

9

Managing Variation in Near-Constraint Systems

by

Kevin L Farrelly

B.S., Material Science and Engineering, Cornell University, 1993

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

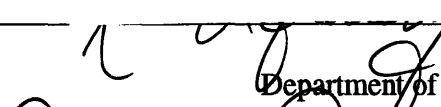
Master of Science in Management
and
Master of Science in Material Science and Engineering

in Conjunction with the
Leaders for Manufacturing Program

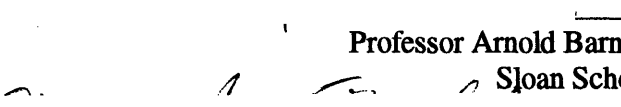
at the Massachusetts Institute of Technology
June 1998

© 1998 Massachusetts Institute of Technology, All rights reserved

Signature of Author _____


Sloan School of Management
Department of Material Science and Engineering
May 8, 1998


Certified by _____



Professor Arnold Barnett, Thesis Advisor
Sloan School of Management

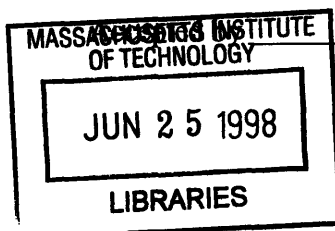
Certified by _____


Professor Eugene Fitzgerald, Thesis Advisor
Department of Material Science and Engineering

Accepted by _____


Linn W Hobbs
John F. Elliot Professor of Materials
Chairman, Departmental Committee on Graduate Students


Larry Abeln, Director of Master's Program
Sloan School of Management



Managing Variation in Near-Constraint Systems

by

Kevin L Farrelly

Submitted to

the Department of Materials Science and Engineering and

the Sloan School of Management

on May 8, 1998

in Partial Fulfillment of the Requirements for the Degrees of

Master of Science in Materials Science and Engineering

and

Master of Science in Management

Abstract

This thesis describes the research done during a Leaders for Manufacturing internship at Intel Corporation's Santa Clara process development facility. In line with theory of constraints, Intel's operational philosophy emphasizes constraint management techniques to maximize factory output. However, there are several systems in the production process that affects factory output other than the factory constraint. These systems, known as "near-constraints," experience availability interruptions that resemble "bottleneck-like" behavior. The focus of this thesis is to understand the impact of near-constraint systems in a high-volume factory and provide some insight on how to manage these systems.

To understand the impact near-constraint systems have on factory output, a statistical approach was utilized. Near-constraint systems become unavailable to run production when an unforeseen interruption occurs to the machine. Because the machine is unavailable to process material, work-in-process (WIP) begins to accumulate behind the system. The length of these interruptions varies considerably. There are occasions when the interruption will go virtually unnoticed and other occasions when the interruption will interfere with the arrival rate of WIP to the factory constraint or the end of the production line. The analysis of this thesis examines the impact one particular near-constraint system has on the WIP flow to the factory constraint.

The results of the statistical analysis show that the length of the near-constraint interruption is a key variable in determining the probability as well as duration of a WIP flow interruption. Additionally, the results of the analysis suggest a methodology to calculate a level of inventory necessary to avoid WIP flow interruptions to critical systems.

Thesis Advisors

Professor Arnold Barnett, Sloan School of Management

Professor Eugene Fitzgerald, Department of Material Science and Engineering

Acknowledgements

There have been many people who have supported me through this experience, but none more than my wife Christine. There are not many people that would have tolerated the disruptions that were required to make this thesis possible, and I will be eternally grateful to her for helping me through this adventure. During the LFM experience she relocated across the country 3 times, gave up her work, and gave birth to my son. She is now, and forever will be, my most valuable asset.

I would also like to thank my children: Cassy and Brandon. They have been a source of happiness during many frustrating moments. Additionally, in order to make my dreams come true, Cassy had to enroll in four different elementary schools over the past two years, and when we move to Texas she will have to enroll in still another school. I know this was very difficult for her – constantly making new friends and having to leave them. I will always be grateful Cassy, and I will try my best to make it up to you.

I would also like to extend a special thanks to my thesis advisors: Professor Anorld Barnett and Professor Eugene Fitzgerald. Their guidance and perspective helped me overcome some difficult roadblocks as well as made this work an incredible learning experience.

I would like to thank all the people at Intel. A special thanks to Dennis Arnow and Don Myers for supervising the project. Also, I would like to thank Sean Cunningham for supporting my work. And Finally, the Breakfast Club (Pete Lancia, Doug Fong, and Chris Cowger) for helping out during times of confusion but mostly for making the internship an excellent experience.

I gratefully acknowledge the support and resources made available to me through the MIT Leaders for Manufacturing program, a partnership between MIT and U.S. Manufacturing. The experience has been nothing short of awesome.

Table of Contents

1. INTRODUCTION	11
2. SEMICONDUCTOR MANUFACTURING AT INTEL CORPORATION	15
2.1 SEMICONDUCTOR MANUFACTURING OVERVIEW	15
2.2 REENTRANT FLOW MANUFACTURING	17
2.3 SEMICONDUCTOR PROCESSING EQUIPMENT	18
2.4 INFINITE BUFFERS	19
2.5 WIP PRODUCTION POLICIES IN SEMICONDUCTOR PROCESSING	20
3. PROBLEM DEFINITION	23
3.1 CONSTRAINT MANAGEMENT	23
3.2 PROCESSING EQUIPMENT RELIABILITY	25
3.3 IMPLICATIONS OF EXCESSIVE DOWNTIME	26
3.4 PROBLEM STATEMENT	27
4. APPROACH	29
4.1 IDENTIFY THE KEY PROCESSING EQUIPMENT	29
4.2 EXTRACTING INVENTORY PROCESSING TIMES	30
4.3 TIME FROM IMPLANTER TO CONSTRAINT	31
4.4 IDENTIFYING INTERRUPTIONS TO IMPLANTER OUTPUT	33
4.5 IDENTIFYING CORRESPONDING INTERRUPTIONS TO WIP ARRIVING AT THE CONSTRAINT	34
4.6 CRITICAL BUFFER SIZE	35
5. RESULTS	39
5.1 THE SHUFFLING OF PRODUCTION	39
5.2 PRODUCTION FLOW SMOOTHING	41
5.3 DETERMINING THE PROBABILITY OF AN INTERRUPTION	43
5.4 DETERMINING THE CRITICAL BUFFER SIZE (CBS)	48
6. FUTURE CONSIDERATION	53
6.1 HOW CAN INTEL USE THIS ANALYSIS IN PRODUCTION?	53
6.2 A SHIFT IN PERSPECTIVE	54
6.3 USING THIS METHODOLOGY WITH DOWNSTREAM SYSTEMS	54
6.4 POTENTIAL ISSUES WITH THE RESULTS	54
6.5 UPDATED INFORMATION IS ESSENTIAL	56

Table of Figures

Figure 1: The process of developing a silicon wafer into devices that are ready to be mounted onto a circuit board.	16
Figure 2: Cross section of a typical semiconductor device.	17
Figure 3: A typical reentrant flow process.	18
Figure 4: Model of how theory of constraints works.	24
Figure 5: Description of operations Upstream from the constraint operation.	28
Figure 6: Distribution of arrival times from the implant operation to the Constraint operation.	32
Figure 7: Data Summary of the Implanter to Constraint distribution.	33
Figure 8: Table of Inventory processed at the implanter and at the constraint.	34
Figure 9: Model of inventory distributions as they travel from the implanter to the constraint.	42
Figure 10: Graph of Implanter Output interruptions vs. Constraint Arrival interruptions.	44
Figure 11: Data summary of implanter output interruptions vs. constraint arrival interruptions.	47
Figure 12: Graphical summary of implanter output interruptions vs. constraint arrival interruptions.	47
Figure 13: Inventory graph to compensate for interruptions at the implanter that are less than 24 hours.	50
Figure 14: Inventory graph to compensate for interruptions at the implanter that are greater than 24 hours.	51

1. Introduction

The research of this thesis was conducted through a six-month internship at Intel Corporation. During the first few weeks of the internship I was told that factory output variation was a constant concern of factory management. One objective of the operations group was to maintain consistent factory output. There appeared to be a constant fight to maintain machine uptime and a balanced production line.* Operations management felt that if you reduce level of variation in production the factory would in turn produce consistent and predictable output.

Intel's manufacturing facilities are state-of-the-art. The manufacturing engineering department, which aid in designing the factory, works with the operations group to determine the factory constraint. This allows the operations group to develop a theory of constraints model when managing operations. Everyone in operations knew which machine in the factory was the constraint. However, there was very little discussion on managing the constraint at the daily operation meetings. There was however, a great deal of discussion about another system: the ion implanter.

This was rather strange to me. According to the theory of constraints as depicted in Goldrat's book The Goal, the factory constraint should be the focus of the factory. The other systems in the factory were supposed to be non-issues. However, meeting after meeting was consumed with discussions about the ion implanter. After several discussions with individuals close to the source, the reason for this somewhat surprising emphasize became clear.

* A balanced line is a manufacturing line were work is distributed evenly throughout the process as opposed to piling up in certain regions.

The constraint is the system with the lowest throughput capacity. Compared to all the other machines in the factory, including the ion-implanter, this machine is the one that limited factory output.

The implanter was however, doing something the constraint was not. The implanter was breaking down frequently and for long and unpredictable durations. These interruptions were causing large amounts of work-in-process (WIP) to pile up behind the machine. Additionally, these interruptions in production caused large gaps in WIP flow throughout the process. Management saw this as an introduction of unwanted variation in the process and felt it required continuous monitoring. They even had a name for systems that behaved in this manner: near-constraints.

The situation escalated over the course of the next several weeks. I recall the situation got to the point where they would call a SWAT* every time the machine went down for any reason. This was a considerable expenditure of internal resources.

After spending some time discussing the issue with several people that worked closely with the situation, it became clear that the impact of these interruptions was not clear. No one knew what was a reasonable time that this system could be down before the interruption affected factory output. For example, if the implanter was unavailable for production for one hour would someone monitoring factory output even notice – probably not. However, if the system was down for a month output would certainly be impacted. So, how long can a near-constraint system be interrupted before it starves the true factory constraint?

The pursuit to answer this question is the research of this thesis. Some considerably long interruptions at the implant operation had no negative repercussions to production. While

* A SWAT was an Intel internal term used to signal the highest level of priority on a system that had failed and the recovery is uncertain. This status is usually reserved for critical systems that have been unavailable for greater than 24 hours.

some comparatively shorter interruptions had a large impact to WIP flow. For this reason I thought that a statistical approach to the problem was in order.

This paper will discuss some of the specifics of the problem that Intel was trying to manage. Chapter 2 provides some background of the semiconductor processing as well as some specific WIP policies used at Intel. Not all this information is necessary to understand the approach or results of the study. The methodology that I have taken is presented in a manner that I hope will be applicable in a variety of cellular manufacturing operations. However, having some appreciation of the industry and especially WIP management policies used in semiconductor processing will certainly aid in the understanding of the analysis. Chapter 3 discusses the specifics of the problem as well as reviews some of the constraint management principals used in the analysis. Chapter 4 discusses the details of how I approached the problem and chapter 5 reviews the results of the analysis. Finally, chapter 6 discusses how the results can be used to manage near-constraint systems in high-volume manufacturing environments.

2. Semiconductor Manufacturing At Intel Corporation

The manufacturing of microprocessors at Intel Corporation is similar to semiconductor manufacturing throughout the industry. In this section, I discuss the issues associated with a semiconductor-manufacturing environment. Processing schemes such as “reentrant flow processing” and infinite buffers are utilized throughout the industry and present some unique processing problems. Additionally, semiconductor WIP management policies are discussed to give the reader some appreciation of how the industry manages these unique processing issues.

Another problem that semiconductor manufacturers face is production flow variation. Mostly due to the way semiconductor-processing equipment behaves, production flow variation issues consume a great deal of internal resources and creates considerable anxiety in the day-to-day operations. This chapter addresses some of the mechanisms that create production flow variation as well as addresses some of the ways to manage it.

2.1 Semiconductor Manufacturing Overview

Before it is a computer chip or a memory device, semiconductors start out as silicon wafer. Silicon wafers, which are typically eight inches in diameter, are released into the production process in batches of 25, known as lots. Each lot is given an identification number so the processing history can be easily tracked. These lots go through extensive processing which takes anywhere from several weeks to several months. As shown in the figure below, the silicon wafers are processed in a manner that will produce many small devices on the wafer’s surface. These devices are depicted as small squares known as die. After the wafers have completed processing, each die is tested for functionality. The devices that are not functional are marked with a spot of ink prior to cutting the wafer. As shown in Figure 1, the die are cut from the wafer and separated. The “good” die are then placed into packages which serve to provide protection and allow them to be easily mounted onto a circuit board.

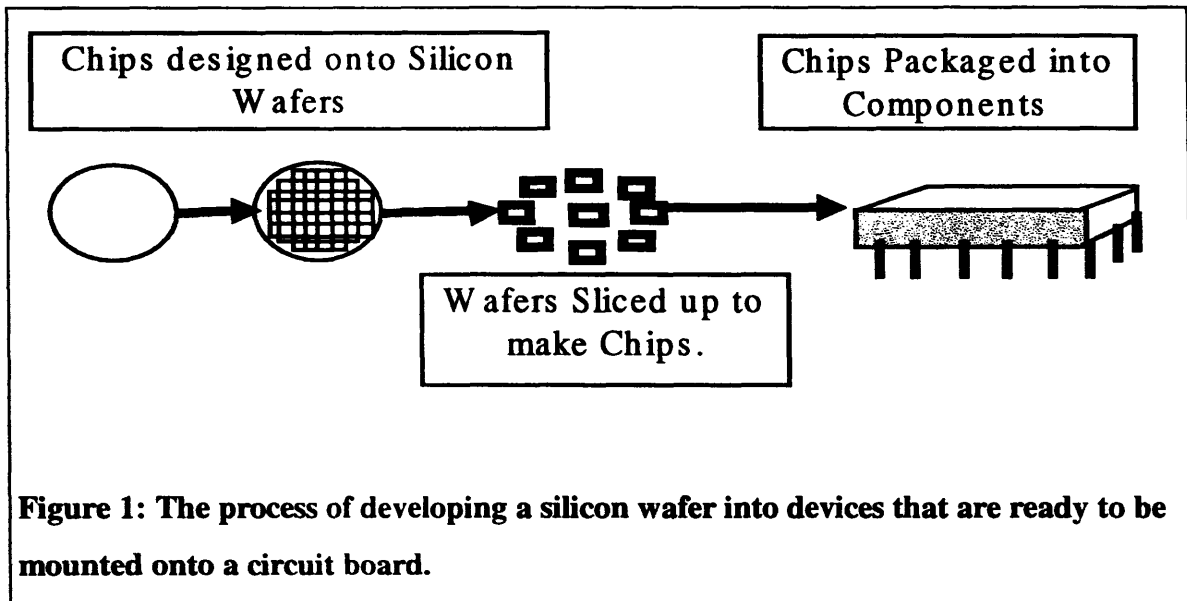


Figure 1: The process of developing a silicon wafer into devices that are ready to be mounted onto a circuit board.

The process of creating these devices on the wafer is traditionally broken down into two parts: front-end and back-end. The first part of the process, the front-end, involves creating the devices on the wafer's surface. This involves processing techniques such as photolithography and ion implantation. Photolithography is the process of depositing a temporary polymer film, known as resist, onto the surface of the wafer, exposing selected areas with light, and then removing these exposed areas. Once the process has been done, the ion implanter accelerates a dopant element towards the surface of the wafer with a force great enough to embed the material into the silicon. However, this will only occur in the areas that were previously exposed to light during the photolithography process. The other areas of the wafer are still protected by resist. These implanted areas create the devices that in turn create the computing properties of the die. During the second half of the process, the back-end, metal and insulation material is deposited on top of the wafer to provide an electrical connection from the outside world to the devices on the wafer's surface. Figure 2 is a schematic of a typical cross-sectioned device after front-end and back-end processing. In this figure, the "W" represents tungsten, which is a common material used to connect metal layers with each other and with the implanted regions.

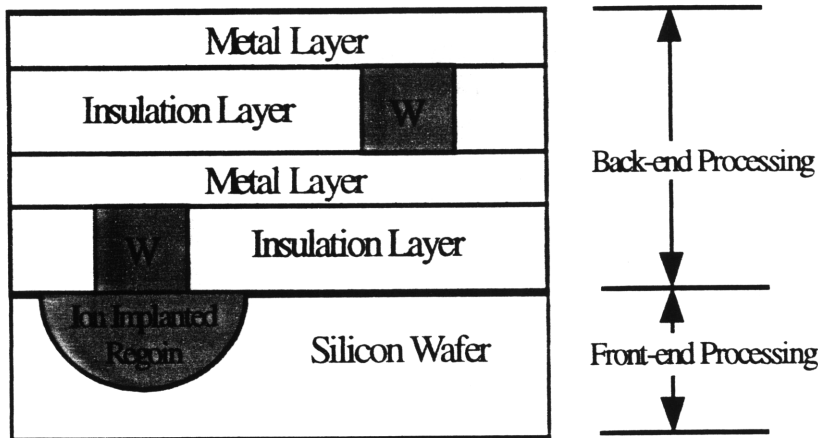
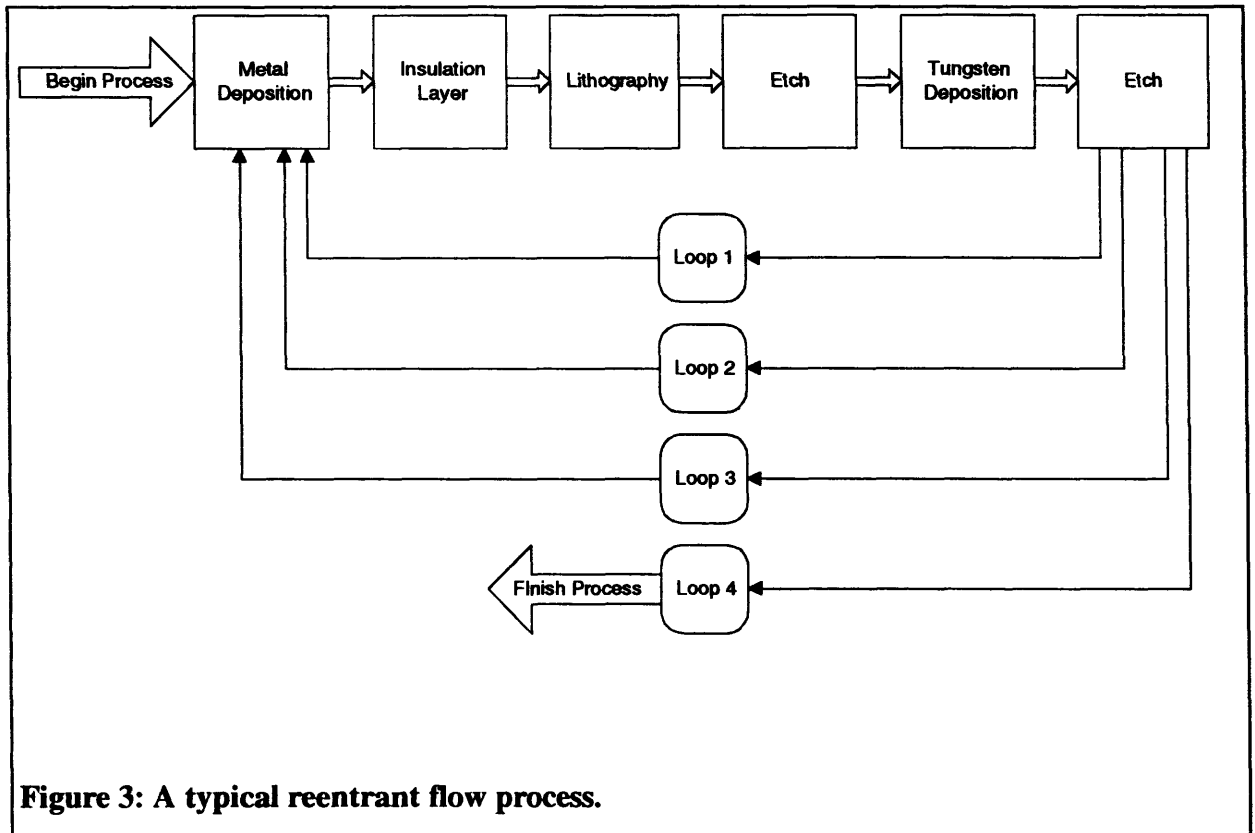


Figure 2: Cross section of a typical semiconductor device.

2.2 Reentrant Flow Manufacturing

Semiconductor manufacturing involves many repetitive processing steps. Although there may be well over a hundred steps in the process, WIP visits some of the same machines several times during the process. For example, a semiconductor component traditionally has several layers of metal as depicted in Figure 2. An insulating layer separates these metal layers. Although there are many processing steps in between the deposition of each metal layer, the same machine is used to deposit each metal layer. This process flow can be visualized as loops constantly flowing back to the same machines for more processing as shown in Figure 3. In this diagram the squares represent processing operations and the arrows indicate the direction that WIP will flow. This process will put four metal layers onto the wafer: once at the beginning, and three more layers when the wafer loops back to the metal deposition step. Since this wafer reenters the same processing steps again and again, it is known as “reentrant flow” processing.



2.3 Semiconductor Processing Equipment

The equipment used in processing wafers in semiconductor manufacturing is similar to equipment used in many other forms of manufacturing. Some of the machines use continuous processing, one lot at a time, other machines use a batch process, several lots of material are processed at once.

In a reentrant flow factory, both types of machines working together have a large impact on the distribution of WIP in the production line. Operators that run machines that process WIP in large batches usually wait for a significant amount of material to accumulate prior to running an operation. This usually creates a “feast or famine” effect

with the machines downstream. Continuous processing machines can usually change the processing operation from one lot to the next without any delays.

Some of the equipment in the factory requires an extensive set up procedure prior to running a particular operation. If this is the case, even the machine that can process lots continuously will only process one operation in a given time interval. This scenario will make a continuous processing machine behave similar to a batch-processing machine.

Regardless if the machine is continuous or batch, processing equipment used in the semiconductor industry is extremely complicated. Many systems are using technologies that have been developed only a few years ago. Even with the high level of preventative maintenance that these systems receive, semiconductor equipment is constantly failing.

Both types of machines working together in the process, along with random failures occurring throughout the production line, create WIP level variation in the production process. It is not uncommon to hear the operations group refer to these large pockets of WIP as “WIP bubbles.” Minimizing these WIP bubbles is a constant goal to maintain a balanced production line.

2.4 Infinite Buffers

Although it may take weeks or even months for one lot to complete its way through the manufacturing line, the lot is only being processed a small portion of the time. Most of the time the lot is waiting to be processed. Similar to many different types of manufacturing, semiconductor manufacturing stores material waiting to be processed in “buffers.” One function of buffers is to decouple processing machines. If one machine in the process flow is interrupted, other machines can continue to process material by taking inventory in and out of buffers. Machines upstream from the interruption can continue to operate until all the buffers from that machine to the interrupted machine are full. This is known as blocking. Likewise, the machines downstream from the buffers can operate until all the buffers from the interrupted machine to the machine of interest are empty. This is known as starvation.

At Intel, the buffers are known as infinite buffers. This term is used because, in essence, the buffers can never fill up, thus they can never block incoming production. However, they can run out of material resulting in starvation of downstream machines.

2.5 WIP Production Policies in Semiconductor Processing

Since many of the processing machines at Intel perform several operations, an artifact of the reentrant flow system, there are many times the operator has to choose which operation to run on the equipment. In order to give some guidelines to this decision, Intel employs several operating rules that are standard to the semiconductor industry: “Back to Front” and “First-In-First-Out” (FIFO).

Back to Front is an operation technique that is common in reentrant flow systems. In the simplest of terms, the Back to Front rule prioritizes WIP that is closest to the end of the production line. For example, if there are four metal layers on the semiconductor device, the wafer will return to the metal deposition operation 4 times, as described in section 2.2. This means the operator for the metal deposition machine may have several different operations to run at any given time: Metal 1, Metal 2, Metal 3, and Metal 4. The Back to Front rule gives priority to the Metal 4 operation since it is the closest to the end of the production line. Then, the Metal 3 operation should be processed followed by the Metal 2 operation and finally the Metal 1 operation.

Within a given operation, the First-in-First-out rule prioritizes the lots to be processed. FIFO says the lot that has arrived to the operation step first should be processed first. Continuing with the above example, let us assume that an operator is going to process the Metal 3 operation. Let us also assume that there are three lots waiting to be processed for Metal 3. The operator should choose to process the lot that has been waiting at the operation the longest.

These procedures have several benefits. First, operators have some guidelines on how to prioritize the work to be processed. Secondly, these procedures ensure that material is not

sitting idle for extended periods of time. But most importantly, these methods will serve as a reasonable way to maintain a balanced line.

3. Problem Definition

Semiconductor manufacturing has many of the same issues as other forms of manufacturing. For example, the theory of constraints is as applicable in a fab as it is in an automotive assembly line. The output of the factory is determined by how the constraint is managed as well as by the reliability of other machines that perform critical operations in the systems. Although the theory of constraints provides a great deal of understanding on how to manage the constraint operation, managing other areas is not as clear.

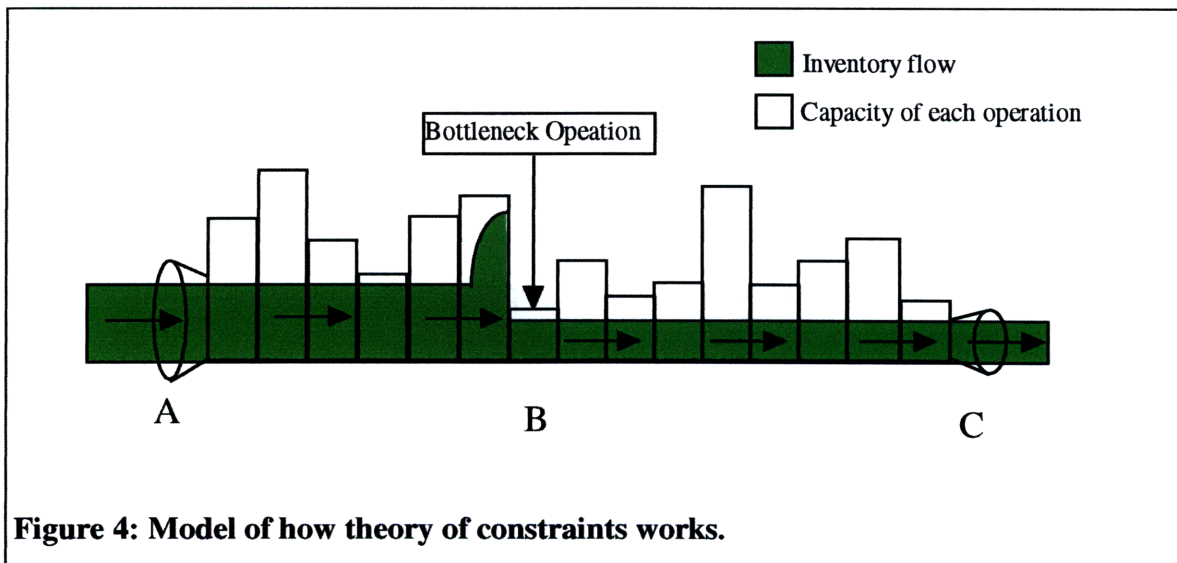
3.1 Constraint Management

During the design stages of Intel factories, the manufacturing engineers attempt to create a balanced line by setting up each operation area to have similar throughput capacity. This is very difficult to do; actually, it is almost impossible to do. Some systems by definition will have more capacity than others. After looking at the completed design plans for the factory, it is a relatively trivial task to determine which operation will be the factory constraint. However, there are other operations in the factory whose throughput capacity is very close to the constraint's (i.e. near-constraints). The throughput capacity of these areas is slightly higher than the constraint's throughput capacity. If these areas experience an excessive amount of unscheduled interruptions, they will become the temporary constraint of the factory.

According to the theory of constraints as depicted in Goldrat's book The Goal, a factory that wants to have consistent output needs to manage the factory by managing the constraint. The factory constraint is considered machine or operation with the lowest throughput capacity. Factory output can not exceed the rate of the rate-limiting step, (i.e. the rate of the constraint). To input material into the factory at a rate greater than the rate that the constraint can process that material will only result in material stacking up in front of the constraint operation.

Figure 4 represents inventory flowing through a simple manufacturing production line. At the very right of the diagram, inventory flow, depicted in green, is flowing into the factory line at point A. Each of the processing operations, depicted by the fifteen rectangles between A and C, have a wide range of capacities, illustrated by the height of the rectangles (also referred to as throughput rate). Inventory can flow easily through any operation if the throughput rate of the operation is greater than the rate that inventory is flowing into the operation. If however, inventory flows into an operation at a rate that exceeds the operation throughput rate, WIP will accumulate in front of the operation.

This event occurs at point B in Figure 4. As shown in the diagram, the flow of inventory from point B to the end of the factory line, point C, is now determined by the rate at which the constraint can process material.



There are four basic states that the constraint can exist. First, the most favorable scenario, the constraint is available to process material and there is sufficient inventory to process. Second, the constraint is not available to process material due to maintenance or possibly some unplanned mishap; however, there is material waiting to be processed. Third, the constraint may be unavailable and there is no inventory to process. This scenario is certainly not desirable, but is not be tragic either. As stated before, machines become

unavailable due to scheduled and unscheduled maintenance and if there is not any WIP to process during this time it may go unnoticed. In the most fortuitous of circumstances a factory manager may be able to take advantage of this situation by planning maintenance for the constraint operation during times of low inventory. The final state of the factory constraint occurs when the constraint is available, but there is no material to process. This is the most devastating of all circumstances. This occurrence can happen when machines upstream from the constraint are interrupted for an extended period of time. This upstream interruption will result in an interruption of inventory flow to the constraint. If there is not sufficient buffer inventory to compensate for the upstream interruption, the constraint will be starved.

3.2 Processing Equipment Reliability

Consistent output of the factory is dependent on several important factors. The first factor, which was discussed above, is the rate of the lowest throughput capacity operation (i.e. the constraint). The second factor involves extended interruptions to machines that feed the constraint or the end of the production line. If these machines go down for too long, they will starve the constraint operation or the end of the production line. The third factor is the frequency and duration of unsynchronized, disruptive events that can occur at any machine in the factory.

Machines run for a length of time with a probability of failure. In literature, this issue usually discussed in terms of mean-time-to-failure (MTTF). Once the machine has failed, the next probability of interest is known as the probability of repair, often referred to as the mean-time-to-repair (MTTR). Machines that are unexpectedly interrupted need to go through a series of events before the root cause of the interruption is corrected. The maintenance technicians diagnose the issue, assess the damage, initiate corrective action, test the repair, and finally qualify the system before the machine can continue processing. Each one of these steps takes time. The speed of each step depends on many variables such as the skills of the technician, the difficulty of diagnosing the issue, the availability of replacement parts, and the potential of causing additional damage to the system during the

repair process. Together, with still further time-consuming issues not mentioned, the time an interrupted system will be repaired and available to process material is probabilistic.

There has been a great deal of work done to model factory output once these factors have been quantified for all machines. However, even the best models can only estimate an average output over a considerable period of time. Calculating exact factory output for any given day is considerably more challenging. Reason being, even in a relatively simple manufacturing line, with only a few machines and buffers, where machines can be in one of two states, operational or under repair, there is an astronomical number of possible states in which the manufacturing line can exist. This makes predicting output for any given moment virtually impossible.

3.3 Implications of Excessive Downtime

The two most common ways to address production variation is to increase the reliability of the machines or increase the size of the inventory in the buffers. The first solution is limited by the available technology. The second solution is an extremely costly.

Ignoring unsynchronized events for a moment, there are typically three cases of machine interruptions, known as machine downtime, which are of concern: interruptions at the constraint, interruptions with machines before the constraint, and interruptions with machines after the constraint. If the constraint is not available to process material, total production of the factory will suffer. As stated earlier, the rate production flows out of the factory is established by the constraint. If the constraint is not producing, the factory's production will decline accordingly.

Machines before the constraint are a different story. They can be down for a period of time without repercussion. However, if they are down for an extended period of time eventually the constraint will run out of material to process as stated in section 3.1. This situation will result in a corresponding decrease in factory production.

There are four variables that determine the impact of an upstream machine's interruption to the constraint: (1) the distance the machine is from of the constraint, (2) the amount of time the machine is down (3) the machine's throughput capacity as well as the throughput capacity of machines between the interruption and the constraint, (4) the amount of buffer inventory between the interrupted machine and the constraint.

The last situation, machines become unavailable after the constraint, is similar to machines before the constraint. In this scenario the concern is not with starving the constraint, but with interrupting the production flow to the end of the manufacturing line. The impact of this situation depends on four variables as well: (1) the distance the machine is from the end of the production line, (2) the amount of time the machine is down (3) the machine's throughput capacity as well as the throughput capacity of downstream machines (4) the amount of buffer inventory between the interrupted machine and the end of the manufacturing line.

3.4 Problem Statement

Although the theory of constraints allows a factory manager to determine the cost of downtime at the constraint, it does not provide a clear answer to determine the cost of interruptions at other machines in the factory.

To determine the cost of an interruption to a system upstream from the constraint we need to understand how and when these interruptions affect the constraint. If the interruption is significantly upstream and the duration of the interruption is short, the flow of WIP to the constraint may go uninterrupted. In this case the cost of the interruption is zero – not accounting for the cost to repair the system. If, on the other hand, the duration of the interruption was excessive or the system was close to the constraint, the interruption may result in constraint starvation – in which case the cost is very high.

Referring to the diagram in Figure 5, if we call the constraint machine “n,” then the first machine upstream from the constraint will be (n-1). It is relatively easy to determine how long machine (n-1) can be down before the interruption will result in starvation to the

constraint. In order to determine an acceptable amount of downtime for machine (n-1) you would simply calculate the amount of inventory between the two systems and the time required by the constraint to process this inventory. If machine (n-1) is not providing output to the buffer by this time, the constraint will be starved.

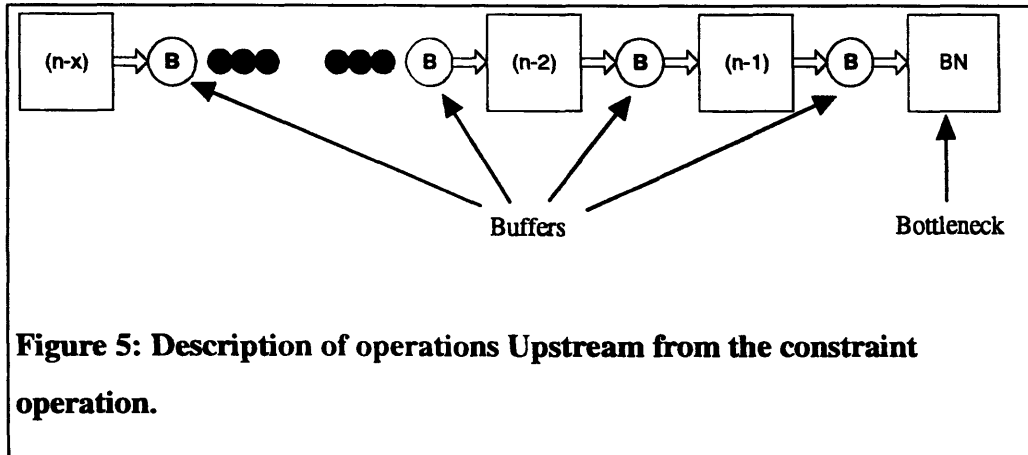


Figure 5: Description of operations Upstream from the constraint operation.

As we move further upstream from the constraint the calculations become far more complicated. To calculate the amount of time machine (n-x) can be down before starving the constraint requires the consideration of all the machines between (n-x) and the constraint. As stated earlier, calculating this with a relative degree of accuracy would almost be impossible due to the probabilistic nature of the entire system.

The focus of this thesis is to determine the probability that an interruption at an (n-x) machine will result in starving the constraint. Additionally, I will attempt to quantify the length of an interruption to WIP flow given and interruption to a (n-x) machine of various lengths. This methodology will also be applicable for machines upstream from the constraint operation. In these cases, the concern would be with interruption of WIP flow to the end of the production line. Finally, I will attempt to assess some of the alternatives available for this situation and demonstrate how possessing a quantitative understanding of the issue provides better decision-making abilities.

4. Approach

In order to assess the impact of an (n-x) machine interruption to the arrival rate of WIP to the constraint, the constraint and the appropriate (n-x) machine need to be identified.

Once the systems are identified, the next step would be to determine how WIP moves from the (n-x) machine to the constraint. If, after assessing a significant amount of data, it is determined that the time for WIP to travel from the (n-x) system to the constraint is probabilistic, a statistical approach may be utilized to determine the impact of an (n-x) machine interruptions.

The next step is to define and determine (n-x) output interruptions and then assess the impact of these interruptions to WIP arrival at the constraint. Finally, if the probability and the impact can be reasonably assessed, a *Critical Buffer Size (CBS)* between the (n-x) system and the constraint may be calculated to compensate for the (n-x) interruptions.

4.1 Identify the Key Processing Equipment

To determine the effect an (n-x) machine interruption may have on the constraint, the constraint needs to be correctly identified. Although this sounds like a trivial point, in many factories the constraint is inconspicuous. Fortunately, as stated earlier, everyone in the operations group knew exactly which operation was the constraint: the photolithography operation.

After interviewing several operation managers about this area, and analyzing the daily operations data, I was reasonably confident that this area was truly the constraint. Due to capacity issues, inventory accumulated quickly behind this operation. The amount of material the factory started each week into production was calibrated by photolithography capacity. Finally, factory output was limited by the operating performance of this machine.

The next step was to determine which (n-x) machine would be the most valuable to assess. I selected the appropriate machine based on three criteria: (1) the machine had to be several operations upstream from the constraint, (2) interruptions at this machine were frequent and of high concern to factory management, (3) it has to be a non-redundant machine (no other machines can be substituted to perform its operations). With this in mind, the ion implanter was a clear candidate; it met all the specified criteria.

4.2 Extracting Inventory Processing Times

In order to assess the behavior of WIP as it flows from the implanter to the constraint, I needed to extract the appropriate data. Fortunately, Intel factories collect an enormous amount of production data. In fact, the major issue initial was to determine which data was relevant for the analysis. As material moves from operation to operation, the time the inventory begins and ends processing is logged. The information that was the most prevalent for my analysis was the time inventory left the upstream machine, the implanter, and the time inventory arrived at the constraint. From the Intel database, the lot number and the times in and out of the two operations were extract.

Since the intent of this project is to provide a reasonable impact analysis of a system interruption during steady state processing, the extracted data had to be filtered. Because Intel's Santa Clara facility is a development site, some of the material processed at these two operations were for development purposes and did not follow the process flow of normal production material. Therefore, lots that were involved with non-standard processing had to be removed from the extracted data.

For example, there were lots in the system that were classified as "Hot Lots." This classification is given to lots that move quickly through the factory in order to qualify products or process changes. Since Hot Lots are processed at an accelerated rate, they did not represent steady state processing and therefore were extracted from the database before the analysis began.

Some of the lots that were processed during the time frame being analyzed were placed on “Hold” at specific operation for experimental purposes. These experiments are done for the purpose of optimizing an existing process step, qualifying new equipment or equipment upgrades, or qualifying a change to the process. In order to perform these experiments, the lot is placed on “Hold” while the equipment is being modified or the process engineer is setting up the experiment. This adds a considerable amount of time to the recorded time that a lot will spend at an operation. Since there were a considerable number of lots that went on Hold it appeared that it was part of Intel’s steady state operations. Therefore, I did not filter these lots from the data. However, it became important during the analysis to be aware of their presence in the factory.

4.3 Time from Implanter to Constraint

The next step in the analysis was to match lot numbers from material processed at the upstream station, ion implantation, with the lot numbers of material processed at the constraint operation. Once this was done, I determined how long it took WIP to move from the ion implant operation to the constraint operation. For each lot, the time the lot arrived at the constraint was subtracted from the time the lot left the implanter.

At this point it became clear that some of the lots took an extremely long time to transition from one station to the other. Since the lot numbers were at hand, I examined the processing history of the lots that took a long time to transition from one station to the next. Indeed many of them did go on Hold for experimentation purposes. As stated above, they were included in the analysis to determine the station to station travel time during steady state processing.

These times, measured in hours, were plotted in the form of a histogram (see Figure 6). The histogram demonstrates that the time from the implanter to the constraint operation is not deterministic, but normally distributed. Thus, a statistical approach to determine an acceptable amount of downtime for (n-x) machines is reasonable. The distribution had a mean time of travel from ion implantation to the constraint of 95 hours and a standard deviation of 25 hours. This information will become important during calculations of the

Critical Buffer Size, but more on that later. The table in Figure 7 gives a more complete summary of the data.

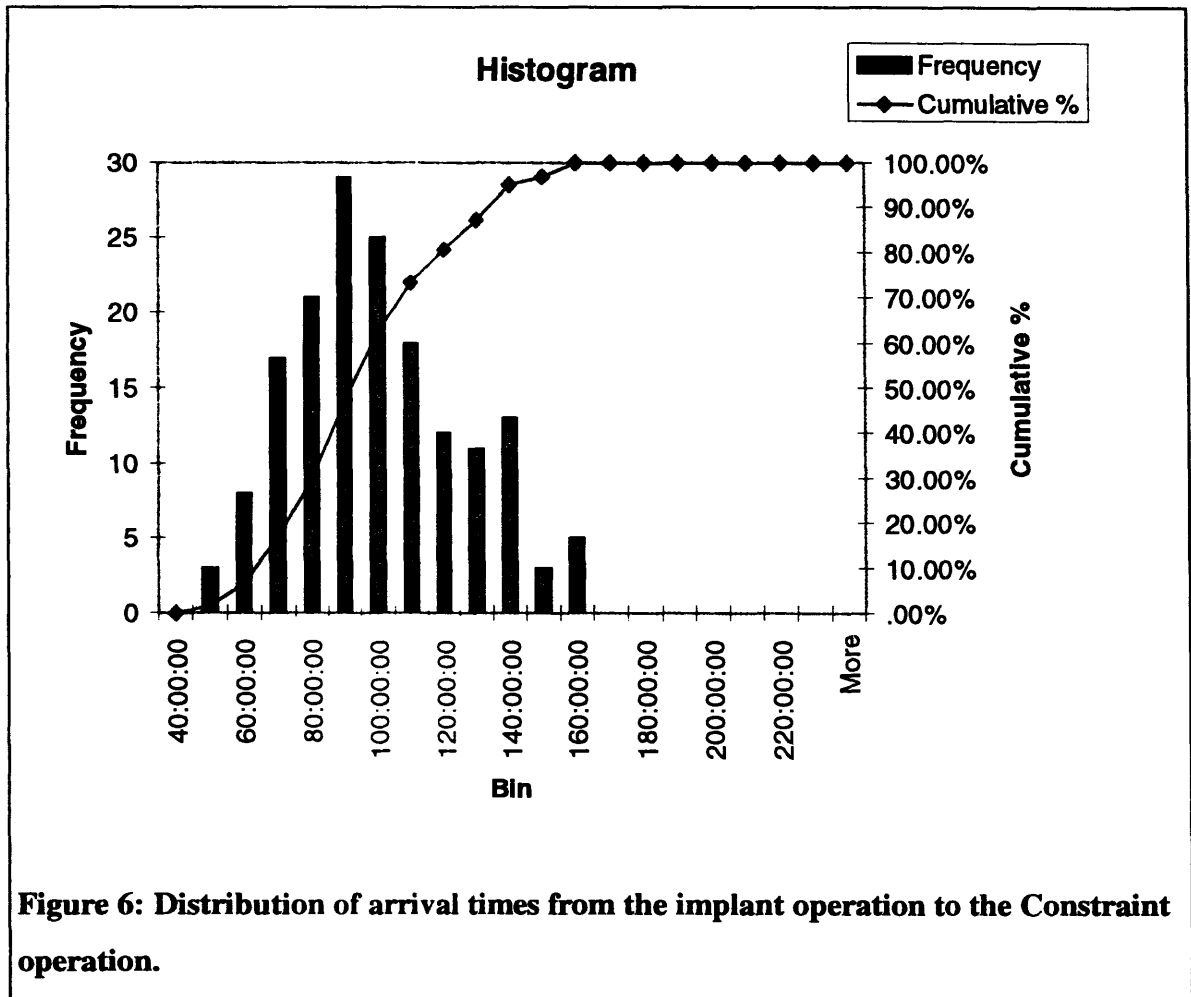


Figure 6: Distribution of arrival times from the implant operation to the Constraint operation.

Data Summary	
Mean	95:26:51
Median	92:53:21
Standard Deviation	25:47:07
Skews	9:23:42
Range	117:19:43
Minimum	41:18:19
Maximum	158:38:02

Figure 7: Data Summary of the Implanter to Constraint distribution.

4.4 Identifying Interruptions to Implanter Output

After the data was filtered, the interruptions to the implanter became visible. Figure 8 shows a table that represents part of the data that was extracted for the analysis. The left-hand side of the table in Figure 8 shows the lots sorted by the time they left the implanter. In the column labeled “Time from last lot released” shows the delta between sequential lot releases. As you look down the column, the interruptions to the WIP flow become obvious. At this point, I grouped lots together in batches that were released without interruption. I colored these groups for easy identification. The reason for this will become evident soon.

Time Lot Moved Out of the Implant Operation	Lot #	Time from last lot released	Lot #	Time Lot Moved Into the Constraint Operation	Time between arrivals to the constraint
7/20/97 6:10	27281370		27281370	7/24/97 6:23	
7/20/97 6:10	27281390	0:00:18	27281360	7/24/97 9:26	3:03:45
7/20/97 6:45	27288240	0:34:17	27281390	7/24/97 12:05	2:38:28
7/20/97 6:52	27281360	0:07:07	27288250	7/25/97 0:00	11:55:09
7/20/97 7:06	27288230	0:13:59	27288240	7/25/97 7:02	7:01:35
7/20/97 9:09	27288210	2:03:17	27288210	7/25/97 18:00	10:58:20
7/20/97 9:24	27278170	0:15:08	27291500	7/25/97 20:34	2:33:47
7/20/97 9:37	27288250	0:12:28	27281240	7/26/97 1:14	4:39:55
7/22/97 10:06	27291500	48:29:54	27278170	7/26/97 2:00	0:46:46
7/22/97 10:38	27281240	0:31:07	27291410	7/26/97 11:29	9:28:54
7/22/97 10:45	27291400	0:07:53	27291400	7/26/97 11:36	0:06:56
7/22/97 10:54	27291410	0:08:11	27298520	7/27/97 17:14	29:37:34
7/24/97 15:33	27298520	52:39:27	27298570	7/27/97 20:49	3:35:10
7/24/97 15:33	27298570	0:00:11	27291450	7/31/97 9:11	84:21:42
7/28/97 10:18	27291440	90:44:59	27291440	7/31/97 13:37	4:26:24
7/28/97 10:18	27291420	0:00:10	27291420	7/31/97 13:42	0:04:33
7/28/97 10:39	27291450	0:20:50	27291490	7/31/97 23:25	9:43:46
7/28/97 10:39	27291510	0:00:08	27291430	8/1/97 4:53	5:27:40
7/28/97 10:59	27291460	0:19:51	27291480	8/1/97 9:09	4:15:48
7/28/97 10:59	27291480	0:00:07	27291470	8/1/97 9:14	0:04:57
7/28/97 11:16	27291490	0:16:35	27291460	8/1/97 9:55	0:41:10
7/29/97 18:06	27291570	30:50:26	27291520	8/1/97 14:13	4:18:14
7/29/97 18:30	27291430	0:23:14	27291510	8/1/97 14:22	0:09:16
7/29/97 18:30	27291470	0:00:06	27291560	8/1/97 16:57	2:34:39
7/29/97 18:47	27291520	0:17:42	27291570	8/1/97 17:04	0:06:56
7/29/97 19:12	27291560	0:24:28	27308680	8/3/97 13:20	44:15:53
7/30/97 19:06	27309030	23:54:22	27309030	8/3/97 19:52	6:32:31
7/30/97 19:06	27291590	0:00:08	27308610	8/3/97 23:36	3:43:13
7/30/97 20:25	27308610	1:19:03	27291540	8/4/97 11:01	11:25:17
7/30/97 20:41	27308680	0:15:38	27291610	8/4/97 20:58	9:57:26
8/1/97 11:09	27291550	38:27:30	27291580	8/4/97 21:01	0:03:04
8/1/97 11:09	27291530	0:00:11	27291550	8/5/97 2:12	5:10:35
8/1/97 11:29	27291540	0:20:01	27309760	8/5/97 10:36	8:23:47
8/1/97 11:29	27291610	0:00:07	27309810	8/5/97 11:02	0:26:22
8/1/97 15:56	27291580	4:27:19	27309780	8/5/97 15:11	4:09:00

Figure 8: Table of Inventory processed at the implanter and at the constraint.

4.5 Identifying Corresponding Interruptions to WIP Arriving at the Constraint

Once interruptions to inventory flow had been identified at the implanter, the next step was to determine how and if these interruptions impacted the flow of inventory arriving to

the constraint operation. Using the same table in Figure 8, the lots were sorted again by arrival time to the constraint. This is shown on the right-hand side of the table. The coloring scheme from the original sorting was kept intact. This made it visually clear how material was shuffled during the travel process from the implant operation to the constraint. I created another column for time between arrivals to the constraint operation and the coloring scheme made identifying implant interruptions with corresponding constraint arrival interruptions very simple.

To determine exactly what constituted an interruption at the constraint I needed more information. I needed to determine how much inventory should arrive at the constraint under steady state operating conditions. According to the supervisor of the constraint operation, an ideal situation would exist if 150 wafers (approximately 6 lots of material) would arrive at the constraint per day or at least one lot arriving every 4 hours. This is equivalent to the number of wafers that are started each day at the beginning of the manufacturing process.

The time of each interruption at implant was then recorded next to the corresponding interruption to constraint arrivals -- regardless if it was greater than 4 hours.

Unfortunately, this process was not as straightforward as I originally hoped. There were several interesting issues and observations made when I was trying to make sense of all this data. However, once the data was plotted, the probability and impact of various implanter interruptions became evident. Chapter 5 presents a more detailed analysis of the data.

4.6 Critical Buffer Size

It is the intent of this research to not only identify the probability and impact of a near-constraint interruption, but also provide some insight on how to avoid or prepare for the consequences of such an interruption. After doing the statistical analysis, I started investigating why some interruption went unnoticed at the constraint and others did not. The answer to this question seemed to be in the distribution of WIP levels between the two systems.

In order for interruptions at the implanter to go unnoticed at the constraint it appears that there needs to be sufficient amount of WIP between the two stations to compensate for the interruption. If for some reason there was a lower than average level of inventory between the two systems, even a small interruption would be felt at the constraint. On the other hand, if there was a large level of inventory between the systems, the interruption may not be felt. This insight led me to determine what level of inventory would be necessary to compensate for various implanter interruptions.

Traditionally, the buffer is considered the storage place between two machines that perform sequential operations in a manufacturing process. However, I would like to treat all the buffers and the machines between the (n-x) tool and the constraint as a large buffer with a distribution of times of when the lots will be available for processing at the constraint. Clearly, lots that are in the buffer before the constraint are available immediately. Lots that are in the production line after the (n-x) will be available for the constraint process within a mean time of 95 hours.

Since inventory requires 4 days on average to travel from the implanter to the constraint, and there should be 150 wafers arriving at the constraint a day, there should be approximately 600 wafers between the two stations at any given time. If an interruption occurs, one would hope that there are enough wafers between the two stations to compensate for the interruption.

For example, supposed there was an extended interruption to the implanter. Under steady state conditions there would be approximately 600 wafers between the implanter and the constraint. However, what would happen if there were 800 wafers between the two stations? Since there is considerable amount of variation within the manufacturing process, this very well may be the case. In this situation the constraint may have more inventory to process than it normally needs. This scenario represents a situation where there may be plenty of inventory to compensate for the interruption.

Alternatively, there could be less than the ideal number of wafers between the two systems. These are the cases where starvation can occur at the constraint.

Therefore, the amount of material that would be required between the two station to maintain continuous processing at the constraint is known as the *Critical Buffer Size* (CBS). On average the CBS should be 600 wafers. But what would the CBS have to be to compensate for output interruption at the upstream station, i.e. the implanter? If the analysis revealed the impact that interruptions at the implanter had on arrivals to the constraint, the CBS would be relatively simple to calculate.

5. Results

Using the approach outlined above, there were several interesting observations made about the inventory flow out of and to the systems of interest. Additionally, the analysis suggests when responding to an interruption at the implanter is warranted and when to let nature take its course. The data also quantifies the level of response necessary to maintain a steady state flow of inventory to the constraint operation.

5.1 The Shuffling of Production

Once the lots had been sorted by time of departure from the ion implantation operation and time of arrival at constraint, an interesting observation was made. There was considerable shuffling in the order of the lots. If the factory worked by the production rules described earlier, this shuffling should not be occurring. All machines are supposed to be processing material with FIFO for each operation it performs. Therefore the first lot released in a batch from implantation should always be processed first at all downstream operations. Seeing this shuffling occur in every batch of material suggests that either the production rules are not being rigidly followed or there has to be something else affecting the order of the lots.

After further investigation, I identified that something else was occurring. Between the two stations there are many processing steps that every lot is required to complete. However, there are also several inspection steps that every lot is not required to complete. These inspection steps either measure critical dimensions, electronic devices parameters, or the level of particles* on the wafers. These testing steps within the manufacturing process are designed to contain issues before they affect a large portion of production. For example, in order to ensure the best possible yield of die on each wafer, the levels of particles need to be kept to a minimum. Thus frequent monitoring of lots allows the

* Particles are small pieces of matter that fall onto the surface of the wafer during production. Human contamination or machine failure usually causes them. Since the dimensions of the devices are in the order of 0.25 microns, even one, very small particle can cause catastrophic failure to a device.

factory to monitor and address any increase in defect levels. Once an increase in defect levels has been identified, the appropriate engineer can take corrective action.

Not every lot is reviewed at these inspection steps for several reasons. First, it is not practical to review every lot. Reviewing every lot would add considerable time to the total cycle time of the process. Secondly, since the throughput of many of the inspection machines is slow, inspecting every wafer of every lot would be expensive. The factory would require additional floor space to support all the extra machines (floor space in a wafer fabrication environment is very expensive). Additionally, the factory would have to purchase additional machines and hire many operators to run them. Therefore, the yield-engineering group determines a sampling frequency that would be adequate to monitor the process and contain any issue within a reasonable response time with minimal loss of factory yield.

This sampling scheme, known internally as “skip lot,” is key to understanding the shuffling of production as it travels through the production line. For example, if one lot out of every four lots is tested, the time required to travel through the line would be longer than lots that were not inspected. My first thought was that the increased cycle time should be equal to the time required to inspect the lot. However, since these lots will lose their position in the FIFO system, the delay will accumulate at subsequent processing steps.

Another factor that influences the shuffling of inventory is lots that are placed on Hold. As mentioned earlier, some of the inventory is placed on hold due to engineering experimentation as well as misprocessing.[†]

[†] Misprocessing is an Intel internal term used to describe occurrences where a mistake was made by the operator or processing machine during normal operations. This usually results in the lot being placed on hold until the process-engineer can disposition the lot accordingly.

5.2 Production Flow Smoothing

Another observation made after reviewing the data was the apparent inventory distribution smoothing. Although lots appear to be released from the implanter in batches, inventory arrived at the constraint somewhat spread out over time. This is best understood with the aid of the illustration in Figure 9. The implant operation releases material in batches illustrated as distributions labeled A, B and C on the top timeline. On the bottom timeline, the inventory that was very close together in time when it left the implant operation arrives at the constraint spread out in time.

In some circumstances, when the interruptions were small or the batches large (interruptions are considered time between the release of batches), the batches of production actually began to overlap each other. I call this the engine and caboose effect. This occurs when the lead lot of one batch, the engine, would catch up to and even pass the trailing lot, the caboose, of the batch in front of it. This is shown on the bottom timeline in the area labeled 2. In other incidents, the time between the releases of production batches resulted in an interruption to production arriving at the constraint. This is shown in Figure 9 in the area labeled 1.

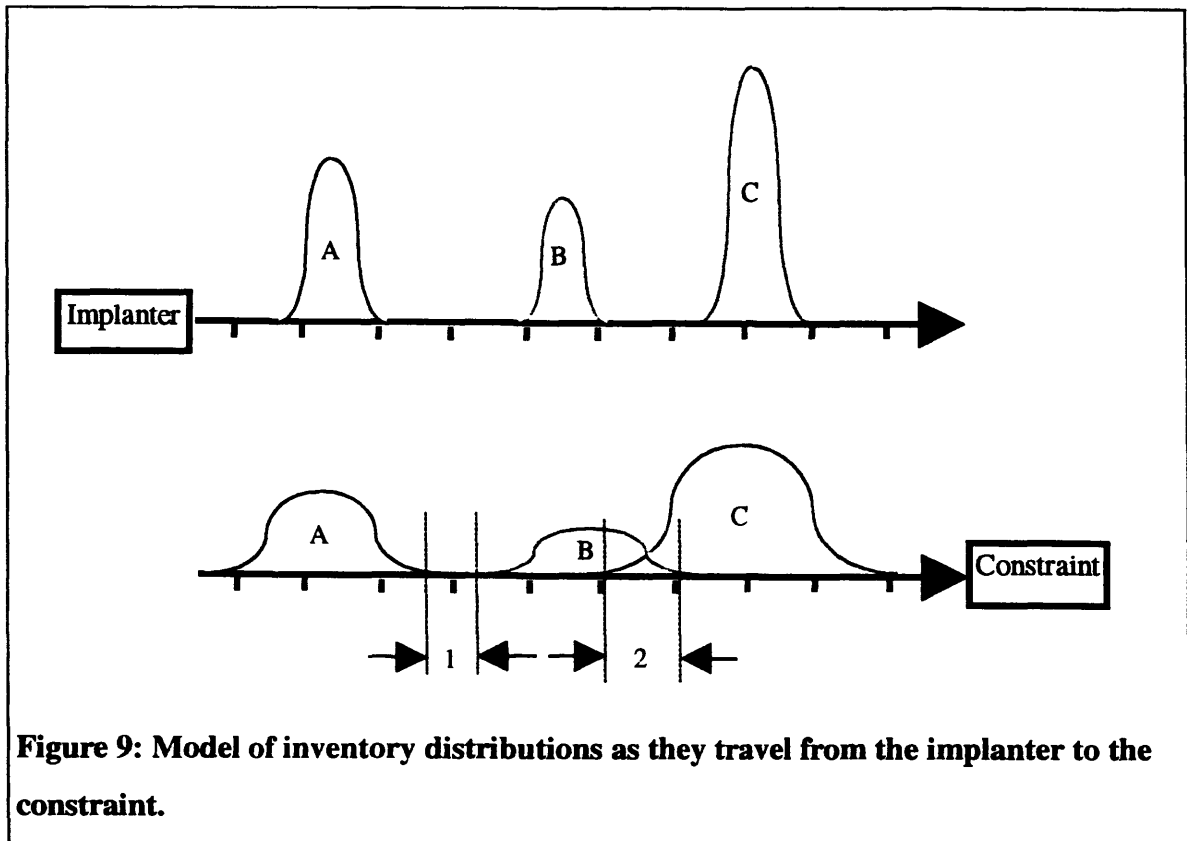


Figure 9: Model of inventory distributions as they travel from the implanter to the constraint.

What was causing this engine and caboose effect? Why did some implanter interruption reduce in magnitude once it reached the constraint? There must be something going on here.

After considering the production flow dynamics, the reasons for this behavior became clear. Imagine if you will that we are talking about the flow of traffic rather than the flow of production. Let us assume that you were driving down your favorite congested highway. Since it is congested, instead of cruising at your typical speed of 65 mph, you had to travel at the reduced speed of 50 mph. To make matters worse, let's assume that the highway department thought that it would be interesting to put a stop light on the highway. Since you are the first person to stop at the light, you look in your rearview mirror and notice a considerable amount of traffic backed up behind you. Once the light turns green you start to take off. However, instead of resuming your speed of 50 mph you accelerate to 65 mph, your preferred travel speed, because there is no in front of you.

Continuing with the highway analogy, let us assume there was a traffic engineer a significant distance down the road from the traffic light that was analyzing the rate that cars were coming down the road. Depending on how far this person was down the road and how long the stoplight was actually red, this person may or may not see the interruption. If you caught up with the last car that made it though the light (this car is still traveling at 50 mph) before you passed the engineer, the engineer may not notice an interruption to the flow. If you had not caught up to the trailing car by the time you passed the engineer, he/she would notice an interruption to the flow. However, since you are traveling at a rate greater than the pack of cars in front of you that did not stop for the light, you are continually closing the interruption gap in the flow. Therefore, the engineer will almost always see a shorter interruption to the flow than the length of time that the light was red, and moreover, the engineer will see smaller interruption the further he/she is from the light.

This situation is similar to the flow of production in the factory. The WIP waiting for the implanter to be repaired will sit for an unspecified period of time. Once the system is fixed this inventory will begin to flow. If the implanter was down for a long time, the systems that the implanter feeds will almost certainly be available to process the first lots out. These first lots, which would normally wait at each station prior to processing, will now flow from station to station without waiting – similar to the cars that were in the front of the line waiting for the traffic light to turn green.

The net result of this behavior would be that interruptions at the implanter would reduce in magnitude as the length of the interruption increase.

5.3 Determining the Probability of an Interruption

Once the data for the interruptions at the implanter was plotted against interruptions to constraint arrivals, it became apparent that not all interruptions are equal. For that matter, if one defines a significant interruption as anything that prevented arrivals for greater than 4 hours, it became apparent that the length of the implanter interruption was significant in

determining the probability and penalty of an interruption to WIP flow to the constraint operation.

Figure 10 represents the data of inventory departure interruptions at the implanter, verse inventory arrival interruptions at the constraint. A one-to-one ratio line is drawn on the graph for ease of reference. Any point that is above the line represents an incident where the penalty of an implant interruption was greater than the length of time the implanter was down. Notice how few points are actually above this line; however some points do exist. The reason the majority of the data is below the 1:1 line is due to the engine and caboose effect described earlier.

However, what about the other occurrences where an interruptions to departures resulted in a “greater than” interruption to arrivals? These interruptions represent the fact that the other operations between the two machines of interest are also interrupted with random occurrences. As described in section 3.2, one of the problems with modeling factory output is factoring in all the unsynchronized interruptions. This data takes all of those instances into account.

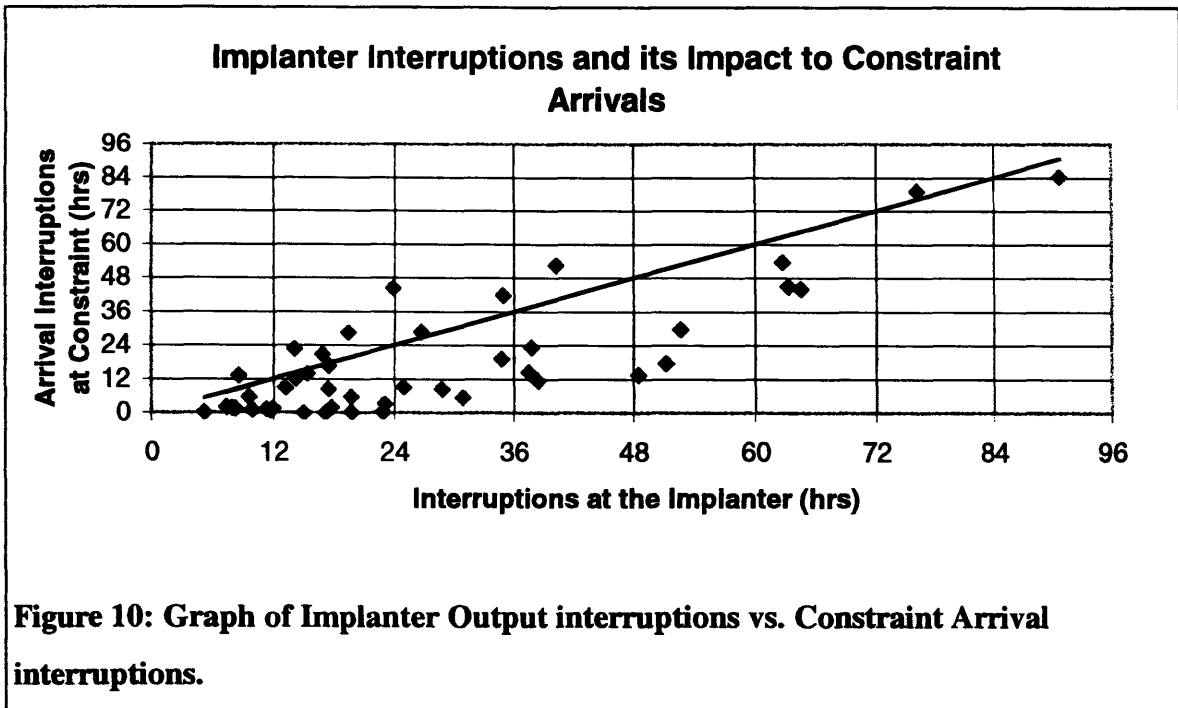


Figure 10: Graph of Implanter Output interruptions vs. Constraint Arrival interruptions.

Referring to the chart above, there are 10 observations for implanter interruptions that lasted between 0 to 12 hours. Of the ten data points, only two actually resulted in an interruption greater than 4 hours; one interruption was 9 hours and the other was 10 hours. The interruption at 10 hours resulted in an approximate 6-hour arrival time interruption at the constraint. This is the typical engine and caboose effect. The 9-hour interruption at the implanter actually resulted in a 13-hour interruption to arrivals to the constraint. Again, this is evidence of unsynchronized interruptions occurring in the factory.

One observation drawn from this data set is that interruptions lasting 12 hours or less have only a 1 in 5 chance of interrupting arrivals to the constraint. In other words, there is an 80% probability that there will be no penalty associated with an interruption less than 12 hours long. However, the ratio of the interruption, given an interruption occurs, will be, on average, 1:1 in terms of an implant departure interruption to a constraint arrival interruption ratio. This 1:1 ratio impact is calculated by averaging the two occurrences out of ten in the 0 to 12-hour category.

From 12 to 24 hours the implanter interruptions start to have a greater impact on constraint arrivals. Of the 16 observations that occurred during this time frame, 6 occurrences had negligible consequences to WIP flow to the constraint. Six of the remaining 10 data points were less than the 1:1 ratio and four were greater than the 1:1 ratio. Therefore, the data suggests that during a 12 to 24-hour interruption at the implanter there is a 62% chance the interruption will be felt at the constraint. The expected interruption ratio, at the constraint, given that the interruption was felt, would be in the order of 1:1. Again this impact ratio is the calculated average of all events that had a significant impact at the constraint.

All interruptions that were greater than 24 hours resulted in a significant impact to WIP flow to the constraint operation. Of the 18 observations that occurred resulting in a greater than 24-hour implanter interruption, four of them resulted in a constraint arrival interruption that was greater than the implanter interruption. However, on average, a

greater than 24-hour interruption at the implanter only resulted in a constraint interruption 0.65 times as large. This is a significant shift from the interruptions that were less than 24 hours.

A summary of the data analysis is given in Figure 11 and Figure 12 below. The left-hand column of the summary table shows the various implanter interruption windows. To the right of that column is the number of observations for the corresponding interruption window. The next column to the right is the probability of a penalty. So for the 0 to 12 hour window there is a 20% chance that an implanter interruption will result in an interruption to the constraint WIP arrival rate. The next column is the expected multiplier given that a penalty will occur. So, if there is an interruption at the implanter that will last for approximately 10 hours and someone was interested in determining the impact that interruption may have to constraint arrivals, they would multiply 10 hours by 1.07 to get the answer of 10.7 hours.

The final column on the far right represents the expected interruption to constraint WIP arrival rate given an interruption will occur for the median time of the time window. So, the median time for the first time window (0 to 12 hours) is 6 hours. For a six-hour interruption where an interruption is anticipated, there will be a 6.42-hour penalty. This column is provided to give the reader an appreciation of the behavior of various length interruptions.

Interruptions (hrs)	# observations	P(penalty)	E(Multiplier penalty)	E(Interruption to arrivals at constraint penalty for the median time) (Hrs.)
0 to 12	10	0.20	1.07	6.42
12 to 24	16	0.62	0.98	17.46
24 to 36	6	1.00	0.63	18.63
36 to 48	4	1.00	0.65	27.3
> 48	8	1.00	0.68	

Figure 11: Data summary of implanter output interruptions vs. constraint arrival interruptions.

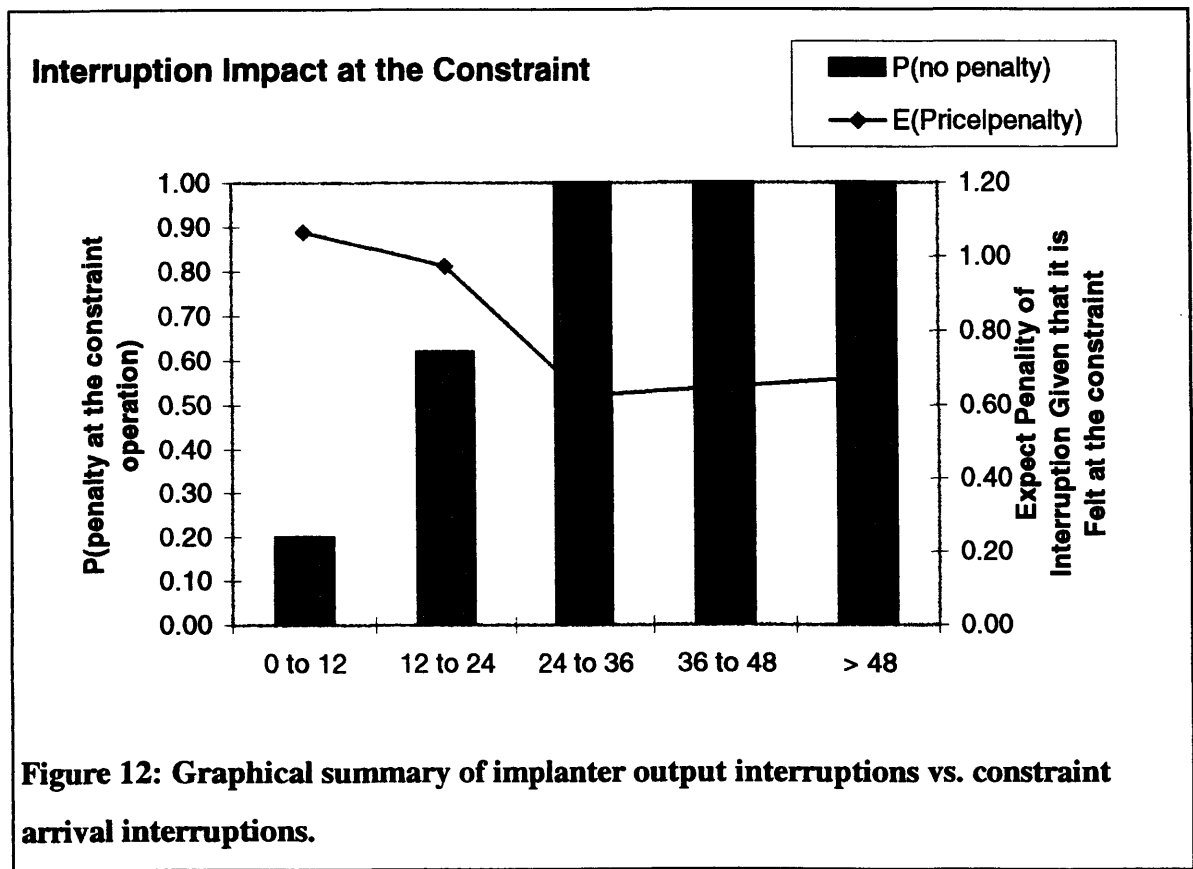


Figure 12: Graphical summary of implanter output interruptions vs. constraint arrival interruptions.

In summary, the shorter the interruption to the implanter, the less likely the interruption will affect WIP flow to the constraint. However, as interruptions to the implanter become longer, the impact of the interruption reduces in magnitude. This is most likely due the engine and caboose effect discussed earlier.

5.4 Determining the Critical Buffer Size (CBS)

Understanding the probability and penalty of an interruption is only half the battle. Remember, constraint starvation occurs when there is no production left to process and there is no production arriving. To completely avoid starvation, we need sufficient inventory between the two stations to compensate for the interruption to arrivals to the constraint. Therefore, the next step in the process is to determine how much inventory between the two stations would be necessary to avoid starvation.

As mentioned earlier, on average it takes a lot (25 wafers) four days to travel from the implanter to the constraint. This means there should to be at least four days worth of production between the two stations. If the constraint is anticipating 150 wafers to arrive at the station every day, there should be, on average, 600 wafers between the two stations at any given time. This is a rough estimate, however it does provide a reasonable starting point for the analysis.

The more important question however, is how much inventory should be between the two stations to avoid constraint starvation in the event of an interruption at the implanter? For Intel, the answer to this question can provide a great deal of insight on how to manage the situation; there will be more on this point in chapter 6.

If an interruption occurs, and the length of the interruption can be predicted with reasonable accuracy, the empirical analysis discussed in the above section can be used to answer the posed question.

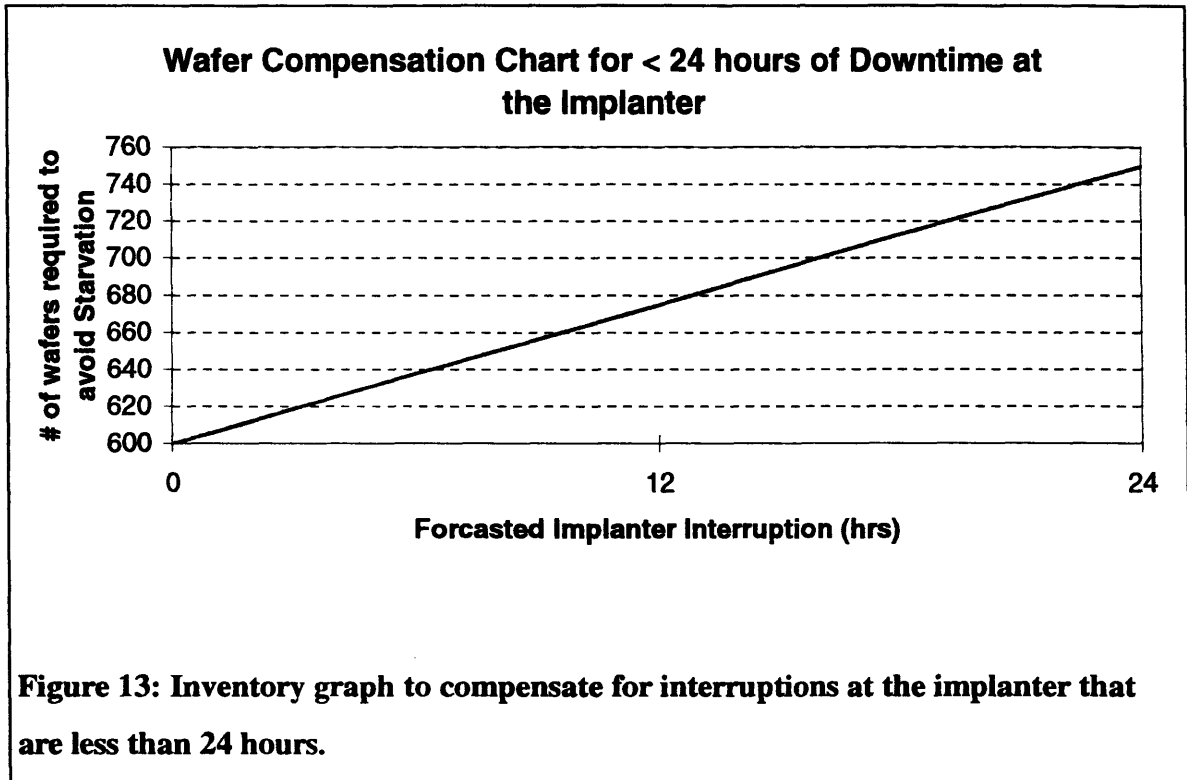
If, for example, an interruption is forecast to last for up to, but no longer than 12 hours, the results of the earlier data suggest that there is only 20% chance that the interruption will be felt at the constraint. During a 12 to 24 hour interruption there is only 62% chance that the interruption will interfere with the constraint arrival rate. One reasonable response would be to do nothing. The odds are in favor of not interrupting the constraint's WIP arrival rate.

If however, the constraint operator knew the number of wafers between the two stations, then the operator can use this information, along with the understanding that a less than 24 hour interruption has a 1:1 effect on constraint arrivals, to make a first order approximation about of the impact. The number of wafers between the implant operation and the constraint is relatively easy to extract from Intel's database. Starting with the premise that there should be at least 600 wafers between the two stations and that the implanter will be down for X hours (where X is < 24 hours), then this simple calculation can be used:

$$\# \text{ Wafers need to Avoid an Interruption} = 6.25 \text{ wafers/hour} * X + 600 \text{ wafers}$$

The 6.25 wafers/hour is the ideal WIP arrival rate for the constraint operation. The X variable is a forecasted time to system recovery that is usually given by the repair technician. For a less than 24 hour interruption, this value can be forecasted relatively accurately.

The graph below shows what to expect. For example, if the implanter is estimates to be unavailable for production for approximately 12 hours, there should be at least 670 wafers in production between the implanter and the constraint to avoid starvation.

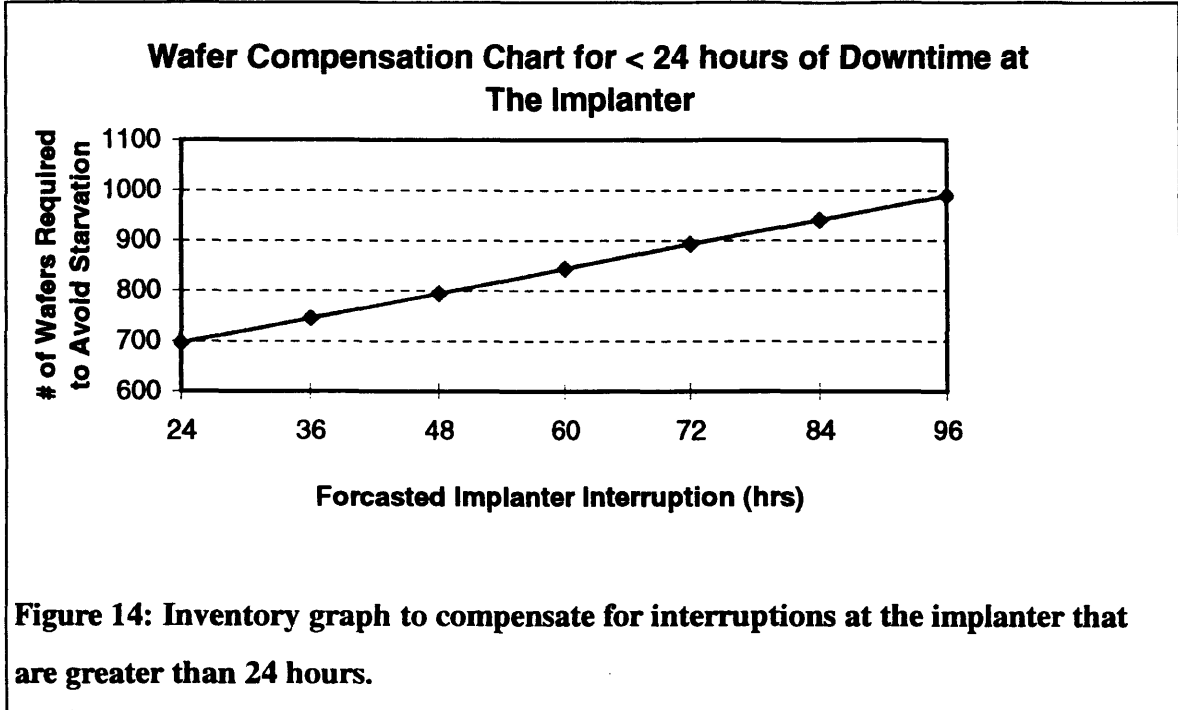


For interruptions that are anticipated to last longer than 24 hours another approach is required. The analysis demonstrated that 100% of the greater than 24-hour interruptions that occurred at the implanter resulted in an interruption to WIP arrivals to the constraint. Also, the interruption to WIP arrivals was not as severe as the interruption to implanter output. The interruptions in arrivals were only 0.65 times as severe as the original issue. For these types of implanter interruptions the following equation can be used to determine the amount of inventory necessary to avoid constraint starvation:

$$\# \text{ Wafers need to Avoid an Interruption} = 6.25 \text{ wafers/hour} * X * 0.65 + 600 \text{ wafers}$$

Again, X represents the number of hours the implanter is forecast to be down and the 6.25 wafers/hour represents the ideal arrival rate for the constraint operation.

The graphical interpretation of this formula is given below.



6. Future Consideration

6.1 How can Intel use this Analysis in Production?

The methodology and information presented in this thesis can be used several different ways. This information gives insight about the probability as well as penalty that an implanter interruption will have on WIP arrivals to the constraint. This information evaluates the production requirements between these two systems necessary to avoid constraint starvation.

Once the actual amount of inventory is known at the onset of an implanter interruption, and the time to complete the repair is forecasted, a first-order assessment can be made on whether the interruption will result in constraint starvation or will go unnoticed. In the event the data suggests the interruption will go unnoticed, the production managers can redirect their resource to other areas of the factory. This fact alone is very valuable. During the time that I started this analysis at Intel, the factory managers were devoting considerable resources to the uptime of this system regardless of how long it was going to be unavailable.

But what do they do if the data suggest that there is a deficiency of production between the two systems given the forecasted time that the implanter will be unavailable? A manager could quickly compare how much inventory exists between the two stations to the charts and then determine how long they can afford the implanter to be down. This piece of information can be used to give the repair technicians a repair time target. If they have the internal resources to achieve this goal, the repair process will just need to be closely monitored. Alternatively, if they do not have the internal resources to repair the system within the specified time, management can call for support from the ion-implant system supplier. This extra level of expertise may provide the resources needed to bring the system back to life.

If the above recommendations are not feasible, all is still not lost. Intel's factories work by a strategy known as "Copy Exactly." This production strategy requires all Intel facilities that are producing the same product to operate exactly the same. These factories will use the same type of equipment, run the same monitors, and have the same production flow. Therefore, there is the potential to ship product to the sister factory for processing in the event that one factory requires some support. Once this production has completed processing of the process steps of concern the production is shipped back to the factory of origin to complete processing. Before this recommendation should be implemented, a feasibility study should be conducted to ensure that the throughput time is actual a benefit.

6.2 A Shift in Perspective

In the past, the focus of near-constraint system interruptions has always been on the inventory that accumulates behind the interrupted machine. When inventory begins to accumulate behind the system, management becomes concerned in anticipation that factory's output goals will not be met. However, the methodology described in this thesis suggests that the focus should sometimes be on the inventory in front of the interrupted system rather than production that is accumulating behind it. Ultimately, management should be concerned with avoiding constraint starvation and not with momentary pockets of interruptions that will quickly be compensated once the interrupted system begins running production.

6.3 Using this Methodology with Downstream Systems

Although the focus of this work is with near-constraint operations several stations upstream from the constraint, the same methodology can be used for operations downstream from the constraint. For these scenarios, the end of the manufacturing line would take the place of the constraint operation: the system that your trying to continually feed.

6.4 Potential Issues with the Results

One of the most glaring issues with the results is the disconnect the two graphs display at the 24-hour mark. If one estimated the implanter would be down for approximately 24 hours, they could potentially look on either graph to determine the constraint system's

WIP arrival rate. Unfortunately, the two graphs give two different answers. For that matter, the graph for less than a 24-hour interruption suggest that there should be at least 750 wafers between the two systems to compensate for a 24-hour interruption. The graph for interruptions greater than 24 hours suggests that 750 wafers are necessary for interruptions that are in the order of 37 hours.

The reason for the disconnect is a direct result of the way the data was divided. Based on the data provided, there were clearly two distinctive and different repercussions associated with interruptions at the constraint: one greater than 24 hours, and the other less than 24 hours. The 24-hour mark appears to be a reasonable dividing point of these two resulting behaviors.

All is not for nothing, even with the disconnect. The results of this work do provide a reasonable first-order approximation of the number of wafers required to avoid constraint starvation.

Forecasting the length of time required to repair the implanter is not an exact science. As the repair to the implanter progresses, the accuracy of the forecast increases. At the beginning of the interruption however, the repair technician can, at best, only estimate the system's downtime with a resolution of 12 hours. Determining the exact hour when the implanter will be available after an interruption is impossible in most cases. Therefore, if the estimate is that the implanter will need extensive repairs that will take longer than 24 hours, the greater than 24-hour CBS graph should be utilized. If the repair seems minor and the repair technician is reasonably confident that the system will be available for processing in less than a day, then the less than 24-hour CBS graph should be employed.

Additionally, not only should the graphs that determine the CBS be considered but also the probabilities of an interruption should be considered. Remember, for an interruption that will be less than 12 hours in length there is only a 1 in 5 chance that it will even be noticed at the constraint. In summary, due to the methodology used to determine

downtime at the implanter, the graphs generated are a reasonable first order approximation of the CBS, even with the disconnect.

6.5 Updated Information Is Essential

As mentioned above, Intel's Santa Clara facility is a development site that is increasing its capacity. Because it is a development site there is continuous improvements being made to the process. Additionally, the amount of inventory started each week is constantly being increased, and the numbers of systems that perform these operations are also being increased. Therefore, in order to use this methodology the graphs need to be updated to reflect the changes in factory conditions. Otherwise, the results may be misleading and actually decrease the quality of the decisions rather than improve them.

7. Conclusions

This thesis demonstrated a statistical approach to assess the impact of near-constraint systems in a factory environment. More specifically, the analysis focused on interruptions to the system and determines the probability of interrupting the flow of inventory to the factory constraint.

The analysis showed that not all interruptions are created equal. The length of the interruption had a great deal of influence of the probability the interruption would be felt at the constraint operation. The longer the interruption the greater the chances are that the interruption would be felt at the constraint.

However, the data also showed that short interruptions that do interrupt the constraint system's WIP arrival rate are usually the same length as the initial interruptions. The longer the near-constraint interruption is however, the proportionally shorter the interruption to the arrivals to the constraint will be.

The purpose of analyzing interruption to constraint arrivals was to obtain an understanding of what could potentially starve the operation. So, the analysis was used to determine the overall level of production necessary to avoid starvation to the constraint operation: this is called the Critical Buffer Size (CBS)

Although this work was done and illustrated in a semiconductor fabrication facility, I believe that the methodology is easily applicable in many manufacturing settings. Additionally, even though this work focused on a near-constraint system that fed the constraint operation, the same statistical procedure could be used for near-constraint systems that are behind the constraint. The major difference would be to focus on the WIP arrival rate to the end of the processing line rather than to the factory constraint.

References

Textbooks:

Hoggs and Ledolter, Applied Statistics for Engineers and Physical Scientists,
Macmillan Publishing Company, New York, 1992.

Nahmias, Steven, Production and Operations Management, Second Edition, Richard
D. Irwin, Inc., Homewood, Illinois, 1993.

Goldratt, Eliyahu M. and Cox, Jeff. The Goal. Croton-on-Hudson, New York:
North River Press, Inc. 1986.

Gershwin, Stanley B., Manufacturing Systems Engineering, Prentice Hall, Englewood
Cliffs, NJ, 1994.

Publications:

Dallery, Yves, and Gershwin, Stanley B., *Manufacturing Flow Line Systems: A Review of
Models and Analysis Results*, MIT, Laboratory for Manufacturing and
Productivity, Report, LFM-91-002. 1991.

Intel Internal Publications:

Cunningham, Calum, *80% Confidence: A Perspective on Predictable Equipment
Performance*, Intel Manufacturing Excellence Conference, 1994.

McMichael, David, *Constraint Management in a Non-Linear Production Fab*, Intel
Manufacturing Excellence Conference, 1994.

Kempf, Karl, *The Concept of Constraint Management Applied to Fab 9.1*, Intel Manufacturing Excellence Conference, 1993.

Menon, Viju, *Constraint Based Policies to Maximize Fab Output*, Intel Manufacturing Excellence Conference, 1995.

Thesis:

Ku, Jason, *Microprocessor Manufacturing Throughput Time Variability*, Massachusetts Institute of Technology Masters Thesis, 1994.

311A-38