A Tactical Planning Model

for a Semiconductor Wafer Fabrication Facility

by

Hemant Taneja

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degrees of Bachelor of Science and Master of Engineering

and

Submitted to the Alfred P. Sloan School of Management in Partial Fulfillment of the Requirements for the Degree of Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 1999

© Hemant Taneja, 1999. All Rights Reserved.

The author hereby grants to M.I.T. permission to reproduce, and distribute publicly paper and electronic copies of this thesis and to grant others the right to do so

Author ....
Electrical Engineering and Computer Science
Alfred P. Sloan School of Management
May 7, 1999

Certified by ....
Stephen C. Graves
Abraham Seigel Professor of Management Science
Thesis Supervisor

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

Accepted by ....
James B. Orlin
E.Pennel Brooks Professor of Management Science
Co-Director, Operations Research Center

Acknowledgments

I am deeply indebted to my thesis advisor, Prof. Stephen C. Graves for guiding me through this entire work. I have learned a great deal about manufacturing operations as his student and more importantly I have learned how to be a better student. I also thank Sumer Johal who supervised my research work at Analog Devices Incorporated. I am highly grateful for the wisdom he imparted to me both as a supervisor and as a great friend.

I express my deepest thanks to Mike Mueller for teaching me a great deal about ADI, for tremendous logistical support and for providing many useful comments during the thesis write-up. I also thank Sam Chow for patiently helping me with all the technical problems I encountered. Charlie Messemer, Marge Hendricks and John Manning were also extremely helpful with their insightful comments all through the duration of this thesis. I also thank Steve Muir for proof-reading my thesis.

I also take this opportunity to thank a few of my friends for their tremendous support and encouragement. First, I thank Amit Dhadwal for numerous insightful discussions in the modeling phase of my thesis and for his constant encouragement throughout my tenure as an MIT student. I also thank my good friend Surya Ganguli for helping me stay up late nights while we both worked on our thesis projects. I also thank Anubha Triphathi for her tremendous love and support throughout.

Last, but not least I would like to thank my family, my parents Shiv and Santosh and my sister Charu, for everything they have done to help me get here. I will be forever indebted to them for their love, support and guidance that they imparted to me all these years.

A Tactical Planning Model for a Wafer Fabrication Facility

by

Hemant Taneja

Submitted to the Department of Electrical Engineering on May 7, 1999, in Partial Fulfillment of the Requirements for the Degrees of Bachelor of Science and Masters of Engineering in Electrical Engineering and Computer Science
and
Submitted to the Alfred P. Sloan School of Management on May 7, 1999, in Partial Fulfillment of the Requirements for the Degree of Master of Science

Abstract

The thesis project describes the development of a model-based framework for analyzing and operating a semiconductor wafer fabrication facility at a tactical level. The fabrication facility is modeled based on a discrete-time continuous flow model of a job shop described by Graves (1986). This model was extended to accommodate the existence of multiple work flows in a fabrication facility. The development of the model was done at the Wilmington wafer fab of Analog Devices Inc. based on a 21 week history of inventory flow and production data. The Wilmington fab was represented by a set of 22 distinct work flows. Each work flow was represented as a sequence of workstations based on the actual processing requirements. The model provides a characterization of the steady state levels of production requirements at and work-in-process in front of each workstation.

Thesis Supervisor: Prof. Stephen C. Graves
Title: Abraham Seigel Professor of Management Science

Table of Contents

1 Issues in Semiconductor Manufacturing 11
1.1 Introduction 11
1.2 Manufacturing Complexities 11
1.3 Literature Survey 18
1.4 Analog Devices Incorporated: Status Quo 20
1.5 Motivation for this Thesis 22
2 Tactical Planning: Model Development 27
2.1 The Basic Model 27
2.2 Wilmington Wafer Fab Model 35
2.3 Conclusion 56
3 Model Application to Fab Operations 57
3.1 Operational impact of Variability in Starts 57
3.2 Impact of Lot sizing 64
3.3 Trade-off between WIP and Lead Times 66
3.4 Impact of Demand fluctuations 69
3.5 Conclusion 73
4 Batching: Simulation Study 75
4.1 Introduction 75
4.2 Simulation Model 75
4.3 Conclusions 84
5 Conclusion 85
5.1 Overview 85
5.2 Tactical Planning Model 85
5.3 Simulation Study 86
5.3 Future Direction 86

Bibliography 90

List of Figures

Figure 1.1: Statistical comparison of batch sizes of fab equipment 12
Figure 1.2: Work flow in Diffusion 14
Figure 1.3: Statistical comparison of recipe times of fab equipment. 16
Figure 1.4: Production control in Wilmington wafer fab 21
Figure 2.1: Description of an *aggregate* workstation 35
Figure 2.2: Histogram of lead times of workstations 39
Figure 2.3: Bar chart of c.v. of daily starts of major flows 46
Figure 2.4: Time series of daily starts of a sample work flow 47
Figure 2.5: Production variability at the equipment 48
Figure 2.6: Histogram of different recipes at a Diffusion furnace 49
Figure 2.7: Histogram of normalized daily production 52
Figure 2.8: Bar chart of normalized cycle times for work flows 53
Figure 3.1: Impact of input variability on production requirements 58
Figure 3.2: Impact of starts variability on production requirements 60
Figure 3.3: Sensitivity of equipment variability to starts variability 62
Figure 3.4: Relationship of production trends and lot sizes 63
Figure 3.5: Ratio of (Workstation Lead time/Workstation Processing time) 65
Figure 3.6: Impact of demand increase on production requirements 71
Figure 3.7: Impact of reducing starts variability on constrained equipment 72
Figure 4.1: Simulation Study: Idle time vs. minimum batch size 79
Figure 4.2: Simulation Study: Wait time vs. minimum batch size 80
Figure 4.3: Simulation Study: Fab cycle time vs. minimum batch size 82
Figure 4.4: Simulation Study: Fab throughput vs. minimum batch size 83

List of Tables

Table 2.1: Representation of a Process flow 38
Table 2.2: Input parameters for Work Flow shown in table 2.1 45
Table 2.3: Estimated Capacities for the Diffusion equipment 50
Table 2.4: Comparison of WIP profile for flow 5 50
Table 2.5: Production Requirements at the Furnaces: Model vs. Actual 54
Table 3.1: Model Output: Reduction of c.v. of starts by 50 percent 59
Table 3.2: Ideal Scenario: Lead Time of Workstations is its Processing time 67
Table 3.3: Impact of reducing variability of low utilization furnaces 68
Table 3.4: Impact of planned lead time reduction on flow cycle times 68
Table 3.5: Model Output: Average Demand of flows 15,20 is doubled 69
Table 3.6: Model Output: Average Demand of flows 15,20 with uniform starts 70

Chapter 1

Issues in Semiconductor Manufacturing

1.1 Introduction

In a little over half a century after the invention of the first transistor, the semiconductor industry has made tremendous advancements. Devices have quickly scaled down in size to sub-micron levels and become increasingly reliable at the same time. Such advancements have come through the development of extremely complicated manufacturing processes. Most of semiconductor fabrication takes place in clean rooms where contaminants are less than one part per million. The manufacturing of a sub micron integrated circuit takes many days of processing at various equipment. The semiconductor industry is a high clock-speed industry and products have short life cycles of high demand. Therefore, it is important for the manufacturing facilities to be able to respond to market conditions quickly and be able to maintain high throughput rates. Operations managers in wafer fabs are in constant struggle to maintain low cycle times and high throughput rates. This research is an effort to develop a model to help understand the operations of a wafer fab and provide insights about how to plan factory operations in response to changing market conditions.

1.2 Manufacturing Complexities

In this section we describe a few complexities of manufacturing in a wafer fabrication facility. Wafers are grouped into lots and routed through several hundred pieces of equipment together for conversion to the same final product. The flow of wafers in the fab is highly variable due to a variety of factors. Processing times are different at each piece of equipment and it is extremely hard to synchronize equipment operations. As a result there is a highly variable flow of work arriving at each equipment. Moreover, equipment break-

downs are very common in a fab which leads to further variability in work flow. Due to these and several other reasons described below, production control in a wafer fab is an extremely hard task.

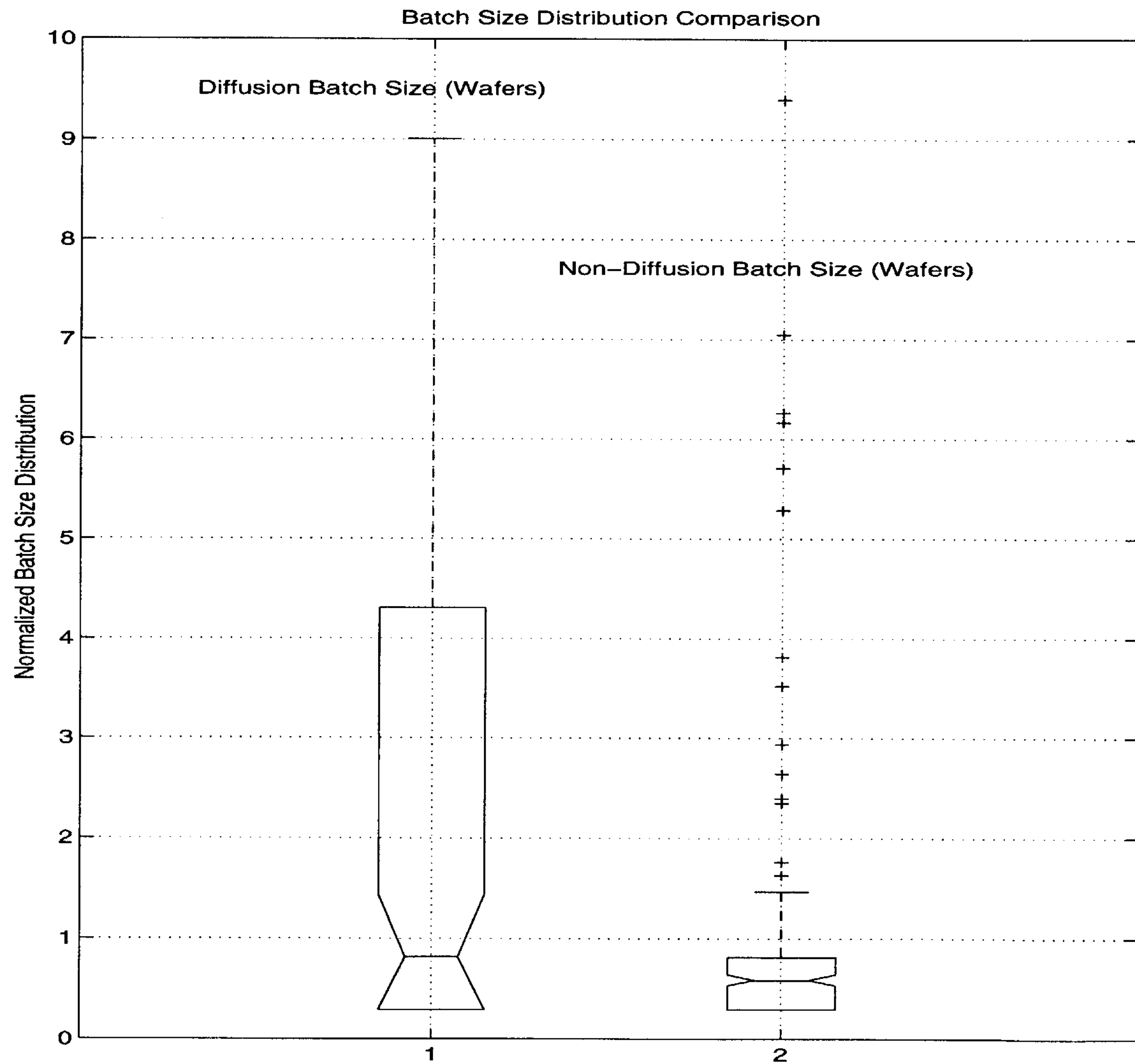


Figure 1.1: Statistical Comparison of batch sizes of equipment in Diffusion processing vs. the rest of the equipment in a wafer fab.

Batch Size of Equipment

Some operations such as diffusion and chemical vapor deposition (CVD) have really long processing times. Therefore, they are usually executed in large batches to prevent them from being bottlenecks. Such batch sizes are typically integer multiples of a lot size. Figure 1.1 does a statistical comparison of the batch sizes of equipment involved in diffusion vs. non-diffusion processes. The batch sizes of the equipment are normalized by the mean batch size of non-diffusion equipment. As can be seen, the batch sizes for the diffusion equipment are can be much larger than those for the non-diffusion equipment.Even though the median batch sizes are relatively similar for the tow categories, many of the diffusion equipment have much larger batch sizes than the rest of the equipment in the fab. However, in a situation where cycle time for the part is very important, the tendency to fully utilize this large batch capacity usually results in long waits for upstream lots. This phenomenon is further amplified by the "re-entrant" nature of the process flows. Due to this, both the On-Time Delivery and the Average Cycle Time of the entire fab suffers. This large mismatch in batching capacity of such operations with the rest of the fab causes a large, intermittent, throughput-based perturbation in the manufacturing flow. The effect is much like a marching troop[1] where a certain subset of troops march with a much longer stride (Cycle time) but at a much slower rate (Throughput rate). Since the rest of the troops (Other Equipment) have their own synchronized rhythm (Throughput Rate), this causes large gaps (Cycle Time Increases) in the ranks of the troops and slows (reduces the Throughput rate) the entire group (Factory) down.

---

1. This example if taken from the "Drum-Buffer-Rope" approach advocated by Constraint Management Principles

Upstream Gating Operations

Typically, there are some steps prior to long operations such as diffusion and chemical vapor deposition (CVD) steps. Some of these steps may involve a "clean" step at a "sink" prior to the actual long cycle. These steps are necessary to prevent contamination or particle defects within a clean room environment. Usually, a sink services multiple furnaces because the throughput rate at the sink is higher than the throughput rate at the subsequent furnaces. This becomes a very important and complex problem in a manufacturing scenario with a diverse product line because the arrival stream of lots at the sink is essentially a set of many sub-streams each with a different processing requirement at the furnaces.

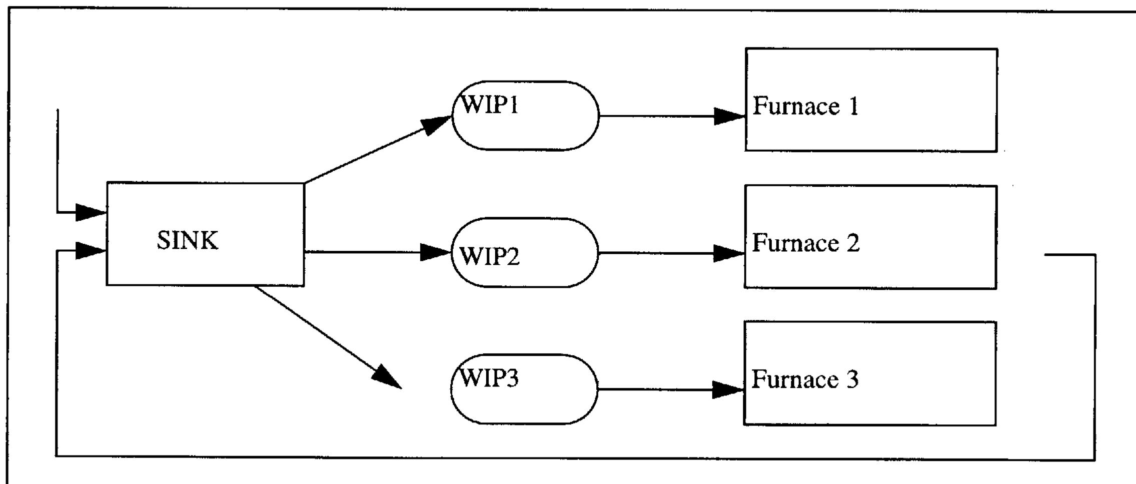


Figure 1.2: Work flow in Diffusion Area

Process Engineering Specifications

The nature of diffusion/CVD operations often requires that the wafer surface be as particle free as possible. This is mainly due to today's semiconductor technology

converging to smaller and higher performing ICs. As devices scale down in size, particle contamination risks increase. However, due to these engineering specifications, the amount of time that a lot can be allowed to wait after it has been "cleaned" at the sink is constrained. Parts which wait longer than this constrained time have to be "re-worked" through the sink operation - thereby further compounding the problem of lost opportunity cost of wait time already incurred by competing parts. Furthermore, this constrained time or the "maximum wait time" is variable for different parts and for different operations for the same part. This issue often conflicts with the need to have a high utilization from the furnace. If the furnace is constantly well-fed to achieve high utilization, one runs the risk of losing sink capacity when waiting wafers have to be re-worked for not meeting the process constraints. This makes the calculation of the optimum "forced wait time" very difficult - especially in the view of unpredictable equipment breakdowns. All these factors make the selection task at the sink - of the "correct" or most "important" lot - extremely difficult to manage.

Significant Mismatch in Processing Times

In addition to a significant mismatch in batch capacity, cycle times in diffusion/CVD are significantly longer from the rest of the fab's operations as shown in figure 1.3. In the figure, the recipe times for all the equipment are normalized by the mean recipe time of the non-diffusion equipment. These longer recipe times make the flow of work highly variable and leads to the development of alternating periods of long queues and WIP starvation for downstream operations. This problem is even more amplified by the re-

entrant nature of the line, where perturbations introduced during one cycle are positively reinforced during each subsequent cycle.

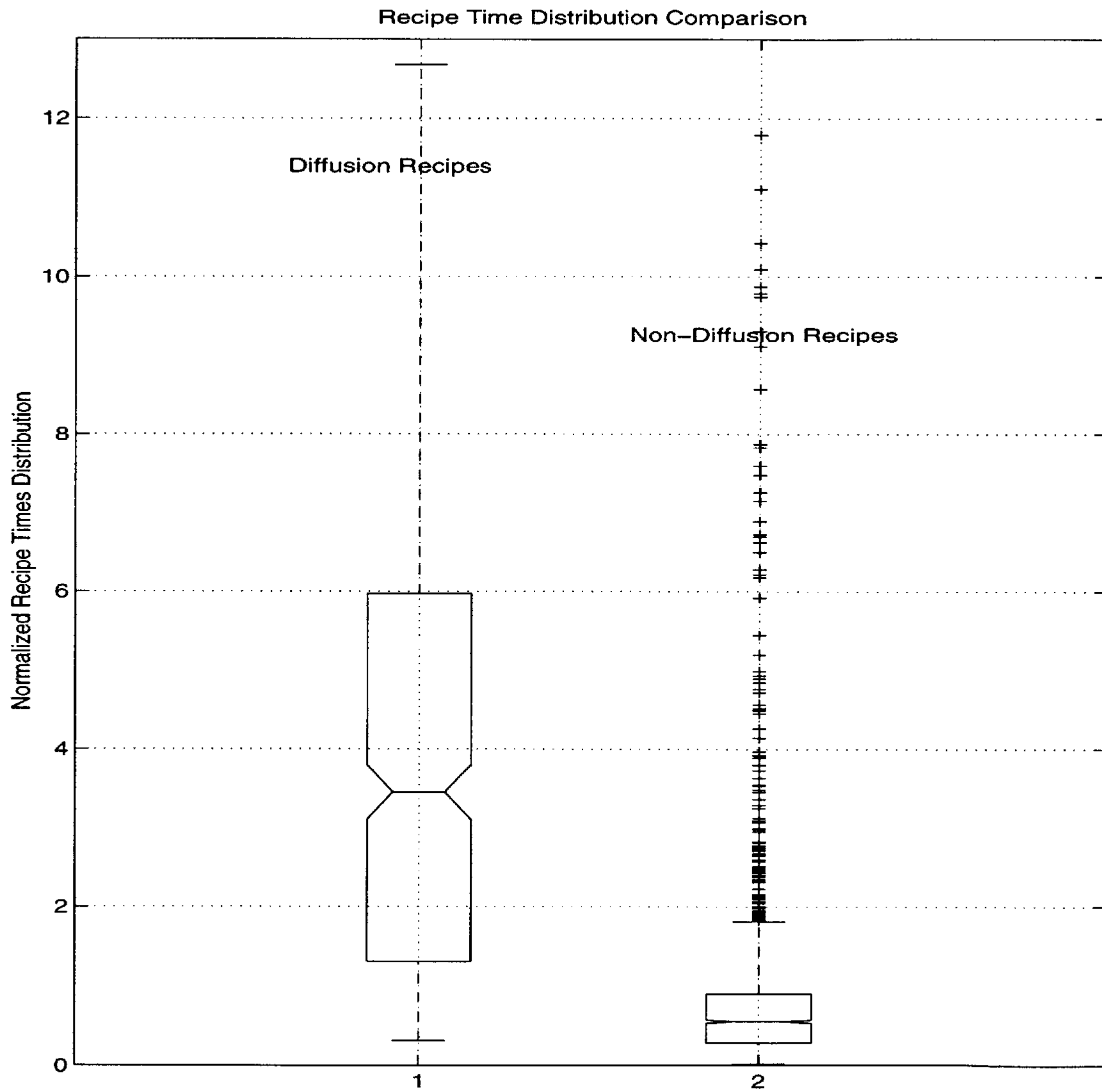


Figure 1.3: Statistical Comparison of recipe times of equipment in Diffusion processing vs. the rest of the equipment in a wafer fab.

Process Diversity and Dynamic Priorities

The inherent diversity of the products/processes in fabs makes the selection criterion at various equipment even more difficult. This is because in order to meet the deadlines for on-time delivery of such a diverse product portfolio, an egalitarian penalty (wait time) absorption rule has to be adopted. This rule equally distributes any delays in processing throughout the various products in the queue. However, sudden changes in demand cannot be met with a static dispatching rule of this sort. Therefore, a *dynamic* prioritizing is essential for a quick shop-floor reaction to changing market conditions. Even if the market conditions are considered pseudo-static for a short time interval (e.g., a few days), optimally utilizing equipment capacity for different "static" priorities composing the work queue becomes a difficult task. On one hand, higher priority lots (or "hot lots") must be run before lots of lower priority. However, since typically the number of "hot lots" is much smaller than the number of "non-hot-lots", equipment utilization and area throughput suffers if such policies are implemented.

Equipment Downtime

Coupled with all the previously discussed phenomenon and the re-entrant nature of the line is the fact that all the equipment in the line suffers from stochastic breakdowns. These breakdowns should not be confused with regular preventive maintenance performed on the equipment which may be scheduled for efficiency. These breakdowns are a stochastic phenomenon which are independent for each machine. They can also be difficult to model using standard statistical modeling techniques and may require additional heuristics procedures for effective modeling. In any case, such breakdowns

affect equipment performance greatly and cause large deviations from expected analytical scheduling/capacity calculations.

1.3 Literature Survey

A review of the production planning and scheduling models in the semiconductor industry is given by Uzsoy *et al.* (1992). They classify the research into the following three areas.

- *Performance evaluation*. Models which are used for understanding the behavior of a given system.
- *Production planning*. Long term aggregate planning with a time horizon of months or weeks.
- *Shop-floor control*, which addresses the questions of how much material to start into the facility and how to control the material once started.

Most of the research on flow shops has focused on studying the impact of lot-sizing, lead times, capacity constraints and in-process inventories. Karmakar (1987) proposes a queueing model to examine the impact of lot-sizing, manufacturing lead time and capacity utilization for batch-manufacturing shops. The model characterizes the lead time as a function of lot sizes assuming exponential processing times. Other work that has focussed on the lot-sizing problem is that of Zipkin (1986), Billington *et al.* (1986) and Dobson *et al.* (1987). Zipkin develops an optimization framework which represents each product by a standard inventory model and each production facility with a standard queueing model. The framework determines optimal lot sizes for a multi-item batch production system based on queueing and inventory considerations. Billington *et al.* (1986) develop a mixed

integer linear programming framework which solves the lot-sizing problem for a facility with a single bottleneck and simultaneously determines the planned lead times for the products. They also develop heuristics that determine good feasible solutions using lagrangian relaxation methods.

Few researchers have used production rates as the parameters controlling shop flow time. Cruickshanks *et al.* (1984) propose a simple production smoothing model that sets production requirements for a period (e.g. a month) based on inventory and upcoming demand contracts. They develop the concept of a *planning window* which is the difference between the *delivery* lead times and the *production* lead times. Production levels are set to smooth out the production over the length of this planning window. Their results indicate that substantial production smoothing could be attained with a fairly short planning window. The final and the most relevant model (to this project) that we discuss here is the tactical planning model (henceforth referred to as TPM) by Graves (1986). This model is based on a queue management system which uses a linear control rule to set production requirements at machines. The flow of work is assumed to be markovian in nature. The model provides a discrete-time continuous flow characterization of a job shop which manufactures a stationary job mix. It focuses on understanding the inter-relationship of production capacity, demand variability and work-in-process inventory. One of the key objectives of the model is to characterize the level of work-in-process inventory that appears in front of the machine and the production requirements, for a given assignment of the lead times. Increasing the lead time of a machine leads to increased *smoothing* of the output stream from that machine but it also leads to high work-in-process levels in front of that machine. The model captures this trade-off between inventory levels and production smoothing and helps determine a reasonable assignment of lead times and production requirements for the equipment. The model quantifies the cost (additional inventory hold-

ing or manufacturing resources) of unpredictable variability in production requirements at a machine.

1.4 Analog Devices Incorporated: Status Quo

Analog Devices Inc. is a semiconductor design and manufacturing company with an established niche in the SLIC (Standard Linear Integrated Circuits) market, which accounts for approximately 70% of the company's revenues. The company also manufactures a wide variety of other products including hardware components used in digital signal processing, telecommunications, and the automotive industry. Analog Devices' diverse product portfolio translates into tremendous process diversity at the level of manufacturing. Currently, the Wilmington Wafer Fab has over 300 different process flows running concurrently which makes production control an extremely complicated problem. The re-entrant nature of the flows, stochastic processing times and stochastic breakdowns of the equipment further add to the complexity.

Current form of production control at the Fab is described in figure 1.4. PROMIS/MES system sets the production requirements at each equipment in the form of a dispatch list. The operators in the fab use this list to carry out production. They simultaneously input the status of each equipment (up/down status, work-in-process, etc.) into the PROMIS system. A 21 week history of the Fab is stored in the PROMIS database. Using datalink, the PROMIS query system, the status of the factory is downloaded into the Scheduler via a relational database. The scheduler has a built-in model of the Fab which is simulated using the factory status to execute dispatch rules every hour. Based on these dispatch rules, the dispatch list is updated every hour. This is a reactive scheduling system which modifies the production requirements at each equipment based on the state of the factory.

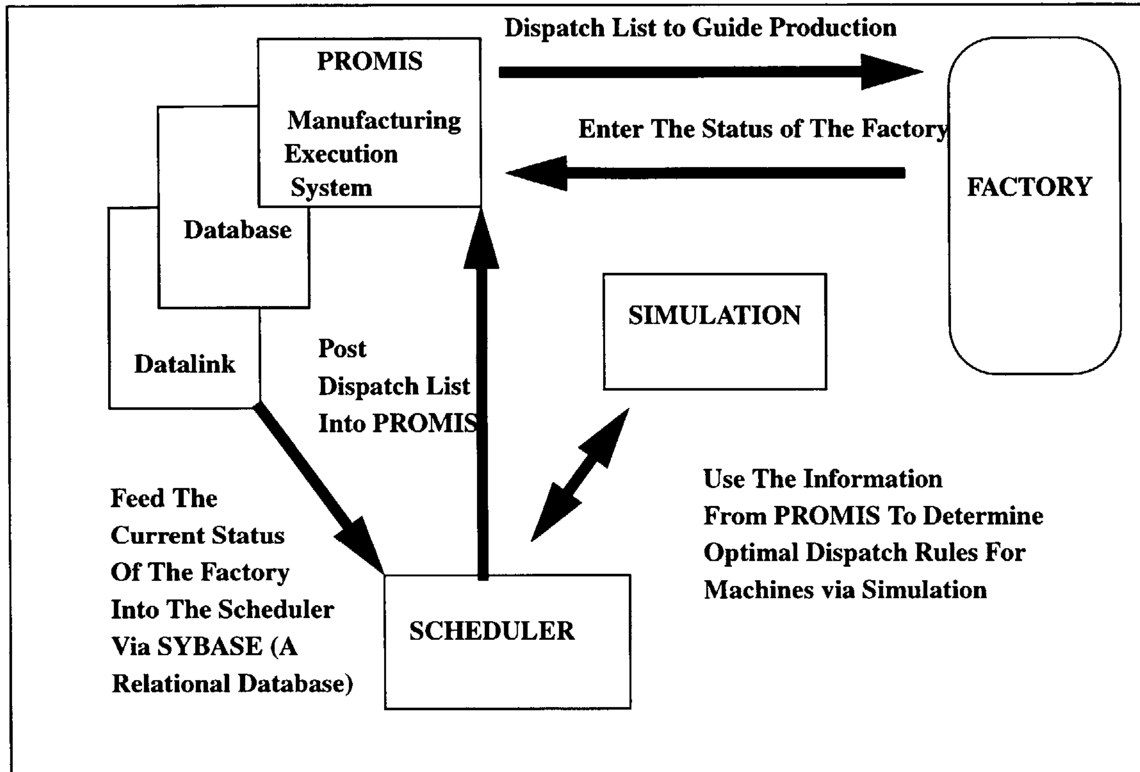


Figure 1.4: Simulation-based Production Control in the Wilmington Wafer Fab.

The production control is exerted in a virtual manner with-in the simulation model. Work-in-process (WIP) levels in the fab are controlled using a *virtual Kanban* system. Under this system, the process flows are divided into a few stages and maximum WIP levels are set for each stage. These maximum levels are dynamically allocated for each process based on the demand characteristics, and equipment status. Wafers move forward in a process flow to a downstream stage only if the WIP in that stage has not reached its maximum level. At each work station, the work-in-process is prioritized and scheduled based on various criteria such as the Critical Ratio of lots (a measure of how late a lot is, relative to its assigned shipment date). The status of kanbans, the priorities of lots and equipment

status are input to the simulation model (via PROMIS). Based on the *virtual* kanban control and work prioritizing mechanisms embedded in the model, lots are scheduled at each equipment. This information is made available to the operators via a dispatch list on the PROMIS system. Although the dispatch list is a mere guide to production, it is closely followed by the operators. This production control mechanism is transparent to the factory floor because the operators only see the dispatch list and all the control is exerted virtually in the scheduler model. Hence, any modifications to this control can be easily implemented without any added resistance from the operators.

The current form of production control with the scheduler has had a tremendous impact on the factory floor. Cycle times have been reduced for all the product flows and work-in-process has been maintained in check as well. However, the system lacks an underlying analytical model of the fab which can help understand the impact of demand fluctuations on the operations of the fab in a proactive manner. In this thesis, we describe the development of a tactical planning model which does just that. This model was developed with the hope of being able to characterize the production requirements and work-in-process profile of the factory for any demand distribution.

1.5 Motivation for this Thesis

As described earlier, wafer fabrication is a long and intricate process. The Wilmington wafer fab manufactures a diverse product portfolio which causes tremendous queueing effects on the shop floor. These queueing effects originate because different products have different processing requirements (*jobs*) at a machine and therefore while the machine processes any one of the jobs, the others have to wait. Manufacturing a typical device involves approximately 1-3 weeks of raw processing time (also called theoretical *cycle time*), but because of these queueing delays, the actual cycle time is 3-6 times the theoreti-

cal cycle time. Buying extra capacity to minimize these delays is usually not feasible due to high capital costs. Therefore traditionally, emphasis is put on high utilization of the equipment via myopic scheduling rules at the machines.

Due to capacity constraints, another goal of production control is to minimize cycle time[1] variability in the fab. The basic idea behind this approach is to stabilize work queues at the machines to minimize the variance in production requirements. The variance in queue lengths arises due to a multitude of reasons. Perhaps the most fundamental of them is the variance in demand itself. The demand process is a non-stationary process with large swings in demand, usually every quarter. Since the demand fluctuations occur every couple of months which is about as long as the cycle time of fab, a steady state production scenario is never attained in the fab and therefore, production requirements often fluctuate dramatically.

Another reason for work flow variability is the high frequency of unpredictable failure of the equipment. Failure of a machine to perform causes work queue to get larger in front of it while the downstream machines deplete their work queues. Jobs that have long processing times (e.g. diffusion recipes) usually increase the work flow variability as well. To attain a reasonable throughput rate, machines performing these jobs usually process many lots at the same time. These machines output large loads for downstream machines at irregular intervals, which causes alternating periods of starvation and long queues. Finally, lot-sizing also has a significant impact on the work flow variability. Larger lot sizes make the work flows less continuous and more lumpy.

The variance in work flows directly impacts the lead times for different products. Average lead time is composed of set-up time, variable production time and waiting time.

---

1. The time between job release and its completion, typically random. According to **Little's law**, Average flow time = (Average WIP) / (Average Throughput)

Waiting times tend to be arbitrarily long in capacity-constrained situations. Some of the waiting time is caused because multiple products wait for processing time on the bottle-neck machines.

Common managerial reaction to the stochasticity of the lead times is to increase the planned lead times to ensure job completion within the lead time. Increasing the lead times results in an increase in work-in-process inventories. Also, long lead times make it harder to determine production schedules because forecasts become inaccurate. It would be extremely valuable for personnel in charge of operations to have access to an analysis tool which can relate the planned lead times with the production requirements.

In this thesis we explore this relationship through an analytical model of the fab. Our hope is to characterize production requirements and work-in-process levels in the fab for a deterministic set of planned lead times. Managers could potentially use such a model to understand the impact of increasing/decreasing lead times on production schedules. Such a model could allow them to answer a variety of questions. For example, if they aggressively reduce the lead times for a process flow, such a model could help them determine if they had enough equipment capacity to handle the additional work loads. Also, it would help to determine where to hold inventory in the fab to reduce the variability in the fab.

Due to the manufacturing complexities described above, it is extremely hard to develop a tractable analytical model to understand fab operations at the level of scheduling. As part of this thesis work, we have developed a model that provides insight into fab operations at the tactical level. The wafer fab model is based on a discrete-time continuous flow job shop model developed in Graves (1986). This model characterizes steady state production and work-in-process levels at each equipment for any demand distribution.

In Chapter 2 we briefly discuss the mathematical details of this model and explain how we developed the model of the fab. The chapter has three main parts: data collection, parameter estimation and model validation.

In Chapter 3 we describe how our model can be used to understand the operations of the fab. We describe how to characterize the impact of changes in demand, as well as changes in the planned lead times for the equipment. We show how to use the model for examining the trade-off between investments in work-in-process inventory and capacity, for setting planned lead times, for determining the benefits of eliminating input variability in the fab, and for understanding the impact of changes in production mix and volume on inventory and capacity requirements.

In Chapter 4 we describe a simulation study which was performed to determine optimal scheduling policies for equipment with large load sizes and long processing times. We will show that emphasis on high utilization of such equipment can lead to an increase in cycle time and a decrease in total throughput of the factory.

Chapter 2

Tactical Planning: Model Development

In this chapter, we describe the development of a tactical planning model for the Wilmington wafer fab, Analog Devices Inc. First, we describe an analytical model that we use to model the wafer fab as a job shop. Then, we describe our representation of the fab in the model and finally, we compare the model with the actual state of the fab using real data.

2.1 The Basic Model

In this section, we provide a summary of the tactical planning model from Graves (1986) and then present an extension from Graves (1988). This model is appropriate for a discrete-parts manufacturing system such as a wafer fab. Graves (1986) presents a linear - systems model for tactical planning for such a production environment. This model provides a characterization of work flows, flow times and production requirements of a production facility.

We first describe how the model represents the behavior of a single workstation. Then we describe how the model links a network of workstations together via the work flow between stations.

The Work Station Model

Each workstation is described by a discrete-time, continuous variable model which involves three random variables.:

A_{it} Amount of work arriving at station i at the start of period t.

Q_{it} Queue of work at workstation i at the start of period t, including the arrival of A_{it}

P_{it} The amount of production completed by workstation i during period t.

Each workstation i is described by two linear relationships:

$$Q_{it} = Q_{i,t-1} + A_{it} - P_{i,t-1} \tag{2.1}$$

and

$$P_{it} = \alpha_i Q_{it} \tag{2.2}$$

Equation (2.1) is a representation of conservation of work flow. Equation (2.2) is a control rule where α_i is the control parameter and $0 < \alpha_i < 1$. This equation assumes that the workstation has enough capacity to accommodate any fluctuations in production requirements. As the queue of work grows longer, the workstation works proportionally harder. α_i is a smoothing parameter as algebraic manipulation of equations (2.1) and (2.2) yield a first order smoothing equation that governs the production at the workstation:

$$P_{it} = \alpha_i A_{it} + (1 - \alpha_i) P_{i,t-1} \tag{2.3}$$

The inverse of the smoothing parameter, $N_i = 1/\alpha_i$, can be interpreted as the planned lead time for the workstation. This assignment of control parameter is intuitive. If the planned lead time for the work-in-process is N days, then $1/N^{th}$ of the work should be finished each day. This assignment of the control parameters is significant because of the importance of planned lead times in most scheduling systems, especially MRP systems such as the one at ADI. These systems use estimates of lead times for the workstations to

project the work flow through the shop. Based on these projections, delivery dates are quoted to the customers.

The Arrival Process Model

The model assumes that work flow displays Markovian behavior. We assume that movement of jobs from one station to the next and release of new jobs occur at the start of a period t. Let A_{ijt} denote the flow of work from workstation j to workstation i at the start of period t. It is given by:

(2.4)

$$A_{ijt} = \phi_{ij} P_{j,t-1} + \varepsilon_{ijt}$$

where ϕ_{ij} is a scalar that denotes the expected amount of work for i per unit of production at j and ε_{ijt} is a random variable that characterizes the variability associated with this work flow.

The total work flow to station i at the start of period t is hence given by:

(2.5)

$$A_{it} = \sum_{j} \phi_{ij} P_{j,t-1} + \varepsilon_{it}$$

where ε_{it} represents the arrivals that are not predicted from the production levels of the previous periods, i.e., new arrivals and noise in the flow from other workstations.

(2.6)

$$\varepsilon_{it} = \sum_{j} (\varepsilon_{ijt}) + N_{it}$$

N_{it} is a random variable, with known mean and variance, and it reports the new work that is released at time t and goes first to station i. In this model, elements of time series $\{\varepsilon_{ijt}\}$ are assumed to be zero mean i.i.d for each workstation.

The complete model of a workstation is given by equations 2.1, 2.2, and 2.5.

Network Model: Single Process Flow

To analyze the work flow in a job shop as described by equations 2.3 and 2.4, vector notation is used to describe a network of n workstations. In equations 2.7 and 2.8, P_t, A_t, and ε_t are n-column vectors, I is the identity matrix, D is the diagonal matrix with $\{\alpha_1, \alpha_2, \ldots. \alpha_n\}$ on the diagonal and Φ is the n-by-n matrix with elements ϕ_{ij}. One implicit assumption that is made here is that all the jobs are of the same type and can be modeled by a single Φ matrix. Later, we show how we have relaxed that assumption in the model.

$$P_t = (I - D)P_{t-1} + DA_t \tag{2.7}$$

$$A_t = \Phi P_{t-1} + \varepsilon_t \tag{2.8}$$

Substituting 2.8 into 2.7, we obtain

$$P_t = (I - D + D\Phi)P_{t-1} + D\varepsilon_t \tag{2.9}$$

Assuming an infinite history of the job shop, we can recursively express 2.9 as a geometric series as shown in 2.10. In order to characterize the joint distribution of the production vector P_t, the noise vector ε_t is assumed to have a mean $\mu = \{\mu_1\ \mu_2, \ldots, \mu_n\}'$ and a

covariance matrix given by $\Sigma = \{\sigma_{ij}\}$. From the definition of ε_{it} given in (2.6) we note that μ_i corresponds to the expected amount of new arrivals at workstation i ($\mu_i = E(N_i)$.)

(2.10)

$$P_t = \sum_{s=0}^{\infty} (I - D + D\Phi)^s D\varepsilon_{t-s}$$

The expectation of the production vector, $\rho = \{\rho_1, \rho_2, ..., \rho_n\}$, is given by

(2.11)

$$\rho = \sum_{s=0}^{\infty} (I - D + D\Phi)^s D\mu = (I - \Phi)^{-1}\mu$$

for which we assume that the spectral radius (maximum absolute eigenvalue) of Φ is less than 1. Graves (1986) has a brief discussion on why this assumption is necessary for the existence of steady state results.

Equation 2.12 describes the covariance matrix S, of the production vector P_t. The S matrix gives the production variances at the individual workstations as well as the covariances across pairs of workstations.

(2.12)

$$S = \sum_{s=0}^{\infty} (I - D + D\Phi)^s D\Sigma D(I - D + D\Phi)'^s$$

To approximate the infinite series in equation 2.12, we can use the recursion as shown in equation 2.13. Repeated application of this recursion can be used to approximate the S matrix.

$$S_{2n} = S_n + (I - D + D\Phi)^n S_n (I - D + D\Phi)'^n \tag{2.13}$$

Equations 2.14 and 2.15 describe how to characterize the queue lengths at each workstation.

$$Q_t = D^{-1} P_t \tag{2.14}$$

$$E(Q_t) = D^{-1} \rho \qquad and \qquad Var(Q_t) = D^{-1} S D^{-1} \tag{2.15}$$

Network Model: Multiple Flows

In complex manufacturing scenarios such as the one at ADI, it is not a fair assumption to treat all the jobs at a workstation as identical. Graves (1986) briefly describes how to relax this assumption when there are a few distinct types of jobs with different process flows and production requirements. To relax this assumption, we need to identify different work flow matrices (Φ_k) for each job type k. Therefore, each workstation has a separate queue of work for each part type and a separate control rule to determine production levels for each job type. Therefore equation 2.1 can be restated as

$$P_{ikt} = \alpha_{ik} Q_{ikt} \qquad and \qquad P_{it} = \sum_k P_{ikt} \tag{2.16}$$

where Q_{ikt} is the queue of work in front of station **i** for job type **k** and α_{ik} is the control parameter for job type **k**.

Consideration of Lot Sizing

The basic model described above assumes a continuous work flow but that is often not a fair assumption to make, especially for a wafer fab scenario where work flows in *lots* from one workstation to the next. Large lot sizes make the work flow irregular and cause lumpiness in the flow. In this section, we describe how Graves (1988) addresses how to capture the impact of lot sizing in a job shop.

Suppose that the workstation j does production in lots of size m_j, where m_j is expressed in units of work content at station j. Let p_{ij} be the probability that any lot completed at workstation j generates a lot arrival to workstation i. Here, Graves (1988) assumes that p_{ij} is less than 1. Also, let L_{jt} be the random variable defined as the number of lots completed at workstation j during period t. L_{jt} is given by:

$$L_{jt} = \frac{P_{jt}}{m_j}$$

For a given realization of L_{jt}, let L_{ijt} be the number of lots completed at workstation j that go to workstation i for the next process step. L_{ijt} is assumed to be a random variable with p_{ij} probability of success, and L_{jt} trials. Expectation of this random variable is given by:

$$E(L_{ij}) = \frac{p_{ij}}{m_j} E(P_j) \tag{2.17}$$

Here, Graves (1988) assumes that P_j/m_j is an integer which is inconsistent with the basic model in which the work flow and production are assumed to be continuous. How-

ever, it is a reasonable assumption when E(Pj)>>m_j. As shown in Graves (1988), the variance of L_{ijt} is given by

(2.18)

$$\sigma^2(L_{ij}) = \left(\frac{p_{ij}}{m_j}\right)^2 \sigma^2(P_j) + \left(\frac{p_{ij}}{m_j}\right)(1 - p_{ij})E(P_j)$$

This description of a *lumpy* work flow can be incorporated in the arrival process model to quantify the impact of lotsizing on work flow variability. Every lot L_{ijt} that goes to workstation j from workstation i has a $m\phi_{ij}$/pij units of work. For the rest of the analysis, we look at steady state and therefore drop the time subscript. Then, the variance of the arrival stream from workstation j to the workstation i is given by:

(2.19)

$$A_{i_j} = \left(\phi_{ij}\frac{m_j}{p_{ij}}\right)^2 \sigma^2(L_{ij})$$

In steady state, the variability of the arrival stream, $\sigma^2(A_i)$ is given by:

$$\sigma^2(A_i) = \sum_j \left(\phi_{ij}\frac{m_j}{p_{ij}}\right)^2 \sigma^2(L_{ij}) + \sigma^2(\varepsilon_i)$$

Here, we assume[1] that the lots arriving to workstation i from each workstation j are independent of each other, i.e. the time series for arrivals to work station is not correlated. Then, substituting for the variance of L_{ij} from equation 2.18, the variability of the arrival stream is:

1. In our wafer fab model, this assumption is OK because almost always, workstation i only gets lots from a single workstation j.

(2.20)

$$\sigma^2(A_i) = \sum_j (\phi_{ij})^2 \left[\sigma^2(P_j) + \left(\frac{m_j}{p_{ij}}\right)(1 - p_{ij})E(P_j)\right] + \sigma^2(\varepsilon_i)$$

As can be seen in equation 2.18, the production variability at workstation i is the sum of three components. First, there is the variance due to production variability at each workstation, given by: $(\phi_{ij})^2\sigma^2(P_j)$. Second, there is variance due to lotsizing at the upstream workstations which is given by: $(\phi_{ij})^2(1\text{-}p_{ij})(m_j/p_{ij})E(Pj)$. Finally, there is variance due to the random noise.

Now, in order to compute the impact of lot sizing on production variability, let us first develop a relationship for the variance of production at a workstation, and the variance of arrival stream of its work. Recall from equation 2.3, P_{it}, random variable for production at workstation i during time period t, is given as:

$$P_{it} = \alpha_i A_{it} + (1 - \alpha_i)P_{i,\, t-1}$$

By repeated substitution, P_{it} can be given by:

$$P_{it} = \sum_{k=0}^{\infty} \alpha_i (1 - \alpha_i)^k A_{i,\, t-k}$$

Here, let us assume that the time series of work arriving to a workstation i is not correlated. Then in steady state, we drop the time subscript, and the variance of production at workstation i is given by:

$$\sigma^2(P_i) = \frac{\alpha_i}{(2 - \alpha_i)}\sigma^2(A_i)$$

Substitution for $\sigma^2(A_i)$ in the expression for production variance, we get:

$$\sigma^2(P_i) = \frac{\alpha_i}{(2-\alpha_i)}\left(\sum_j (\phi_{ij})^2\left[\sigma^2(P_j) + \left(\frac{m_j}{p_{ij}}\right)(1-p_{ij})E(P_j)\right] + \sigma^2(\varepsilon_i)\right) \qquad (2.21)$$

From equation 2.22, it can be seen that the impact of lot sizing on production variability is given by:

$$\frac{\alpha_i}{(2-\alpha_i)}\sum_j (\phi_{ij})^2\left(\frac{m_j}{p_{ij}}\right)(1-p_{ij})E(P_j)$$

2.2 Wilmington Wafer Fab Model

In this section we describe how we developed a tactical planning model to characterize the operational behavior of the Wilmington wafer fab, Analog Devices Inc.

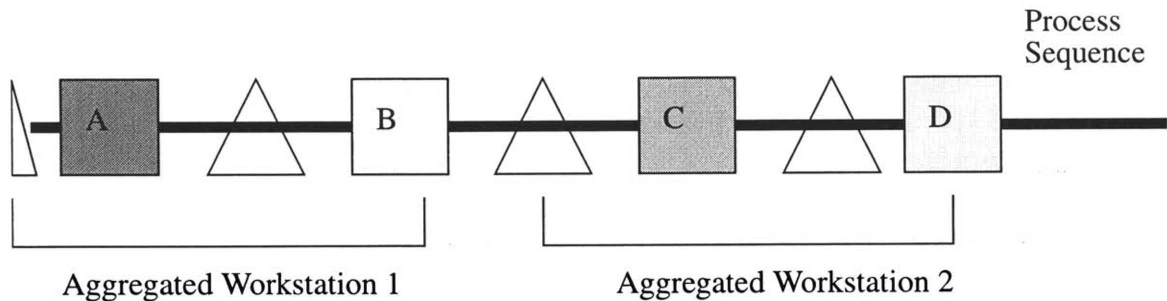


Figure 2.1: Segments of Process flows are lumped and treated as aggregated workstations for simplicity.

Model Development

As we mentioned previously, there are over 300 distinct process flows (henceforth referred to as routings) that exist in the wafer fab. Moreover, each of these routings consists of a few hundred distinct processing steps (*recipes*). We thought that this level of

detail was unnecessary for understanding the fab operations at a tactical level and it might actually provide misleading results as well. Therefore, we decided to aggregate the process flows in order to reduce the complexity of our model. At the Wilmington wafer fab, the routings are grouped into 22 families based on the similarity of their process sequences. In our model, we decided to treat all the routings within a family as identical and thus, reduced the number of work flows in our model from over 300 to only 22. We carefully scrutinized all the routings within a family and chose a *representative* routing which best represented the process sequence of all routings in that family.

One of the assumptions in the tactical planning model is that movement of inventory from one workstation to another as well as the arrival of new jobs into the shop occurs at the beginning of a time period. This time period should be long enough so that each workstation can finish a few jobs during the period. At the same time, it is essential that this time period is short enough so that it would be highly unlikely for a job to move through two successive workstations in one period. Unfortunately, it was not possible to find such a time period because while certain pieces of equipment in the fab take only a few minutes to perform a recipe, certain others take many hours to perform a recipe. To overcome this difficulty, we decided to aggregate multiple processing steps and treat them as one workstation in our model. This aggregation scheme is described pictorially in figure 2.1. In the example shown, the first two machines are aggregated as workstation 1 and the last two machines are treated as workstation 2. The work-in-process for workstation 1 represents the sum of the work-in-process in front of machines A and B. Likewise, the work-in-process for workstation 2 represents the sum of the work-in-process in front of machines C

and D. Also, the throughputs of workstation 1 and 2 represent the throughputs of machines B and D respectively.

There was no intuitive way to determine which set of machines in a process sequence should be grouped together as one workstation. We spent a lot of time talking to various fab personnel to determine which type of equipment induced a lot of variability in the downstream production requirements. Based on these talks, we identified two main causes for this variability. First, if an equipment takes a long time to finish a job, then the downstream equipment faces alternating periods of large work-in-process and no work-in-process. Second, if an equipment takes large batch sizes, then also it generates large amounts of work-in-process at irregular intervals for the downstream equipment. We tried to capture these sources of variability in our model by having such equipment at the boundaries of our *aggregate workstations* (henceforth referred to as workstations). We chose 8 diffusion furnaces to determine the workstations for each routing. We identified each workstation by the furnace which marked the end of the segment that it represents. These furnaces were chosen because they process as many as 8 lots simultaneously and recipes at these furnaces were as long as 24 hours. The choice of diffusion furnaces was reasonable because they take large load sizes and they also take a long time to perform recipes, thus contributing heavily to the work flow variability.

The nature of semiconductor manufacturing is such that the diffusion steps are usually quite a few processing steps away from each other in any given processing sequence. Therefore, it is reasonable to assume an underlying time period for inventory transfers between the workstations in our model. Most workstations have cumulative raw processing times of at least a few hours. Since, the cycle time for the process flows is X*(the

actual processing time) where X is some constant usually greater than 4, the actual time it takes for jobs to move from one workstation to the next is X times those cumulative processing times. Based on this analysis, we chose 1-day as the underlying time period for our model. In figure 2.2, we show a histogram of planned lead times for the workstations that appear in our 22 routings when X = 4. The figure indicates that most workstations have planned lead times of over a day in our representation. We can see that there are 26 workstations that have lead times of less than or equal to a day and 115 workstations have lead times longer than one day.

As described in chapter 1, routings in a wafer fab are highly re-entrant in nature. Therefore, some equipment appears multiple times within a routing. However, the processing requirements on a given equipment vary depending upon the stage in the routing. Therefore, if two lots of the same routing arrive at an equipment and one of the lots was further down the routing than the other, then most probably the two lots can not be processed simultaneously by the equipment. We found this to be always the case for the diffusion furnaces that we chose to define our workstations.

Table 2.1: Representation of a Process Flow

Workstation (j)	*Furnace #*	*Recipe #*	*Step #*	*Processing Time: hours*
1	7	1	4	7
2	8	1	50	40
3	4	1	145	33
4	4	2	204	21
5	2	1	206	4
6	4	3	209	2
7	2	2	238	12
8	3	1	240	5

To clarify our modeling approach, we describe in detail our representation of one of the work flows in the fab as shown in table 2.1. Recall from before, each work flow is a sequence of steps and each workstation in our model represents the series of steps after a diffusion recipe, up to and including the next diffusion recipe. Each furnace in the flow is designated by a 2-touple (x,y) where x represents the furnace index and y represents the recipe index.

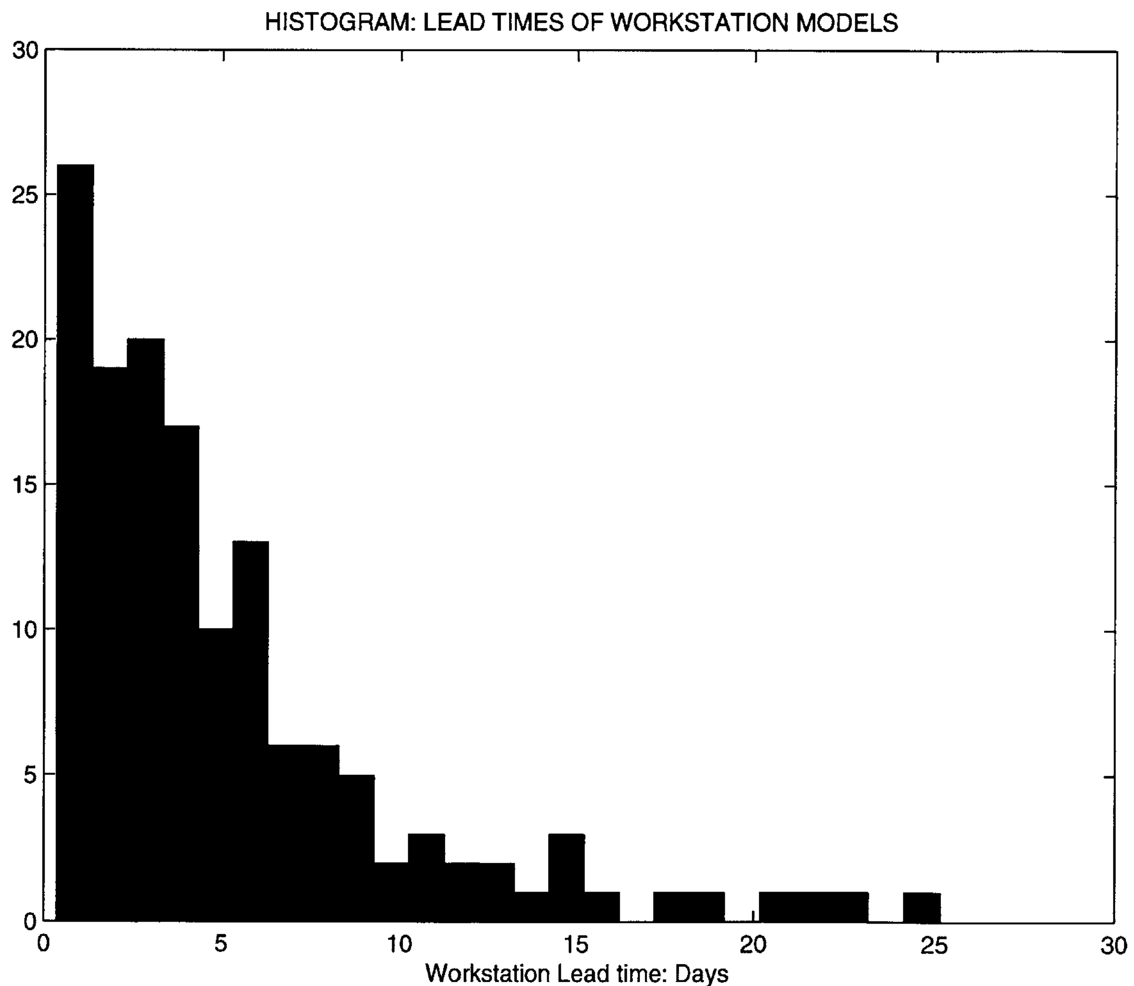


Figure 2.2: Lead times of most of the 'aggregated' workstations in the model are longer than a day.

From the set of 8 furnaces that we chose to represent the fab, only furnaces 2,3,4,7 and 8 appear in this particular process flow. The table indicates that there are 8 different 2-touples (furnace #, recipe #) in this work flow and therefore in our model this flow is modeled by 8 workstations. In the work flow, furnace 4 appears three times but the recipes are different for each of those times and therefore each appearance of furnace 4 is treated as a separate workstation. The first workstation, given by 2-touple (7,1), represents the first 4 steps of this work flow which take 7 hours of theoretical processing time. Similarly, the second workstation represents steps 5 to 50 of the work flow which take approximately 40 hours of raw processing time and so on. Note that workstations 3 and 4 are both at furnace 4 but since they have different processing requirements at that furnace, they are treated separately. In this work flow representation, the assumption of a 1-day underlying time period is reasonable for all workstations except the workstation 6 (steps 207-209). This is because actual lead time of these workstations is usually at least 4-6 times the theoretical lead times given in table 2.1.

The work flow matrix (Φ) for this example in our model assumes that one wafer at a workstation produces one wafer of work for the downstream workstation. The Φ matrix is given by an 8x8 matrix as shown below:

$$\Phi = \begin{bmatrix} 0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&1&0 \end{bmatrix}$$

The elements of this Φ matrix indicate the expected number of wafers workstation j yields for workstation i. For example, one wafer of work at workstation 3 produces 1

wafer of work for workstation 4 ($\Phi_{43} = 1$). Since work leaves the system from workstation 8, Φ_{j8} is 0 for all j.

In our computations, the mean and variance of production at a furnace are given by the sums of means and variances of production at all the workstations representing that furnace. For example, mean production at furnace 4 is given by the sum of mean production at workstations 3,4 and 6. Therefore, on average if 2 lots are released into the fab for this work flow every day, then in steady state workstations 3,4 and 6 will produce 2 lots every day. Hence, furnace 4 will have average production requirements of 6 lots of this work flow. Total production requirements at furnace 4 will similarly be the sum of production requirements of all the work flows

.

Parameter Estimation

In this section we describe our efforts to characterize the aggregate behavior of production operations in the Wilmington wafer fab. We collected information on all the process flows for approximately 21 weeks. For each day, we obtained three measures from the wafer fab: (a) the aggregate starts for each family, (b) the aggregate work-in-process (WIP) of each family in front of each piece of equipment; and (c) the aggregate output at each piece of equipment. These measures were collected in terms of number of wafers.

There are a few qualifications of this data that need to be mentioned.

First, the scheduler was the source of data on work-in-process and starts whereas PROMIS/MES system was the source of data on aggregate production. A CRON script was written to periodically capture WIP snapshots of the fab from the scheduler for 21

weeks. The WIP snapshot was captured at midnight every day. The data on aggregate production was downloaded from PROMIS for the same 21 weeks. Initially, the scheduler was also used to obtain production data but it provided erroneous results for the production because its information is based on hourly snapshots of the fab that are extracted from the PROMIS. Therefore, we resorted to extracting the production data from PROMIS.

Second, three separate times the CRON script failed to extract data in the 21 week period due to scheduler (see section 1.4) shutdowns. Those three days were ignored in the time series analysis of WIP and starts. However, we still expect this analysis to be reasonably accurate.

Third, work-in-process which is put on hold for engineering evaluations was not captured in the aggregate WIP data. This is because the evaluation steps are not logged into the PROMIS system and hence information on work-in-process at these steps could not be captured. We expect the calculated levels of total WIP on the floor to be lower than the actual levels because of this discrepancy.

Fourth, the starts for 10 of the 22 families were minimal during the 21 week period and hence they were ignored in the time series analysis.

Finally, over the 21 week period, the fab was not in a steady state situation. There were changes in the production mix caused by a continuous increase in the six-inch wafer production. Also, there were ongoing efforts to improve scheduling through out the fab; for example, new scheduling rules were implemented in the Diffusion sector during this period.

We characterized the distribution of starts into the fab based on the 21 week history of the daily starts that we collected. We assumed that it was a reasonably long period to

justify the assumption of having infinite history of the fab available to us. For each family, we characterized the demand by the mean and variance of its daily starts over the 21 week history. The demand (μ) vector (equation 2.11) for each family in the model was a nx1 vector where n was the number of workstations in the work flow. All the entries of this vector were set to zero except the first entry. This entry, which represents the first workstation in the work flow was set to the mean demand for that family. Thus, in our model we assume that all the work of a particular routing is started at the first workstation in that routing and no work is introduced in the later stages of the routing. We constructed the covariance (Σ) matrix for each family (equation 2.12) in a similar manner. This matrix is a nxn matrix in our model with all zeros except the Σ_{11} which was set to the variance of the starts for that family. Here, we assume that the only variability in the work flow is introduced in the very beginning. In a fab scenario, there are other factors such as yield and rework that exaggerate the work flow variability. We did not capture these second order effects in our model.

Equations 2.17-2.23 show how we captured the impact of lot sizing in our model. In order to use this approach, we needed two parameters. First, we needed to know the lot size m for each family which was made readily available to us by the personnel in the fab. Second, we needed to estimate the parameters p_{ij} for each flow. Recall from before that p_{ij} is the probability that a lot completed at workstation j moves next to workstation i. At first glance, it seems that this parameter should be set to one because if a lot has been processed, it should move downstream. However, often some wafers in the completed lots do not meet the desired specifications and are therefore reworked. When a few wafers from a lot have to be reworked, the entire lot has to wait before it is transferred to the downstream

workstation. To determine the p_{ij} values, we looked at time series of percent of lots reworked at the equipment. The data suggested that about 3-7 percent of the lots at most of the key equipment such as photo-lithography and diffusion stations were reworked. We talked to several personnel from the operations department in the fab and they suggested that it was reasonable to assume a X percent rework rate through out the work flow. Based on their suggestion, we set the p_{ij} to (1-X/100) if there was a work flow from j to i. Note, this formulation also suggests that p_{jj} should be X/100 indicating that X percent of the lots leaving workstation j return to it for rework. We ignore the impact of rework on the variability of production and treat rework as a loss (p_{jj} = 0). We decided to ignore rework because for small flows (e.g. p_{jj} = X/100), our formulation considerably over-estimates the impact of lotsizing on production variability. Since we do not estimate the variability due to lot sizing in conjunction with other sources of variability, as shown in equation 2.21, it becomes unbounded as the flow of work from j to i becomes minimal (p_{ij} goes to 0).

We now describe how we estimated the planned lead times for different workstations in each work flow. Consistent with the assumptions of our underlying model, we also assume that the fab is in steady state. Therefore, we assume that each workstation has enough capacity to meet the production requirements for the current demand. To further clarify this assumption, let's refer back to table 2.1. In the work flow depicted in table 2.2, if the starts are 40 wafers each day, then we assume that in steady state, workstations 1-8 are capable of producing 40 wafers each day. We compute the lead time for each workstation in a very intuitive manner based on the daily production requirements and the amount of work-in-process in front of it. For example if a workstation is required to process 80 wafers each day and on average, it has 300 wafers of work-in-process sitting in front of it,

then the lead time for the workstation is roughly 4 days in steady state. The average work-in-process levels in front of each workstation were determined from the 21 weeks history.

Table 2.2: Input Parameters for Work Flow shown in table 1.

Workstation (j)	*Furnace #*	*Recipe #*	*Step #*	*Planned Lead Time*	*Demand* μ_j	*Cov* Σ_{jj}	$p_{j+1,j}$
1	7	1	1-4	1	45	1369	1-X/100
2	8	1	5-50	13	0	0	1-X/100
3	4	1	51-145	21	0	0	1-X/100
4	4	2	146-204	12	0	0	1-X/100
5	2	1	205-206	1	0	0	1-X/100
6	4	3	207-209	1	0	0	1-X/100
7	2	2	210-238	5	0	0	1-X/100
8	3	1	239-240	2	0	0	0

Table 2.2 shows the input parameters for the work flow shown in table 2.1. As mentioned before, there are 8 workstations representing this flow. The planned lead times were determined using the procedure described in the earlier section on parameter estimation. Workstations 2-4 have very long lead times which is reasonable because each represents a large number of processing steps as shown in the table. The estimated cycle time for this flow, which is the sum of planned lead times for each workstation, is 56 days. As can be seen, wafers for this work flow are started in furnace 7 at an average rate of 45 wafers/day ($\mu_j = 45$). The standard deviation of the starts for this flow is 37 and hence the first entry in the covariance matrix is the square of this value ($\Sigma_{11} = 1369$). The rest of the entries in the covariance matrix are all zeros. The last column in table 2 provides the $p_{j+1,j}$ values which are used to determine the impact of lot sizing on work flow variability (see equation 2.21).

The model described above was implemented using MATLAB[1] and Awk software. The routing files were read using Awk scripts and work flows for each routing file were

generated using Awk scripts. Steady state demand and WIP profiles were also generated using Awk scripts. All this information was read into a MATLAB application to characterize the operational behavior.

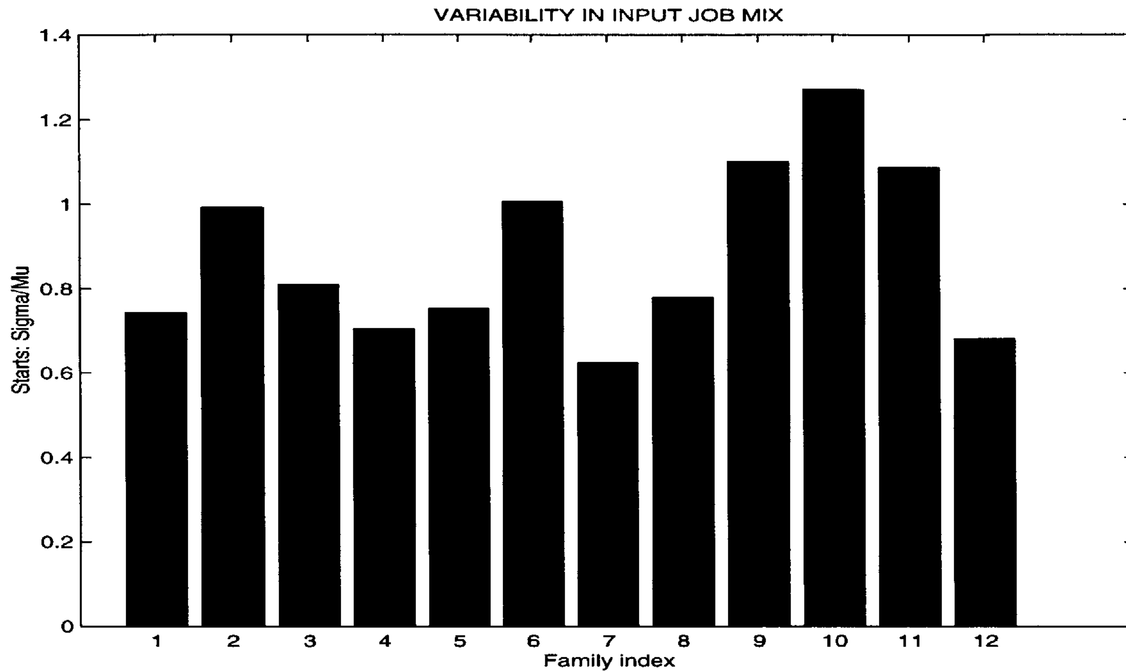


Figure 2.3: Coefficient of variation of the input job mix

Model Validation

To obtain insight into the behavior of the process flows from the data that we collected, we performed various time series analysis on these aggregated flows. We plotted the time series and computed summary statistics for daily starts, production output and WIP for these sectors. We report the normalized statistics in this thesis. Figure 2.3 shows a bar graph of the ratio of standard deviation of the starts to the mean of the starts for each

1. MATLAB is a licensed product of Mathworks Inc.

family for the 21 week period. This ratio is a measure of the variability in production requirements caused by the fluctuations in the demand process. As can be seen in the figure, the σ/μ ratios are greater than one-half for all 12 flows shown. We only show the coefficient of variation for 12 process flows because the other 10 flows had minimal demand

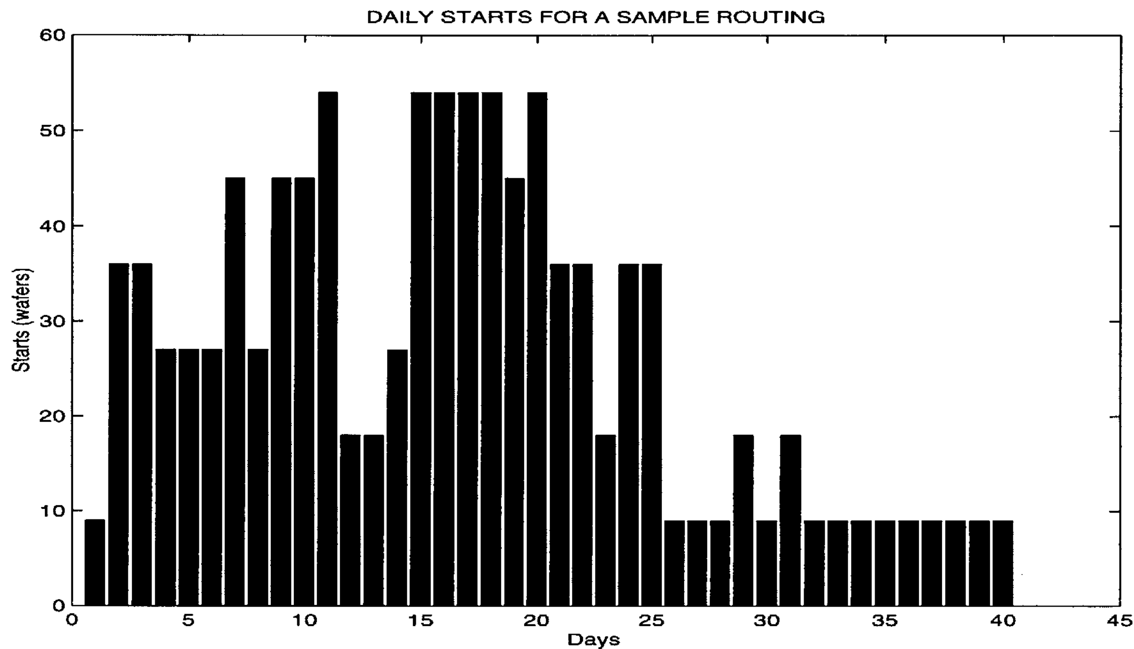


Figure 2.4: Time Series of daily starts of a sample work flow

during the period that we collected data. The σ/μ ratio for the total starts in the fab was 0.32. The ratio is lower for total starts because even though the number of lots started each day varies greatly for a family, the total number of lots started each day in the fab is much less variable. The starts into the fab have such a high σ/μ ratio because wafers are started in multiples of lots, each lot with as many as 20 and as few as 9 wafers. Figure 2.4 shows a bar chart of time series of starts for one of the process flows. This figure shows starts for that particular family were in lots of 9 wafers. As can be seen from this example, if the

number of lots released into the factory is increased (or decreased) by one, there is a large change in production requirements.

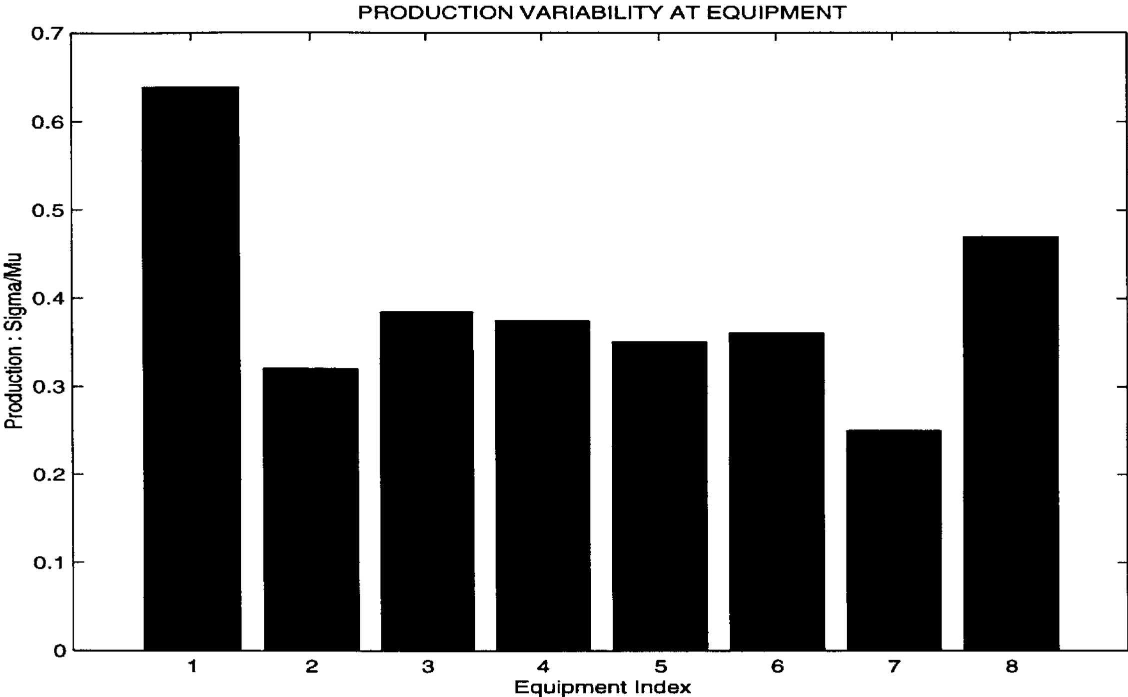


Figure 2.5: Coefficient of Variation of Production at the furnace equipment: Real Data

We expected large variations in production output in all the workstations because of this high variance in the starts that percolates down the shop floor. The behavior of production at the furnaces that define the boundaries of the workstations is plotted in figure 2.5. The σ/μ ratio of daily production at these machines ranges between 0.3 and 0.6 which is quite significant but not as high as the variance of the starts. It is our belief that the variability in daily production of the workstations is not so high because of the presence of work-in-process. WIP decouples the production workstations and dampens the variability in production requirements. The variability of production requirements is further damp-

ened by the capacity constraints across these workstations which put an upper bound on how much work any workstation can generate for its downstream workstation.

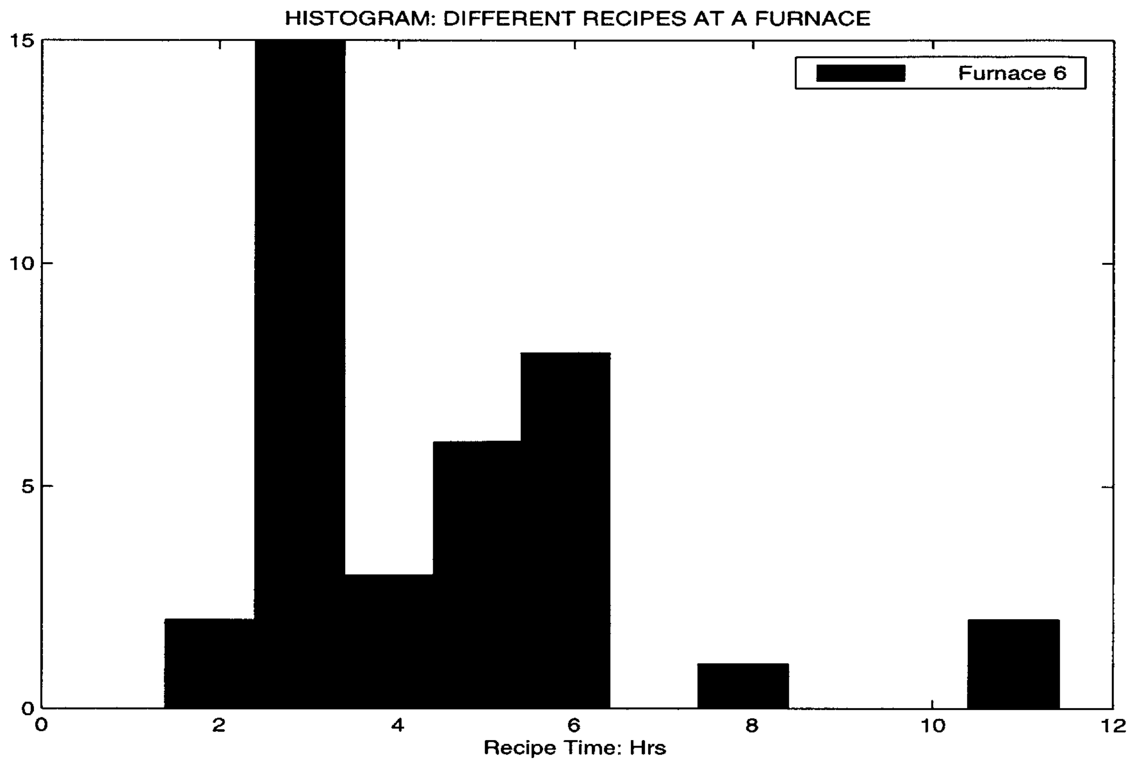


Figure 2.6: Different recipes require different processing times at the same furnace

We now describe how we estimated the capacity of the diffusion furnaces. Different recipes that are performed at a given furnace usually have different processing times associated with them. In figure 2.6, we show a histogram of processing times of different recipes at furnace 6. As can be seen, different jobs at this furnace can take anywhere from 2 to 11 hours. To estimate the capacity of each furnace, we first determined the mean

Table 2.3: Estimated Capacities for the Diffusion equipment

Furnace Index	ρ *(hrs)*	*b (wafers)*	*Furnace Qty.*	*Theoretical Capacity*	*Max. Production*
1	2	96	2	2300	400
2	3	150	2	2400	1240
3	5	122	2	1170	850
4	2	72	2	1730	1300
5	8	184	4	2200	400
6	7	184	3	2650	820
7	5	138	3	2000	1390
8	12	184	4	1470	550

Table 2.4: Comparison of WIP Profile (Actual Vs. Model) for Flow 5

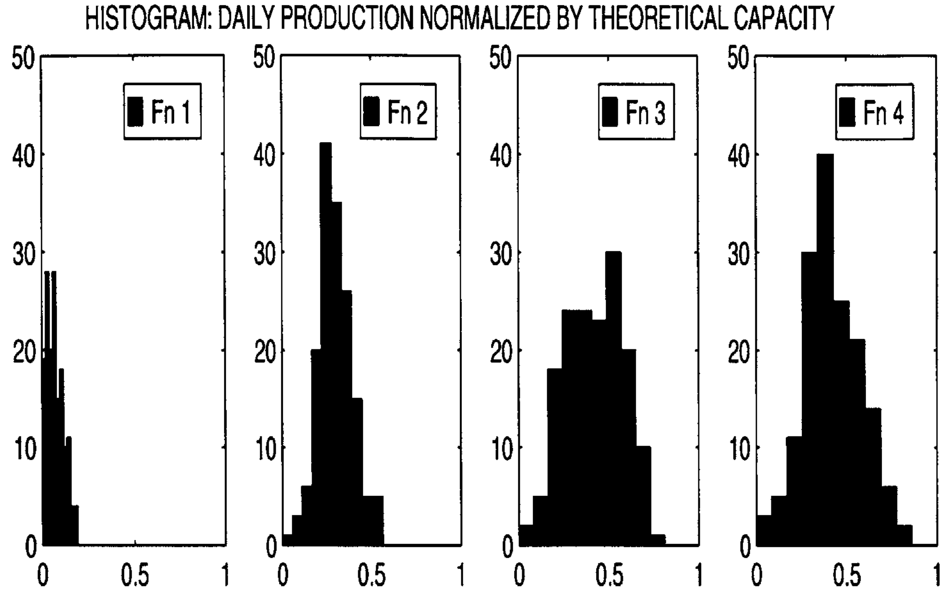
Workstation	*Furnace*	*Lead Time*	μ	*(Model) WIP*	*(Actual) WIP*
1	7	2	58	116	131
2	7	8	0	463	436
3	4	20	0	1158	1177
4	2	1	0	58	30
5	4	1	0	58	52
6	2	4	0	232	227
7	3	1	0	58	42
8	2	1	0	58	53

processing time (ρ) of all the recipes at that furnace. For any given furnace, ρ was determined simply by taking an average of the processing times (in hours) of all the different recipes that are performed at that furnace. We didn't take a weighted average (based on starts) because the starts mix changes from day to day. The personnel in the fab were much more confortable with a direct average. We also determined the maximum batch size (b) of each furnace. The theoretical capacity of the furnace was thus defined as ($b*24/\rho$) wafers.

For example a furnace with batch size of 100 wafers and a average recipe time of 3 hours is capable of making 8 runs in a day. Therefore, it has a theoretical capacity of 800 wafers.

Table 2.3 shows the estimated theoretical capacities of all the furnaces that were used to define the workstations in our model. In order to determine which furnaces were capacity-constrained, we looked at the daily production at all the furnaces. We averaged the 10 largest outputs of each furnace and denoted that the maximum production achieved by that furnace under current demand conditions. As can be seen in table 2.3, furnaces 3, 4 and 7 used over 70% of their theoretical capacities. Figure 2.7 shows a histogram of the daily outputs for all the furnaces normalized by their theoretical capacities. All the furnaces except 3,4,and 7 have production requirements well below their theoretical capacities. Furnace 1 has many days with minimal output which indicates that it either breaks down frequently or it is starved frequently.

In table 2.4, we compare the outputs of our model to actual data for one of the process flows. Using the modeling approach, this flow was represented by 8 workstations as shown in the table. Workstations 1 and 2 both represent work flow sequences ending in furnace 7 but they are treated as distinct workstations because they represent different processing requirements. The lead times for each workstation were determined empirically for each workstation as described in the section of parameter estimation. Based on these lead times, the total cycle time for this work flow according to our model is approximately 38 days. We determined the actual cycle time of this work flow to be 42 days based on real data. One of the reasons for error in cycle time estimates is that we assume there is an underlying period of one day. As can be seen in table 2.1, workstation 6 given by 2-touple



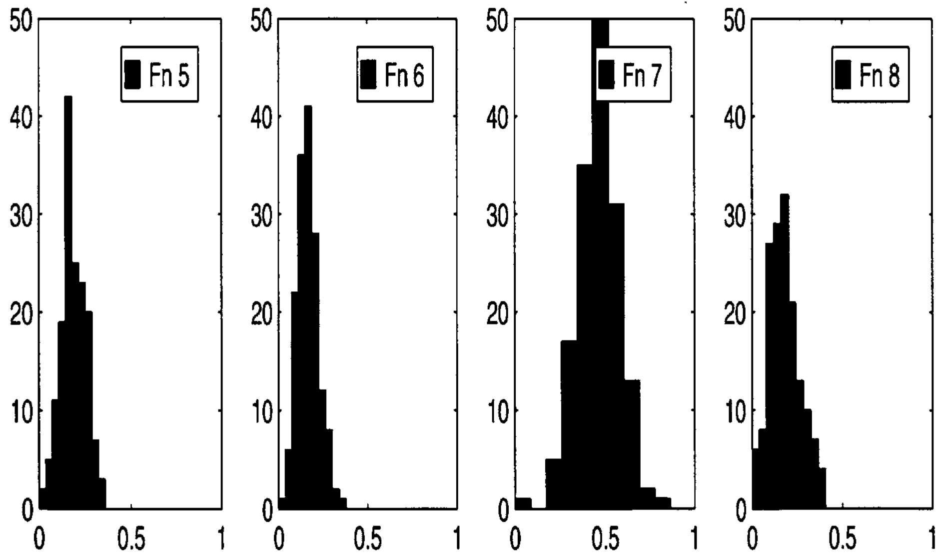


Figure 2.7: Capacity Analysis of Diffusion Furnaces

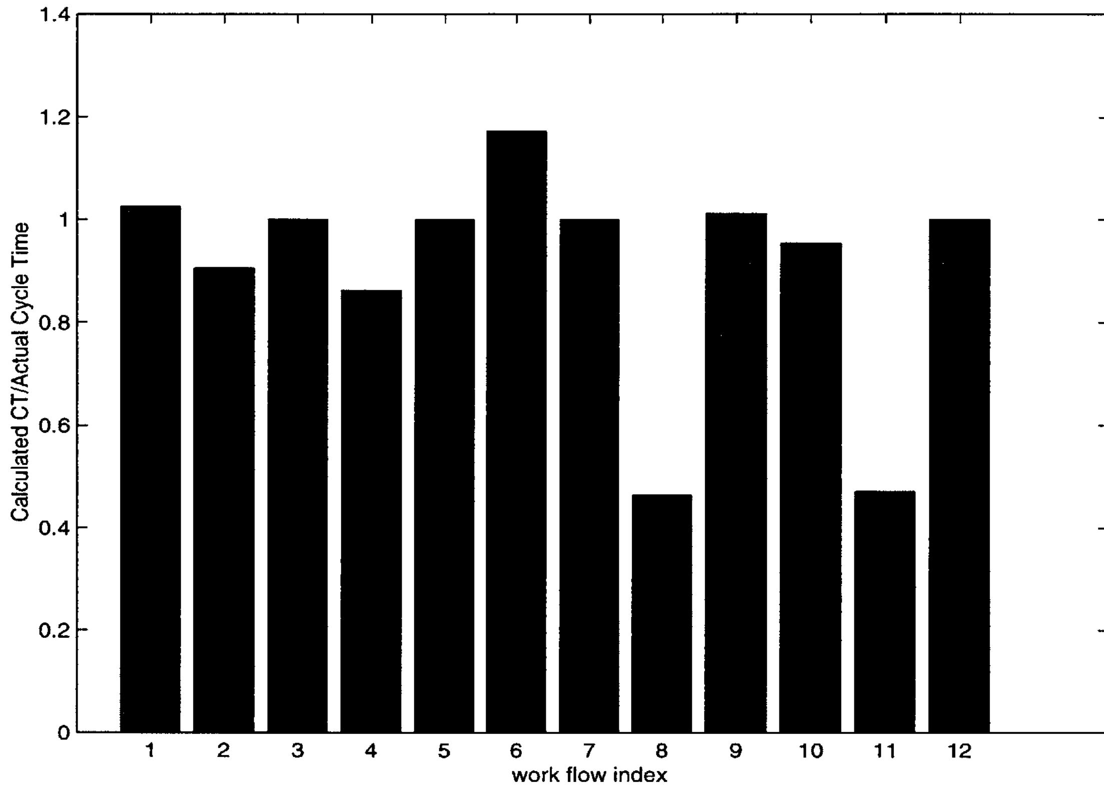


Figure 2.8: Analysis of flow cycle times: Actual Vs. Model

(4,3) represents only 2 hours of processing time but in out model it has a planned lead time of 1 day. Another reason for error in cycle time estimates is that the tail end of process sequence of any flow (after the last diffusion step) is not captured in our model. Hence, our model does not account for the time it takes to finish a job after its last diffusion step and under estimates the actual cycle time. We present our cycle time estimates for all the major flows (12 of the 22 we considered) in figure 2.8. The cycle time estimates are fairly accurate for all flows except flows 15 and 20 which represent the 6" wafer routings. The main reason for underestimation of their cycle times is that these two flows are still relatively

new and therefore the work-in-process for these flows is held for extensive scrutiny at several points in their process sequence. These delays are not captured in our model and therefore, our estimates of the lead times for these flow are considerably lower than in reality. Aggregating the WIP for all the flows, our estimate of the total work-in-process in the fab was within 7 percent of the actual work-in-process in the fab.

Table 2.5: Production Requirements at the Furnaces: Model Vs. Actual

	Model		*Actual*	
Furnace	*E(Prod.)*	*σ(Prod.)*	*E(Prod.)*	*σ(Prod.)*
1	152	41	163	104
2	741	288	720	231
3	425	150	489	188
4	672	263	728	273
5	397	184	410	144
6	608	295	446	161
7	728	262	928	231
8	245	113	268	126

One of the outputs of the model is the steady state levels of work-in-process at each workstation. We report these numbers in table 2.4 for work flow 5 and compare them to the actual average amount of work-in-process in front of each workstation. We report the outputs of the model on the rest of the work flows in the appendix. We think that the model was quite successful in determining cycle times for all but 2 of the flows.

Based on the outputs of the model for each work flow, we characterized the production requirements at the furnaces as shown in table 2.5. We assumed that the production at each workstation was independent of the production at other workstations. Based on this assumption, we aggregated the means and variances of all the workstations that ended in

the same furnace to determine the mean and variance of production at that furnace. For example, as shown in the table, furnace 1 has production requirements with a mean of 152 wafers/day and a standard deviation of 33 wafers a day. We compared these estimates to the mean and standard deviation of the actual production requirements at the furnace during the 21 week history that we obtained. As can be seen, the expected requirements as projected by the model are roughly within 10 percent of the actual mean requirements at the furnaces. The exception to that is at furnaces 6 and 7. The model is off by over 20 percent in predicting the average production requirements at these furnaces. The estimates of production variability determined by the model are less accurate than the estimates for expected production of workstations. For example, for furnace 1, the model estimated the standard deviation to be 41wafers whereas real data reveals that the variability was about 104 wafers. The estimates of all the other furnaces were within 35% of the actual variability determined from production time series.

This discrepancy in the outputs of the model and the actual data can be attributed to several reasons. First, all the process flows in a given family are not exactly similar to the one that we chose as the representative flow for that family. For example some flows in a family might have more workstations than the others. Separately accounting for each flow would achieve more accuracy in determining the mean production at the furnaces but the disadvantage is that it makes the model very complicated and hard to analyze. We did not account for yield loss in our model due to the unavailability of accurate data on percent yield for each flow. Incorporating the impact of yield would reduce the ϕ_{ij} values, thereby making the production estimates lower in our model as expected. We also didn't account for the impact of rework in our model. Rework increases the production requirements and

adds to the variability of the work flow as well. The impact of the existence of rework is amplified by the re-entrant nature of the work flows. Another key issue that is not captured in our model is the impact of breakdowns. Workstations that are highly unreliable will have large variability in their outputs which can not be captured in our model. For example, as shown in table 2.5, our model significantly underestimates the variability of outputs of furnace 1. We suspect that this is because this furnace is unreliable. Referring back to figure 2.7, we can see that furnace 1 has many days of minimal throughput which was caused by equipment downtimes.

2.3 Conclusion

In this chapter we described how we developed the tactical planning model for studying the operations of the fab. We think this model can be put to use in determining the changes in production requirements that are caused by fluctuations in demand. A comparison with the real fab data indicates that the model provides a reasonable characterization of production requirements, lead times and WIP profiles of different work flows in the Wilmington wafer fab. In the next chapter, we will describe how we used this model to understand fab operations.

Chapter 3

Model Application to Fab Operations

In the previous chapter, we described how we developed a tactical planning model for the Wilmington wafer fab. This linear flow model of the wafer fab assumes that each workstation operates according to a linear control rule, which is parameterized by a planned lead time. The model characterizes the aggregate production flows that are necessary to achieve the planned lead times and meet the production requirements. In this chapter, we will demonstrate the use of the model for examining the trade-off between investments in work-in-process inventory and capacity, for setting planned lead times, for determining the benefits of eliminating input variability in the fab, and for understanding the impact of changes in production mix and volume on inventory and capacity requirements. Then we will show how this model can be used to anticipate the change in the production requirements of the fab due to demand fluctuations.

3.1 Operational impact of Variability in Starts

In the Wilmington wafer fab, the number of lots released for processing (starts) is constant from week to week in a period with relatively stationary demand conditions. Once, there is a change in demand forecast, this starts level is appropriately adjusted. For example, for a particular family total number of lots released into the fab might be 20 for the first 10 weeks and then due to a sudden increase in demand for that part, the weekly starts might be raised to 40 lots each week. Even though the weekly starts are constant for many weeks at a time, the variability in the number of lots started each day is still quite high. Therefore, within a week, during some days large number of lots are released into

the fab while during some other days very few lots are released into the fab. Figure 2.3 shows that the coefficient of variation of daily starts of all the major flows range from 0.6 to 1.3.

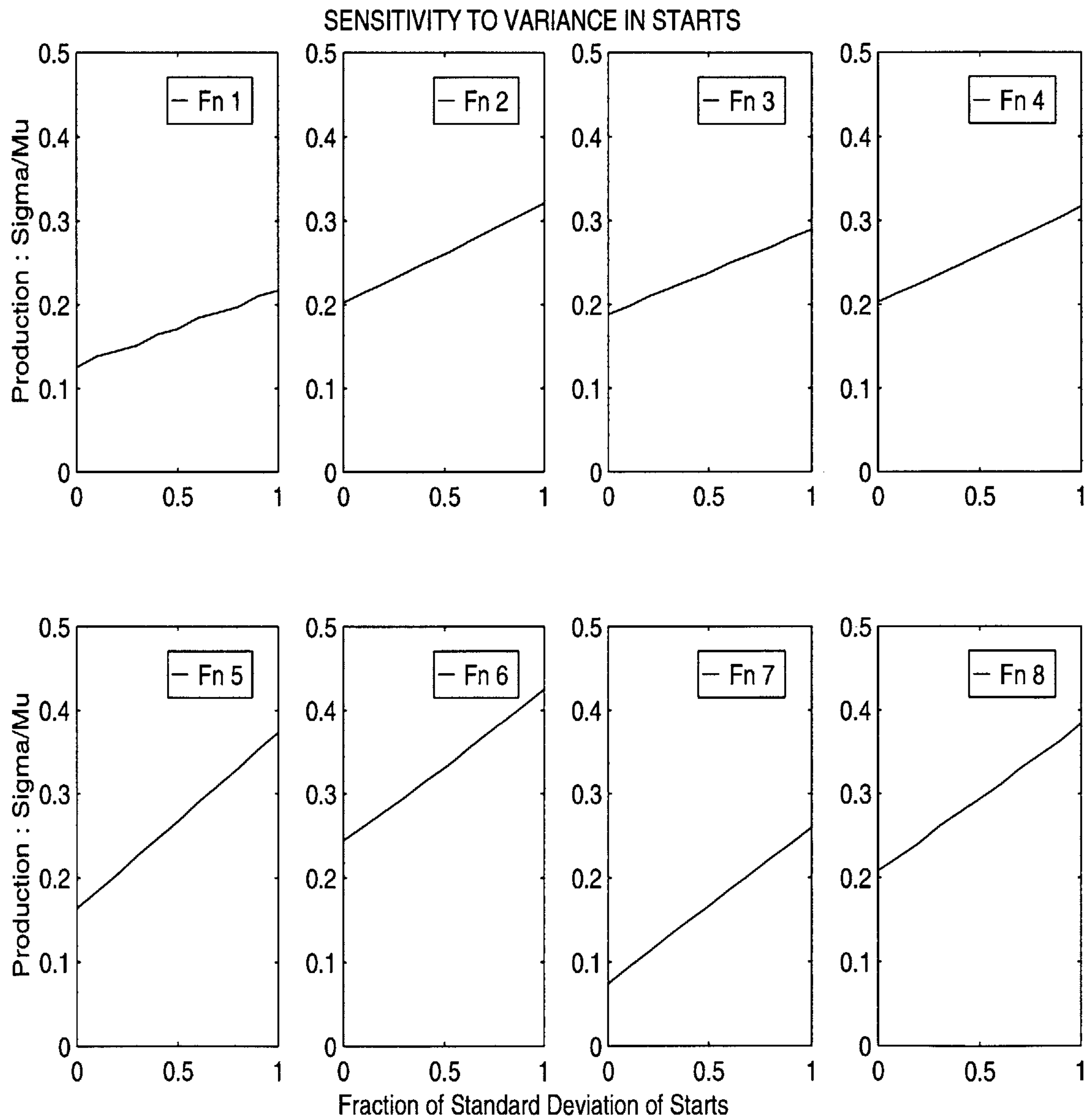


Figure 3.1: Impact of reducing the variability of lots released into the wafer fab on production variability

We now show how to use the model to characterize the impact of the variance of the starts. Figure 3.1 shows a quantitative relationship between the variability of starts and production variability of the diffusion furnaces. We gradually reduced the coefficient of variation of the starts from its actual value[1] to zero for all work flows and computed the production variability at each furnace for various intermediate values. When the variability of the starts is set to zero, we get a characterization of the variability inherent to fab operations when no external variability is introduced into the work flow. As can be seen in the figure, the variability in production at the furnaces reduces proportional to the decrease in variability of the starts. For most furnaces, the production c.v. is nearly halved as the variability of the starts is reduced from its current levels to zero. The remaining variability is due to the lumpiness in the work flow which is caused by large lot sizes.

Table 3.1: Model Output: Reduction of c.v. of starts by 50 percent

Furnace	*E(P)*	*reduced c.v.* $\sigma(P)$	*actual c.v.* $\sigma(P)$	*Theoretical Capacity*
1	152	33	41	2300
2	741	243	288	2400
3	425	128	150	1170
4	**672**	**223**	**263**	**1730**
5	397	143	184	2200
6	608	239	295	2650
7	728	194	262	1400
8	245	92	113	1470

Table 3.1 shows the decrease in production variability when the input variance of all the flows is reduced to half of their current values. As can be seen, σ(P) is reduced for

1. Based on real data - See Chapter 2

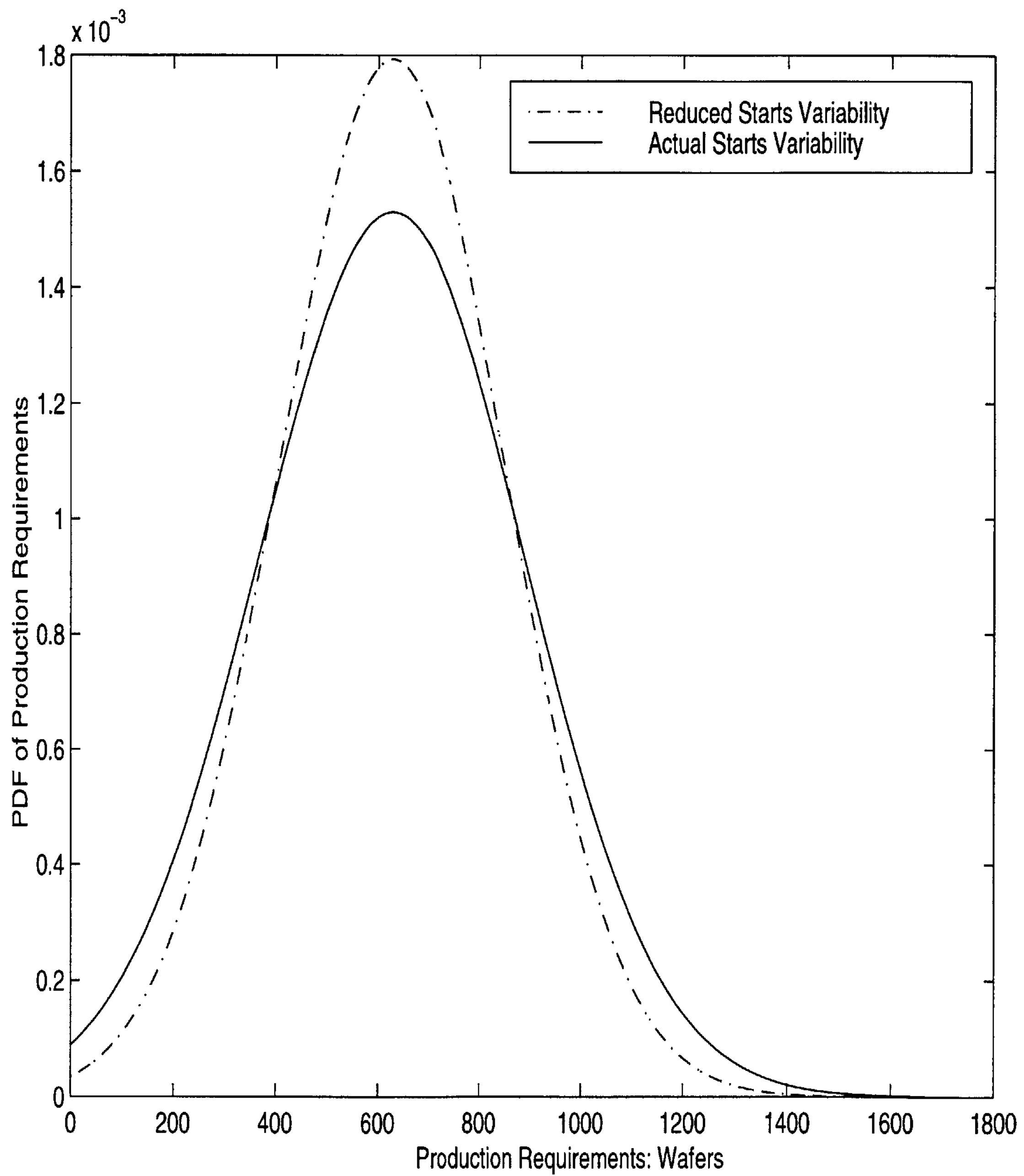


Figure 3.2: Impact of reducing starts variability by 50% on production variability of Furnace 4

all the furnaces by over 10 percent. The reduction of production c.v. of the workstations can be used to evaluate the impact of maintaining a less variable starts release policy on capacity requirements of the workstations. In figure 3.2, we show the impact of reduction in input variability on the production requirements at furnace 4. We assume that the PDF of production requirements at the furnace can be characterized by a gaussian distribution bounded by 0 and 1730[1] with mean and variance given by the model. We assume that a workstation can not produce in excess of 60% of its theoretical capacity to account for capacity loss due to set-up, maintenance and breakdown periods. For this characterization of production at the workstations and the current level of variance of starts, there is approximately a 6 percent chance that furnace 4 would have a capacity shortfall. However, by reducing the variance of starts by half, there is only a 3 percent chance there will be a shortage of capacity at the furnace as shown in figure 3.2. This suggests that maintaining a more uniform starts mix would smooths the work flow considerably and reduces the frequency of days when the workstations have to be extremely busy.

The model can be used to characterize the sensitivity of workstation operations to variability in individual flows also. In practice it might not be feasible to reduce the variance of starts of all the routings. Therefore, we looked at the impact of reducing the variance of only the dominant work flows. To gain further insight, we reduced the variance of starts of flows 1, and 5 and did a similar analysis. In figure 3.3, we show the impact of reduction of variability of starts on smoothing the production requirements of the workstations. We reduced the variability of starts from the actual down to zero in increments of 25 percent. The results of this experiment indicate that for a 50 percent reduction in starts

---

1. maximum capacity of furnace 4: See table 2.3 in Chapter 2.

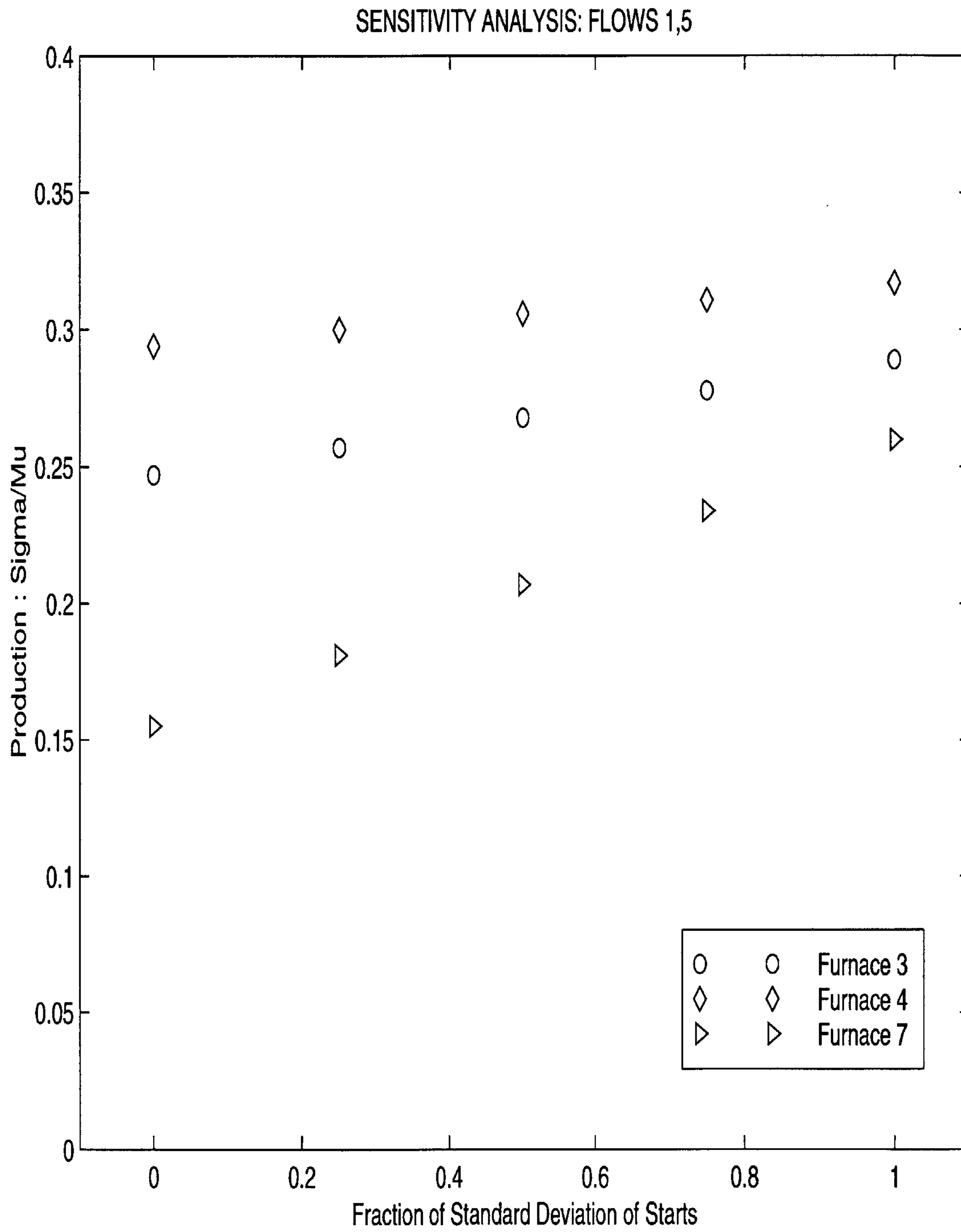


Figure 3.3: Sensitivity of equipment variability to starts variability

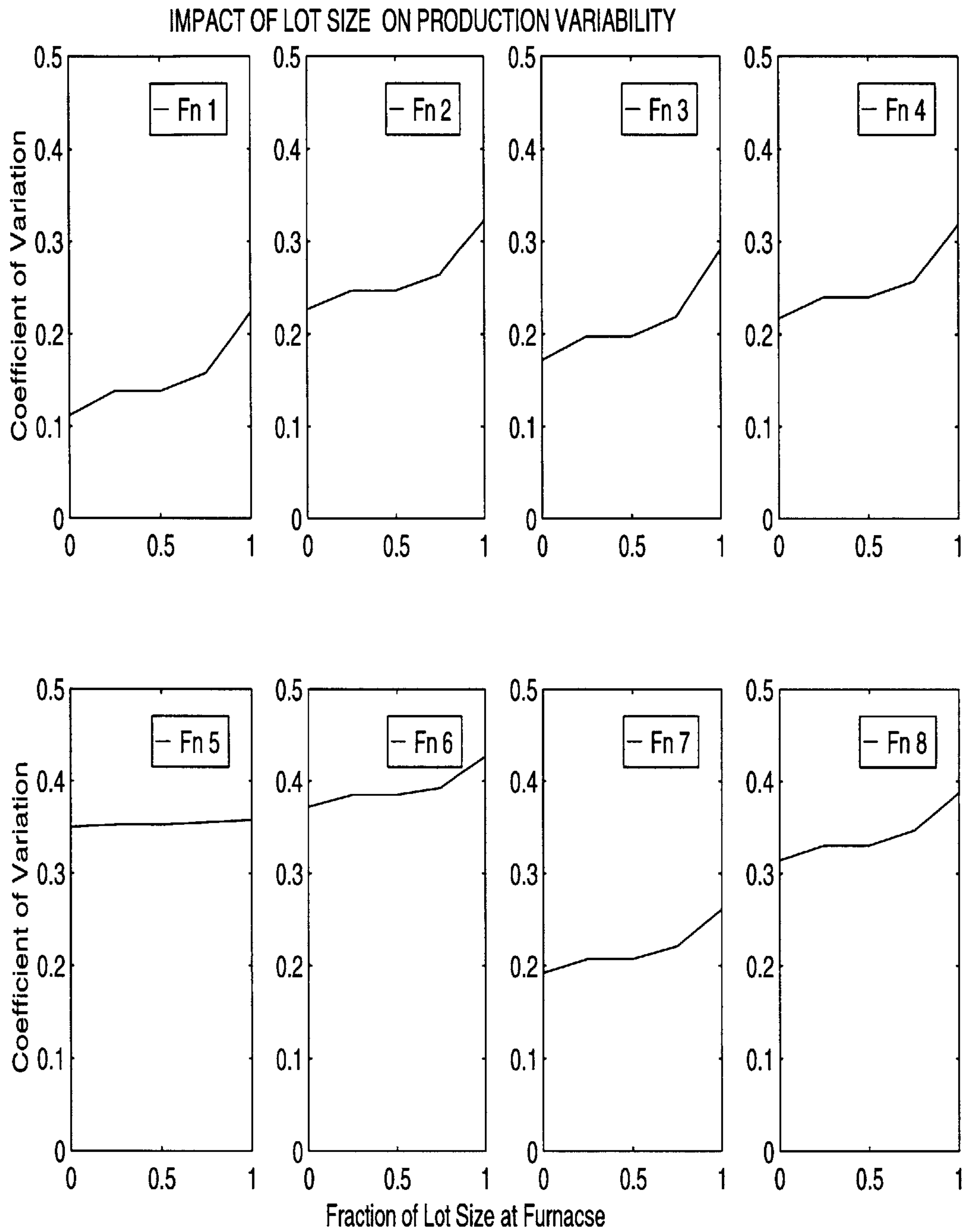


Figure 3.4: Trend of production variability of workstations as lot size is reduced for all families

variability, while the c.v. of furnace 7 is reduced by 0.06, the c.v. of furnaces 3 and 4 show minimal reduction. This is because furnace 7 processes many more recipes in these flows than furnaces 3 and 4. Therefore, to smooth the production at a constrained equipment, it is important to identify the work flows in which the furnace appears quite frequently. A 50% reduction of the standard deviation of starts of 2 major flows (1,5) lead to a reduction in probability of shortfall from 6 percent to 4 percent. By smoothing the input stream of lots for only two flows, we obtain two-thirds of the level of production smoothing that we attain by smoothing the input stream of all 12 major flows.

These few exercises show how this model can be a useful tool to identify the work flows that need to have less variation in starts to permit production smoothing at the critical equipment. For a certain demand portfolio, if certain workstations are in a capacity constrained situations, this model can help identify the routings that need to have smoother input into the fab to reduce the frequency of peak loads at these workstations. The model helps determine the sensitivity of workstations to the release policies for different work flows.

3.2 Impact of Lot Sizing

In this section, we quantify the impact of lot sizing on production variability. As mentioned before, a larger lot size leads to a less regular flow and makes it highly variable. We reduced the lot size of different families gradually and determined the impact on production variance due to the reduction. In figure 3.4, we plot the relationship of production variability to the size of the lots of the families. As the lot size is decreased, the flows become more regular. Although it is evident that a smaller lot size is better for production

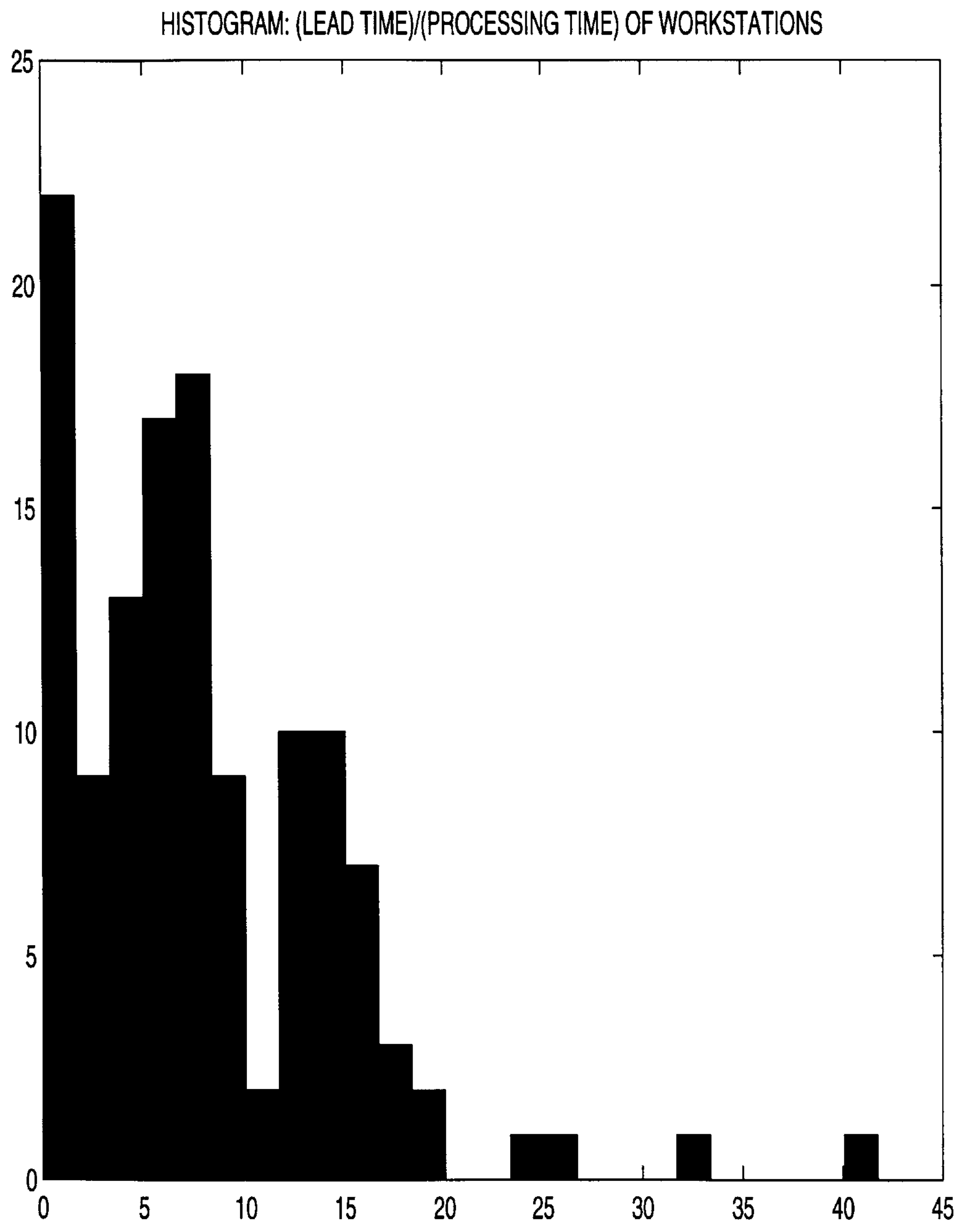


Figure 3.5: Ratio of (Workstation Lead Time/Workstation Processing time)

smoothing, it is not always the best decision to reduce the lot sizes. This is because equipment with long processing times have large batch sizes to ensure a reasonable throughput. If the lot sizes were made too small, then the throughput of these equipment will be considerable smaller or the wait times of lots will increase significantly (see Chapter 4 for more details). This section provides a quantitative relationship of lot size and production c.v. for different equipment. It can be used to determine the sensitivity of a workstation to the lot size of a particular work flow.

3.3 Trade-off between WIP and Lead Times

In a capacity constrained wafer fab, there are two common approaches that can be taken to meet the production requirements. The most intuitive approach is to invest in additional capacity but often this is not very practical due to the high cost of capital investment in wafer fabs. Therefore, the preferred way (if possible) is to increase the efficiency of the operations through better performance on metrics such as equipment utilization. The improvements in such criteria require a change in the profile of WIP in the fab. In this section, we show how this model can be used to capture the impact of changing the WIP profile on production requirements in the fab. The bottleneck workstations[1] are identified in figure 3.5. The workstations with a high (Lead Time)/(Proc. Time) ratio can be the workstations with an opportunity for most reduction in lead times. The model can be used to evaluate the implications of reducing their planned lead times.

1. Workstations which have a large (Lead Time)/processing Time) ratio.

Table 3.2: Scenario: Lead Time of Workstation is its Processing Time

Furnace	*E(P)*	$\sigma(P)$	*E(WIP)*	$\sigma(WIP)$
1	152	79	390	148
2	741	535	844	484
3	425	278	490	238
4	672	506	933	532
5	397	356	402	504
6	608	599	642	511
7	728	412	730	354
8	245	229	348	255

Table 3.2 shows the production distribution and profile of the work in process when the planned lead times for each workstation are set to be the actual processing time of that workstation. The processing time of a workstation is the sum of processing times of all the recipes that are represented by it. As expected, the variability in production in this scenario is much higher then in current practice because the total WIP in the system is reduced from 24000 wafers to 4779 wafers. In this case there is not enough WIP in front of the workstations to absorb the variability of the arrival stream from the upstream workstations. Therefore, the variability of the work flow percolates down the routings and gets exaggerated by the re-entrant nature of the routings and by the stochastic behavior of equipment.

The model provides a framework to analyze the trade-off in increasing lead times and inventory holding costs. The capacity analysis of the furnaces showed that furnaces 1,2,5,6 and 8 had excess capacity available. These furnaces should be able to handle a more variable work stream. Hence, there is an opportunity of reducing planned lead times of these workstations without encountering any capacity shortfalls. We reduced the planned lead times for all the workstations defined by these furnaces by half and compared

the impact of additional variability on production operations. The results are shown in tables 3.3 and 3.4. As shown in table 3.3, the variances for the furnaces go up slightly but the work in process is reduced from 24000 to approximately 19000 wafers. The impact of reduction in planned lead time on flow times is shown in table 3.4. As can be seen, reducing the lead times by half for the low utilization furnaces leads to large reductions in cycle time at the expense of minimal increase in production variability.

Table 3.3: Impact of reducing variability of low utilization furnaces.

Furnace	*E(P)*	*σ(P)*	*original σ (P)*
1	152	42	41
2	741	288	288
3	425	147	150
4	672	265	263
5	397	199	184
6	608	313	295
7	728	262	262
8	245	123	113

Table 3.4: Impact of planned lead time reduction on of flow cycle times

Flow	*Model CT*	*Model CT Reduced Lead Times*	*Percent Reduced*
1	41	27	34
5	38	38	0
6	56	56	0
9	60	39	35
12	68	68	0
13	68	68	0
14	56	47	16
15	56	47	16
17	81	62	23
18	42	36	14
20	70	60	14
22	56	50	11

This exercise shows how this model can be used to identify the places in work flow where work in process can be reduced to reduce cycle times without causing a significant impact on the variability of the work flow. The model can also be used to evaluate the impact of increasing planned lead times at capacity constrained workstations by comparing the consequent increase in work in process and the decrease in production variability.

Table 3.5: Model Output: Average Demand of flows 15,20 is doubled.

	increased demand		*original demand*	
Furnace	*E(P)*	$\sigma(P)$	E(*P*)	σ *(P)*
1	152	41	152	41
2	**897**	**324**	**741**	**288**
3	425	150	425	150
4	672	263	672	263
5	**679**	**259**	**397**	**184**
6	**998**	**398**	**608**	**295**
7	728	262	728	262
8	**299**	**125**	**245**	**113**

3.4 Impact of Demand Fluctuations

The tactical planning model provides an estimate for mean and variation of production for a stationary job mix. A wafer fab is constantly faced with changing market conditions and therefore, the steady state assumptions of the model are not very appropriate. However, it can be used to determine the variation in production requirements once the job mix changes. Although the model does not capture the dynamic impact of changing demand on operations, it gives an idea of production requirements for the new demand.

For example, currently the Wilmington wafer fab is ramping up its production of the 6" process flows. The model can be used to quantify the increase in work loads at the workstations due to an increase in 6" wafer starts. Table 3.5 shows the output of the model when the 6" starts (flows 15 and 20) are twice as much as the current starts. For this analysis, we doubled both the mean and the variance of the starts for these work flows. As can be seen, a two-fold increase of 6" wafer starts causes the mean production for furnace 6 to significantly increase from 608 wafers/day to 998 wafers/day. In figure 3.6, we show the shift in the distribution of production at this furnace. With the current demand levels, there is minimal chance that the furnace would be in a capacity constrained situation. However, when the demand for 6" parts is doubled this percentage jumps to 7 percent. In such a situation, the arrival stream to furnace 6 should be smoothed by increasing the planned lead time for workstations defined by furnace 6. Increasing the planned lead time for those workstations reduces the variance of production distribution as shown previously, which reduces the chance that the workstation will have to work extremely hard.

Table 3.6: Model Output: Uniform Starts Vs. Variable Starts: Increased Demand

	variable starts		*uniform starts*	
Furnace	*E(P)*	$\sigma(P)$	E(P)	$\sigma(P)$
1	152	41	152	41
2	**897**	**324**	**897**	**276**
3	425	150	425	150
4	672	263	672	263
5	**679**	**259**	**679**	**154**
6	**998**	**398**	**998**	**252**
7	728	262	728	262
8	**299**	**125**	**299**	**109**

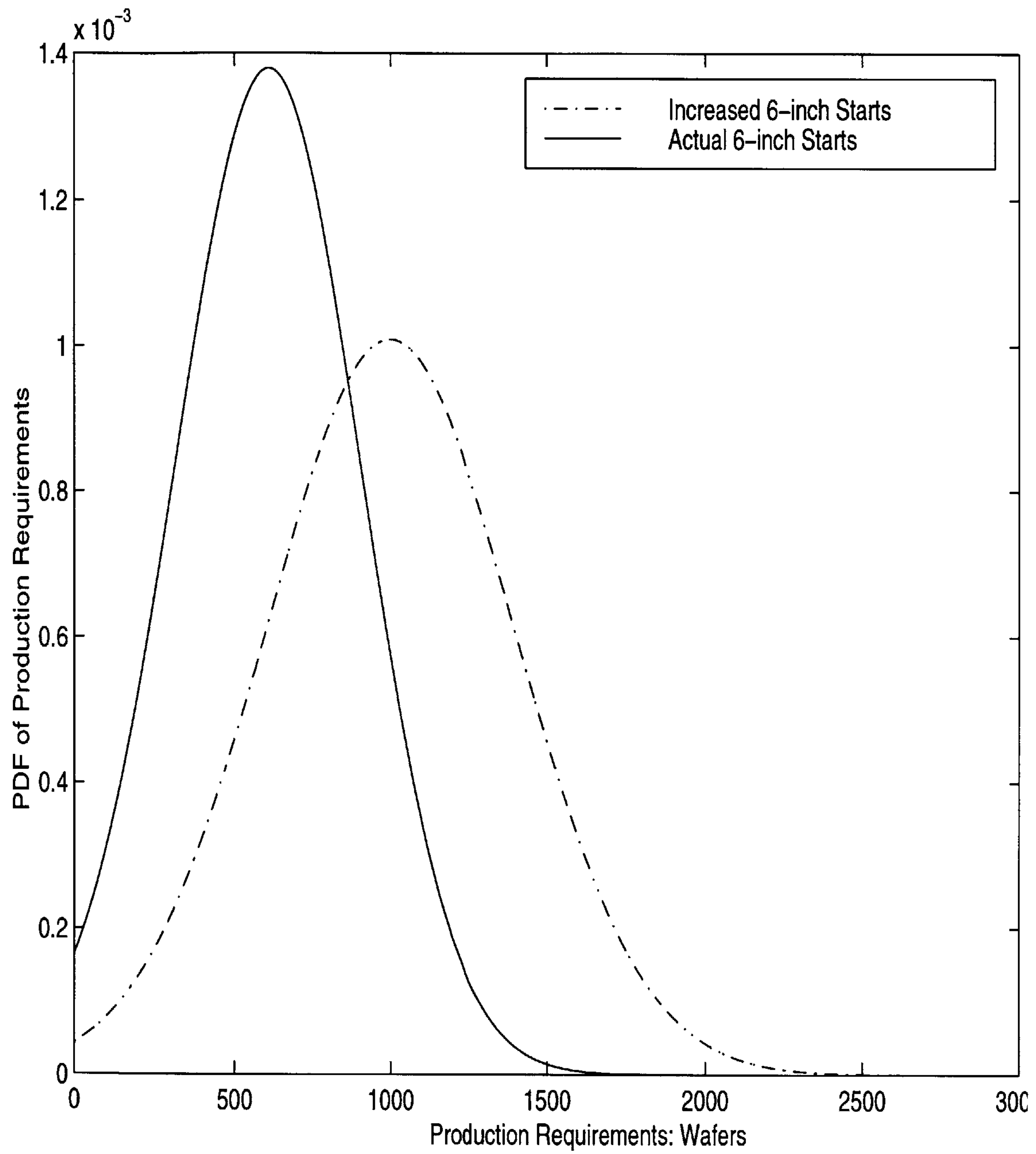


Figure 3.6: Shift in PDF of Production Requirements at furnace 6 due to demand increase of flows 15,20

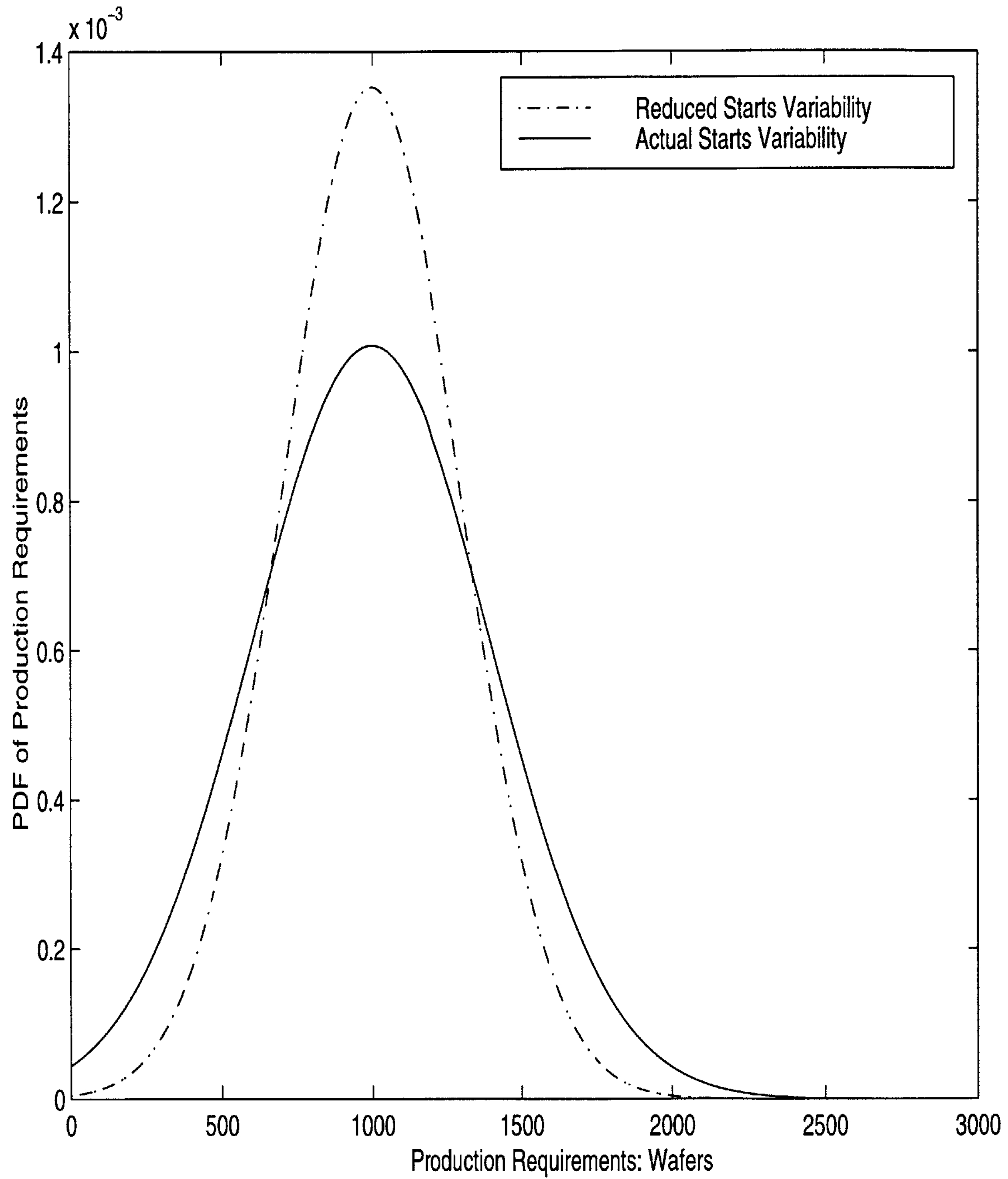


Figure 3.7: Reduction in c.v. of starts for 6" flows helps reduce the variability of production at furnace 6 when demand for flows 15,20 is doubled.

In this scenario, the model allows us to evaluate various options for meeting the production requirements at furnace 6. Various options are to reduce the variability of the starts of these 6" flows or to do production smoothing by increasing planned lead times for workstations with furnace 6. We present the analysis of one such scenario where we release constant starts into the fab for flows 15 and 20. The model output for constant 6" lot releases into the factory is compared to the model output for variable 6" lot releases in table 3.6. The standard deviation of production distribution is reduced by 150 wafers/day for furnace 6 with this release policy. Figure 3.7 helps quantify the results of this analysis in terms of capacity shortfall at furnace 6. As shown in the figure, the distribution of production is a lot tighter around the mean for the case of constant releases. The percentage of days with capacity shortfall is reduced from 7 to 2 percent.

Another approach for meeting the increased demand would have been to increase the planned lead times for workstations with furnace 6 to smooth the arrival stream of lots into those workstations. As shown in the above example, the tactical planning model can be a useful tool for determining the impact of changing demand conditions on fab operations. The model can be useful in evaluating the trade-off between work-in-process inventory and capacity of the equipment in the fab.

3.5 Conclusion

The tactical planning model that we developed under this thesis project gives a reasonable characterization of operations in a wafer fab with multiple flows. However, there is a serious limitation to the usability of the model as it currently exists. Our model looks

at 22 flows[1] and characterizes fab operations based on these 22 flows which is an over simplified view of the fab. Here, we provide an example to describe where this model falls short. We assume there are 28 starting lots per week for some family. We further assume that there are 2 different routings under this family each with starts of 14 lots per week. In this scenario, according to our model the best release policy would be to release 2 lots of each routing every day. However, in practice this approach will adversely impact the throughput of the fab. Because of different processing requirements, only lots of the same routing can be processed at a given time. Hence, the utilization of the equipment is forced to be very low. Instead, if 4 lots of each routing are released every other day (on alternate days), then the utilization of the equipment is much higher at the expense of a more variable arrival stream of each routing. General sentiment in wafer fabs is to maximize the throughput which might require release policies which are not recommended by our model. Although it is obvious that running equipment at really low utilization levels is not good for the performance of the fab, we show in the next chapter that an emphasis on high utilization might not be the best strategy for enhancing fab performance either. We will show that a strategy that focuses on increasing the throughput of the furnaces by running large batch sizes might not always be the best due to the large wait times it might create.

1. Note that the starts for 10 of those flows was minimal.

Chapter 4

Batching: A Simulation Study

4.1 Introduction

So far in this thesis, our focus has been on understanding fab operations at a tactical level. We have tried to develop a model-based framework for characterizing the impact of variability in starts and of lot sizing on the production requirements at the diffusion equipment. Our model evaluates the trade-off of investments in capacity vs. work-in-process inventory at a tactical level and describes how to model the fab using a linear control rule. Such a model suggests WIP levels at each workstation to handle excess/shortage of capacity at the workstations with variations in starts.

In this chapter, we present a brief simulation study that looks at fab operations at the scheduling level. As explained in section 1.2, diffusion/CVD equipment has long recipe times and to maintain reasonable throughput rates these pieces of equipment typically have large load sizes as well. Here, we try to understand the impact of batching at such equipment on fab throughput and cycle time. Traditionally, operators emphasize the high utilization of such equipment to maintain high levels of throughput. However, even though the equipment utilization is higher, the cycle time of the fab increases dramatically due to such an operations strategy. Our simulation study quantifies the impact of equipment utilization on throughput and cycle time. We conducted this study on a simulation model that is used by ADI operations personnel for capacity planning. This model was developed using a discrete-event simulation model developed by Auto Simulations, Inc. (ASI)

4.2 Simulation Model of Wilmington Wafer Fab

In this section, we briefly describe how we developed the simulation model of the wafer

fab using ASI simulation software. Then, we will describe the set of simulations that we ran and finally, we will discuss our results.

Model Development

The ASI simulation package is a discrete event simulation tool which is used to model the flow of work in a factory. The simulation package takes as inputs, the operational characteristics of a factory, the demand portfolio, and a description of process flows, and it simulates the work flow through the factory for a period of time determined by the user. The model inputs to the simulation package consisted of a set of files that were generated based on historical data and desired operation of the fab. We used the same 22 process flows in this simulation model that were used in the development of the tactical planning model. Each process flow was represented by a text file which contained information about each processing step such as the equipment required for that step, the processing time at the equipment, and any engineering specifications (see section 1.2). One of the input files contained information about the equipment performance such as *mean time to repair* (MTTR) and *mean time between failure* (MTBF), as well as the load size of each piece of equipment. Similarly, another input file contained information about the equipment required, setup-changeover requirements and mean processing time for all the recipes. Finally, one of the input files provides information about the time and quantity of lots started for each family for the duration of the simulation.

The ASI package uses these input files and simulates production in the wafer fab based on the starts file. The package simulates the release of lots into the fab based on the starts schedule. It provides a detailed profile of work in process in front of each equipment by lot ID. A recipe is run on the equipment only if equipment is available. The equipment can be unavailable if another recipe is being performed on it, or if the equipment is down.

The equipment downtimes are modeled as exponential random variables based on their MTTR which were determined from historical data. Similarly the frequency with which an equipment breaks down is determined from historical data. Once a recipe is run on an equipment, the duration of a recipe is modeled as a gaussian random variable given by the mean processing time and variance based on historical data.

The model simulates the flow of lots down their respective process flows in the manner described above. The outputs of the model include the cycle time for each lot, the throughput of each equipment and throughput of the entire fab.

Experiment Setup

We conducted a simulation study to understand how batching impacts the cycle time and throughput of the fab. In each experiment, we established a minimum utilization requirement for the diffusion furnaces. For example, a minimum utilization requirement of 0.25 at a furnace implied that a set of lots was only run on that furnace if it used at least 25 percent of the furnace load size. For further clarity, let us take an example. Suppose a furnace can take up to 20 lots at a time. Then, if we impose a minimum utilization requirement of 0.25 at that furnace, it implies that a recipe would be run at that furnace only if there are at least 5 lots waiting to be processed with that recipe[1]. This scheduling rule can lead to really long wait times for lots with recipes which do not have much demand. To avoid such a scenario, we designated a threshold wait time for all lots. If a lot has waited for the threshold time, it is given maximum priority. The threshold time was kept constant as a certain fixed percentage of the "process engineering time constraint" mentioned earlier as the maximum wait time allowed between immediate upstream operation and the diffusion/CVD equipment (see Chapter 1).

1. Note that those five lots could be from different families but as long as the recipe for them is the same, they can be batched together.

Each simulation run consisted of five simulation runs, each 365 days long with random, normally distributed seeds and the same inputs. The starts for each family were based on the actual starts in the fab. Five such experiments were conducted for different values of the minimum batch size as a fraction of the load size of the furnaces. In each experiment, the same batch minimum fraction was used for all the relevant diffusion/CVD equipment in the fab simulation model. For each experiment, throughput and cycle time trends were observed.

Results

The results of the simulation experiments are shown in figures 4.1-4.4. In figure 4.1, we show how the equipment idle time varies as the minimum batch size is increased at the diffusion furnaces. As can be clearly seen, the idle time of the furnaces increases as the minimum batch size is increased. This is an intuitive result because as the minimum batch size is increased, more lots of the same recipe have to be present in the queue and therefore, idle time of the furnaces is increased. A similar intuitive argument can be used to explain the behavior of waiting time for lots as the minimum batch size is increased at the furnaces. It can be seen in figure 4.2 that a minimum batch size of 0.25*(load size of furnace) helps achieve the smallest waiting time for the lots (on average). As the minimum batch size is further lowered, the wait time for lots increases because of low capacity utilization. As the minimum batch size is lowered, the throughput rate of the furnace becomes much smaller than the arrival time of lots at the furnace thereby causing longer wait times. As the minimum batch size is made too large, the long idle times of the furnaces (as seen in figure 4.1) cause the wait times of the lots to increase.

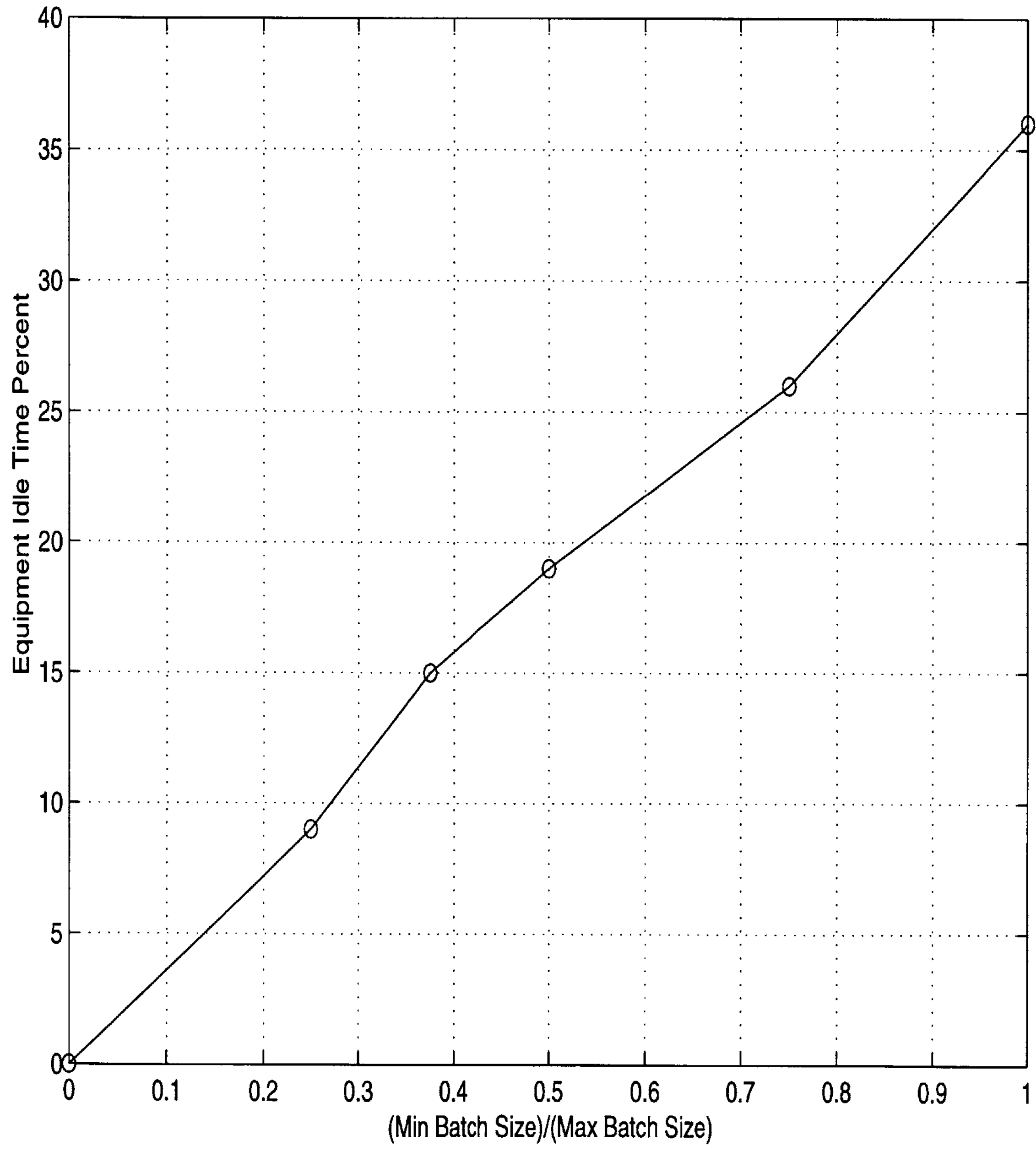


Figure 4.1: Minimum Batch Size Vs. Percent Idle Time at Furnaces

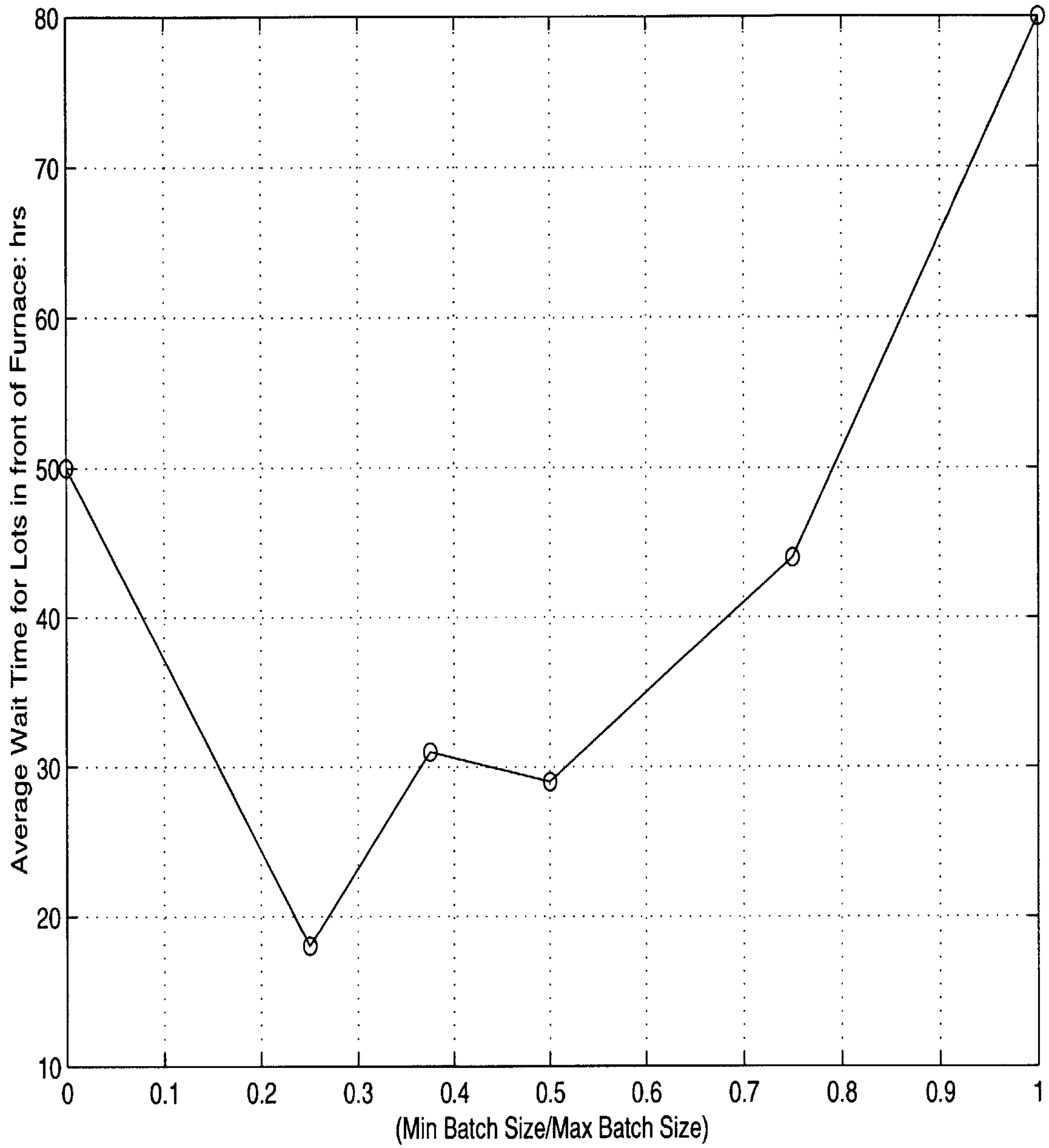


Figure 4.2: Minimum batch size vs. Average Wait time for lots at the Furnaces

The results shown in figures 4.1 and 4.2 show that reducing the minimum batch size at the furnaces increases the wait time for the lots even though it reduces the idle time of the equipment. Making the minimum batch size too small or too large can also have an

adverse impact on the throughput of the fab. Making the minimum batch size really small is not the best scheduling decision because it leads to a low capacity utilization of the equipment. Even though reducing the minimum batch size allows us to use the equipment for longer periods, the throughput rate of the equipment suffers significantly. For diffusion/CVD equipment it is important to maintain high throughput rates because recipes have long processing times at such equipment. If this equipment is under-utilized then the downstream equipment will face long periods of starvation as well. Making the batch size too large also reduces the throughput rate of the fab because the wait times for lots are extremely long and therefore even though each run of the furnace is at high percent utilization, the frequency of runs is very low. In brief, a small minimum batch size leads to low equipment utilization while a large minimum batch size leads to long wait times. Either way, the throughput of the fab suffers significantly.

In figures 4.3 and 4.4, we show that for a given demand portfolio, the minimum batch size should be set to a certain optimal value. This optimal minimum batch size is determined based on the trade-off of large wait times for lots vs. low equipment utilization. Figure 4.3 shows the impact of the size of the minimum batch size on the cycle time of the fab. It can be seen that a minimum batch size of 0.25*(load size of furnace) helps achieve largest reduction in the cycle time of the fab. When the minimum batch size is smaller than this optimal value, the reduction in cycle time is not as high. This is because when the minimum batch size is too small, the production trends are similar to the current state[1]. At the same time, if the batch size is made larger than this optimal value, the cycle time reductions are not as high because idle time of diffusion/

1. Current state implies a minimum batch size of 0

CVD furnaces increases. Furthermore, if the minimum batch size is made too large, then the cycle times become even higher than the current state because of the extremely long equipment idle times.

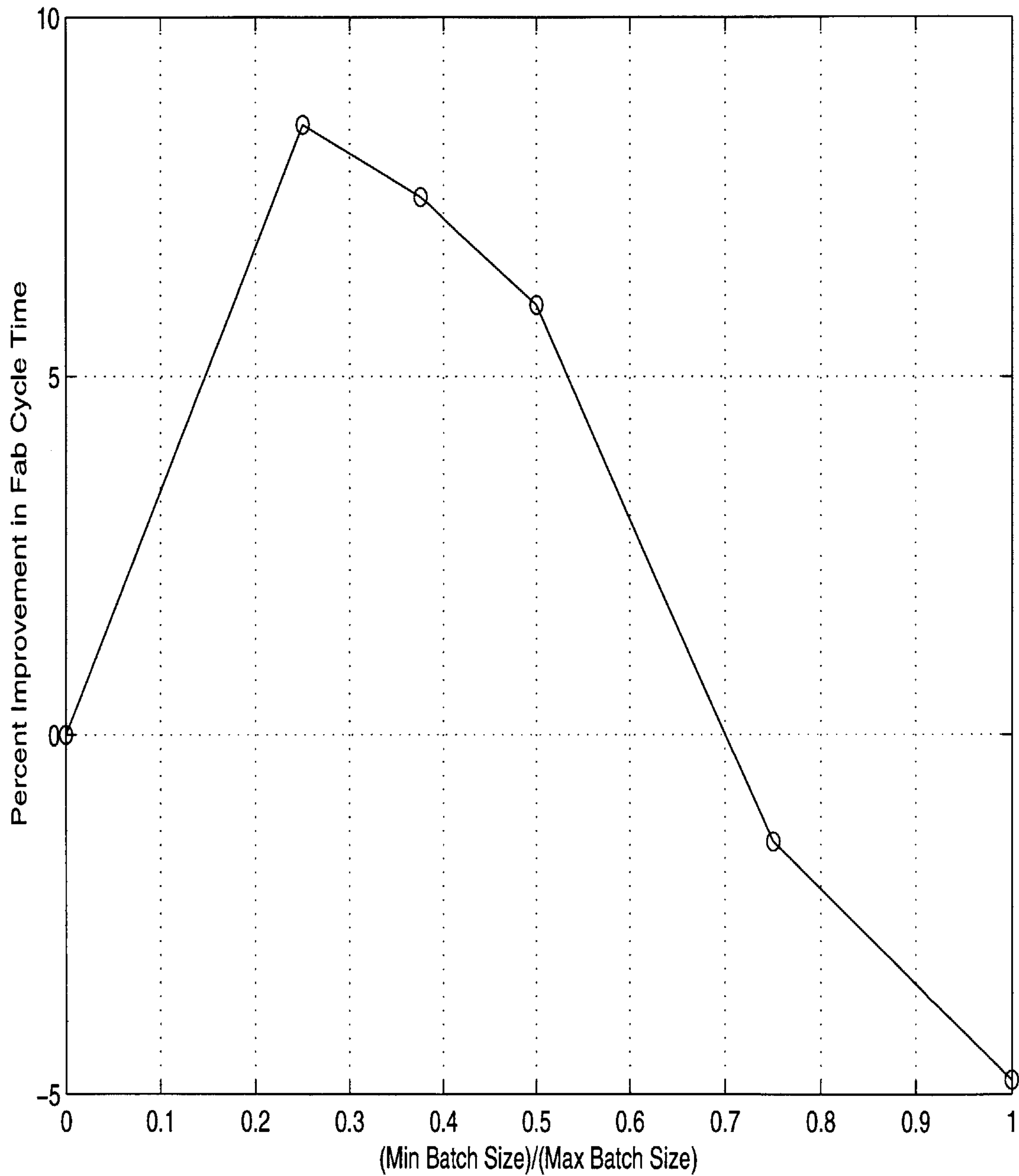


Figure 4.3: Reduction in Cycle time vs. Minimum Batch Size at furnaces

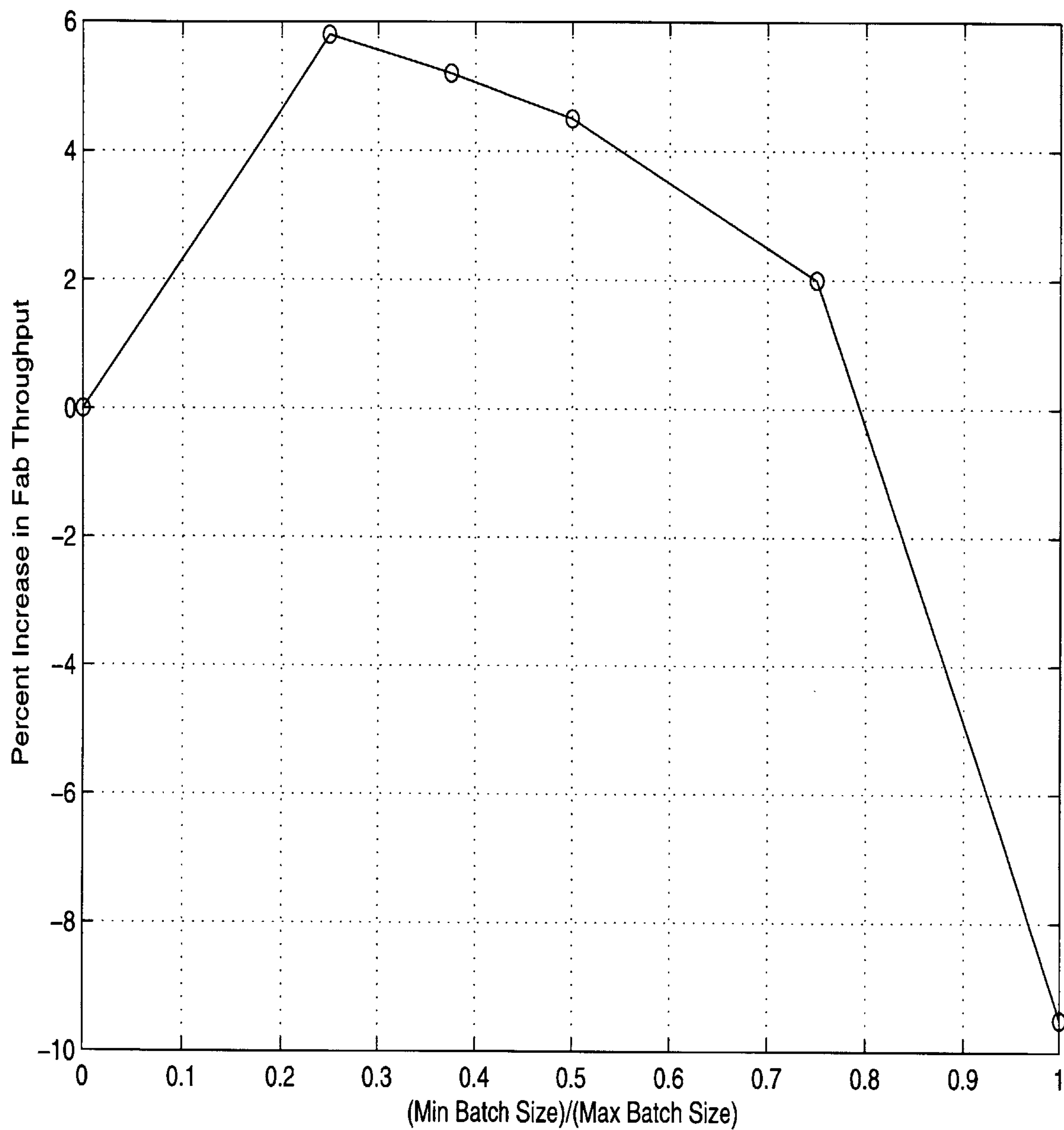


Figure 4.4: Percent Increase in Fab shipments vs. Minimum batch size of furnaces

Figure 4.4 shows the impact of size of the minimum batch size on the throughput of the fab. The throughput suffers if the minimum batch size is too small because of low utilization of the furnaces which causes starvation at downstream equipment as well.

Similarly, it suffers if the minimum batch size is made too large because of increased wait times and the low frequency of runs at the furnaces. The optimal minimum batch size for maximum improvements in throughput was also 0.25*(load size of furnace).

4.3 Conclusion

In this chapter, we described a simulation study that was conducted to understand the impact of equipment utilization on cycle time and throughput of the factory. The case analysis was done using the scenarios at Analog Devices' Wilmington Wafer Fab. We show that a simulation model can be used to determine optimal dispatch policies at such operations and gain insight into the interdependencies of the equipment utilization and performance metrics. The simulation results suggest that diffusion/CVD equipment should be operated at an optimal minimum batch size which is based on demand. This minimum batch size is obtained from a simulation based optimization of cycle time and throughput rate of the entire fab. Too small a minimum batch size decreases the throughput of the factory due to low equipment utilization and too large a minimum batch size increases the cycle time of the fab due to large wait times for lots at the furnaces. The model determines the optimum utilization level against these two performance metrics.

Chapter 5

Conclusion

5.1 Overview

This thesis work was finished over the course of a year. We have developed a discrete time continuous flow model of the Wilmington wafer fab using a job shop model developed by Graves (1986). We also conducted some simulation experiments to understand the relationship of equipment utilization to cycle time and throughput performance. Here we briefly describe our work and mention some further research work that could be pursued.

5.2 Tactical Planning Model

Our intent at the beginning of the project was to develop a model that provides insight into the operational behavior of the fab. We wanted to quantify the inter-relationships among the work flow variability, input variability, and lead times of the fab at a tactical level. We modeled the work flow variability as the sum of three components: input variability, variability due to lot sizing and random variability. For simplicity, we chose 22 different work flows[1] to represent all the work flows and empirically determined the lead times for each work flow. Each work flow was represented by the furnace process steps and demand-based production requirements were characterized at the furnaces for each work flow. We estimated model parameters based on a 21 week history of production and WIP data. With the exception of 2 furnaces the production requirements determined by the model reasonably correlated with the historical data. We used the model to obtain the relationships of starts, lot sizes, and planned lead times to production requirements. We also showed how this model can be used to allow production smoothing in the fab.

1. 10 of the 22 flows had negligible demand.

5.3 Simulation Study

We conducted a short simulation study to understand the impact of equipment utilization on performance metrics like throughput and cycle time. We used a discrete-event factory simulation model that was developed outside the scope of this thesis. We pre-set the utilization levels of the diffusion/CVD equipment and simulated production for a year's duration. Our results indicated that high utilization of equipment wasn't necessarily the best strategy for improving fab performance. In our experiments, a batch minimum size of 0.25*(maximum load size) of the diffusion equipment provided the highest throughput and lowest cycle time for the fab.

5.4 Future Direction

The operations personnel in Wilmington were skeptical about the applicability of the model for a few reasons that were discussed in chapter 3. They liked the modeling approach undertaken in the thesis and were confident that such a model would be of great value in wafer fabs with a few process flows. However, they felt that in the Wilmington wafer fab scenario where there are over 300 process flows, this model might not represent reality very accurately. Their main reservation against the model was that it suggested uniform starts release as the best possible strategy for the fab. However, in practice they found that such an approach leads to a really high cycle time due to low equipment utilization. However, they felt that the model could be of value in determining expected production requirements with changes in starts and identifying which equipment would be facing capacity constraints with such changes.

The ADI personnel will be evaluating the use of this model as an aid towards the implementation of a pull system for inventory control in the fab. The basic idea of a kanban system is that when a machine starts processing a lot, a signal percolates back to the upstream machine to start another lot. This system helps regulate the total inventory in the

system. We think the tactical planning model could be of value in determining the appropriate design of a kanban system for the fab. Recall that each workstation is really a series of process steps (at different equipment). By setting the kanbans at the boundaries defined by the workstations of the tactical planning model, we can use the model to determine the amount of WIP that should be in each kanban. It is simply given by the expected queue lengths for each workstation (see section 2.2). The advantages of this approach could be tested using a simulation based approach.

References

[1] Billington, P. J., J.O. McClain, and L. J. Thomas. 1983. Mathematical Programming Approaches To Capacity-Constrained MRP Systems: Review, Formulation And Problem Reduction. *Mgmt.Sci.* 29, 1126-1141

[2] Billington, P. J., J.O. McClain, and L. J. Thomas. 1986. Heuristics For Multilevel Lot-Sizing With A Bottleneck. *Mgmt Sci.* 32, 989-1006.

[3] Bispo, C. F., and S. Tayur. 1997. Managing Simple Re-entrant Flow Lines I,II & III. *GSIA working paper*, CMU, PA.

[4] Chen, H. et. al. 1987. Empirical Evaluation Of A Queueing Network Model For Semiconductor Wafer Fabrication. *Oper. Res.* 36, 202-215.

[5] Cruickshanks, A. B., R. D. Drescher, and S. C. Graves. 1984. A Study of Production Smoothing In A Job Shop Environment. *Mgmt. Sci.* 30, 368-381.

[6] Deleersnyder, J., T. J. Hodgson, H. Muller, and P. J. O'Grady. 1989. Kanban Controlled Pull Systems: An Analytic Approach. *Mgmt. Sci.* 35, 1079-1091.

[7] Dobson, G., U. S. Karmakar, and J. L. Rummel. 1987. Batching To Minimize Flow Times On One Machine. *Mgmt. Sci.* 33, 784-799.

[8] Fine, C. H., and S. C. Graves. 1989. A Tactical Planning Model For Manufacturing Subcomponents Of Mainframe Computers. *J. Mfg. Oper. Mgmt.* 2, 4-34.

[9] Graves, S. C. 1986. A Tactical Planning Model For A Job Shop. *Oper. Res.* 34, 522-533.

[10] Graves, S. C. 1983. Scheduling Of Re-Entrant Flow Shops. *J. Oper. Mgmt.* 3, 197-207.

[11] Graves, S. C. 1988. Extensions To A Tactical Planning Model For A Job Shop. *Proceedings of the 27th IEEE Conference on Decision and Control, Austin, Texas,* December 1988.

[12] Groeneelt, H., and U. S. Karmakar. 1988. A Dynamic Kanban System Case Study. *Prod. Inv. Mgmt. J.* Second Quarter, 46-50.

[13] Karmakar, U. S. 1987. Lot Sizes, Lead Times And In-Process Inventories. *Mgmt. Sci.* 33, 409-418

[14] Karmakar, U. S. 1987. Lot-Sizing And Sequencing Delays. *Mgmt. Sci.* 33, 419-423.

[15] Mascolo, M. D. 1996. Analysis of a synchronization station for the performance evaluation of a kanban system with general arrival process of demands. *Eur. J. Oper Res.* 89, 147-163.

[16] Segerstedt, A. 1996. A capacity-constrained multi-level inventory and production control problem.*Int. J. Prod. Eco.* 45, 449-461.

[17] Uzsoy, R., C-Y Lee, and L. A. Martin-Vega. 1992. *IIE trans.* 24, 4-60.

[18] Yanagawa, Y. and S. Miyazaki. 1994. An optimal operation planning for the fixed quantity withdrawal Kanban system with variable lead times. *Int. J. Prod. Eco.* 33. 163-168.

8-71-92