(17)

Analysis of Control Mechanisms in a Re-entrant Manufacturing System

By

Paul Gifford

B.S. Mechanical Engineering
University of California at Los Angeles, 1991

Submitted to the Department of Mechanical Engineering and the Sloan School of Management in partial fulfillment of the requirements for the degrees of
Master of Science in Mechanical Engineering and Master of Science in Management

in conjunction with the Leaders for Manufacturing Program

at the
Massachusetts Institute of Technology

June 1997

Signature of Author _____
Department of Mechanical Engineering
MIT Sloan School of Management
May 9, 1997

Certified by _____
David Cochran
Assistant Professor of Mechanical Engineering
Thesis Supervisor

Certified by _____
Larry M. Wein,
Professor of Management
Thesis Supervisor

Accepted by _____
Ain Sonin, Chairman
Department Committee on Graduate Students

Accepted by _____
Jeffrey Barks, Associate Dean
Sloan Masters and Bachelors Programs

Analysis of Control Mechanisms in a Re-entrant Manufacturing System

by
Paul E. Gifford

Submitted to
the Department of Mechanical Engineering
and the MIT Sloan School of Management
on May 9, 1997 in partial fulfillment of the requirements for the degrees of

Master of Science in Mechanical Engineering
and Master of Science in Management

ABSTRACT

Modern semiconductor wafer fabrication facilities need to run at high levels of utilization in order to absorb the huge investments made in facilities and equipment. Simultaneously these facilities desire to minimize the cycle times of products in the factory for a variety of reasons including improved customer response, increased yield, and support of product development efforts. However, probabilistic models argue against the full realization of both of these goals at the same time.

Digital Equipment's Fab 6 facility had implemented a series of factory rules aimed at achieving these dual purposes. However this facility faced some unique challenges based on dual the mission of the fab to make multiple products and to cooperate with product development direction. Available publications and theory were considered inadequate to evaluate the effectiveness of the policies in use.

This study was conducted to assess the contribution to cycle time and cycle time variability of three operational policies in use in Fab 6. The three policies include start policy, dispatch policy and kanban-type control system. Through a design of experiments style investigation conducted on a simulation model of the factory, it will be shown that the fab has opportunities to improve its cycle time performance by adopting different policies. Moreover, it will be demonstrated that certain combinations of policies do not work well together and should be avoided.

Thesis Supervisors: Assistant Professor David Cochran, Mechanical Engineering
Professor Larry Wein, Management

Acknowledgments

I gratefully acknowledge the support and resources made available through the Leaders for Manufacturing Program, a partnership between the Massachusetts Institute of Technology and major United States manufacturing companies.

This research project would not have been successful without the resources and support of the Fab 6 organization of Digital Equipment Corporation. In particular, I would like to thank Ellen Mager and Karl Koch for their direction and input.

Professors Dave Cochran and Larry Wein provided important research insights during the internship period and valuable writing direction during the thesis phase. This paper would not exist without their assistance.

My MIT experience could not have happened without the prayerful and loving support of my parents Duane and Marilyn Gifford and my new parents Hae-Yong and Song-Cha Park. Nor would I have survived the hot summers and cold winters without the unwavering commitment of my dear wife, Kyoung.

Table of Contents

Chapter 1- Introduction

Motivation

Alpha Microprocessor Strategy

In the early 1990's, Digital Equipment predicted that Microsoft's Windows NT operating system (NT) would become the operating system of choice for powerful desktop computers connected to large networks. If the market adopted NT, a processor that could run it could also run the myriad of applications software packages expected to be available. Digital believed that it could produce a microprocessor that would run the NT operating system at higher levels of performance than those produced by rivals like Intel, Cyrix, IBM, and AMD[1]. Digital worked to design the fastest microprocessor possible, and settled on a 64-bit, reduced instruction set computing (RISC) architecture that the company named Alpha. Digital hoped to produce high performance personal computers and servers based on the Alpha chip at prices rivaling computers based on Intel's Pentium Pro processor but that would offer significant performance advantage.

Fab 6 to Produce Alpha

In order to produce the very complex Alpha microprocessor and the other semiconductor products being offered and planned, Digital Equipment needed a state-of-the-art wafer fabrication plant, often referred to as a fab. Digital broke ground for a new facility in Hudson, Massachusetts in 1992, and began operating the factory called Fab 6 in early 1994. Fab 6 was hailed by many measures as a model for fab building techniques[2]. Nevertheless, Digital's investment to build Fab 6 and fill it with the necessary processing equipment was mammoth. Digital would need to successfully execute its Alpha strategy in order to recoup the $550 million spent on plant and equipment, plus the additional millions invested in product development.

Move to Mass Market

When Microsoft released version 4.0 of Windows NT in 1996, it became clear that Digital's gamble on the operating system was correct. This was the first fully 32-bit version. Finally, an operating system was available that could harness the advanced computing power of the Alpha microprocessor and run the newest generation of applications software packages that were beginning to become available.

In the latter half of 1996 and early 1997, other elements of Digital's Alpha strategy fell into place. First, Digital signed agreements with the Korean semiconductor giant, Samsung, allowing Samsung to build and market Alpha chips. This was seen as a way to "broaden the chip's use and establish it as an industry standard."[3] Digital also began to reduce prices on Alpha processors in order to make them more attractive[4,5]. While these moves were necessary, Digital had work to do in its fab to be successful.

Focus on Output and Cycle Time

Digital's designs for Alpha and its other chips used in networking, multimedia, and other advanced applications were generally considered technically excellent. But to compete in the personal computer and consumer electronics markets, these products would have to be produced cost competitively. Digital believed that to offer significant performance advantages at equivalent cost terms of its competitors, it should focus on operating its Fab 6 at a supreme performance levels. This would require Fab 6 to focus on increasing output and reducing cycle time. Furthermore, the facility would also have to support continuing development of products and processes.

Output

In order to achieve competitive unit costs, Digital would have produce a huge number of chips in Fab 6 to absorb the huge fixed costs of the plant and equipment. It is estimated that for a modern fabrication facility between 65% and 75% of the product costs are attributed to fixed costs of plant and equipment and other fixed overhead[6,7]. Achieving a maximum production rate of salable chips from the factory would contribute significant toward success by spreading fixed costs over a larger number of chips.

The number of good chips available for sale depends on three things: the production rate of full wafers, the line yield, and probe yield. From eighty to as many as several hundred semiconductor chips can be built simultaneously onto an eight inch diameter silicon wafer. A greater output of wafers implies a greater output of chips. Near the end of the wafer fabrication process, the wafers are tested to verify that important steps have been completed properly. Line yield measures the fraction of wafers started into the factory that pass this test. Finally, each chip on passing wafers is individually tested and probe yield measures the proportion of chips that pass rigorous electrical testing.

Cycle time

Fab 6 also needed to focus on reducing cycle time. While Alpha was the most strategic and important product being produced, Fab 6 was producing other semiconductor products that were important sources of revenue to Digital. Fab 6 needed to make these other products in order to reduce the fixed costs being allocated to Alpha. Also, enhancements in design and process were continually being suggested for Alpha and other products. If Digital's semiconductor products were to be successful, they would need to be constantly improved. This required producing experimental versions of both Alpha and the other products. The fab would have to be able to

quickly run these new versions to see if the hoped for enhancements would materialize. Shorter cycle times also let Digital respond more quickly to changes in demand or special customer situations.

Besides being short, cycle times also would have to become more predictable. Between the factory and the customer sat a finished goods inventory known as a die bank. This bank was desired to be as small as reasonably possible once the fab became capable of volume production and processes became more settled. However, greater unpredictability of cycle times of chips through the fab required the holding of greater quantities in die bank. In order to reduce the number of chips stored in the die bank and be able to supply anticipated customer demand, cycle times would have to become more predictable as well as simply shorter.

Problem - Misapplied Control Policies

In early 1996, Fab 6 sought help in order to compete on both cost and cycle time. A well known consulting company generated a series of recommendations intended to increase the output while decreasing the cycle times of wafers in Fab 6. The recommendations in concert with other policies put in place by Digital management formed what will be referred to in this paper as *control mechanisms and policies*. More detail will be provided later, but these are operational rules that guide decision making in the factory. However, there were several concerns regarding the control mechanisms instituted in Fab 6.

First, the proposed control mechanisms were not generally used in the semiconductor industry. This created some anxiety and uncertainty for Fab 6 personnel since performance benchmarking with other members of the industry consortium, Sematech, became difficult. These mechanisms had not been as well studied as others with which people were familiar and there was a great deal

of concern about their ability to work. Nevertheless, the proposed mechanisms were implemented in Fab 6.

Once implemented, it was apparent that the system as designed did not work as expected. The original system imposed very strict controls on the level of in-process inventories. When one machine, in the relatively immature factory, experienced a failure, upstream machines would quickly become blocked and downstream machines would quickly run out of material to process. Several changes to the system, including making less stringent the inventory level, were required for the fab to perform at an acceptable level.

Pondering the initial failure lead to the question that is the basis for this paper: What control mechanisms provide the strongest performance for a multiple product, production-development semiconductor wafer fabrication facility?

Solution Method

In order to determine what control mechanisms would provide the best results for Fab 6, a computer simulation of the fab was built and run under a variety of different scenarios. A design of experiments approach was used to guide the selection of factory conditions to be modeled. The results of the simulation runs were then analyzed to suggest what start policy, dispatch, and kanban design would provide the best throughput and cycle time performance.

Preview of remaining chapters

The remainder of this thesis will address the question raised above. Chapter two will provide a discussion of semiconductor fabrication including the most important processing steps will be followed by a short history of Digital Equipment and the Fab 6 facility as background for subsequent chapters. Chapter three will introduce a set of theoretical frameworks including

Theory of Constraints, queuing theory, and inventory policy within which to view the operations of a semiconductor facility and will also review relevant literature in these areas. In Chapter four, the computer model used to analyze the operations of Fab 6 will be presented. The model will account for the products being made, the tool set as it exists in the facility, and the unique features that necessitated a simulation. The chapter will close with a presentation of the experiment methodology. Chapter five contains an analysis of the data generated by the Mansim model. Statistical and design of experiment tools will be used to analyze several aspects of operational performance. Finally, chapter six will contain conclusions based on the analysis and a set of recommendations.

Chapter 2 - Background

Before beginning a detailed discussion of the operation of a fab, some background may put the rest of the discussion into context. First, the process of creating a semiconductor product from concept to completed chip will be discussed. This chapter will also contain a brief history of the Digital Equipment Corporation and what events led it into semiconductor production.

Semiconductor Manufacturing

By some estimates there are 350 billion semiconductor chips in use in the world today in applications like personal computers, cars, home appliances, and thousands of others[8]. Each chip has thousands, if not millions of electrical circuits on its surface. A semiconductor chip is a many level stack of electrically conducting, partially conducting, and insulating materials built on a silicon foundation in such a way as to provide useful functions. Although there are several different technologies for the design and manufacture of semiconductors, all chips come into being through a set of fairly standard steps including design, fabrication, and packaging. These will be described with a particular emphasis on fabrication.

Design and preparation

Logic and circuit design are the first steps in chip production. Engineers, using sophisticated computer aided design software to aid and verify, organize logic gates to translate desired functions into circuit schematics. Then the actual circuits must be designed. First, physical elements like transistors, resistors, diodes, and capacitors must be specified that will carry out the desired logic. Finally the circuits must be arranged on the chip. This can be incredibly difficult since modern chips contain millions of transistors and other elements that must be arranged

carefully. It is desirable that elements be close together so the chip can be small, but when components are too close together they may interfere with one another.

Once the circuits have been designed, wafers and masks can be prepared. Most of the chips produced today are formed on silicon wafers, though other semiconducting materials like gallium arsenide are used for special applications. To produce silicon wafers, a small crystal is dipped into a vat of molten silicon laced with trace elements like boron or phosphorous for the electrical properties they provide. As the seed crystal is slowly turned and pulled from the melt, atoms bond to the growing ingot. After several hours, the ingot will have grown to several meters in length. Diamond blade saws slice the ingot into wafers less than one millimeter thick before the wafers are polished to the perfectly flat state required for fabrication.

Meanwhile, masks, the templates for the production process can be prepared. A very thin coat of chromium is deposited onto a glass plate several inches on a side. An electron beam then removes the chromium from areas where it does not belong. A mask will have to be made for each layer on the chip meaning that a dozen or more masks is required for each type of chip. Several sets of masks may be produced so each lithography tool in the fab has a mask set.

Wafer Fabrication

The fabrication of modern semiconductor chips is one of the most technologically demanding manufacturing processes known to man[9]. The circuits that have been designed get built up as lots (usually 25 wafers to a lot) of wafers experience a long sequence of four basic types of steps.

Deposition

In deposition steps, material is added to the surface of the wafer. Both metal layers and insulating layers are deposited on the surface. Most modern deposition occurs in a chamber that

has been evacuated of all air. Material can be deposited on the surface by "raining" small bits of matter onto the surface. This type of process, known as Physical Vapor Deposition (PVD) is commonly used to deposit metals like aluminum that wire all the devices to form useful circuits.

Another technique to achieve deposition is Chemical Vapor Deposition(CVD). In a CVD process, reactive gasses are introduced near the surface of the wafer. Chemical reactions take place as the gasses interact with each other and the wafer forming the desired material on the wafer surface. This process can be used to create both oxide insulators and conducting metals like tungsten.

Etch

Etch refers to a process where material is removed from the wafer. Several different physical mechanisms may cause the removal. Some etch processes submerge wafers in acidic solutions. The acid can be chosen to remove only certain material or may react with all it encounters. In other etch processes, the wafer is subjected to a bombardment of charged ions which strike the surface with such force that small bits of matter are dislodged from the surface.

Photolithography

The photolithographic steps in semiconductor manufacturing form the pattern for the other processes. A light sensitive polymer, photo resist, is applied to the surface of the wafer before the wafer enters a machine called a stepper. The stepper shines light through the mask onto a small portion of the surface of the wafer. When the light is off, the wafer "steps" or is moved so a new part of the wafer is available for exposure. In this manner the chip pattern is repeated across the entire wafer surface. Where light hits the wafer the photo resist wafer hardens and

material underneath is protected. Soft, undeveloped resist can be rinsed away and the uncovered areas now may be subjected to one of the other processes.

Diffusion & Implant

Diffusion and implant steps are used to change material properties of the silicon. In the early days of semiconductor manufacturing, large furnaces were used to diffuse impurities into the silicon to change the electrical properties of the devices. Today, impurities are introduced by machines called implanters which shoot small impurity atoms at the wafers to drive them beneath the surface. In some cases, furnaces are used to grow high purity oxide and nitride layers on the silicon.

Test & Packaging

Testing happens at several stages, depending upon a variety of factors. Usually the entire wafer is quickly tested to make sure that no critical steps were missed. Then the individual chips on the wafer are tested on wafers passing the initial test. Next, the wafers are cut up into the individual chips which are called die at this point. Good die are then placed into packages. Plastic or ceramic packaging provides protection, facilitates electrical connection to the greater system, and may aid in heat transfer during operation. Once chips have been packaged they may be tested once again and marked in preparation for shipping.

Special Challenges

There are a couple of features of semiconductor manufacturing that make it different from other types of manufacturing enterprises. First, the manufacture of chips is very capital intensive. A modern fabrication facility can cost as much as $2 billion. Expensive construction techniques and environmental control systems are required to build fabs that provide the extraordinarily

clean environment required to fabricate semiconductor products. These fabs must be filled with production equipment which is becoming more and more expensive as chip dimensions and tolerances become smaller requiring tools that are mechanically, optically, and thermally precise. Unlike in many industries, these factories and tools have a useful life of only five to ten years, meaning that depreciation of capital assets are a large portion of operating costs.

The re-entrant nature of semiconductor manufacturing also sets it apart. Re-entrance means that each wafer will visit a processing station many times. In order to fashion the layers that make up a modern microprocessor, for example, a wafer visits the photolithography processing equipment more than a dozen times. This makes analyzing and controlling the operation much more difficult than more traditional assembly line manufacture that is familiar in so many industries.

A third challenging factor is the shortness of technological supremacy. Semiconductor technology is changing so fast that a product that is state-of-the-art one day will typically be eclipsed in performance just a few months later. Anyone who has recently purchased a personal computer recently fully understands this phenomena. A leading edge product will tend be more profitable than the product it overtakes, but it is certain that an improved version is just around the corner. Therefore, it is advantageous to make a product just before a customer wants it and avoid building large inventories from which to satisfy demand.

Digital Equipment

Digital Equipment was founded in 1957 by two MIT-trained engineers, Ken Olsen and Harlan Anderson[10]. Both had worked with Dr. Jay Forrester at MIT's Lincoln Labs on the Whirlwind project which was among the world's first digital computers. Olsen was convinced of two things, that the age of interactive computing, where a user would provide input and receive

output directly from a computer instead of through a stack of punch cards, was coming, and that IBM, at the time the definitive computer industry leader, would not be the company to instigate the revolution. With $70,000 of venture capital and a stern warning not to use the use the word *computer* in their company name, the pair rented 8,000 square feet of an old mill building in Maynard, Massachusetts and began building logic modules. By 1988, Digital Equipment Corporation had grown into an $11 billion company employing 120,000 people worldwide.

Through its first three decades of existence, Digital Equipment experienced great success. The original notion, that computer users wanted more direct interaction with computers proved true, and Digital pioneered then dominated the mini-computer market. Its first computer product, the Programmed Data Processor (PDP)-1 was followed up with wildly successful products, the PDP-8 and eventually the PDP-11. These products were well designed and offered computing power at very competitive points to technically capable customers. In 1977 Digital introduced one of its last great product families - VAX and a corresponding strategy. VAX was an architecture allowing Digital to offer a large number of products at various performance and price points, but with a common interface and ability to run the same software. Almost as significant, Digital created the ability to connect VAX machines into networks, pioneering the now common computer network.

Some have credited the strength of Digital's product design for its phenomenal ascendance. In support of a large cadre of talented and dedicated engineers was a novel management structure and a motivated founder. Ken Olsen ran the company in a very hands-on fashion until he was ousted in 1992. As interested as Olsen was in the projects progressing in the company, he knew that Digital needed a structure that would provide autonomy and oversight at the same time. The

now familiar matrix approach to management provided that. Product lines were given autonomy to develop products and create markets. At the same time, the engineering and manufacturing functions that served them created a system of checks and balances. These two factors combined to give Digital a culture that few companies could ever match.

Though Digital experienced small setbacks throughout its first years, it experienced particularly hard times in the 1980s and 1990s. For one thing, Digital, like other large system vendors failed to recognize the emergence of the PC as an alternative computing platform to mainframe systems. Olsen predicted "The personal computer will fall flat on it's face in business." Nevertheless, Digital made several attempts to enter the market without success. Digital, perhaps in a reflection on its founders' beliefs, has generally been thought to have weaknesses in marketing and software. Since Olsen was replaced by Bob Palmer, a longtime Digital engineer, the company has made headlines with public failures again in the PC marketplace and with massive layoffs that have cut the company to its current employment of 59,100.

Whatever the reasons for earlier failures, Digital's current strategy is fairly clear. The company's home page on the world wide web presents the strategy: Connectivity, and clearly the Alpha microprocessor is a significant part[11]. Digital aims to sell Alpha microprocessors for use in high performance desktop machines running Microsoft's Windows NT operating system connected on networks to computers running Alpha processors controlling mammoth databases which can be stored in high speed random access memory (RAM) and serve up data through Alpha microprocessor based worldwide web servers[12].

Chapter 3 - Theoretical Framework

In attempting to understand the complex challenges of semiconductor fabrication, three frameworks can be very useful. Those frameworks which will be discussed below include Theory of Constraints, Queuing Theory, and Inventory Policy.

Theory of Constraints

In his 1984 book, *The Goal,* Eli Goldratt changed much of the thinking of industrial management by introducing a methodology that has come to be called Theory of Constraints (TOC) [13]. TOC first clarifies the primary goal of a factory - to make money. Next, TOC helps a factory identify bottlenecks, the resource in the factory with the least capacity to produce (achieve the goal). Important ideas from TOC will be expanded and related to semiconductor manufacturing below.

The Goal

Goldratt proposes that the goal of a factory is to make money now and in the future. On reflection, this is not so startling, however, he also proposes a simple accounting scheme against which the goal can be measured. In this new accounting, **T** stands for throughput, and is measured as the rate at which the factory generates money through sales. This definition does not count items produced and stockpiled as throughput. **I**, inventory, measures all the money that the factory has invested in things which it intends to or could sell. **O** for operational expense, represents all the money the system spends turning inventory into throughput.

The fictional characters in *The Goal* use these new accounting definitions to look at some familiar quantities like profit and return on investment. First, profit can be calculated as T-O, or total sales minus all costs to achieve those sales. Return on investment can be calculated as

Profit / Inventory or (T-O)/I. Obviously an enterprise wants to achieve both high profit and a high return on investment. Goldratt concludes that in order to reach the goal, a factory must increase T while simultaneously decreasing O and I.

In his Master's thesis, Menon discusses Goldratt's accounting formulation for a semiconductor fab and suggests that a factory ought to most carefully manage throughput[14]. Of the three, T, O, and I, Throughput is the quantity over which the factory has most control. In many cases simply producing more wafers and selling the chips is the easiest way of increasing throughput. In contrast, a factory does not have as much control over either inventory or operational expense because of the capital intensive nature of semiconductor manufacturing. The inventory of a fab includes the undepreciated value the clean facility and process tooling. Since modern fab facilities cost billions of dollars to construct and outfit, the capital dwarfs the amount of money tied up in other items that make up I. Because of the relatively short useful life of a fab, a significant portion of operational expense is accounted for in depreciation. Estimates are that 65% - 75% of the operational cost is related to depreciation of fixed assets and to overhead. While there are other operational costs such as salaries and raw materials, once a fab has been built and equipped the largest fraction of these costs have largely been set for the life of the fab.

There are several ways that a fab can increase its throughput measurement. It can optimize the mix of products that it produces in favor of more profitable chips. This requires strong development and marketing functions. The fab can also focus on yields so that more chips that reach the end of the process are available to be sold. Typically, yields are improved as equipment and process engineers stabilize and improve process parameters. Finally, a fab can

improve throughput by producing as many wafers on it's equipment as possible. This approach requires a focus on factory operations.

Bottlenecks

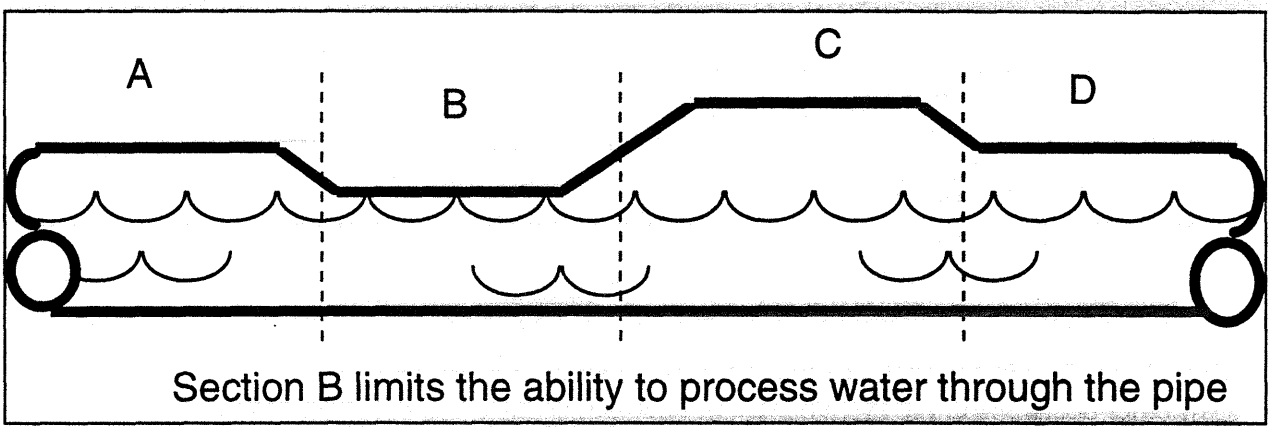

Figure 1 - Bottleneck Illustration

One other significant contribution that Goldratt has made is the acknowledgment that a production system can produce only to the rate of its slowest resource. This concept is often illustrated by considering a pipe like that in Figure 1. The pipe is narrowest at section B, and it is this section that dictates the maximum amount of fluid that can exit at the end.

Goldratt uses the term bottleneck to refer to any resource that keeps the system from achieving the goal. As important as acknowledging the bottleneck, Goldratt gives a methodology for coping with the bottleneck. His five step process is as follows:

1) Identify the bottleneck.

   In simple cases the bottleneck resource will be obvious because it has a slower processing rate than other resources. In cases where it is not readily obvious, the bottleneck may present itself because materials needing processing pile up in front of it.

2) Exploit the bottleneck.

The output of the entire system is being dictated by the output of the bottleneck. Achieving maximum system output requires that the bottleneck resource be used to its fullest extent. Any time that the bottleneck is idle or down, the system output will decrease. Similarly, any time that the bottleneck is idle waiting for materials to process, productive capacity is being squandered.

3) Subordinate everything else to the above.

All other decisions in the system must not get in the way of keeping the bottleneck busy. For example, lunch breaks might be scheduled so that there is always a crew attending to keeping the bottleneck tool working. Similarly, work might be scheduled in such a fashion to minimize the amount of time that a bottleneck tool spends setting up.

4) Elevate the system's constraint.

Look for ways to increase the output capability of the bottleneck. This might include speeding up a processing rate or using other equipment to perform operations that are currently being done on the bottleneck resource.

5) Return to step 1.

Queuing Theory

Queuing theory is another field that provides useful framework for understanding some of the workings of a semiconductor fabrication facility. Two results will be particularly relevant. First, we will arrive at a relationship between utilization and cycle time and then a well known relationship between cycle time and inventory.

Utilization - Cycle Time

Elementary queuing analysis consists of calculating performance measures of a server. Bank ATMs, interstate highway tool booths, and fast food cash registers are often modeled as servers in queuing systems. Items (customers) arrive at the server according to some stochastic process at a rate λ units/time. The server processes each item in M amount of time, where M is also a stochastic quantity. Alternatively, $M = 1/\mu$, where μ is the service rate. When a system's arrival and service rates are known, the system utilization can be calculated. Utilization, ρ, which measures the fraction of time that a server spends providing service is simply, $\rho = \lambda \times M$. With these three we can calculate some system performance measures, like how many items are in line waiting to be served or how long a part will wait in line to be served.

One very useful relationship that queuing theory provides is a relationship between W, the length of time an item will spend in the system (waiting time + processing time) and the utilization of the server. The relationship $W = \left(\frac{\rho}{2(1-\rho)} + 1\right)M$ holds when arrivals are perfectly random (Poisson distributed) and the server takes exactly M time to process each item[15]. This relationship, depicted graphically in Figure 2, is such that as the server becomes more utilized, the cycle time grows in a non-linear manner.

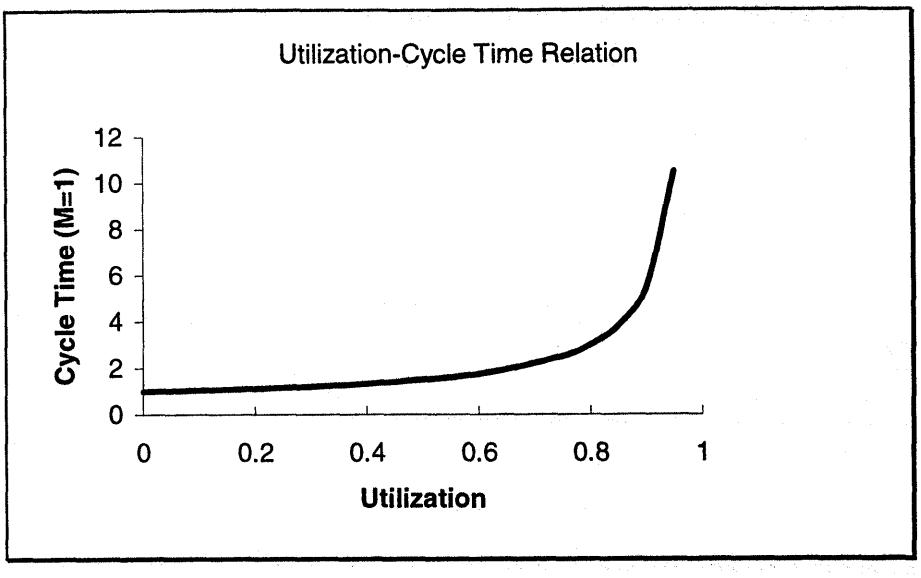

Figure 2 -Utilization - Cycle Time

Variability

The relationship just derived can be further developed. If the requirement that arrivals be Poisson and that service times be constant is relaxed, we gain insight into the effects of variability of both the arrival and service processes. Consider an arrival process with a mean arrival rate of λ and a variance of σ_a^2. Also, the server services items in a mean time of M and a variance of σ_s^2. The new expression for the amount of time a part will spend in the system becomes:

$$W = \left[\frac{\rho}{2(1-\rho)} \times \left((\lambda\sigma_a)^2 + (\mu\sigma_s)^2\right) + 1\right] M$$

This expression tells us that greater variability in either the arrival process or processing time will increase cycle time at any utilization level. Figure 3 below shows the cycle time vs. utilization curve for different levels of variation.

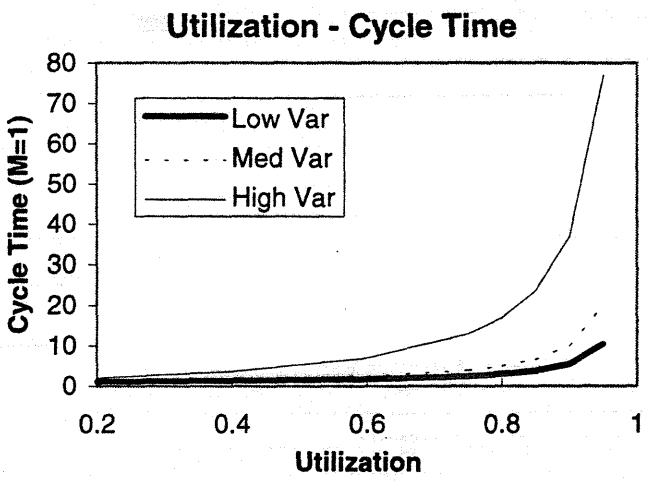

Figure 3 - Increasing Variability Increases Cycle Time

Little's Law

To this point, queuing analysis has provided a relationship between cycle time and utilization. John Little of the Sloan School of Management used queuing to publish an important result relating cycle time to inventory which has borne his name since. He was able to show that $L=\lambda*W$ where L is the number of items in a system, λ is the mean arrival rate, and W is the total amount of time spent in the system. He was able to demonstrate that the relationship holds regardless the arrival distribution.

This result is frequently used in manufacturing because it indicates that the amount of inventory in a factory is related to the cycle time of the product through the factory.

Inventory Policy

Inventory policy provides a framework for considering how a firm should hold the minimum possible inventory while satisfying customer demand. One very common model for such a system can be seen below in Figure 4.

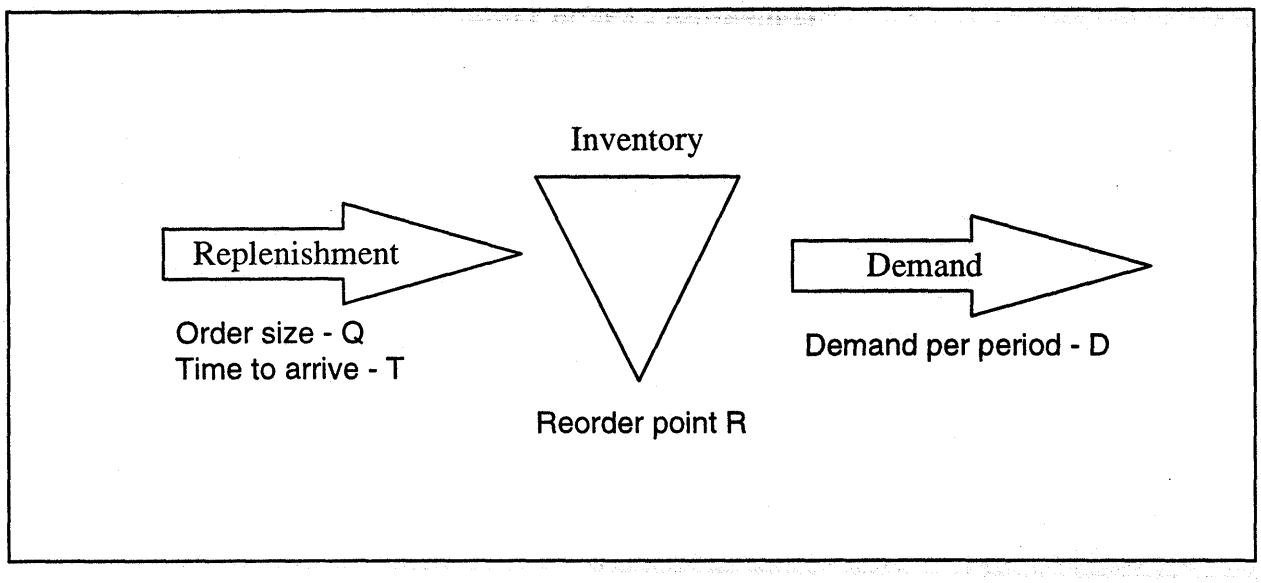

Figure 4 - Order Point - Order Quantity Model

Figure 4 demonstrates a Order Point - Order Quantity model. Customers order D items each period which are shipped from the inventory. Each period, the inventory level is checked and when inventory dips below R units an order is placed. Every order placed is for Q units and the order will arrive T periods after it is placed, since the model assumes zero order placement and communication time. The objective of such a system is to fill demand without holding an exorbitant amount of inventory. D and T are imposed leaving management to choose Q and R. Q is often calculated using the economic order quantity (EOQ) formula which in principle is a cost minimizing quantity and R is chosen to assure a certain level of customer service.

In the case that D and T are deterministic, if $R > DT$, all customer orders can be fulfilled. A graph of inventory over time will have a saw tooth shape. In practice, however, both the demand rate, D, and the replenishment time, T, are stochastic. Nahmias expands on the simple model described above to account for the variability[16]. Once Q has been determined by other means, an R must be calculated that supports the service level the system wishes to provide. For the first part of the formulation, we will assume that the replenishment time is deterministic.

Consider an OP-OQ system that has reached the re-order level, R. Before the order arrives, the system will experience T periods each with a mean demand of D and demand variance of σ_d^2. Therefore over the replenishment epoch the system will experience TD mean demand with a demand variance of $T\sigma_d^2$.

With this observation it is possible to make an intelligent assessment of R. The order point should certainly be chosen to be greater than the average demand over the replenishment time. But it is also reasonable to select R to protect against some of the variability in demand over the replenishment lead time. Begin by choosing a service level, the percentage of orders that should be filled from inventory during the replenishment time (back orders are not allowed). Consult a standard normal table and convert the service level into z, the number of standard deviations of demand necessary to hold to insure the service level. Now calculate R as follows:

$$R = DT + z\sigma_d\sqrt{T}.$$

Having developed a first assessment for R it is possible to relax the requirement that the replenishment time is fixed. Allow T to be represented by a distribution of replenishment time with mean T and variance σ_t^2. Now the demand over a replacement period can be represented by a distribution with mean DT and variance $T\sigma_d^2 + D^2\sigma_t^2$. The order point can be calculated using the reasoning as before:

$$R = DT + z\sqrt{[(T\sigma_d^2 + D^2\sigma_t^2)]}$$

The key result that will be useful later is that R is an increasing function of σ_t, the lead time variability. That is, more inventory must be held to protect against greater uncertainty in replenishment time. This conclusion is confirmed by Kumar and Arora who find that service

levels in a system where R is set without considering the variability of replenishment time experience service levels lower than would expected[17].

Synthesis

Now that the three frameworks have been discussed, it is possible to show how they relate to the manufacture of semiconductors.

Theory of Constraints thinking says that the goal of our fab is to make money. Intense capital requirements drive the fab to strive for maximum throughput to achieve the goal. Within the fab exists some piece of processing equipment that can produce less than all others and the output of this resource constrains the output of the total factory. Once we have identified this bottleneck and performed the five steps for upgrading its capabilities, the simplest thing to do is keep it as busy as possible.

However, queuing theory predicts that as utilization of a server increases, the amount of time that an item spends in queue and in process increases non-linearly. Therefore, we should expect to see a line of wafers in front of the bottleneck toolset in a fab. Furthermore, semiconductor equipment is so expensive that many fabs run several of their toolsets at nearly full utilization.

A semiconductor fab can be considered as a very large system of queues and servers. Chen demonstrates that modeling a fab as such gives results which are quite close to those actually observed in a real factory[18]. When we combine the fact that several tool sets are run at high levels of utilization with queuing knowledge, we find that as utilization levels increase, cycle time at each station will increase.

This increase in cycle time has several consequences. First, Little's law predicts that for a given start rate, a longer cycle time implies a larger in-process inventory. While throughput is the quantity of most concern to the fab, an increase in inventory is undesirable. But an increased cycle time may also affect T directly. The longer a wafer spends in the fab, the more chance it has of being contaminated by damaging particulate material. This will affect the line and probe yields. While such a cycle time - yield relationship has never been conclusively proven, it is the subject of several investigations [19,20].

Finally, fabs supply most customer demand from a finished goods inventory. The more variable cycle time is through the fab, the more inventory must be kept, which is undesirable. With the rapid pace of technological innovation, many chips are rendered obsolete very quickly. Therefore, it is always preferable to fewer finished chips as long as current demand can be adequately filled.

The operational challenge facing the fab boils down to the following:

Operate the factory in such a manner that keeps the bottleneck resource busy as much as possible while keeping cycle times through the factory short and predictable.

Control Policies and Mechanisms

Control policies and mechanisms are an answer to the challenge presented above. They are the rules and systems that guide the day to day operations within a fab. Two basic types of rules have been studied and are commonly used. The first type of rule guides when a lot of wafers is allowed to begin. The second type guides how wafers travel through the manufacturing process. A control mechanism, for the purpose of this paper is a system, more complex than just a set of

rules, that guides the internal movement of lots in the factory. Both policies and mechanisms will be discussed along with published results pertinent to each.

Start Policies

Lot start policies dictate conditions under which a new lot of wafers will be allowed to begin their journey through the production process. Several different methods are available and include random starts, uniform starts, Workload Regulating starts, Conwip starts, Starvation Avoidance, and others. Random and uniform start policies fall into open loop policies, that is lots are started without regard for the state of the factory or any other mitigating factor. Workload regulating, Conwip, and Starvation Avoidance policies are closed loop policies which take into account certain factors before allowing wafers to be started into the process.

The open loop policies are fairly simple to understand and implement. Under random starts, lots are released into the factory with no predetermined pattern. It might be that at the beginning of each hour, a coin flip determines whether or not to start another lot. In this simple case, the inter lot release times follow a binomial distribution and on average one lot is started every two hours. In practice, factories do not intentionally use this policy, but it is useful to consider the performance of other start policies in relation to this one. One simple and commonly used open loop policy is the uniform start policy. According to this policy, one lot is started every H hours.

Closed loop start policies, on the other hand, assess the state of the factory and allow a new lot to start only if the factory is in one of several allowable states. In his paper on control policies, Glassey argues that any closed loop policy should outperform an open loop policy[21]. Just as it would be difficult to drive a car without a speedometer, it is not easy to run a factory without some feedback. However, different policies exert control based on different measures.

The Conwip policy tabulates C, the number of lots already in the factory. If C is below some predetermined level, a new lot is allowed to start. In this way, the total work in process in the factory can be held constant. At some point it should be that as any lot exits the factory, a new lot is allowed to begin. The name Conwip comes from the nature of this policy to maintain a constant work in process.

Wein proposes an alternate start policy termed Workload Regulating which is one way to implement the drum-buffer-rope concepts of synchronous manufacturing in a re-entrant environment[22]. Under Workload Regulating, the total number of hours the bottleneck tool would have to work to process all the lots in the factory is calculated. A new lot is allowed to begin when this falls below some appropriate level. A workload parameter can be found which allows enough material in the factory to keep the bottleneck busy but that will not put so many lots in that wafers simply sit in queues waiting for processing. Wein demonstrates that Workload Regulating reduces cycle time by 35% over random starts and that start policy is more important to fabrication cycle time than dispatch policy. The Starvation Avoidance policy is similar to Workload Regulating in that new lots are started in such a fashion as to not allow the bottleneck tool to ever run out of material to process.

Dispatch Policies

Dispatch policies are one type that control the progress of material through steps of the process. When several items are in a queue waiting for service, dispatch policies assist the server to decide which lot to serve next.

A myriad of dispatch policies exist. Chang, Sueyoshi, and Sullivan discuss and compare the performance of nearly forty different dispatch policies in a job shop[23]. Their list of rules was

derived from the list of 113 compiled and categorized by Panwalker and Ishkander[24]. The choice of which lot to serve next can be determined by a simple priority organization. It may be that lots can be prioritized by processing time, due date, setup cost and time required, or order of arrival. More complex dispatch policies combine two or more of these simple criteria or may be based on entirely different factors.

Both Wein and Kumar have studied the effects that dispatch policies have on semiconductor fabrication cycle time. Wein tested several different dispatch policies but determined that start policy affects cycle time more than dispatch policy.

Kumar develops a set of dispatch policies that are entitled Least Slack policies[25]. These policies combine information about when an item is due and how much processing remains to be done. There are policies in the set that are specifically designed to reduce the variation of cycle time and reduce the mean cycle time. Kumar tests these policies in a semiconductor facility and concludes that these policies indeed perform their designed function.

Control Mechanism

Various control mechanisms have been an important part of manufacturing systems for the last couple decades. Perhaps the most familiar is the kanban type. Kanban, literally translated from Japanese, means visible record. A card is usually transferred from one production center to another as a signal for the receiving center to produce and deliver product to the sending center[26].

In many modern production environments, the term kanban has been used more generally to describe pull-based manufacturing systems. However, for a variety of reasons including the process complexity and length, kanban systems have generally not been adopted in

semiconductor facilities. In fact, Bonvik models a semiconductor facility and suggests that a kanban system provides less than optimal performance[27].

Nevertheless, it is possible to implement certain pull-based control systems in a semiconductor fab. The re-entrant nature, that a wafer will visit the same machine multiple times, allows for two different alignments that have been termed Flow Based Control Mechanism (FBCM) and Tool Based Control Mechanism (TBCM). These terms are somewhat the invention of this author who has chosen to use them rather than referring to the systems as "kanban". Neither of these systems fit the strict definition of a kanban system, and therefore both will be referred to as control mechanisms.

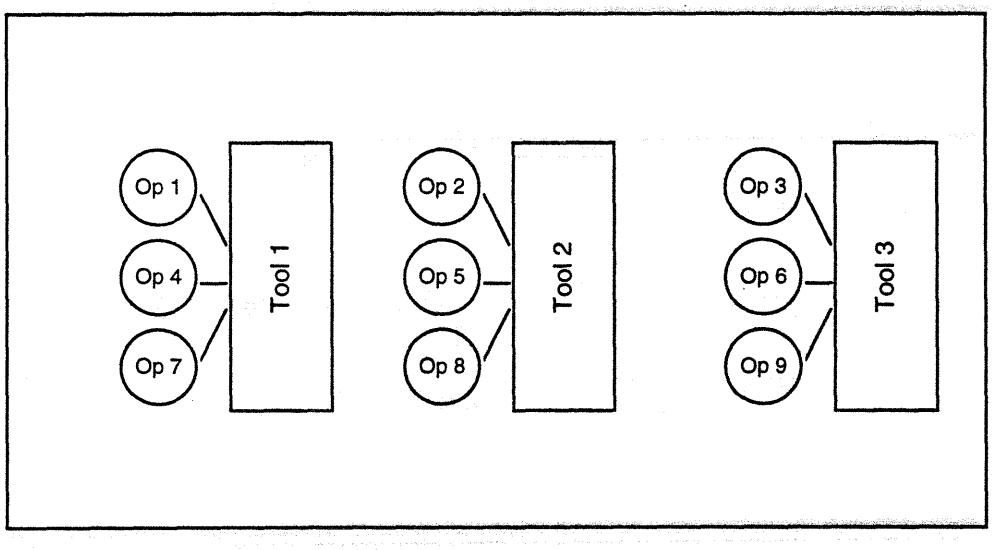

Figure 5 - Schematic Fab process flow

In Figure 5, a very simple wafer fab has been represented. This fab has three tools and products complete nine operations. These operations must be completed in order. Assume that the factory is heavily loaded so that there are several lots of wafers waiting to be processed at each operation and that each tool is capable of operating on only one lot at a time.

Flow based

A flow based control mechanism divides the total production sequence into a series of control zones and establishes an upper limit on the material allowed in each zone. For example, operations 1, 2, and 3 might be joined to form a control zone(zone A). Similarly, operations 4, 5, and 6 could be joined into a control zone (zone B) as could operations 7, 8, and 9 (zone C). Any lot waiting for or being processed at operation 1, 2, or 3 is considered in zone A. Next, a limit to the number of lots of wafers allowed in each zone will be set. It will be demonstrated later how to use Little's law to calculate control parameters. Finally, the control zones interact as the factory is in operation. A lot in zone B completes operations 4 and 5 when tools 1 and 2 give it priority. However, the lot will sit in front of tool 3 waiting to complete operation 6 until the following two conditions are met: Zone 3 signals it is needs the lot (by having fewer lots in the zone than the control parameter allows) and the lot has high enough priority in the dispatching system of tool 3 to be processed.

For a flow based control mechanism to operate properly, every operation must be included in exactly one control zone and every control parameter must be one or greater. However, there are a couple of extreme cases of flow based mechanisms to consider. First, if each operation constitutes its own control zone, this FBCM be the familiar kanban type of control. Next, if all operations are in one zone we would find a case where the total in process inventory of the factory has a pre-set limit. This is similar to the Conwip described earlier.

Tool based

In a re-entrant system like that in Figure 5, it is also possible to create what will be called a tool based control mechanism. Suppose that all operations performed on Tool 1 are grouped together. The operations performed on Tool 2 form a second group, and those of Tool 3 form a third. If a

maximum limit to the number of lots allowed in each group is set, a tool based control mechanism exists. In this case, a lot of wafers will be allowed to process on Tool 1 only if the number of lots waiting and processing at Tool 2 is less than the maximum allowed. The TBCM either allows or disallows processing while the dispatch policy determines which lot to process.

Why might a tool based control mechanism be preferable to a flow based mechanism? Suppose Tool 3 fails for a long period of time under a FBCM. If Tool 2 continues to process material in operations 2, 5, and 8, a significant amount of material will be waiting in front of Tool 3. Even when Tool 3 returns to operation many lots will be waiting for a long time before the tool catches back up to the rest of the tools. However in a TBCM, control parameters are set on the tools themselves and this situation cannot occur. However, there is an analogous peril contained in a TBCM. It could happen that the entire buffer at Tool 3 contains material waiting to be processed at operation 3. This mass of material moves through the system in a wave and is sometimes described as a "pig moving through a snake." A FBCM disallows this situation by setting limits on inventory in process sequence.

A reasonable fab manager or industrial engineer might want to know which of these systems perform better on the key parameters of throughput and cycle time. A set of tools based on Markovian analysis has been developed to analyze transfer lines and yield closed form solutions to their performance[28]. However, these tools are not adequate to cope with the complexity of re-entrance. Current theory is inadequate to answer these questions and simulation is the only known method for determining performance characteristics of re-entrant systems.

Chapter 4 - Fab Model

Digital Equipment's Fab 6 faced operational challenges in 1996. In order to be profitable, the fab needed to produce a large number of chips. Moreover, the fab was producing several different types of chips and preparing to make still more types. It was therefore necessary to reduce the cycle time of wafers through the fab. Additionally, a reduced variation in cycle time was desired as this would allow a smaller finished goods inventory. At this time, Digital began investigating whether it was using the most appropriate start policy, dispatch policy, and control mechanism for its situation.

Overview

A simulation model was conceived and built in order to understand the performance characteristics of the control policies and mechanisms chosen for Fab 6 and to suggest others that might be better. This model accounted for unique features of Digital's fab including the several products that were built in the factory, the flow based control mechanism and start and dispatch policies that had been installed, the tool set used in the fab, and a special allowance for experimentation that the factory needed in its development role. The major source of uncertainty in the model comes from the tool repair and failure, personnel policy was not considered. The model was built in a discrete event simulation software package Mansim/X v3.8. A short description of the products, tool set, and experimental method are described below.

It is appropriate to mention at this point that the model describes a predicted future state of the factory. The product requirements and tool set were drawn from business and technical plans that were based on a consensus estimate of what was in store for the factory approximately 6-12 months from the time of the experiment. Given that the simulation is intended to compare policy

implications rather than perfectly predict factory performance, precision in input parameters is not absolutely required. The purpose for making this statement is to highlight that although simulation results are being presented without disguising mechanisms, it should be clear to the reader that no Digital proprietary data is being divulged.

Products

The model accounts for the four most common process routings used in the fab. It is clear from Table 1 below that these products are operationally very different as well as technically different. Previous published studies mentioned earlier have tended to model only single product fabs. It was decided that a multi product fab simulation would be more appropriate given the situation.

Name	Volume	Process Steps	Description
C5L	43%	302	High volume, well understood product, simple technically
C5WF	9%	507	Very unusual process, optimized for certain performance characteristics
C64	25%	384	Replacement process for C5L
C66	23%	496	Newest, more complex version of C64

Table 1- Fab 6 Products

Tools

The tools used in Fab 6 were represented in the Mansim model. The table below shows the important characteristics that were modeled and clarifying discussion is below.

Process Tools

General Tool Information						Product Information (Mean Process Time in hours)								Performance Information (% time spent)			
Tool	# Machines	Mode	Batch size	MTBF (hours)	MTTR (hours)	C5L Visits	C5L MPT	C5WF Visits	C5WF MPT	C64 Visits	C64 MPT	C66 Visits	C66 MPT	Down	PM	Busy	Total
A-SINK.D	2	BS		230	10	11		12		12		12		4%	0%	#DIV/0!	#DIV/0!
A-SINK.F	2	BS		135	5	9		11		11		11		4%	0%	#DIV/0!	#DIV/0!
A-SINK.I	1	BS		69	6	2		3		3		3		8%	3%	#DIV/0!	#DIV/0!
ANNEAL.A	2	B	4	545	5	3	3.00	7	3.00	5	3.00	7	3.00	1%	5%	48%	54%
AL-PVD	3	S	1	60	7	4	1.15	6	1.15	5	1.15	7	1.15	10%	13%	53%	76%
ANN-RTP	2	S	1	175	7			2	1.85	1	1.85	1	1.85	4%	3%	16%	23%
CLN-OX	6	B	4	500	8	7	5.55	5	5.90	5	5.90	5	5.90	2%	5%	38%	45%
CMP.A	2	S	1	45	3	1	2.63	1	2.63	1	2.63	1	2.63	6%	11%	36%	53%
CMP.B	3	S	1	42	5	2	2.25	5.00	2.25	3	2.25	5	2.25	11%	9%	65%	84%
CMP-SCRUB	2	S	1	250	5	3	0.67	6	0.67	4	0.67	6	0.67	2%	5%	38%	45%
DEP-A	2	S	1	69	10	2	1.20	5	1.20	3	1.20	5	1.20	13%	13%	52%	77%
DEP-B	2	S	1	70	6	2	2.60	2	4.00	1	4.00	1	4.00	8%	8%	67%	83%
DEP-E	4	S	1	90	9	7	1.43	12	1.00	8	1.00	12	1.00	9%	9%	69%	88%
DRY-STRIP.A	2	S	1	300	9	7	0.52	9	0.52	9	0.52	9	0.52	3%	3%	57%	63%
DRY-STRIP.B	3	S	1	300	4	9	0.60	21	0.60	14	0.60	20	0.60	1%	3%	75%	78%
DUV	3	S	1	70	7	4	1.42	6	1.42	4	1.42	6	1.42	9%	8%	60%	77%
ETCH-A	2	S	1	80	3	3	1.46	3	1.46	3	1.46	3	1.46	4%	10%	60%	74%
ETCH-D	4	S	1	80	10	2	2.00	5	2.00	3	2.00	5	2.00	11%	7%	43%	62%
ETCH-E	2	S	1	300	3	2	1.50	2	1.50	5	1.50	7	1.50	1%	7%	78%	86%
GATE-OX	2	B	4	176	9	2	5.08	3	5.06	3	5.06	3	5.06	5%	8%	44%	57%
HIMPLANT	2	B	2	100	8	2	2.58	3	2.06	3	2.06	3	2.06	7%	7%	39%	54%
LITHO	7	S	1	100	8	12	1.20	24	1.20	19	1.20	23	1.20	7%	5%	80%	93%
METALETCH	4	S	1	110	4	3	1.72	12	1.72	5	1.72	7	1.72	4%	9%	60%	73%
MIMPLANT	2	B	2	40	3	7	0.93	7	0.93	7	0.93	7	0.93	7%	10%	44%	62%
NITRIDE	2	B	4	200	20	1	6.50	2	6.34	2	6.34	2	6.34	9%	13%	34%	55%
POLYETCH	2	S	1	250	7	2	1.50	2	1.50	2	1.50	2	1.50	3%	7%	41%	51%
POLYTUBE	2	B	4	475	25	2	4.25	2	4.25	2	4.25	2	4.25	5%	13%	29%	47%
SAL-RTP	2	S	1	240	8	2	1.85	2	1.85	2	1.85	2	1.85	3%	5%	51%	59%
S-SINK.A	2	BS		60	20	17		45		26		38		25%	0%	#DIV/0!	#DIV/0!
TINPVD	2	S	1	45	5	3	1.00	6	1.00	4	1.00	6	1.00	10%	9%	57%	76%
UVOVEN	2	B	1	100	5	2	1.67	2	1.67	2	1.67	2	1.67	5%	5%	46%	56%
WCVD	2	S	1	65	10	3	1.00	6	1.00	4	1.00	6	1.00	13%	7%	57%	77%
WETCH	3	S	1	150	6	3	1.50	6	1.50	4	1.50	6	1.50	4%	11%	57%	72%

Metrology Tools

General Tool Information						Process Information (Mean Process Time in hours)								Performance Information (% time spent)			
Tool	# Machines	Mode	Batch size	MTBF (hours)	MTTR (hours)	C5L Visits	C5L MPT	C5WF Visits	C5WF MPT	C64 Visits	C64 MPT	C66 Visits	C66 MPT	Down	PM	Busy	Total
AUTOPROBE	4	G	1	20.7	2	1	5.0	1	5.0	1	5.0	1	5.0	9%		34%	43%
CD-SEM.A	2	S	1	60	6	10	0.3	14	0.3	12	0.3	14	0.3	9%		48%	57%
CD-SEM.B	1	S	1	60	7	7	0.3	17	0.3	11	0.3	17	0.3	10%		90%	101%
DEF-INSP.A	2	G	1	200	2	2	0.5	2	0.5	2	0.5	2	0.5	1%		15%	16%
ELLIPSE.A	2	S	1	38	2	12	0.3	13	0.3	13	0.3	13	0.3	5%		53%	58%
IMP-DOSE	2	S	1	200	15	6	0.3	7	0.3	7	0.3	7	0.3	7%		27%	34%
M-THICK	2	G	1	240	5	8	0.4	15	0.4	11	0.4	15	0.4	2%		58%	60%
OPT-INSP.A	4	G	1	250	2	22	0.3	43	0.3	35	0.3	47	0.3	1%		74%	75%
OPT-INSP.B	5	G	1	250	2	4	0.4	4	0.4	4	0.4	4	0.4	1%		9%	10%
OVERLAY	3	G	1	140	3	12	0.4	20	0.4	16	0.4	20	0.4	2%		52%	55%
PAT-PART.A	4	G	1	38	2	15	0.5	26	0.5	19	0.5	26	0.5	5%		62%	67%
PROFILER	2	G	1	800	16	6	0.2	12	0.2	8	0.2	12	0.2	2%		19%	21%
THICKNESS.A	4	G	1	724	9	18	0.4	31	0.5	20	0.5	30	0.5	1%		73%	74%
THICKNESS.B	2	G	1	540	9	6	1.0	12	1.0	8	1.0	12	1.0	2%		108%	110%
UPAT-PART.A	7	G	1	38	2	28	0.4	40	0.3	30	0.3	38	0.3	5%		44%	49%
UPAT-PART.B	2	G	1	38	2	2	0.5	2	0.5	2	0.5	2	0.5	5%		14%	19%

Table 2 - Fab 6 Tools

Tool Classes

Two classes of tools have been modeled, process and metrology. Process tools are the ones that perform the basic etch, deposition, photo, and furnace steps discussed earlier. Metrology tools perform a wide variety of measurement tasks on the wafers; measuring layer thickness, layer

uniformity levels, stress levels in deposited layers, the accuracy with which one layer lines up with another, and numbers of surface defects and contamination. A significant portion of the total steps in each product are metrology steps. This reflects the fact that Fab 6 is not strictly a production fab, that processes are being developed and refined all the time.

Processing Modes

Different tools process lots in different ways and Mansim permits some of the differences to be modeled. Serial and General Purpose tools process one lot of wafers at a time. Several very complex tools in the fab were modeled as an equivalent set of Serial type tools. In contrast, Batch tools perform a process on several batches of wafers at the same time. Very often, long processing time tools like diffusion furnaces are operated in a batch mode. A Batch Sequential tool operates on more than one lot of wafers at a time and it has the possibility of simultaneously processing several different batches of lots. Many of the etch sinks can best be modeled as batch sequential tools. Two lots of wafers are loaded into the first acid bath where they sit for 10 minutes. Then they are carried to a rinse tank where they might sit for another 12 minutes. In the meantime, two new lots of wafers can be put into the first acid bath.

Modeling conventions

All tools have been modeled to have exponentially distributed failure and repair times. This is a fairly common practice in simulation models and appeared justified after reviewing failure data generated by tools in the fab. The parameter values were gathered from the tools already in use in the fab. It was assumed that all identical tools in a set had the same failure and repair performance.

The model also attempted to account for time that process tools spend in a state other than "up" or "down." Because of the development activity that is ongoing in Fab 6, there are many times when engineers will request time on a tool to explore certain settings or conditions or to run an experimental process. There are also times when maintenance personnel perform preventative maintenance checks on the equipment. Records had been kept detailing how much time had been allocated for such activity. In an attempt to account for such incidents each tool was subjected to three Preventative Maintenance (PM) incidents per week. All time spent in these other states was modeled as preventative maintenance even if other types of activities were being performed. Discussions with industrial engineers suggested that modeling PM activity as normally distributed was more appropriate than exponential.

The four rightmost columns on the table summarize performance information of each tool in the model. Failure and repair information allow calculation of the amount of time a tool will spend in a down condition - unable to process wafers. The amount of time spent in PM state is listed in the next column. With knowledge of processing times and number of process steps the amount of time a tool will spend processing is calculated in the third column. The final column sums the previous three. This summation might be considered utilization - the portion of time that the tools is doing something.

There are two items to notice about this final column. First, it is difficult to calculate processing time for batch sequential tools because there are complex sets of rules that dictate what types of batches can be run together and these rules do not easily lend themselves to codification. More

importantly, notice that most tool sets have utilization levels in the 65% to 85%[1] range. The most utilized tool is LITHO which is a set of photolithography equipment. This was considered the fab bottleneck for the simulation.

Experiment

Once the fab had been modeled in software, two factorial experiments were run to ascertain the performance implications of each policy set. Figure 6 displays the entire experiment in graphical format. Fortunately, the Mansim software had built-in functions sufficiently powerful enough to handle all the experimental points necessary and no external programming was required.

1 A glance at the table suggests that CD-SEM.B and THICKNESS.B are utilized more than 100%, which is impossible. However CD-SEM.A and CD-SEM.B tools are identical and the true utilization is the weighted average (2 × 57% + 101%)÷3 = 72%. The same logic applies to THICKNESS.A and THICKNESS.B.

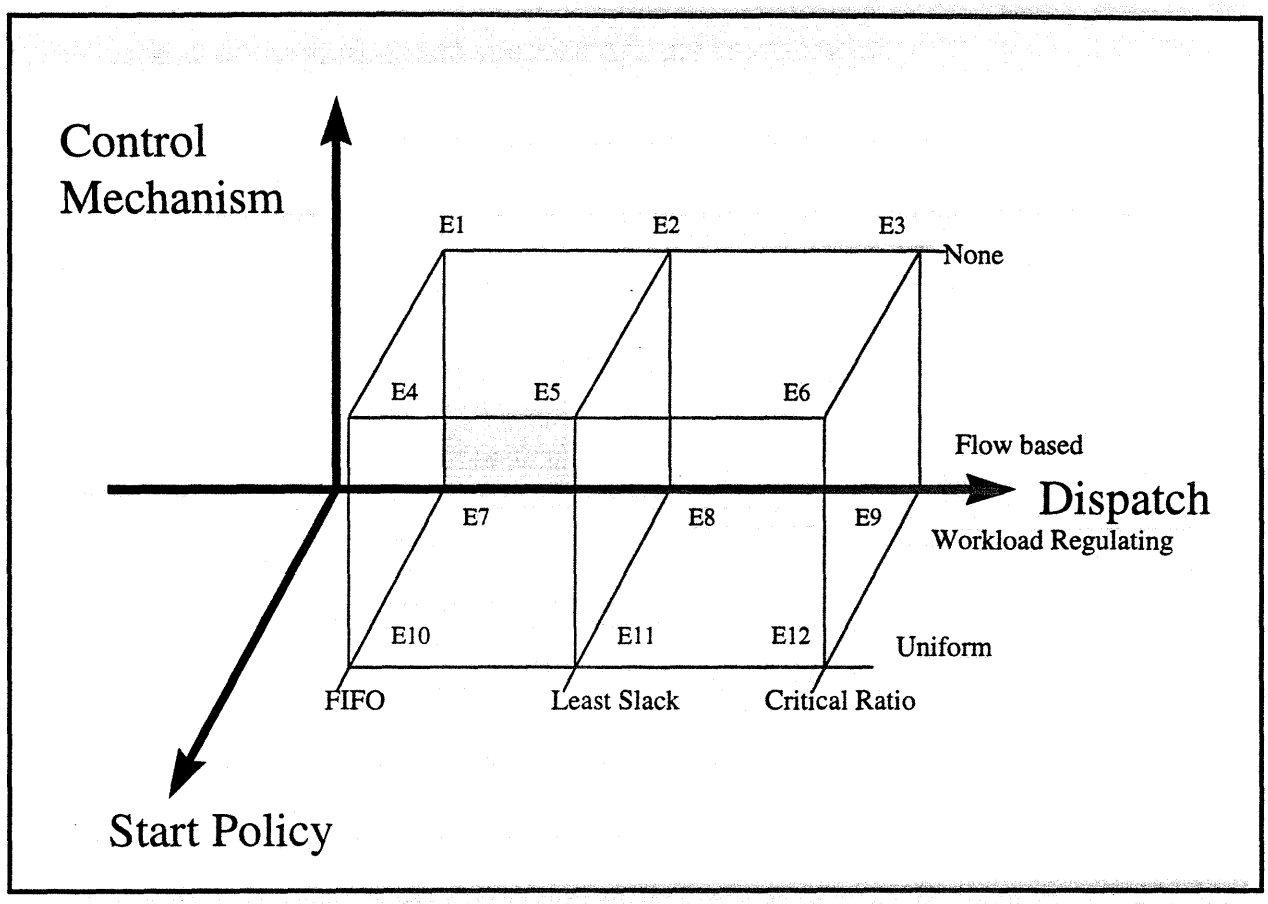

Figure 6 - Graphical Representation of Experiment

Method

For each of the 12 experiment points, 15 replications were run. It was determined that 15 replications were sufficient to provide measures of statistical confidence.

Several input factors at each of the 12 experiment points were manipulated so that the output of each factory scenario had statistically indistinguishable throughput. If the output of each scenario is the same, that requires that the tools be equally busy in each scenario. The variables available for consideration become cycle time and standard deviation of cycle time. The analysis of these variables will be done in Chapter 5 and allows us to consider whether these policies and mechanisms perform their intended functions.

Each simulation run began with the factory completely empty and a period of one year was simulated. It was readily observable that the factory would reach a steady state, where the level of in process material stabilized, after a simulated two month period, in most cases. However, the simulation ran for three more simulated months before data gathering began. Statistics were collected on only the last 60% of the year, or a simulated 7.2 month period.

The Mansim software accepted a set of seed values that would generate the patterns of random failures and repairs of tools in the factory. Fifteen sets of five digit prime numbers was generated and used to seed the randomization routine. The replications in all 12 experimental points used the same seed sets.

Factors

Lot Starts

Both uniform and Workload Regulating start policies were considered. The factory was currently planning to operate under Uniform start but wanted to consider the alternate.

Under uniform starts, the simulation was simply fed the desired input rate as a parameter. Under the assumption that a steady state condition would be achieved, the desired output rate should equal the input rate.

In order to use Workload Regulating start policy a workload parameter was needed. The parameter is the number of hours of work on the bottleneck tool allowed to be in the factory. When the workload in the factory drops below the parameter, a new lot is allowed to start in the factory.

Below is a sample of how the parameter was determined.

Bottleneck Tool: Mean processing time 1 hour/visit, 10 visits per lot

Other processing required: 90 hours

Target Cycle time :200 hours, target output rate = .1 lot/hour

By Little's Law: $L=\lambda*W$ or $L=.1*200 = 20$ lots in the fab

Assume: Work is spread evenly, i.e. 2 lots on each layer

Workload Parameter(est.) = 2 lots * 1hour/lot/visit * sum(10,9,8,7,6,5,4,3,2,1) = 110

In some cases, the calculated workload parameter was too small to allow adequate throughput and had to be adjusted upward.

Control Mechanisms

The original intent of this project was to compare cycle time performance of fabs with flow based control mechanisms and tool based mechanisms to a fab with no such mechanism. However, results of simulation runs with tool based mechanisms failed to yield acceptable results. After considerable unfruitful efforts to understand the problem, it was decided to proceed and omit this analysis. Fab 6 was operating with a flow based mechanism and it was determined that the comparison between flow based and no control mechanism would suffice.

The design of the flow based mechanism required breaking the sequence of operations into control zones and then sizing the acceptable in-process inventory level for each zone based on the time required to complete the operations in the zone and the types of tools in each zone.

Since the factory was already using a flow based mechanism, modeling the zones was straightforward. Other than minor differences arising in unusual situations, the set of control zones defined in the Mansim model match the control zones already in use in Fab 6.

One difference exists between the system as implemented in Fab 6 and what Mansim allowed. In the factory all products shared the same zones. That is, a zone dubbed Metal1 was created and allowed to have no more than 8 lots in it regardless the product mix. However, Mansim required separate zones for each product. The table below shows the number of control zones by product.

Product	Process Steps	Control Zones	Steps/Zone (average)
C5L	302	32	9.4
C5WF	507	45	11.3
C64	384	40	9.6
C66	496	50	9.9

Table 3- Fab 6 FBCM Control Zones

Once the zones were created, analysis was required to determine how to size each zone based on time spent in the zone and other factors dictated by the type of tools used to perform the operations in the zone. A simulation run was performed with no control mechanism present which returned the amount of time spent waiting for and performing each operation. Little's law was used to then calculate an appropriate number of lots to allow in each zone. If a lot spent 40 hours waiting for and performing the operations in a zone and a desired output rate was .1 lot/hour, 4 lots would be set as the limit. However, some zones contained batch processing equipment and the number of allowed lots was set to min(calculated size, batch size).

Once the control system was modeled in Mansim, the calculated number of lots allowed in each zone was verified. An industrial engineer developed a prediction of how many "tags" would be required when the factory was fully loaded. The number predicted was quite close to what was modeled in Mansim. Differences existed mainly because of the inability to model the system as

it actually exists. For example, the IE model might assign 10 tags to Furnace1 zone where one of the batch tools was. However to model in Mansim, I modeled each product as allowed to have 4 lots in the product separate Furnace1 zones. Therefore, on this operation the model will have 6 more tags than the IE model would suggest.

Dispatch

Three dispatch methods were tested. The factory was currently using First-In-First-Out (FIFO) but was considering a switch to Critical Ratio. In addition, the Least Slack dispatch policy was tested.

First-In-First-Out (FIFO) dispatching is conceptually easy for most readers to understand. The factory was currently using FIFO because of its fairness. Since the factory was processing material for both the production and development organizations a FIFO policy gave no preference to material of interest to either organization. An analysis of cycle time for lots actually run in the factory showed that cycle time was highly variable. The variability was not exclusively attributable to FIFO dispatching since the newness of the fab often dictated that lots be put on hold while technical questions are being addressed.

Both Critical Ratio and Least Slack dispatching methods were built into Mansim, and were tested in the experiment. Both methods seek to have all products complete processing equally early, however Critical Ratio attempts to have all lots complete proportionally early while Least Slack seeks to have all lots complete equally in an absolute fashion. A description of the calculation method makes this point clearer.

When a queue exists in front of a tool every lot in that queue can be represented by two parameters, T_r and P_r. When a lot is released into the factory it is assigned a desired completion

time and T_r is the amount of time remaining until the lot is due. P_r is the amount of processing time required to complete. The Least Slack parameter $T_r - P_r$ therefore can be calculated. Similarly, the Critical Ratio parameter T_r / P_r can also be calculated. If Least Slack dispatching were being used, the lot with the smallest value of $T_r - P_r$ would be run first. A negative value would suggest that the lot will be late since the remaining processing time is greater than the allowed time until the lot is due. Least Slack will have the effect that lots will all complete their processing more or less the same number of hours late or early as all others.

If Critical Ratio dispatching were being used, again, the lot with the smallest value of T_r / P_r would be chosen for processing next. Notice that values of $T_r / P_r > 1$ represent lots that have more time remaining until due than processing time. Lots where $1 < T_r / P_r < 0$ are lots not yet due but whose processing time is less than the amount of remaining time until due. Lots where T_r / P_r is negative represent lots that are already late. Critical Ratio has the effect that all lots exit the system proportionally on time. For example, all lots exit 5% early, which means that a product with a 20 day expected cycle time will actually exit in 19 days or 1 day early, while a product with a 60 day expected cycle time will exit in 57 days.

Chapter 5 - Analysis

This question motivated the building of a fab model and will guide the analysis of the model output: What are the best control mechanisms for a multi-product, hybrid production-development semiconductor wafer fabrication facility?

This chapter will present the results of analysis of the data that was generated by the running of the Mansim/X 3.8 model of Fab 6. Recall that earlier in this paper it was determined that a fab is supremely concerned about throughput but that fab cycle time and the variability are also key performance indicators, particularly in a leading edge fab. Therefore, the analysis in this chapter will begin by looking at these three items in turn. However, some questions exist even after a thorough simulation experiment and this chapter will close with a discussion of issues that are not completely closed.

Background

Exploring three factors lead to 12 distinct combinations. The combinations were demonstrated graphically in the previous chapter, but for the sake of clarity, will be enumerated below.

Experiment #	*Kanban*	*Start Policy*	*Dispatch*
E1	None	Uniform	FIFO
E2	None	Uniform	Least Slack
E3	None	Uniform	Critical Ratio
E4	None	Workload	FIFO
E5	None	Workload	Least Slack
E6	None	Workload	Critical Ratio
E7	Flow-Based	Uniform	FIFO
E8	Flow-Based	Uniform	Least Slack
E9	Flow-Based	Uniform	Critical Ratio
E10	Flow-Based	Workload	FIFO
E11	Flow-Based	Workload	Least Slack
E12	Flow-Based	Workload	Critical Ratio

Table 4 - Experiment

Throughput

One objective in the experiment was to run all scenarios at the same throughput. If this is accomplished, the simulated tools are equally busy in all cases and any differences in cycle times can be attributed to effects of control mechanisms and not the non-linear cycle time-utilization relationship.

Experiment #	*Throughput Mean* (wafers/week)	*Throughput Standard Deviation*	
E1	1057.00	7.02	
E2	1060.53	5.62	
E3	1055.93	4.65	
E4	1058.93	8.75	
E5	1057.53	8.31	
E6	1054.07	8.48	
E7	1054.20	10.04	
E8	1059.27	11.28	
E9	1054.20	9.62	
E10	1053.08	15.21	13 data points
E10*	1044.07	27.92	all 15 data
E11	1054.00	9.37	
E12	1056.67	10.07	

Table 5 - Throughput

For reasons that are not clear, two of the replications of the experiment E10 would not run properly. In these cases, one product would not run in the factory resulting in a very low output and correspondingly low cycle times for the other products. From here on, E10 will refer to the 13 replications that ran successfully.

A casual glance at the data suggests that the objective of simulating a fab at equivalent loading was achieved. However, we can construct confidence intervals as proof.

Let μ_i represent the true throughput of experiment i. If the throughput quantities are the same for experiments i and j, then $(\mu_i - \mu_j) = 0$. Rather than knowing μ_i, however, we have samples drawn

from the distribution that makes up μ_i. With the data available, we can construct a confidence interval for (μ_i - μ_j) and if it contains 0, then we can be reasonably confident that the throughputs from the two experiments are the same.

The limits of a 95% confidence interval for (μ_i - μ_j) will be calculated using the following equation:

$$\bar{i} - \bar{j} \pm 1.96\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}$$

where s is the sample standard deviation and n is the number of samples.

We can apply this result to compare experiment 1 to experiment 2.

$CI_{1\text{-}2}$ $= 1057.0 - 1060.53 \pm 1.96\sqrt{[(7.02^2 + 5.62^2)/15]}$

$= -3.53 \pm 4.55$

0 is contained in the interval which allows us to say with reasonable confidence that these two fabs have the same output.

Mean Cycle Time

Because the 12 fabs simulated were equivalently busy, the cycle time performance should be affected only by the different control policies in each of the scenarios. The table below exhibits the mean cycle time and the 95th percentile cycle time performance for each of the fabs. Recall that a shorter cycle time is preferable.

Experiment #	*Mean Cycle Time (hours)* (±95% Confidence Interval)	*95th Percentile Cycle Time (hours)*
E1	718.21 (±6.51)	783.85
E2	724.51 (±8.12)	744.43
E3	694.59 (±3.61)	718.93
E4	724.60 (±2.73)	786.49
E5	723.13 (±2.73)	748.24
E6	697.92 (±1.00)	713.14
E7	727.41 (±10.79)	792.84
E8	811.29 (±11.63)	855.71
E9	765.99 (±23.83)	827.73
E10	702.81 (±10.17)	771.09
E11	780.27 (±8.39)	830.28
E12	901.66 (±10.57)	960.48

Table 6 - Mean Cycle Time Results

This analysis would suggest that the policy set used in experiment E3 would provide the best performance for this fab, with experiment E6 being very close. Notice that both these policy sets have the common elements of Critical Ratio dispatching and no flow based control mechanism. Also remember that the Fab 6 had been planning to use the policy set represented by experiment E7. Choosing either E3 or E6 reduces mean cycle time by approximately 4.5% over E7.

This experiment was designed so that it could be analyzed as a traditional factorial experiment. Factorial experimentation allows us to determine the contribution of each dependent variable (factors- start policy, etc.) on the independent variable (mean cycle time). In the ensuing

analysis, the factorial experiment involving Critical Ratio vs FIFO dispatching will be analyzed (E1, E3, E4, E6, E7, E9, E10, E12).

Factor	*Effect (hours)*	*t value*
Mean Cycle Time	741.64	
Start Policy Workload=1, Uniform = -1	+30.2	16.7
Dispatch CR = 1, FIFO = -1	+46.8	26
Control Mechanism None = 1, FBCM = -1	-65.6	36
Start-Dispatch Interaction	+39.3	21
Start- Control Mechanism Interaction	-25.3	14
Dispatch- Control Mechanism Interaction	-71.9	39
Start-Dispatch- Control Mechanism Interaction	-40.8	22.7

Table 7 - Mean Cycle Time Factors

$t_{.975,110} = 1.980$, therefore all effects are significant.

The main effects of the Start and Control Mechanism factors are not what would be expected. The Workload Regulating start policy has a positive (increasing) effect on cycle time, when the policy was chosen because in other situations it has been seen to reduce cycle time. Also notice that the effect of not adopting a Flow Based Control Mechanism has a negative (decreasing) effect on cycle time. However, the primary purpose of the FBCM was to reduce cycle time.

The other result that we did not predict is the magnitude and sign of the Control Mechanism - Dispatch interaction factor. This implies that some synergistic effect occurs between these two factors. After reviewing the cycle time variation, we will come back to postulate what might be happening to cause this.

Cycle Time Distribution

We will now analyze the standard deviation of cycle time in the same fashion that we analyzed the mean cycle time. However, the standard deviation for each product will be analyzed separately. The table below contains the data for all 12 experiments and the lowest value (most desirable) has been highlighted.

	Pro duct			
Experiment #	*c5l*	*c5wf*	*c64*	*c66*
E1	28.85	67.95	39.75	50.01
E2	15.03	16.25	15.48	15.17
E3	11.63	16.50	15.15	19.64
E4	27.15	66.78	36.40	45.85
E5	13.45	15.87	14.35	14.39
E6	8.25	9.65	9.71	10.41
E7	36.49	56.33	36.32	42.95
E8	27.55	25.55	26.77	25.51
E9	29.89	52.75	36.20	52.45
E10	35.47	61.64	40.42	48.75
E11	30.82	31.19	31.14	30.71
E12	28.94	47.19	36.24	49.35

Table 8 - Cycle Time Variation Results

Experiment E6 has the lowest cycle time standard deviation values for all four products. Again, we can use analysis of factorial experiment tools to analyze the contribution of the three factors to cycle time variation.

Factor	*C5l SDCT* Effect (hours)	t value		*C5WF SDCT* Effect (hours)	t value
Mean Effect	25.83			47.35	
Start Policy Workload=1, Uniform = -1	-1.76	2.76		-2.06	2.04
Dispatch CR = 1, FIFO = -1	-12.31	19.3		-31.65	31.3
Control Mechanism None = 1, FBCM = -1	-13.73	21.51		-14.25	14.1
Start-Dispatch Interaction	-0.39	.62		-4.13	4.1
Start- Control Mechanism Interaction	-0.78	1.21		-1.94	1.92
Dispatch- Control Mechanism Interaction	-5.75	9.0		-22.63	22.4
Start-Dispatch- Control Mechanism Interaction	-0.44	.68		1.29	1.3

Table 9 - Cycle Time Variation Factors I

Factor	*C64 SDCT* Effect (hours)	t value		*C66 SDCT* Effect (hours)	t value
Mean Effect	31.27			39.93	
Start Policy Workload=1, Uniform = -1	-1.16256	-1.53778		-2.67346	-2.30471
Dispatch CR = 1, FIFO = -1	-13.9008	-18.3873		-13.9265	-12.0056
Control Mechanism None = 1, FBCM = -1	-12.0441	-15.9314		-16.8965	-14.566
Start-Dispatch Interaction	-1.53744	-2.03365		-3.48987	-3.00851
Start- Control Mechanism Interaction	-3.2341	-4.27791		-4.02654	-3.47115
Dispatch- Control Mechanism Interaction	-11.7492	-15.5413		-18.9801	-16.3622
Start-Dispatch- Control Mechanism Interaction	0.494103	0.653575		0.956538	0.824602

Table 10 - Cycle Time Variation Factors II

$t_{.975,110} = 1.980$, significant effects have been highlighted.

For all four products, the Dispatch, Control Mechanism, and Dispatch- Control Mechanism Interaction effects are statistically significant and the values of these effects are large. It is not

unexpected that dispatch effect is large and negative, one purpose of the Critical Ratio dispatch rules is to make all wafers equally early (or late), leading to very low cycle time variation. It was not obvious at the beginning that a FBCM would increase variation to the extent found. But as with mean cycle time, the least expected finding is the large Dispatch- Control Mechanism interaction.

Discussion

The analysis has surfaced three subjects for discussion. Before moving to the final chapter where recommendations are presented, I would like to briefly comment on these.

There is one effect that does not show up in the above analysis, which I would term *preference*. In the fab scenarios using FIFO, the more simple C5L product tended to be "preferred" over the other product types. That is, cycle times for C5L were shorter on a proportional basis than for the other products.

Experiment # (Dispatch)	*C5L Actual/Theoretical cycle time*	*C5WF Actual/Theoretical cycle time*	*C64 Actual/Theoretical cycle time*	*C66 Actual/Theoretical cycle time*
E1 (FIFO)	1.92	2.22	2.04	2.09
E2 (LS)	2.06	2.05	2.06	2.05
E3 (CR)	1.96	1.98	1.97	1.97

Table 11 - Preference

However, in scenarios employing either the Least Slack or Critical Ratio dispatching methods, all products are equally proportionally on time or late. In some sense, we should not be surprised, that is what the policies were designed to do. However, the previous work in this area was looking at single product fabs and at the cycle time standard deviation of only one product. This result would confirm that the policies behave as would be desired in a multiple product fab.

Mean cycle time aggregates the individual cycle time results of four products. While it is true that the fab would like as small a mean cycle time as possible, a shorter cycle time for newer, more complex products is also desirable. These are the products that are more likely to be undergoing process experimentation, and quick achievement of results can be fed back into the fab in the form of design or process changes.

The second item to discuss is the unexpected performance of the Workload Regulating start policy. For this simulated fab, a Workload Regulating policy actually increases cycle time and only slightly decreases cycle time variation. Since the Workload Regulating policy takes account of only the bottleneck took, the order and frequency with which products visit the bottleneck may determine the effectiveness of this policy.

The third, and most obvious issue is the large interaction effect that occurs between the choice of dispatch policy and flow based control mechanism on both the mean and variation of cycle time. At the outset of this experiment, there was no intuition that would have predicted this result. Unfortunately, neither theory nor simulation is not helpful in understanding the cause of such interaction. In the absence of proof, a couple possible explanations might be offered.

A brief discussion at Digital hinted at one possibility, over constraint. Both the dispatch policy and a flow based control mechanism provide rules that govern how wafers move through the process, although in slightly different ways. Both the Critical Ratio and Flow Based Control Mechanism are dynamic policies, that is the system state is evaluated each time a decision is called for. However, neither is required to account for the other as a decision is called for. The postulate of over constraint suggests that each system may be exerting control in such a fashion that negates some of the control effect of the other.

It should be said that this experiment was run on a digital computer, and that the kanban feature of Mansim was not as thoroughly documented as many of the others. In fact, during a call to tech support, an employee expressed surprise that the feature was even being used. The author does not call into question the skill of the Mansim programmers, but would like to make the point that we used a digital representation of a control mechanism and that it is not inconceivable that the actual implementation would vary subtly. We do know that the modeled version of the FBCM was not identical to the actual implementation. It is not implausible that the resulting performance could be different as well.

Chapter 6 - Conclusions

In concluding this work I would like to offer a recommendation and a challenge.

Recommendation for fab policies

Based on the analysis conducted in the previous chapter, I would like to offer a recommendation that Fab 6 consider switching its set of control policies to those that correspond with experiment E6; workload regulating starts, critical ratio dispatch, and no FBCM. Such a switch should decrease cycle time about 5% and cycle time variation about 70% over the currently chosen set of policies. This suggestion was made to the fab during the research internship, and there exist some legitimate concerns relating to the dissolution of the FBCM that has been in operation. Nevertheless, this set of policies is quite close to the set of policies advocated by Kumar in his paper on re-entrant systems and similar to Wein's approach. It is believed that many fabs currently operate in a similar fashion and that benchmarking data should be available through Sematech or other industry sources.

There is an important caveat to the above recommendation. Any implementation plan that seeks to change one control policy at a time should not allow the critical ratio dispatching method to be in place at a time when the FBCM is still in operation. Previous analysis demonstrated that the combination of these two policies detrimentally affects both cycle time and cycle time variability. At one point the fab had considered switching to critical ratio dispatching and continuing to use the FCBM. Needless to say, the analysis would suggest that this course of action is unwise.

Challenge

There exists an assumption in this paper that Professor David Cochran brought into the open at several points during the internship and writing period that leads me to issue a challenge to Digital's staff. The assumption is that all factors other than control policy are unchangeable. My challenge to the fab staff is to consider the improvements in fab performance that could be attained by improving tool reliability and downtime, both of which were considered fixed inputs to the Mansim model.

In order to complete the project in the time allotted, it was necessary to limit the bounds of the problem under consideration. Therefore, it was decided to build a model and allow control policies to be the only independent variable. While this approach is appropriate for an academic paper, the fab must not consider the performance of its tools or people to be stationary. Additional sources of improvement should be pursued and in fact are by the Total Productive Maintenance (TPM) efforts that were recently started in earnest at Fab 6.

TPM is too complex a subject to adequately cover in the short space available. It does however provide the framework necessary for an organization to dramatically improve the productive levels of the equipment it owns by forming cross-functional teams and improving the skills and communications abilities of the members. As an example of one activity that was performed during the internship period, engineers, operators, and maintenance personnel together performed a thorough inspection of a piece of equipment in the fab. Within just an hour nearly, three hundred items had been identified needing correction, ranging from dirt and dust needing cleaning to missing screws and loose wires. The truly amazing thing was that the teams pulled together and solved most of these items within two weeks. Admittedly, this activity by itself did

not enhance the productivity of the equipment in question, but the teams became more acquainted with their equipment and with working together.

The Fab 6 TPM efforts will attempt to improve three of the inputs to the model that have heretofore been considered fixed. First, the teams will seek to reduce the mean time to repair and lengthen the mean time to fail. In this way, the tool will spend less time in a "down" state. Through process optimization techniques the mean processing time may be able to be shortened, and the tools can spend less time processing each lot. Finally a review of procedures may allow shorter preventative maintenance actions. The sum of these effects is to increase the effective capacity of the tools. Obviously, these efforts should begin with the most capacity constrained or bottleneck tools.

References

1 Judge, P., "Digital's Struggle to Save Its Alpha Chip," *BusinessWeek*, December 30, 1996, p. 44.

2 Cheyney, T., "CFM Plays a Pivotal Role in Yield Engineering Team at Digital's Fab 6," *MicroContamination*, August 1996, pp. 59-66.

3 Bloomberg Business News, "Samsung Licensed to Market Digital's Alpha Microprocessors," *New York Times*, June 19, 1996, p. D5.

4 "DEC to Cut Chip Prices to Pursue Mass Market," *ClieNT Server NEWS*, October 25, 1996.

5 Digital Equipment, "Alpha is Launched into the Volume Windows NT PC Market with Low Cost 21164PC Microprocessor [On-line]," Available: http://www.digital.com/semiconductor/press-pc64.htm.

6 Gwennap, L., "Estimating IC Manufacturing Costs," *Microprocessor Report*, August 2, 1993, pp. 12-16.

7 Competitive Semiconductor Manufacturing Program, "The Competitive Semiconductor Manufacturing Survey: Second Report on the Results of the Main Phase," (ed. R. Leachman), ESRC/CSM-08, University of California, Berkeley, September 16, 1994, p. 130.

8 Port, O., *et al.*, "The Silicon Age? It's Just Dawning," *BusinessWeek*, December 9, 1996, pp. 148-152.

9 Sematech, "Flow Control in Semiconductor Manufacturing: A Survey and Projection of Needs," Technology Transfer Number 91110757A-GEN, November 1991.

10 Rifkin, G. and G. Harrar, The Ultimate Entrepreneur: The Story of Ken Olsen and Digital Equipment Corporation. Chicago: Contemporary Books, 1988.

11 Digital Equipment, "The Digital Report: Connecting With Tomorrow [On-line]," Available: http://www.digital.com.

12 Digital Equipment, 1996 Annual Report.

13 Goldratt, E., and J. Cox, The Goal. Great Barrington: North River Press, 1992.

14 Menon V., "A Constraint-based Systems Approach to Line Yield Improvements in Semiconductor Wafer Fabrication," (unpublished Master's thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge) 1994.

15 Graves, S., Class notes for MIT course 15.762, spring 1996.

16 Nahmias S., Production and Operations Analysis. Burr Ridge, IL: Richard Irwin, 1993, pp.266-267.

17 Arora, S., and S. Kumar, "Effects of Inventory Miscount and Non-inclusion of Lead Time Variability on Inventory System Performance," *IIE Transactions*, vol. 24, no 2, May 1992, p. 96.

18 Chen, H., *et al.*, "Empirical Evaluation of a Queuing Network Model for Semiconductor Wafer Fabrication," *Operations Research*, vol. 36, no. 2, March 1988, pp. 202-215.

19 Wein, L., "On the Relationship Between Yield and Cycle Time in Semiconductor Wafer Fabrication," *IEEE Transactions on Semiconductor Manufacturing*, vol. 5, no. 2, May 1992, pp. 156-158.

20 Cunningham, S., and J. Shanthikumar, "Empirical Results on the Relationship Between Probe Yield and Cycle Time in Semiconductor Wafer Fabrication," ESRC 94-15/CSM-11, University of California, Berkeley, September 1994.

21 Glassey, C., and G. Resendes, "Closed-loop Job Release Control for VLSI Circuit Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 1, no. 1, February 1988, pp. 36-46.

22 Wein, L., "Scheduling Semiconductor Wafer Fabrication," *IEEE Transactions on Semiconductor Manufacturing*, vol. 1, no. 3, August 1988, pp.115-130.

23 Chang, Y., T. Sueyoshi, and R. Sullivan, "Ranking Dispatching Rules by Data Envelopment Analysis in Job Shop Environment," *IIE Transactions*, vol. 28, 1996, pp. 631-642.

24 Panwalker, S., and W. Ishkander, "A Survey of Scheduling Rules," *Operations Research*, vol. 25, 1977, pp. 45-61.

25 Lu, S., D. Ramaswamy, and P. Kumar, "Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing Plants," *IEEE Transactions on Semiconductor Manufacturing*, vol. 7, no. 3, August 1994, pp. 374-385.

26 Schonberger, R., "Applications of Single-Card and Dual-Card Kanban," *Interfaces*, vol. 13, no. 4, August 1983, pp. 56-67.

27 Bonvik, A., "Performance Analysis of Manufacturing Systems Under Hybrid Control Policies," (unpublished Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge) June 1996.

28 Gershwin, S., Manufacturing Systems Engineering. New Jersey: Prentice Hall, 1994.

Other References

Hogg, R., and J. Ledolter, Applied Statistics for Engineers and Physical Scientists. New York: Macmillan, 1992.

5138-17