low, opportunities to improve performance were still seen. Based on preliminary floor observations and feedback from area supervisors, the main focus was placed on examining operator behavior and preventive maintenance procedures. Operator behavior affects mainly queuing time in the area, whereas preventive maintenance operations affect tool availability. Improving availability is important, since in Intel's capacity planning systems tool availability is directly linked with tool utilization: if tool availability goes down, also tool utilization goes down reducing tool capacity. Thus improvements in tool availability would directly result in improvements in tool capacity, at least in theory.

4.3.2.1 Operator Behavior

Lithography tool operator's daily tasks include:

1. Building WIP queue by pulling lots from stockers. The operator has to select which process technology and which product lot to pull. The operator makes the selection based on set run priorities ("hot" priority lots, rework lots, engineering lots), current tool setup (recipe that the tool is currently running, reticles that are available for use), existing WIP queue and expected feeds, as well as supervisor instructions. To find out all the required information, the operator needs to access multiple different web based control tools simultaneously and combine information from multiple different sources. Operators try to avoid introducing long setup times when making their selection.
2. Picking up lots from the stocker and carrying them to tools.
3. Introducing a lot to a tool: "move-in" the lot and place it in the wafer loading port.
4. Exit a lot: "move-out" the lot after the tool has finished processing and carry it from tool loading port to the stocker.

Step 1 was seen as the biggest source of variability. Given instructions are only directional and leave room for the operators to use their own judgment. More experienced operators tend to run the tool smarter, since they are more familiar with combining data from multiple IT control windows and are able to evaluate the effect of their choices on the overall manufacturing flow.

In addition to decision making, operator availability causes variability in the effective process times. Steps 2 to 4 in the operator task list require operator movement. Since operators take care of multiple tools and control terminals simultaneously, lots can end up waiting for the operator at the stocker and at the tool. A study in one of the lithography areas typically controlled by one operator revealed, that the average wait-to-tech time after a lot had finished processing was 7min. Maximum wait time was 1hour and 50min. Thus, on average, wait time at the end of the process alone adds more than 10% to the total lithography process time per layer. Tools have no signal or alarm when a lot finishes processing, and operators often fail to walk around enough to spot the finished lots.

Tools could also be waiting for operators. The batch size for lithography stepper is one wafer, but several wafers at a time can be under processing in the coater and developer modules. Tools can start to unload the next lot (typically 25 wafers per lot) while wafers from the previous lot are still under processing. These operating principles create a so called first wafer effect which means that the effective process time for the first wafer is longer than for the wafers processed in the steady state. The minimum number of lots needed in the wafer loading port to maintain steady state flow is called the critical level. Loading less than the critical level is not desirable, since that would introduce the "first wafer" prolonged process time. Loading more than the critical level is also not desirable, since excess lots could be processed at other tools instead of being tied in the tool queue accumulating queuing time. Thus, in an ideal situation operators would have all their tools loaded exactly up to the critical level. However, in real life operators don't necessarily have time to do this and unnecessary breaks are introduced discontinuing the steady state flow.

Finally, according to operators some of the older tools can suddenly "die" without giving any alarm or notice. The reasons were unclear, but apparently the tools' serial arms which transfer wafers between units freeze. This can only be noticed if an operator is standing next to the tool and checking if the serial arm is moving. Even though problem tools are known, operators do not always remember or have time to check the arm movement.

4.3.2.2 Preventive Maintenance

Preventive maintenance, PM, is done for all lithography tools by specialized equipment technicians. Equipment engineers prepare the schedule based on supplier recommendations and historical tool performance. Currently the schedule includes four types of maintenance procedures done in different intervals. Preventive maintenance operations are tool down-time and thus reduce tool availability; an ideal PM would be as short as possible while still maintaining excellent tool health. The following figure shows average durations for different PM types over a 13 week period by shift:

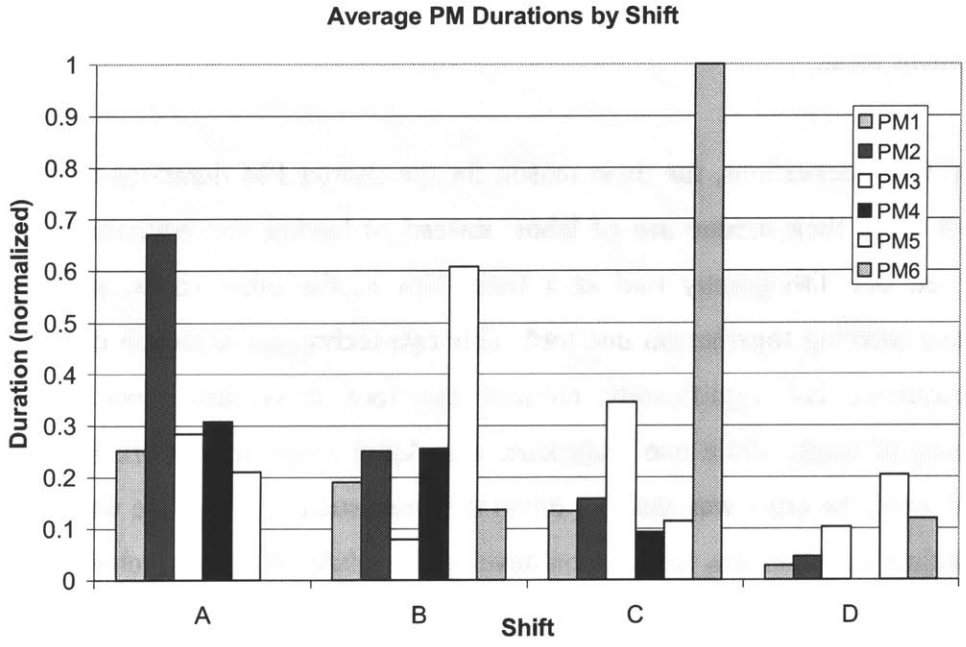

Figure 12. Preventive maintenance durations by shift.

As can be seen from the figure, PM durations varied considerably between shifts though in theory all shifts should go through the same operating procedures with similar labor resources. Shift D was on average the fastest while shifts A and B seemed to be the slowest performers. Also some inconsistencies with the number of PMs were noticed: shift A did not perform any type 6 PMs and shift D any type 4 PMs. One possible explanation for these differences could be that shifts A, B, and C had few problematic PMs that increased the average durations. However, in that case technicians should log in the work as "repair", not as "preventive maintenance". Faulty log in practices in shift D were offered as an explanation for their superior performance by other shifts. Another

explanation was that day shift (A and B) operators get more special work requests from the engineers which disturbs and lengthens their routine PM work.

To understand the true reasons for these inconsistencies, on-the-floor observations were conducted. Due to time limitations, it was decided to focus only on PM type 1. First, the PM procedure was mapped using specifications and direct observations in shifts A, B and C. After the process map was verified with equipment technicians and engineers, PMs were timed and compared against shift D performance. Finally, meetings with technicians from all the shifts were arranged to share the findings and to collect feedback and improvement ideas.

Based on the observations, the main reason for the shorter PM durations in shift D was found out to be their unique use of labor: instead of having one equipment technician working on one lithography tool at a time, like in the other shifts, shift D had 2 technicians working together on one tool. This two-technician approach did not require more resources, but significantly reduced the tool down-time since it enabled overlapping of work: while one technician was doing preparation work like collecting parts and tools, the other was shutting down the work station; while one was working on coater module the other was working on developer module. With one technician a typical PM takes so long that a break is needed in the middle. Breaks introduce non-value added time that unnecessarily lengthens the tool down-time. In the two-technician approach PM is so short that breaks are not needed. In addition, two technicians working together were able to support each other and solve problems faster.

During observations it was noticed that technicians had problems finding needed tools, spare parts and test wafers. Much time seemed to be wasted in non-value adding activities. The parts ordering system was not automated and missing parts occasionally delayed PM procedures. The user interface for PM specifications was cumbersome and difficult to use. Specifications were outdated and tool testing procedures failed to meet current technical challenges in the area. It was also noticed that some tool types in the lithography area suffered from the so called Waddington effect. The Waddington effect is

increased unscheduled downtime immediately following a PM event, and it indicates that PM schedules and procedures are actually causing problems. A complete list of observed problems, their priority as well as recommendations are given in Appendix I.

A business case study showed that if shifts A and B were to reduce their PM1 and PM2 durations down to shift C level, so not even to achieve the best case performance of shift D, a significant number of hours per week could be saved between the two shifts. These savings would translate into increased equipment technician resources and reduced tool down times, which in turn would increase tool availabilities and area capacity. For the lithography area increased capacity (if there is WIP to process) would result in decreased wafer costs. Improvements would also reduce the time technicians have to spend on routine PM work, and leave more time available for productivity improvement activities such as training, knowledge sharing and experimentation. Fixing problems would not only eliminated waste-time but also improve technicians' work morale.

4.4 ANALYSIS RESULTS AND RECOMMENDATIONS

The main results from the lithography variability analysis were:

1. The area reduces overall manufacturing flow variability for the main process technology, since variations in the departing material are smaller than variations in the arriving material.
2. Process time variations are small.
3. Area utilization is high. Hence lithography steps introduce non-value adding queuing time in to the process increasing total wafer cycle time. The area could be a local bottleneck.
4. Variations in arrival times are the major source of overall area variability.
5. Variations in tool availabilities are not a major concern in the area.
6. Preventive maintenance operations have plenty of room for improvement.
7. For lower volume process technologies flow variability through the area stays constant, thus material experiences less queuing time. However, inconsistent material flows increase tool specific arrival variations and disturb processing of main process technology products.

The main recommendations for the lithography area were:

1. Reduce arrival time variations by
   - improving tool-operation dedication matrix
   - improving tool health in upstream operations to increase availabilities
   - synchronizing batch sizes between lithography and its feeder areas
2. Reduce process time variations caused by operator decision making by automating the queue building process. Improve tool signaling to reduce wait-to-tech time after processing.
3. Improving preventive maintenance operations for example by
   - sharing the best-known-methods between shifts
   - updating preventive maintenance schedule and specifications to match the current technical challenges in the area
   - removing equipment technicians from running operations

The final and most important recommendation is to educate operators and supervisors on factory physics principles, especially on topics related to variability, and to engage them in improvement brainstorming sessions to better utilize their knowledge of the area. Any long term changes will be impossible to sustain unless the staff has been actively engaged in the improvement process and understands the reasons behind changes as well as implications for the whole factory.

Finally, it is good to keep in mind, that while Factory Physics helps the organization to understand the importance of variability, the analysis methods were not created specifically for semiconductor fabrication facilities. Due to the highly complicated manufacturing flow, several assumptions and generalizations had to be made in order to analyze the data. Results might, therefore, be somewhat misleading and should be interpreted carefully. However, the approach helps to identify the most important variability drivers and problem areas.

5 MANUFACTURING METRICS

Manufacturing metrics provide tools to evaluate the performance of a manufacturing system. Some typical manufacturing metrics include cycle time, throughput, utilization, inventory turns, setup times, yield, and order fulfillment. M.G. Brown describes the most important characters of a functional metrics system with the following [21]:

"It is alright to have hundreds or even thousands of metrics in your organization's database. It is just that no individual should have to focus on more than a few major ones. Along with having a reasonable number of metrics, another key to success is to select measures that are linked to your key success factors. If you are serious about running your organization with a specific set of values, it is also important that you have metrics in your scorecard that tell you how well you live by your professed values. Measures need to be derived from your strategy and from an analysis of the key business factors you need to concentrate on to ensure that you achieve your vision."

It will be very challenging to achieve lean manufacturing or to reduce manufacturing variability as long as the daily floor level metrics are not aligned with the vision and do not provide the right set of data. Additionally, the performance of a manufacturing line is often also part of the performance evaluation criteria for individual employees in the manufacturing organization. Thus the daily metrics are the main motivational factor and influence how people behave on the floor level. Wrong metrics can promote behavior that aims at optimizing local performance, weakening thus the overall performance of a manufacturing line. Even if the metrics are measuring the right things, the impact on overall performance can remain weak unless employees are taught the relationships of individual performance measures and higher level factory performance.

Metrics are also used to compare different fabrication facilities against each other. Benchmarking key competitors and overall industry performance is a way to set goals for

[21] Mark Graham Brown: Keeping Score—Using the Right Metrics to Drive World-Class Performance. Productivity Press 1996, Portland

the organization. The key performance metrics in the semiconductor industry is days per mask layer. Days per mask layer describes the time in days it takes for a facility to fabricate one layer on the multi-layered integrated circuit wafer. This metric takes into account the fact that fabs can manufacture different types of dies with different numbers of layers. However, the metric does not take into account the loading factor: a factory that is almost empty of WIP can make one layer on the wafer much faster than a factory that is fully loaded with inventory. Also high product mix is not accounted for.

5.1 METRICS AT FAB11

While industry benchmarking metrics might have some inaccuracies, they still help organizations to evaluate performance against competitors. Intel Fab11 has not been entirely satisfied with their performance, which is why they are looking for improvements for example by focusing on understanding and implementing Factory Physics methodologies.

Table 3 lists some of the most important metrics currently in use in Fab11 together with related challenges.

Table 3. The main performance metrics used in Fab11.

Current Metrics	Main user group	Comments
Output: wafers out	Operators and supervisors daily on tool level; manufacturing managers and factory management longer term on plant level	Floor level metrics "moves" → push vs. pull → WIP bubbles
Cycle Time: WIP turns	Manufacturing managers, factory management	WIP turns or absolute CT days cannot be used to compare factories with different loadings
Cost: wafer cost	Factory management	Drives high utilization → less flexibility Depreciation dominant over variable cash items → hard to notice improvements in wafer cost Driven either by starts or outs depending on group
Quality metrics: e.g. yield	Manufacturing engineers and managers	Not the biggest concern, under control

The two most important daily metrics for area supervisors and operators on the floor were moves, i.e. output per workstation, and inventory levels. Both values were compared against daily goals that were calculated automatically based on a complicated system on desired WIP turn value, wafer start levels and estimated cycle times throughout different work stations. The main goals for the supervisors and their shifts were to achieve at minimum the required outs, preferably even exceed the goal level, and to keep the work station inventory levels below the goal. In addition to outs and inventory at the station, cycle time on a lot level was also used to define run plans for the operators. Material to be processed was prioritized partly based on lot age (cycle time vs. goal).

Manufacturing managers were following output levels and cycle time on a plant level over a slightly longer time horizon. A convenient daily metric was inventory level graphs which helped manufacturing managers to see the movement of WIP bubbles (variations in WIP levels) and thus problem areas in the factory at a glance. Wafer cost, cycle time and WIP turns were the main tools for the factory management to compare Fab11 performance against other Intel facilities and competition.

As is typical for a semiconductor facility, there were several manufacturing quality metrics in use in Fab11. However, quality metrics were not seen as a major concern in metrics alignment.

5.1.1 Weaknesses in Current Metrics

Output

The daily output metric used by supervisors to guide their area had some inherent problems. First, since the output was measured only at the end of a shift, the consistency of material flow was not captured. A work station could be processing most of the material either in the beginning or in the end of the shift, thus causing a variable output flow for the next station downstream. Second, exceeding the output goal was not penalized, and due to lack of a "pull" system the level of inventory at downstream stations was not included in the output goals calculation. Thus a workstation could be processing too much material, congesting the station downstream and causing WIP bubbles. Third, output measures were documented and used to compare daily and weekly

performance of different shifts and tool areas. This motivated the operators to behave in ways that optimized the local by shift performance instead of the global performance of the whole manufacturing line. Operators might have chosen to run products that required the least setup times, or ignore a service alarm that appeared close to the end of the shift in order to maximize the output quantity for their shift. At the same time they "dumped" non-value adding work to the next shift and deviated from the run plan causing excess queuing time for neglected products.

Cycle Time

WIP turns were used mainly to represent cycle time and to compare different Intel factories against each other and against competition. The WIP turn metric is typically defined as throughput divided by average inventory; the higher the number the better, since the less WIP is needed to achieve desired throughput, the less variability and queuing time is induced into the system. However, this metric tells nothing about the speed of the factory. In Intel systems WIP turns was defined using activity steps and cycle time. However, a factory that had a great WIP turn performance could have had many more activity steps in their processes than a factory manufacturing the same product with a more sophisticated process flow. Thus, a WIP turn metric does not encourage implementing technical improvements in order to reduce the number of process steps. WIP turns or absolute cycle time should not be used for comparing purposes, since they do not take into account factory loading or product mix complexity.

Wafer Cost

As was mentioned in Chapter 3.3, one of the challenges in semiconductor manufacturing comes from the high capital expenditures which drive high tool utilizations through wafer cost measures. Wafer costs are formed of fixed and variable costs, but due to the dominant role of fixed capital costs any improvements in the variable cost portion are hard to notice. Due to industry conditions, cost is overall a dominant business driver, which makes it hard to implement any improvements that would result in lower tool utilizations. Additionally, some parts of the Intel organization use the number of starts to calculate wafer costs, i.e. reduction in the number of starts would result in increased per

wafer cost. Thus, even though reducing start levels might in certain circumstances benefit the overall factory performance through decreased loading and queuing time, reductions in start levels are extremely hard to implement due to the fact that certain management groups have their individual performance measures tied in with wafer cost metrics. The fact that in other organizations wafer costs are calculated using the output from the factory instead of starts creates even more conflicting drivers for operational principles and goals.

As a summary, the main weaknesses in the current metrics are that they

- do not measure variability
- do not capture the relationship between loading (WIP level) and cycle time
- point out problems only after they occur (WIP bubbles)
- promote wrong operator behavior (hard for operations to interpret current metrics)
- cannot be used to compare different factories

5.2 METRICS RECOMMENDATIONS

The following table shows a set of metrics recommendations based on literature and company internal meetings. These metrics should not in all cases replace the old ones, but they could be used as complimentary to improve the weaknesses in the existing metrics portfolio.

Table 4. Examples of improved manufacturing metrics.

Focus Area	Metric / Tool	Goal
Cycle time & Variability	Operational Curve	- Depends on strategic CT goals; limit for high variability = 1.33
Cycle time & Loading	LACTE = Load Adjusted Cycle Time Efficiency	- $(t_0/CT) * u \geq 30\%$ (industry benchmark)
Flow variability	Arrival variations vs. departure variations	- Even flow with $(c_a - c_d) = 0$ - Steady outs flow during shift
WIP management	Inventory Variability Compliance to run plan	- "Minimum Inventory Variability" - Meet outs goal (not exceed), run the right material

Cycle time and variability relationship can be monitored with a so called operational curve that shows the relationship between factory utilization (bottleneck loading), cycle time and variability. An example of an operational curve (similar to the one in Figure 4) is presented in the following Figure:

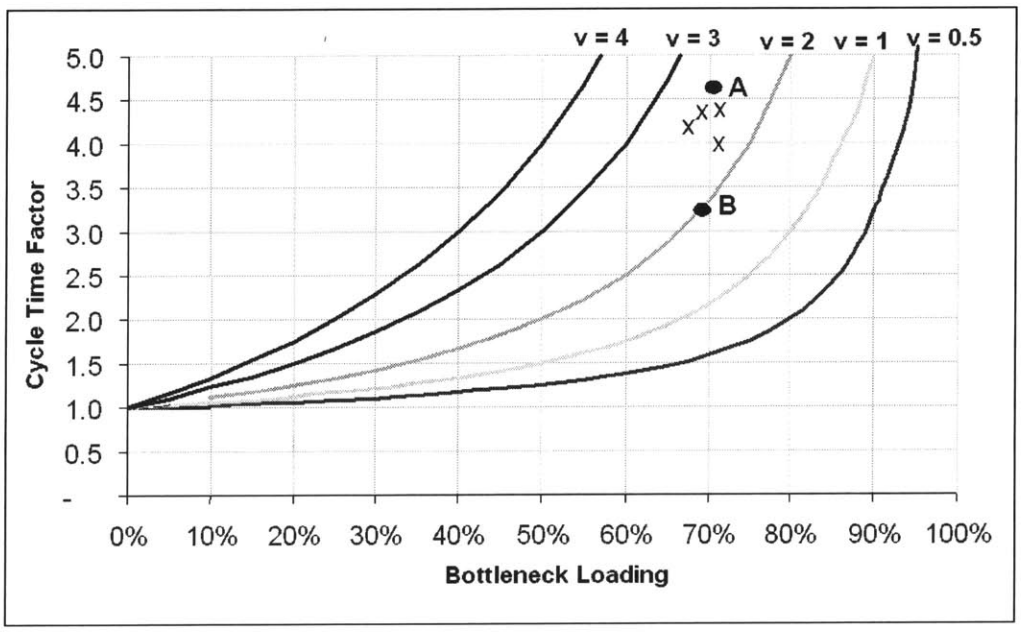

Figure 13. An example of using an operational curve to track performance and goals.

An operational curve helps the organization to understand how performance is developing over time and what is required to achieve a desired cycle time and/or utilization. Figure 13 illustrates an imaginary situation where a manufacturing facility was at point A with variability of 2.5, cycle time factor of 4.5 (true cycle time divided by theoretical cycle time), and bottle neck utilization of 70%. Without a change in utilization the organization wanted to achieve point B with cycle time factor of 3.3. To achieve their goal the organization has to reduce variability down to 2. Based on most recent performance measurements they are on the right path. More information on operational curves and how to create them can be found in Aurand and Miller (1997) [22].

[22] Stevens S. Aurand, Peter J. Miller: "The Operating Curve: A Method to Measure and Benchmark Manufacturing Line Productivity". IEEE/SEMI Advanced Semiconductor Manufacturing Conference 1997, pp. 391-397.

As was mentioned earlier, using real cycle times or cycle time factors to compare facilities does not take into account different levels of loadings. The problem could be solved by using a load adjusted cycle time efficiency (LACTE) number, which multiplies the cycle time efficiency (theoretical cycle time divided by true cycle time) by utilization. World class manufacturing facilities have cycle time efficiencies of around 30%. When taking into account utilization, for example 70%, load adjusted cycle time efficiency becomes only around 20%. The biggest problem with this metric is that you would need to know the bottle neck utilization for a facility. As discussed in Chapter 3.3 bottle neck recognition can be very challenging. Additionally, even LACTE does not fully take into account the differences in production mix.

Currently Intel has no metrics to measure variability in the manufacturing flow. The following graph shows a proposal developed in-house that could potentially be used for this purpose. The graph plots departure variations for 50 consecutive operations in one segment of the manufacturing line. As was explained in Chapter 3.2.1, departure variability from a tool becomes the arrival variability to the next tool downstream. Thus the segment graph compares the relationship of arrival and departure variabilities for individual operations: each dot represents departure variability and the dot before represents arrival variability for an operation.

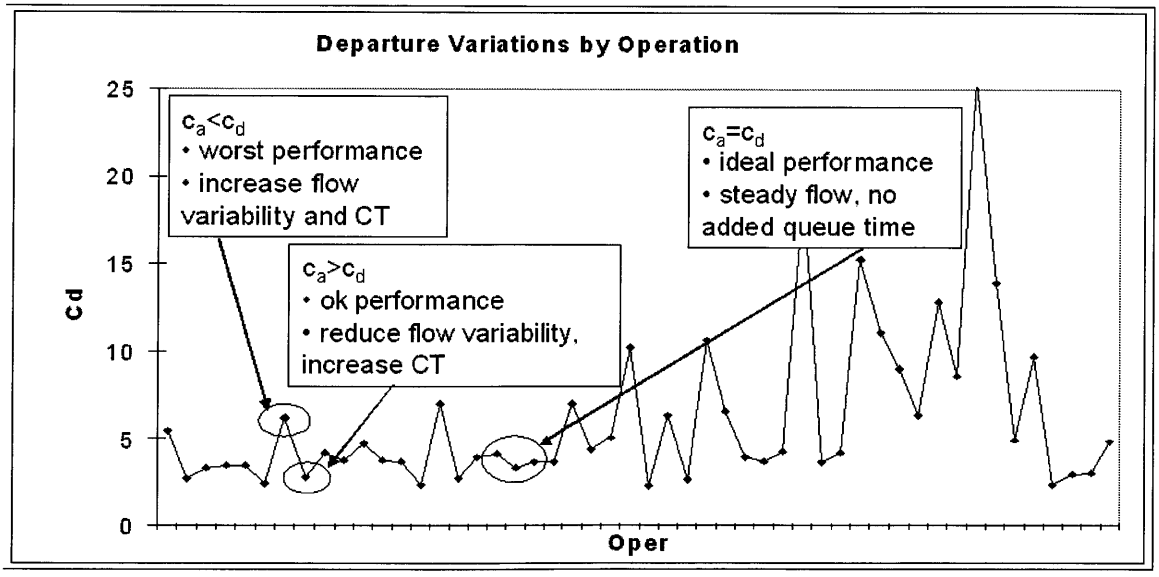

Figure 14. A method developed in-house to measure manufacturing flow variability.

The leftmost circled operation illustrates a situation where variability in the arriving material is smaller than variability in the departing material. This operation represents the worst kind of performance. Manufacturing flow variability is increased which means high tool utilizations and high process time variations. Thus the operation creates queuing time and as a result increases the overall cycle time. The next operation downstream (to the right) represents a case of medium performance: departure variability is smaller than arrival variability. This was the situation in most of the lithography operations studied in this work. Flow variability is reduced, but the behavior still indicates high tool utilizations resulting in queuing time and increases in overall cycle time. The three circled operations in the right represent the ideal case when manufacturing flow remains fairly steady; arrival variability equals departure variability. This type of behavior indicates that tools have lower utilizations creating little or no queuing time for the system. The challenges for using this kind of metric are largely the same as the challenges faced during the variability analysis phase of this thesis work: how to access data in the right form and how to simplify the highly complicated manufacturing flow by using acceptable assumptions. Finally, as is the case for all the proposed metrics, a big challenge is how to implement an automated recording system in the factory.

As was discussed in the previous chapter, one of the weaknesses in the current daily output measure is the fact that it does not monitor how consistently material is processed during a shift. Another company internal recommendation for monitoring variability in the daily manufacturing flow was a system that records outs in regular short time intervals over a shift duration and compares the output to a desired trend. At the same time any overproduction, outs exceeding goal, could be captured. In addition to production rate, compliance to run plan should be monitored. The idea of consistent production is illustrated in Figure 15.

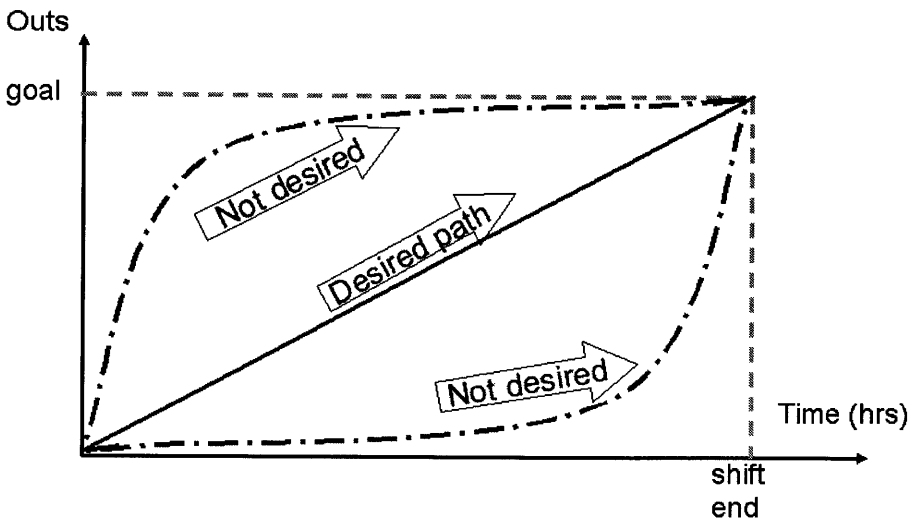

Figure 15. The preferred manufacturing flow at a work station.

Currently Fab11 has very poor WIP flow control systems. Only the last segments of the line use a pull system which signals the needs from the end of the line based on promised customer shipments. The rest of the factory is more or less functioning on a starts based push system. Though WIP management is too wide of a topic to be covered in this work, it is important to highlight the current lack of proper metrics controlling WIP movements in the factory. Inventory level graphs show where WIP bubbles are located and thus give an indication of where efforts to solve problems should be placed. However, they do not capture the effect on variability. As Collins and Hoppensteadt (1997) highlight in their paper, maintaining steady inventory levels decreases the inventory level standard deviation which in turn results in decreased average WIP (characteristic behavior for an exponential distribution)[23]. As stated in Little's law, reduced WIP levels would decrease cycle time, assuming throughput remains constant. A method for controlling variability in inventory levels, "Minimum Inventory Variability" or MIV, is presented in the same paper. According to the MIV method, material release (run) policy should be made based on WIP profile and consider inventory status at own tool and at the next tool downstream. The goal is to maintain inventory at each station close to an average WIP level in order to minimize variations. The idea is illustrated in the following figure.

[23] D.W.Collins, F.C. Hoppensteadt: "Implementation of Minimum Inventory Variability Scheduling 1-Step Ahead Policy in a Large Semiconductor Manufacturing Facility", 1997 IEEE

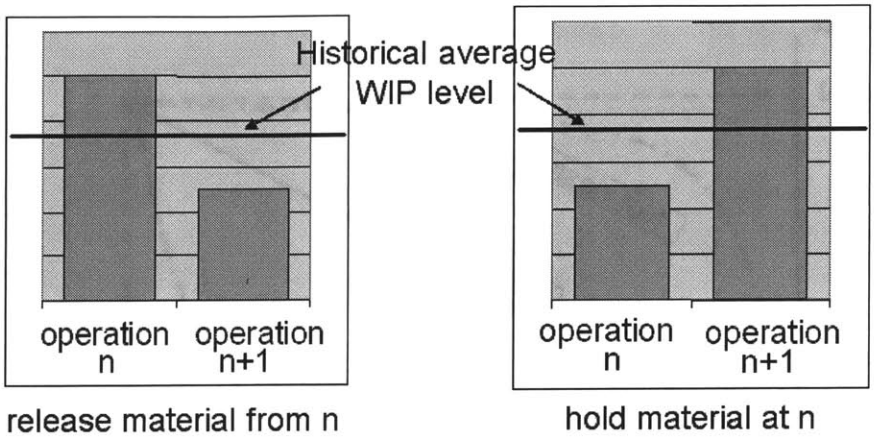

Figure 16. Concept of "Minimum Inventory Variability" by Collins and Hoppensteadt (1997).

Currently lots are run and prioritized based on urgency labeling, total passed cycle time per lot, tool setups and send-ahead requirements, as well as supervisor instructions. Minimum Inventory Variability policy could be used together with these policies in areas where there is no pull system from the end of the line. Another potential tool for inventory control has been presented by Leachman, Kang and Lin (2002)[24].

As was noted earlier in conjunction with the variability analysis in Chapter 4.2.1, some of the basic manufacturing parameters in Fab11, such as equipment availability and utilization, were formulated differently from the Factory Physics approach. The starting point for Factory Physics definitions is operations based, where as for Intel it is cost based. As a result, the data coming out from Intel systems is not ready as such to be inserted into Factory Physics formulas, but either the data or the formulas need to be modified before actual analysis. As these basic parameters are deeply woven into the systems including company wide capacity planning, correct handling of data becomes critical.

No matter what metrics are implemented, it is important to remember that in this case the implementation must happen from top down. Management's approval and support are

[24] Robert C. Leachman, Jeenyoung Kang, and Vincent Lin: "SLIM: Short Cycle Time and Low Inventory in Manufacturing at Samsung Electronics". Interfaces, Vol. 32, No. 1, January-February 2002, pp.61-77

crucial for strategic organizational alignment and successful change implementation. The metrics presented in Table 4 are important for the implementation of Factory Physics methods, since variability which is one of the key concepts is not accounted for in today's metrics. Not only would the right metrics provide the right kind of data to analyze the process, but they would also help to guide operator behavior. Without proper metrics alignment, operators and area supervisors do not have the right incentives to run production under Factory Physics principles. In addition, it is important to understand how incentives influence various cultural and political aspects that could potentially work against implementing Factory Physics. These issues will be discussed in more detail in the following chapter.

6 ORGANIZATIONAL CHALLENGES

Organizational culture and political structure are key determinants in how changes get implemented. Although it is hard to get a full understanding of these aspects during the short duration of a thesis study, the following paragraphs try to highlight some of the most apparent organizational challenges for implementing Factory Physics methods.

This thesis study and the overall Factory Physics project supports Intel values: risk taking, quality, great place to work, discipline, results orientation, and customer orientation. Difficulties arise because for Fab11 Factory Physics presents a radically new way of looking at operations management. In many occasions, Factory Physics approaches are counterintuitive compared to the current operational principles. In addition, current metrics do not fully support Factory Physics methods. Thus a considerable amount of organizational learning is required. Moreover, the concepts are not easy. It took the management level team several months before they properly understood the related methodologies and were ready to start applying some of the concepts in factory environment. And last but not least, implementing change is a battle at Intel. These features make it challenging to successfully implement Factory Physics principles in to Fab11 manufacturing operations.

6.1 CULTURAL FACTORS

Copy Exactly

One cannot talk about Intel culture without mentioning "Copy Exactly". Copy Exactly means duplicating everything (process flow, equipment set, suppliers, plumbing, clean room, and training methodologies) from the development plant to the volume-manufacturing plant with the help of a "Virtual Factory" organization. The methodology was first established by early Intel executives who wanted to prevent the separation of product development and manufacturing. Today, the method significantly shortens the production learning curves and enables high-volume factories to become online quickly; hence, decreasing time to market and increasing yields. Copy Exactly reduces corporate

risks since products are not dependent on only one manufacturing location. Similarly, customers only need to qualify products at the development facility. [25, 26]

Even though Copy Exactly improves knowledge sharing and minimizes risks in the highly sensitive manufacturing process, it also makes change implementation more bureaucratic. Any changes to the floor level operations require several approval rounds and extensive pre and post verification studies. This process would have to be followed for example when wanting to change the lithography maintenance operations. Even though changes could potentially improve the factory performance, they might never get implemented, since the time and effort required from equipment technicians and engineers to get the change approval conflict with daily tasks and performance metrics.

Copy Exactly provides a rigorous approach to process development. However, surprisingly, the same discipline does not extend to standard working procedures. Although documentations exist, they are not necessarily followed or kept up to date. Moreover, the principles how supervisors run their areas or operators their equipment are not standardized, but there are significant differences between factories, shifts and even individual operators within one shift. The lack of "copy" in working procedures does not support Factory Physics approach, but leads to increased variability and conflicting priorities among supervisors.

Documentations and bureaucracy do not necessarily have to take away flexibility from the system. As was noticed by Spear and Bowen (1999) in their study on the Toyota production system, it was the rigid specifications that made flexibility and creativity possible[27]. However, the key difference is that in the Toyota system the rigorous change implementation was made as part of the every day work, according to continuous improvement principles.

---

[25] Intel Virtual Press Room: http://www.intel.com/pressroom/kits/manufacturing/copy_exactly_bkgrnd.htm
[26] "Intel Labs (A): Photolithography Strategy in Crisis", Harvard Business School Case # 9-600-032, 1999
[27] S. Spear, H.K. Bowen: Decoding the Toyota Production System, Harvard Business Review, September – October 1999, pp. 97-106

Local vs. Global Competition

At Fab11 shifts and equipment areas have traditionally been treated as individual competing units. Functional teams are evaluated based on their output performance during one shift. Area performances are then compared daily and weekly between all four shifts. Similarly, in the individual level, knowledge is a way for the operators to feel empowered and special. Instead of sharing their best known methods people are tempted to keep technical skills to themselves to receive better than average performance reviews. This attitude has partly been reinforced by increased job insecurity caused by recent layoffs.

Though Factory Physics focuses on individual manufacturing units, the concepts require also a great deal of cooperation and knowledge sharing. The performance of an individual functional area should be as effective as possible while it is operating, but when and how much is produced should depend on the surrounding units up and down stream. Thus Factory Physics aims at improving the performance for the manufacturing flow as a whole, even when it sometimes means weaker performance for individual areas. However, the individualistic company culture supported by the metrics system strongly promotes local rather than global performance optimization. In such an environment it is hard to achieve standardized working procedures, knowledge sharing or cross training.

Just as individuals and teams compete within the Fab11 organization, fabs also compete against each other on the corporation level. Metrics like cost effectiveness, quality and customer service determine which fabs get to produce the next new products. Flash is a strategic rather than profit creating product, and due to the waterfall principle flash fabs tend to be at the bottom of the food chain. However, this gives flash fabs a special incentive to focus on operational excellence. Fab11 is the first one to try to implement the Factory Physics methods on a larger scale. A flash factory in Colorado is trying to reach the same goal, but instead of taking a very quantitative hands-on approach, they chose to have a higher level top-down lean implementation approach aiming in the first place at

changing the organizational culture[28]. The Fab11 approach might offer faster wins, since it uses the language understood in the Intel organization: data. However, implementing new methods will still require a shift in the organizational culture, at which point learnings from the Colorado experience will be valuable. The "not-invented-here" attitude arising from the competitive fab landscape creates resistance when factories are trying to adopt methods created in other sites.

6.2 STRATEGIC DESIGN AND POLITICAL TENSIONS

The Factory Physics task force team at Fab11 crosses several organizational boundaries including members from manufacturing, engineering, virtual factory, and finance. Multifunctional structure ensures that Factory Physics methods are thoroughly evaluated and understood by all organizational functions. More importantly, support from these various functions is crucial in order to accomplish successful implementation. Though all functions support the same methods, their motivations might differ. In short, manufacturing wants to reach faster cycle time; engineering wants to improve tool availabilities; virtual factory wants to ensure that the results are usable in other Intel factories; and finance wants to reduce wafer costs and inventories. At the current stage these goals are strongly conflicting, forcing production managers to balance their daily operational decisions between various stakeholder requirements. Factory Physics is seen as a tool that can help to achieve several of these goals simultaneously; the method is, therefore, supported by the different organizational entities.

However, Fab11 organizational structure creates also challenges for implementing Factory Physics methods. As was mentioned earlier, significant training is required before implementation can take place. Manufacturing organizations consist of several vertical layers and horizontal functional areas which are all equally important for day-to-day operations. Which layers or functional groups to train first? Should some levels receive more in depth training than others? Due to the high number of re-entry operations change implementation needs to happen simultaneously in all functional areas of the manufacturing line. Otherwise global optimization will not work and improvements from

[28] For more information on Lean Implementation at Intel see LFM Thesis by Roy Wildeman (2004) and Jason Connally (2005)

areas using Factory Physics are neutralized in areas that operate under traditional methods.

Additionally, current metrics create political tensions in the current system. This was especially observed between engineering and manufacturing organizations. Engineers work in an office environment with normal working hours. Floor level operators and area supervisors on the other hand work in shifts (night or day) and spend most of their time in the cleanroom. As a generalization, floor level employees feel that engineers work in isolation "in their tower" and do not spend enough time on the factory floor where the real knowledge is. Engineers again feel that operators have nothing to offer since they do not understand the importance of the rigorous science that lies behind manufacturing. While similar attitudes are common in many manufacturing organizations, the behaviors can better be understood by looking at the current incentive system. Lack of metrics alignment creates tensions, as can be shown with the following two examples:

Equipment technicians face a virtual matrix organization in their daily work. Equipment technicians work in the cleanroom and are responsible for equipment maintenance. They are organizationally under shift supervisors in manufacturing. However, their most important daily point of contact is the equipment engineer, who comes from the engineering organization and is responsible for long term tool health. Shift supervisors' main focus is to achieve their daily output goals. On the other hand, equipment engineers' main focus is to follow the maintenance schedule in order to maintain production quality. Both goals are perfectly aligned with individual performance metrics, but as a result the technicians face conflicting requests: an equipment engineer might want the technician to perform a scheduled maintenance operation, while at the same time an area supervisor wants to postpone the maintenance and have the technician running production. Under current metrics my recommendations for improving maintenance operations serve only equipment engineering. Thus manufacturing supervisors have no interest to support the effort, even when they might understand the long term benefits.

Another example is a situation when the engineering department wants to have test wafers processed to get important information for a critical quality study they are performing. However, for the supervisor and the team on the floor this means increased manufacturing flow variations and product runs that might require extra attention and setups, but that are not counted towards the daily output goal. Thus there is no incentive to run the test wafers.

Engineers are unable to implement Factory Physics methods alone. They need to make sure that manufacturing staff is included starting from the planning phase to get buy-in and to take advantage of the enormous amount of knowledge operators have on operations. Similarly, manufacturing staff needs to be trained on the theoretical concepts to help them understand why the engineering department is requesting changes. Simultaneously the metrics system needs to be aligned.

7 SUMMARY

The flash industry is a challenging environment for Intel. The company is no longer the dominant player with significant power over the whole supply chain. Instead, the company is facing intense rivalry from competitors who can manufacture flash memory more cost effectively. Though the importance of flash to Intel might be in the ability to complement its platform strategy rather than to generate high profits, any in-house flash manufacturing still needs to remain cost effective.

Fab11 has realized what challenges lie ahead and what should be changed in order for the facility to stay competitive. The only way to reduce throughput times with the current pressures on tool utilizations is to attack variability. Factory Physics provides a framework approach to analyze variability in a data driven operational environment. However, the method has not been designed for a semiconductor facility, which lengthens the analysis and implementation phases. Additionally, the influence of metrics on organizational culture and on change resistance should not be ignored.

7.1 THESIS CONCLUSIONS AND KEY LESSONS LEARNED

This thesis work developed a process to analyze variability in manufacturing flow. Though the method was used here to study the lithography area, the principles are applicable to other functional areas in the factory. The main conclusions from the analysis showed that the lithography area reduced the overall manufacturing flow variability. The main source of variability was arriving material flow; process time variations at lithography tools were minor. However, this type of behavior indicates high tool utilizations, and thus increased queuing time for the total product throughput time. Since lithography is a critical part of the re-entry loop in semiconductor manufacturing, the global performance of the factory could benefit from improvements in the area.

The main recommendations focused on three areas affecting variability:

- reduce variability in the arriving material flow by for example focusing on tool-operation allocations

- reduce variability in process times by standardizing operator decision making
- reduce tool utilizations by improving preventive maintenance processes

Additionally, the alignment of manufacturing performance metrics with variability reduction goals should be strengthened by introducing new variability monitors and by educating manufacturing staff. Current output focused goals create a culture that promotes local instead of global performance optimization. Similarly, misalignment in cost metrics extends to the top levels of the organization. Unless these metrics are aligned, the underlying culture and political tensions will cause significant resistance for changing the basic operational principles in the organization.

7.2 SUGGESTIONS FOR FUTURE WORK

When this study was finished the Factory Physics task force was only starting to implement their new knowledge into a production environment. A natural continuation for this work would thus be to implement some of the recommendations and see if the lithography area is able to improve its performance. Additionally, many of the recommendations brought up in this work should be further investigated and developed before actual implementation:

- One of the major recommendations for reducing arrival flow variability was to optimize tool-operation dedications; in other words, deciding which tools should be running which products, if these dedications should be fixed, and how changes in capacity should be accounted for. Due to expected capacity changes in the area, the effort was not pursued further in this thesis work. Nevertheless, having a better understanding of the dedications and having a tool that would take away the operator decision making would be very valuable for the lithography area.
- Preventive maintenance was another area with identified potential for improvement. Tools are continuously aging, but maintenance schedules have not been revised in a long time. One suggestion, therefore, would be to use Total Productive Maintenance (TPM) methods to study the maintenance schedule and to reduce the number of unexpected tool breakdowns currently disturbing production in the litho area.

APPENDIX I

Complete list of problems and recommendations based on preventive maintenance observations

	P=primary S=secondary priority	Issue	Current state	Problem	Causes	Improvement Ideas
Materials	P	Test wafers	Wafers brought to manuf. when needed, no lot tracking on the floor	Systems show that several test wafers on the floor so cannot proc in new lots, but no idea where wafers are	Time wasted in searching for test wafers and organizing lots; no accountability	Locked strorages for each shift, contol and accountability over own test wafers
	P	Tool boxes	Share tool boxes; boxes messy and incomplete; nowhere to place tools during PM	Techs need to look for missing tools, chemicals, keys etc; tools not conveniently available during PM	Wasted time; frustration	Personal tool boxes, vendor tool boxes good; nominated tool box reps who make sure boxes have required tools (create list); smarter carts that have more room to place tools on
	P	Replace part ordering	Replace part cleaning outsourced, before done in-house	Parts need to be ordered manually, min 1 hr delivery time; deliveries sometimes late; parts no longer dedicated to certain tools	Wasted time due to late deliveries; no understanding why process was outsourced; bowl-track dedication eliminated some coater setup steps	Automatic ordering system based on PM schedule; previous shift orders bowls for next shift
	P	New / harvested parts	First priority is to get missing parts from harvested, ordering new parts secondary method; no quality checks	Harvested parts arrive often faulted or broken; pumps and pump filters have problems with reorder points, parts not available even if wings shows them	Time wasted in waiting for right or functioning parts, replacing faulty harvested parts; expedite orders more costly	Better control systems for harvested parts and new parts inventory; better quality checks by supplier
	S	Quality monitor tools	Number of monitor tools reduced, tool locations change	Techs cannot always find free tools, need to walk long distances with sensitive test wafers and wait	Wasted time, decrased test wafer reliability	Check if the testing capacity sufficient; create rules for tool locations
PM Procedure	P	2 techs' PM	Most shifts allocate one tech for one tool	PM time on average x hrs; includes breaks, waiting for parts etc.	Tool down time decreases availability; unecessarily lengthy PMs with breaks	Copy shD method: two techs work simultaneously with one tool; realistically down time could be reduced by 50%
	P	PM frequency	Litho has different PM frequency from sister area even though almost identical tools	Engineers do not talk to eachother; litho makes more PMs?	Wasted time for unecessary PM activities	Improve communication between sister area and litho; share BKMs on PM processes; standardize across tool areas
	P	PM scheduling	Not all shifts make PM schedules in advance, most decide in the beginning of the shift based on WIP level and tools up	No visibility, longer term planning or plan sharing between shifts; some techs end up doing more PMs than others	Time spent every day for scheduling; hard to divide work evenly; previous shift cannot support in handover; no knowledge on holidays, trainings etc.	Standardized scheduling tool for all shifts; create schedule for whole week before shift starts; select responsible techs to create fair schedule
	P	Specifications	Techs do not always follow specifications or check lists; memorize PM steps by heart	Specs do not match with check lists; specs are out of date and complicated to use, refere to documents that do not exist	No trust or respect for specs; documents don't seem to have any role; techs have created their own PM check list orders	Specs should be simplified and updated based on BKMs from tech feedback; all relevant information should be included in one document with clarifying pictures; keep updating every 6 months
	P	Wafer handling audit	Currently done as separate step?	Wafer handling audits do not quarantee future performance for tools	Audits useless since tool could fail the next minute; only one moment in time check	Include wafer handling audits in quarterlies
	P	PM 2 process	In PM schedule	Some steps seen as unecessary	This PM not needed in combination with PM 1 and in-line monitors	Combine PM 2 with other PMs
	P	Stepper - track coordination	Weak coordination between Stepper and track PMs; techs can do PM for one of the two units only	PM scheduling, planning and execution do not involve coordination between track and stepper; techs not able to help when one unit rans into trouble	One module always ends up waiting for the other; total link PM time increases	Better coordination, e.g. change checklist so that track monitors run before stepper
	S	Tool logging	Suspicion that shifts log PMs in different ways	False belief that logging short PMs makes shift or self look better; uncertainty what logging parameter to use	Major engineering or capacity decisions might be made based on unrealistic data	Emphasize the importance of correct log-ins; create more options to cover abnormal situations e.g. waiting for link
Other Procedures	P	Cross training / running operations	Techs run ops and do cross training in addition to maintenance work	No time to adapt to the new work load (headcount) or changing technologies (resist); no time for knowledge sharing or experimental learning; not all shifts operate under same principles	Increased stress, dissatisfaction; increased risk of major excursions	Free techs from running ops for now to give time for adjustment; encourage knowledge sharing; standardize principles across shifts
	P	Area goals	Differences in the way sups run their areas; what goals try to achieve	Techs in some shifts run ops to meet output goals, in some shifts do purely maintenance work to keep tools running	No standardization between shifts; variability in performance; unfair treatment for techs	Better standardization of goals and principles between shifts
	P	Parts order rights	Only some level 3 and 4 techs can order parts (1-2 per shift)	Techs without account rights need to find an authorized tech to order parts; do not know when parts arrive	Wasted time; complicated communication loops; no clear understanding why change was made and on which basis autorized techs were selected	More account authorizations; alternatively give order responsibility to lower level techs to free up more time for higher level techs or use automated order system
	S	Feedback loops	Ideas from the floor are not taken seriously; engineers dictate what to do	Techs' ideas and knowledge not captured	Frustration; feeling of disrespect	Establish a solid process for tech-to-engineering feedback; create culture of tech and PM work respect
	S	Engineering requests	Requests from engineers often added into normal PM	Lengthens PM process; some engineering data collection seen as useless	Tool down time increased; tech frustration	Better scheduling for engineering requests; reduce the number of requests if possible
	S	Tech levels	Levels 1-4	High level tech not used for their best abilities; no time for training	Hihgly skilled techs feel frustrated, unappreciated	More clear principles on tasks for different levels