14

Introducing Pull Methodologies in a Semiconductor Fab

By

Jason Walker Connally

Bachelors of Science in Mechanical Engineering, Texas A&M University (1998)

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

Master of Business Administration
Master of Science in Mechanical Engineering

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

Signature of Author____
May 6, 2005
MIT Sloan School of Management
Department of Mechanical Engineering

Certified by____
Stephen Graves, Thesis Advisor
Abraham J. Siegel Professor of Management

Certified by____
Daniel Whitney, Thesis Advisor
Senior Lecturer, MIT Engineering Systems Division

Accepted by____
David Capodilupo
Executive Director of Masters Program
Sloan School of Management

Accepted by______________________
Lalit Anand
Chairman, Committee on Graduate Studies
Department of Mechanical Engineering

Introducing Pull Methodologies in a Semiconductor Fab

By

Jason Walker Connally

Submitted to the Sloan School of Management and the Department of Mechanical Engineering on May 6, 2005 in partial fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Mechanical Engineering

Abstract

Semiconductor manufacturing is a highly complex and re-entrant process. In a fabrication facility, hundreds of decisions are made during a production shift regarding how shared tool capacity will be prioritized. These decisions contribute to how balanced or unbalanced the manufacturing line will be. Characteristics of an unbalanced line are large WIP bubbles, long queue times, and expediting. A balanced line has less WIP bubbles, shorter queues, and WIP is positioned at all points throughout the line to be in position to meet the demand forecasted.

This thesis focuses on work performed at Intel's Fab 23 to improve the process for assigning production priorities through the introduction of pull methodologies. Existing processes and tools were studied, and an improved methodology and decision-support tool was developed to aid operations managers in driving towards a balanced line. An experiment was also designed and executed in production to test the methods and tools developed.

Target cycle time data was used along with throughput goals to construct a target inventory profile throughout the line. Actual inventory was then compared to the ideal "balanced" profile to determine where WIP deficits and surpluses existed. Using this information, the operations managers had objective metrics that could be used in determining which operations should receive priority.

Significant externalities inhibited performance during the experiment, preventing measurable improvements in line balance and cycle time from being realized. However, these externalities were known prior to experimentation, and a decision was made to learn from the experiment. The tool proved helpful in promoting consistency across shifts in how the factory was run. There were many anecdotal examples of the decision-support tool driving more intelligent priority decisions than operations managers would have made without the tool.

Thesis Supervisor: Stephen Graves
Title: Abraham J. Siegel Professor of Management

Thesis Supervisor: Daniel Whitney
Title: Senior Lecturer, MIT Engineering Systems Division

This page is intentionally left blank

Acknowledgements

I would like to express my sincere appreciation to everyone at Intel's Fab 23 who made my time there a terrific experience. Specifically, I would like to thank my supervisor, Clay Carlson, for his support and guidance during my internship. I would also like to thank fellow intern, Roy Wildeman (LFM '05), for showing me the ropes at Fab 23 and providing advice on project direction. Others I owe special thanks to are Morgan Burke, Brad Weaver, Jim Beuerle, Joe Lewis, Clint Thielemier, John Reschar, Aaron Darnell, Allan Rowe, and Mike Westbrook. Thanks to all of the Operations Managers who dedicated their time to working with me and educating me on the manufacturing processes. Without their patience and willingness to experiment with something new, the project would not have been possible.

I would also like to acknowledge the staff, faculty, and students of the Leaders for Manufacturing (LFM) Program. The last two years have been truly incredible and have exceeded all expectations. I would like to give special thanks to my thesis advisors, Steve Graves and Dan Whitney for their guidance during the internship and writing of this thesis.

I would also like to thank my co-workers and sponsors at Dell for their support. I am extremely grateful for this opportunity and look forward to returning to work with such an outstanding group of people.

I would also like to thank my parents and family for their encouragement, support, and prayers.

I am especially thankful to my wonderful wife, Corinne, for her unwavering support, encouragement, and love. I also wish to thank my daughter, Lauren, who brightens my every day. The last two years have been quite an adventure, and I am eternally grateful to Corinne and Lauren for their patience and sacrifices along the way. This thesis is dedicated to them.

This page is intentionally left blank

Table of Contents

Table of Figures and Tables

Table of Equations

This page is intentionally left blank

1. Introduction

One would be hard pressed to imagine a more complex manufacturing process than the fabrication of a semiconductor. Feature dimensions in some of the newest micro-processors are as small as 60 nm. That is 60 billionths of a meter! However, it is not just the *process technology* required to produce these wondrous products that is complex. Attempting to *optimally control the flow of inventory* through the manufacturing process is no small task.

The fabrication of hundred of circuits onto a silicon wafer is a very re-entrant process. Wafers will return to the same toolsets many times throughout the manufacturing process to have different operations performed. This means that when a shared tool is available to process a lot of wafers, there are wafers at multiple points in the manufacturing process to choose from. There are often multiple products to choose between as well. Priority decisions have to be made regarding which operations get performed on which wafers. The decisions made by technicians at the tool level are influenced by the operational goals set by management at the beginning of every shift.

This thesis is the culmination of a 6.5 month internship at Intel's Fab 23 flash manufacturing facility in Colorado Springs. At Fab 23, an Operations Manager in a role referred to as "Captain of the Ship" reviews inventory, equipment, and line status at the beginning of every shift and sets high-level production priorities and goals for wafer movement through the line. This process is called "goaling the line". Goaling affects what decisions will be made on the floor when technicians select lots for processing at shared toolsets. Poor goaling decisions will result in processing wafers that are less urgent than other available wafers. This can lead to unbalanced production flow, starved bottlenecks, WIP bubbles, missed shipments, and unsatisfied customers.

The process of goaling the line was observed to have significant variability across shifts. With different shifts applying different WIP management methodologies to the line it is difficult for the factory to keep momentum. There is a need for more than WIP management "rules of thumb". Factory performance can benefit from making goaling decisions more intelligent and less subjective.

This thesis will describe the planning and implementation of an improved goaling methodology and decision support tool. Besides improving the goaling decisions, another key goal is to promote consistency across shift in the WIP management methodology applied. The methodology described will introduce aspects of "pull" manufacturing to the goaling process. The principles discussed could readily be applied to any re-entrant manufacturing process.

1.1 Thesis Outline

This thesis is divided into six chapters.

Chapter 2 gives an overview of the flash memory industry as well as Fab 23 and the lean manufacturing initiative currently underway.

Chapter 3 describes WIP management as practiced at Fab 23. The decision making process for setting WIP priorities will be discussed along with the variability that exists between shifts.

Building on the previous chapter, *Chapter 4* will discuss major gaps in the existing processes and present opportunities for targeted improvements. The improvement opportunities discussed here provide the basis for the experiment conducted.

Chapter 5 provides an overview of the experiment and implementation. Both technical details and implementation details will be discussed along with reflection and feedback received.

Finally, *Chapter 6* will present conclusions and lessons learned during the internship. Hypotheses for improving the WIP experiment will be presented, as well as potential areas for continuing progress towards the ideal state for WIP management.

2. Background

Fab 23 manufactures a single flash memory technology and was recently selected to produce a key chip in the Centrino wireless platform. The fab and its employees have endured through a major industry downturn and are now looking to lean manufacturing principles to help compete successfully in the extremely competitive industry of flash memory. This chapter presents an overview of the flash memory market as well as Fab 23 and its lean initiative.

2.1 Flash memory overview

Flash memory is a type of EEPROM[1] device that can retain information even when power is removed from the device. The difference between flash memory and other types of EEPROM is the ability to erase and write to blocks of memory very quickly [1]. This has made flash memory the memory of choice for devices such as cell phones, PDA's, digital cameras, and MP3 players. As these markets have experienced tremendous growth this decade, so has the flash memory market. In 2004 flash memory was a $15.6 billion market [2].

The two types of flash memory available in the market today are NAND and NOR. Intel invented NOR memory in 1988 and Toshiba invented NAND memory in 1989 [1]. As of the writing of this thesis, Intel only manufactured NOR memory. Generally, NOR memory offers better data integrity and faster read speeds than NAND. However, erasing and writing to a NOR device takes several seconds. NAND devices generally offer faster write speeds, lower cost per byte, and come in larger capacities [1]. These characteristics led to the use of NOR devices for storing program and executable code and use of NAND devices when write-speed and increased memory are a necessity. Thus, NOR has generally been the choice for cell phones and PDA's, while NAND is the leading memory choice for digital cameras and MP3 players [3].

After introducing NOR memory, Intel was the number one flash memory supplier for more than a decade, until the third quarter of 2003. From the second to the third quarter of 2003, Intel fell from the number one flash memory supplier to number four. Tremendous growth in the NAND

---

[1] Electrically Erasable Programmable Read Only Memory

market vaulted Samsung and Toshiba to the top two spots respectively [4]. Additionally, Intel even gave up the number one spot in the NOR market when AMD and Fujitsu merged their flash memory operations into a jointly owned subsidiary called Spansion. In response, Intel launched an aggressive price competition and regained the number one spot in the NOR market and number two spot in the overall flash memory market by the third quarter of 2004. Samsung remains the number one overall flash memory supplier overall [5].

Competition in the flash memory market is only going to get more intense. The markets for NOR and NAND had previously been somewhat segmented. More than 50% of the NOR market comes from cell phones, however with the increased storage requirements of multimedia phones, NAND is beginning to make inroads [6]. Similarly, NOR suppliers are improving their products and beginning to market them for products that have traditionally been NAND based [7].

Revenue from the sale of NAND flash memory is expected to pass NOR revenues for the first time this year. The growth of its traditional base in digital cameras, music players, memory cards, combined with its entry into the cell phone architectures should continue to propel sales growth at a rate faster than NOR [8]. As recent as 2002, NAND only held 10% of the flash memory share [9].

The intensification of competition only increases the importance of flexible, responsive, low-cost manufacturing. As flash memory is a commodity business, these are the dimensions upon which differentiation will occur.

2.2 Fab 23 Background

Purchased from Rockwell International in February of 2000, Intel started up Fab 23 in early 2001 [10, 11]. The Colorado Springs fab ramped up production and yields faster than any 200mm fab in Intel's history, and the future looked bright [11]. $1.5 billion was invested in Fab 23, and it was slated to be a dedicated high-volume flash memory fab.

Soon, after Fab 23 opened in 2001, the worldwide chip industry entered a major downturn, and Intel found itself with excess capacity. The future of the just-opened fab was uncertain and it was urgent that costs be significantly reduced. Aggressive goal-setting and cost-cutting led to significant cost reductions as operational costs per wafer were cut by over 40%.

The successful execution of these cost-cutting strategies was key to surviving the downturn. However, the intensifying competition in the flash industry requires that more be done to continually reduce costs and increase manufacturing flexibility. Management looked to lean manufacturing as a way to instill a culture of continuous improvement and empower the workforce. Lean had a successful track record in other industries, and Fab 23 was selected as Intel's pilot site to explore implementation in semiconductor manufacturing.

2.3 Lean Manufacturing Initiative: mX (Manufacturing eXcellence)

Fab 23 is in the midst of a lean-transformation in which the entire manufacturing organization is being taught lean manufacturing principles and being challenged to apply them to continuous improvement and waste elimination. Spurred by the success other industries have had with lean, Fab 23 is leading Intel into uncharted waters as it tests the applicability of lean principles to the unique environment found in semiconductor manufacturing. In 2003, the fab began working with the Lean Learning Center (http://www.leanlearningcenter.com) to build its implementation and execution strategy around lean. Fab 23's name for this initiative is mX: manufacturing Excellence.

The goal of mX is to empower employees at every level to systematically eliminate waste and continuously improve processes towards an ideal state. The technicians on the manufacturing floor are seen as a huge untapped resource to generate ideas and drive improvements. If the lean transformation is to be successful, the rules and philosophy of lean manufacturing must be internalized by everyone in the fab, from the technician on the floor up to the fab manager.

The Lean Learning Center has rephrased the four rules of lean[2] espoused by Steven Spear and Kent Bowen in their article "Decoding the DNA of the Toyota Production System" [12]:

---

[2] In addition to the Bowen / Spear article, another great paper for understanding lean as Fab 23 is attempting to apply it is "Beyond Lean" by Jamie Flinchbaugh at the Lean Learning Center [13].

The Rules:

1. Structure every activity
2. Clearly connect every customer-supplier
3. Specify and simplify every flow
4. Improve through experimentation at the point of activity toward the ideal state

The focus Rule #4 places on experimentation was helpful in overcoming inertia and gaining support to test the improved goaling processes developed during the internship. Experimentation is framed as being a way to investigate new processes and not permanent changes. The expectation is that experiments will be a series of steps towards the ideal state.

Roy Wildeman (LFM '05) conducted a previous internship at Fab 23 that focused on applying lean manufacturing principles in semiconductor manufacturing. Roy's thesis, "An Application of Lean Principles within a Semiconductor Manufacturing Environment", is a terrific resource for anyone interested in more detail on the challenges and opportunities in putting lean manufacturing methodologies into action in a fab environment [14].

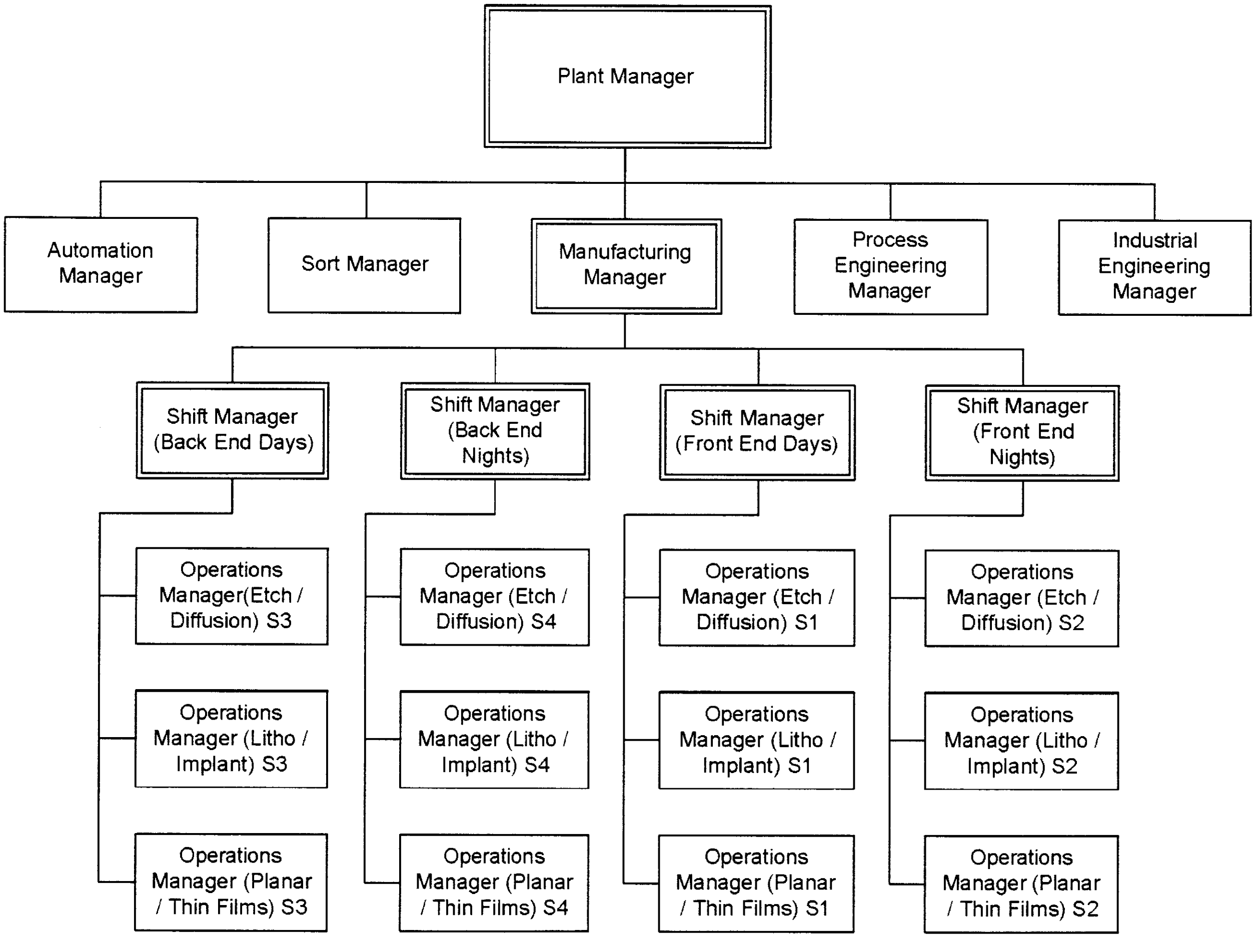
2.4 Organization Background

The factory runs 24-hours a day, 7-days a week. Manufacturing operations is divided into 4 shifts to support production. 1 day & 1 night shift support the front end of the week, and 1 day & 1 night shift support the backend of the week. Changeover from the front end of the week to the backend of the week occurs on Wednesday.

See Figure 1 for a partial organizational chart for Fab 23. The plant manager is in charge of all operations, and reporting to him are senior level managers for Automation, Sort Manufacturing, Fab Manufacturing, Process Engineering, Industrial Engineering, Quality, HR, and Finance. The org chart in Figure 1 zooms in on the Fab Manufacturing portion of the organization.

Each of the four shifts has a Shift Manager (SM) who reports to the senior Manufacturing Manager. Each SM has three to four Operations Managers (OM) directly reporting to them. Operations Managers are over functional areas on the floor and have 20-25 Manufacturing Technicians (MT) who report to them. These MT's operate the tools in the fab and process the wafers. A senior MT in each functional area is dubbed an Area Coordinator (AC) and helps coordinate activities and sits in on the start-up meetings at the beginning of every shift.

Figure 1 – Organizational Structure at Fab 23



This page is intentionally left blank

3. WIP Management

Before discussing specific WIP management practices at Fab 23 or in semiconductor manufacturing in general, it is beneficial to have a basic understanding of the manufacturing process itself. This chapter will give a high level overview of the semiconductor manufacturing process, and then discuss specific details of the WIP management methodology and process at Fab 23.

3.1 Semiconductor Manufacturing Overview

Semiconductor manufacturing is a highly re-entrant process. Wafers will return to the same tool or tools multiple times throughout the manufacturing process to have different operations performed. Hundreds of decisions are made every shift regarding which operations and wafers will be given priority to run on shared toolsets.

The main functional areas of the fab are:

- Diffusion – dopants introduced into wafer through thermal diffusion
- Implant – dopant atoms accelerated to high speeds and driven into wafer
- Lithography – photoresistant mask is applied, exposed, and developed to create circuitry layer patterns on wafer.
- Thin Films – thin film of material is deposited onto wafer surface. Metal connects layers and isolating material prevents shorts.
- Etch – removes or "etches" away material from wafer
- Planar – wafers are electromechanically polished to smooth wafer surface

Once the wafers are fabricated, they are shipped to a Sort facility where they are tested. Fabrication and Sort facilities are typically co-located. Each die is tested for speed and functionality at Sort. Defective die are noted either physically with ink or virtually in wafer database.

After the die have been "sorted" they are sent to an Assembly facility where they are physically separated and assembled into usable packages. The package protects the chip and allows it to communicate with outside circuitry. Before the final package is shipped to customers, they are sent to a test facility for final electrical tests. Assembly facilities are typically co-located with final test facilities.

3.2 WIP Management at Fab 23

Intel is well known for its Copy Exactly methodology[3]. One might think that this would lead to all Intel fabs treating WIP Management in the same way. However, Copy Exactly applies to the process flow, equipment, suppliers, and training methodologies, but not to how WIP priorities are set [15]. There are several recognized WIP philosophies in practice across Intel's virtual factory network. This section will focus on the WIP management practices of Fab 23. However, as will be discussed, the variability present across shifts means that Fab 23's methodology in practice does not fit neatly into one standard system.

Any data related to demand, capacity, inventory levels, or cycle time has been disguised or is strictly for illustrative purposes.

3.2.1 Wafer Starts

Wafer starts are defined as the number of wafers introduced into the line over a given period of time. A "push" methodology is employed when determining how many wafers will be started. Knowing the target cycle time for a wafer moving through the fabrication process, the factory will start as many wafers in the current week as are required to meet forecasted demand. For example, if the target cycle time to fabricate a wafer is eight weeks, then the factory will simply start as many wafers in the current week as are forecasted to be required from the fab in eight weeks.

---

[3] Intel introduced "Copy Exactly" in the 1980's to ensure that the exact same process developed in development fabs was transferred to every Fab responsible for producing a technology [15].

The distribution of WIP in the fab is not given strong consideration when determining wafer starts. There may be equipment issues that have caused the fab to miss ship schedules or accumulate WIP at certain points in the line. However, wafers will still be pushed into the fab according to what the fab-out schedule dictates. Because a pull methodology is not used, it would theoretically be possible that WIP could continue to rise unbounded. Realistically of course, there would be a physical limit to how many wafers management would tolerate in the fab, and starts could be decreased to avoid an unmanageable WIP level [16]. However, in practice WIP levels and distribution do not play a major role in wafer start decisions.

3.2.2 Back-to-Front Philosophy

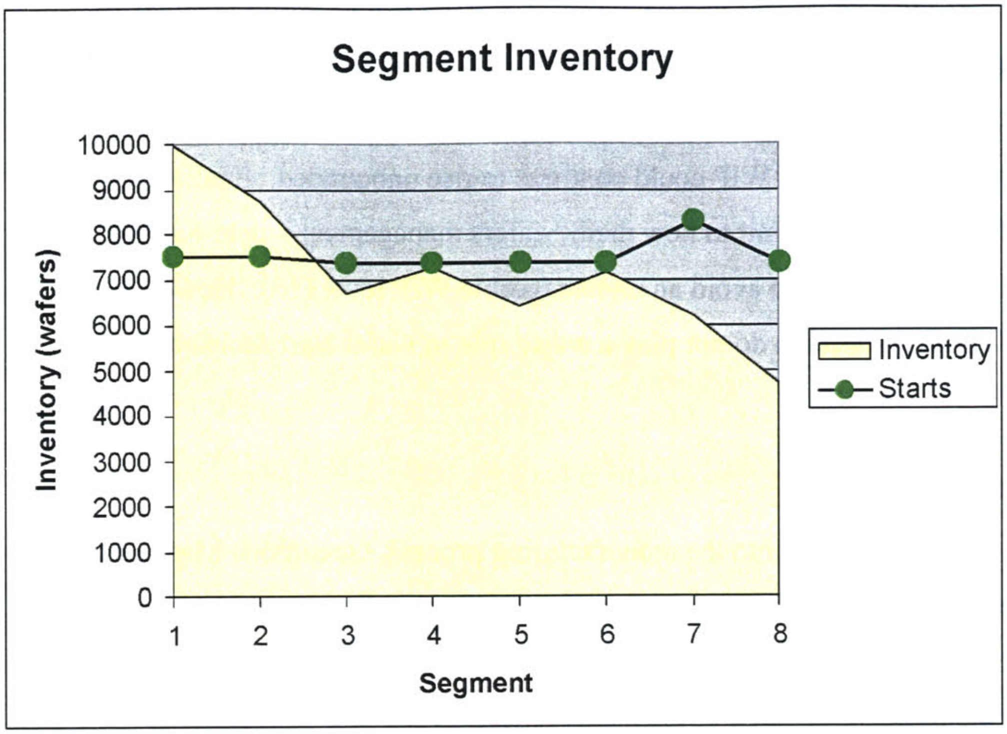
Once wafers have been introduced into the manufacturing process, Operations Managers and Manufacturing Technicians make the tactical decisions for how wafers should be moved throughout the process. The various fabs in Intel's Virtual Factory network employ different WIP management methodologies. One of the major aspects of the WIP management methodology employed by Fab 23 is referred to as "back-to-front".

"Back-to-front" dictates that priority is generally given to wafers that are nearest to shipping. The re-entrant nature of the manufacturing process means that many tools and operations have inventory waiting in buffer to be processed that are at different stages of the manufacturing process. When choosing which operations will be given priority and run first, preference will be given to wafers that are furthest along in the manufacturing process.

3.2.3 Evaluating line balance

For purposes of observing WIP distribution in the line, the process is divided up into week-long segments. If the target cycle time for producing a wafer is 8 weeks, the process operations will be divided into 8 segments. Segment 1 represents operations that a wafer should move through in the first week of the process, segment 2 represents operations that a wafer should move through in the second week of the process, and so on. The segment graph shown in Figure 2 depicts the gross number of wafers in each segment of the line at a particular moment in time.

Figure 2 – Starts vs Inventory by Segment



When the fab is operating in a steady-state environment (in that the number of wafers required out each week is relatively constant), the general phenomena of being "heavy" on inventory in the front of the line and "light" at the back of the line is commonly observed. This is symptomatic of operating in a "back-to-front" mentality in that wafers at the back of the line are given priority and spend less time in queue.

The green dots in Figure 2 represent how many wafers were started in past weeks that correspond to that segment of the line. The starts associated with the inventory in Segment 1 correspond to the number of wafers that were started in the most current work week. The number of starts associated with the inventory in Segment 8 corresponds to the number of wafers that were started 8 weeks ago. If wafers consistently moved through the manufacturing process according to the target cycle time of each operation, the historical starts level would be the same as the inventory in the corresponding segment.

Figure 3 represents a theoretical wafer start schedule by work week. Varying start levels are used to better illustrate how start levels correspond to target inventory levels by segment (Figure 5).

Figure 4 depicts the fab-out schedule over the target cycle time to fabricate a wafer. Eight weeks was chosen for illustrative purposes. Looking at Figure 3 and Figure 4 together, one can see how wafer starts are driven by the fab out schedule (discussed in section 3.2.1). If 5600 wafers are required out in WW9 (Figure 4), then 5600 wafers should have been started 8 weeks earlier in WW1 (Figure 3).

Figure 5 shows what the target inventories in each segment would look like assuming that wafers consistently marched through the fabrication process according the their target cycle times. This is essentially just a mirror image of the wafer starts (Figure 3) and fab-out (Figure 4) graphs. If we started 5600 in WW1, then at the beginning of WW9 we would expect to have 5600 wafers in Segment 8. If we started 2100 wafers in WW5, then at the beginning of WW9 we would expect to have 2100 Wafers in Segment 4.

Figure 3 – Historical Wafer Starts by Work Week

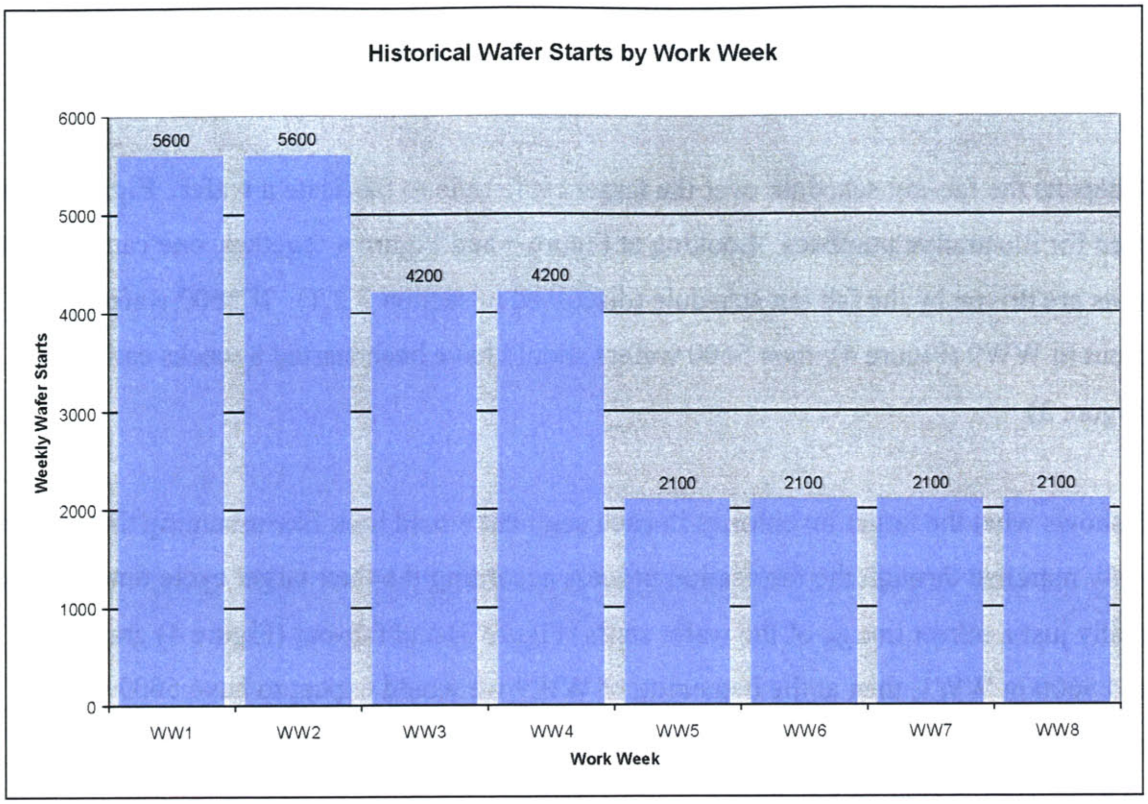


Figure 4 – Fab Out Schedule over Next 8 Weeks

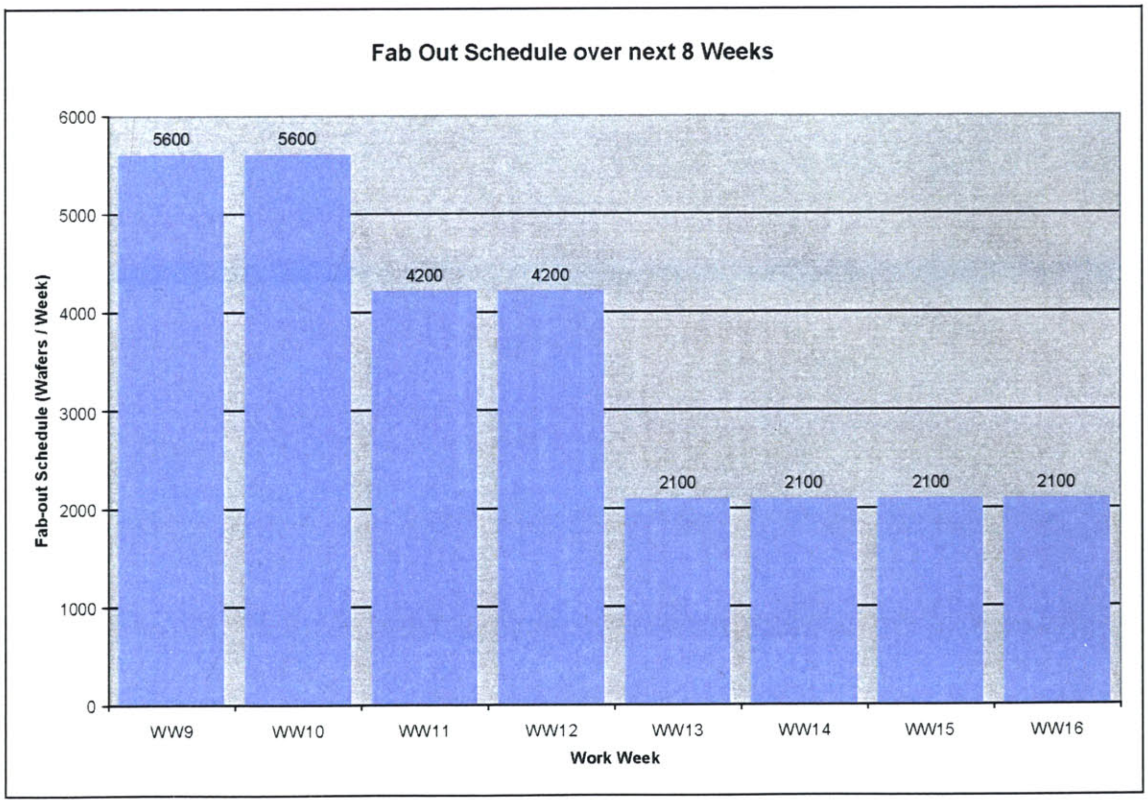
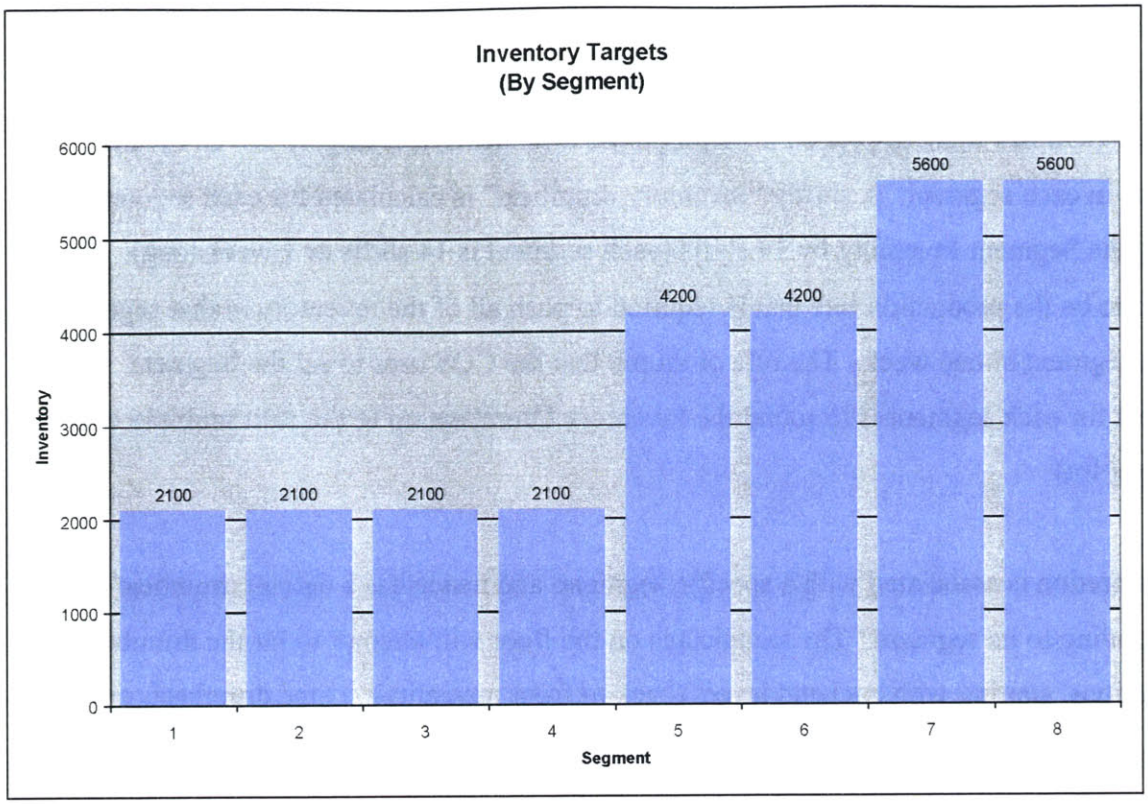


Figure 5 – Inventory Targets by Segment



The line is thought to be "balanced" when the actual inventory in each segment is equal to what the corresponding starts were for that segment[4].

3.2.4 Goaling the Fab

As described in Section 2.4, each shift has several Operations Managers. Every week, each shift designates one OM to occupy the role of "Captain of the Ship" (COS). The COS is responsible for "goaling" the factory. Goaling involves assigning priorities to segments of the line and also assigning target goals for how many wafers certain operations should process in that shift.

---

[4] This would result in the green dots in Figure 2 (representing wafer starts) tracking exactly with inventory levels in each segment.

3.2.4.1 Drumbeats

The first step in goaling the line is to assign a drumbeat to each segment of the line. Drumbeats represent a goal for how many wafers per shift each operation in a segment should process. While the COS has some discretion in setting these drumbeats, it is largely driven by the inventory in each segment. A shiftly "inventory drumbeat" is calculated for each segment by dividing the Segment Inventory by 14 shifts (each segment is 14 shifts or 1 week long). This is assumed to be the production rate that is required to push all of the inventory in that segment to the next segment in one week. The rule of thumb that the COS uses to set the Segment Drumbeat for each segment is to round the Inventory Drumbeat up to the next multiple of 25 (25 wafers per lot).

Every operation is associated with a specific segment, and hence has a default drumbeat corresponding to its segment. The technicians on the floor will attempt to hit the drumbeat for all operations, starting with backend layers (back to front mentality). Once drumbeats are hit, technicians will continue to process wafers, giving preference to backend operations. As a result, operations towards the end of the line typically output wafers at a rate higher than the inventory drumbeat.

3.2.4.2 Operation Specific Goals

Once segment drumbeats have been set, the COS will set specific output goals for certain operations. These goals will override any drumbeat that is associated with that operation. When an operation is goaled, it will be flagged to the technicians as a priority through their tool interface, and progress towards the goal will be tracked throughout the shift. Every time a lot is processed through that goaled operation, the "percent to goal" will be updated so that the MT can monitor progress. The operation specific goals help communicate priorities to the floor. If a re-entrant tool has wafers waiting for several operations, goaled operations will get priority.

The COS uses a web-based tool to set the operation specific goals. The tool displays all of the operations associated with producing a wafer in sequential order, along with the following information:

- Inventory currently at the operation (In Queue and In Process)
- Beginning of shift Inventory
- Operation's Drumbeat
- Number of wafers processed through operation in current shift
- Notes – Captain can enter notes for a specific operation that techs on the floor can see.

A major driver for selecting which areas of the line to goal is based on where WIP bubbles are. The COS will review the segment inventory graphs to identify which segments have excess inventory. This judgment is made by comparing actual inventory levels to what the corresponding starts levels dictate the target inventory should be. The COS also has the ability to drill down into a segment and view the inventory profile by "days" rather than "weeks". This facilitates identification of local WIP bubbles within the segment, even if the segment inventory is in-line with starts.

The segment drumbeats and operation specific goals are reviewed with Area Coordinators at the beginning of every shift. Goals are set to be feasibly achieved within one shift of operation. AC's will give feedback to the COS as to when goals may not be feasible due to tool or inventory constraints, or when the goal is too conservative and they believe they can process more WIP through a particular step.

3.2.4.3 Product Specific WIP Management

Wafers are primarily managed on a gross wafer basis throughout most of the line. The drumbeats and operation specific goals communicate how many wafers should move through an operation in one shift, but do not communicate how many of what product to process. Towards the end of the line, fab personnel will start focusing on moving particular products if they are behind schedule. When a particular product is deemed by planning to be behind schedule, the COS will manually communicate these product priorities to AC's and MT's to get the WIP moving. The tool used to communicate how far ahead or behind schedule a product is referred to as the "pull station". However, this is really a misnomer in that the product is expedited and "pushed" further along in the process. If a true pull system were in place, WIP would always be prioritized to have the right amount of product at all points throughout the line.

3.3 Variability Across Shifts

Although the "official" WIP management policy at Fab 23 espouses setting segment drumbeats and operating in a "back-to-front" mode, Captains execute differently upon these guidelines. This section discusses the variability across shifts in how the goaling process is implemented and how WIP management rules-of-thumb are applied. The variability between the shifts was readily apparent and was a major factor in determining the direction of the project.

3.3.1 Selection of COS

The shift on the front-end of the week had only experienced OM's acting as Captain of the Ship. The COS responsibility was not rotated among all of the OM's. In contrast, the backend shift rotated the COS role across all OM's, and even had senior manufacturing technicians responsible for goaling the line from time to time.

3.3.2 Setting Drumbeats

Another difference between shifts dealt with the setting of drumbeats. The front-end shifts would reset the drumbeats at the beginning of every shift (twice per day) based on the current inventory in each segment. However, the backend shifts were not adjusting the segment drumbeats at all during the week. They would take the drumbeat that was handed to them on Wednesday and stick with it through Saturday. Individual operations would be goaled, but segment drumbeats would not be adjusted at all. This would result in the backend of the week overcorrecting on what the last shift on the front-end was trying to execute to. After discovering this discrepancy, corrective action was taken and the backend commenced to reset the drumbeats at the beginning of every shift.

3.3.3 Different tools used by different Captains

There are several tools available to the Captain to assist in decision making. Captains who attempt to manage WIP in an effort to balance the line often rely heavily on detailed segment graphs. The detailed segment graph depicts the line divided up into day-long segments and displays how much WIP is sitting in each day. Any day with a large WIP bubble will be looked at more closely to select particular operations that should be goaled. Other Captains who focus more on back-to-front priority setting will not utilize the detailed segment graphs and will instead scan the web-based goaling tool, starting from the back of the line looking for WIP bubbles to goal.

3.3.4 Competing WIP Management Philosophies

Although the stated WIP management policy stressed "back-to-front" priority-setting, not all shifts adhered to this method to the same degree.

- Some Captains operate fairly strictly on a "back-to-front" methodology. When selecting operations to goal, they start from the back of the line looking for WIP to goal and work their way to the front.
- Other Captains give more effort to achieving overall line balance. They proactively look for areas of the line with excess inventory that feed areas with low inventory. These high-inventory areas are given priority.
- Finally, other Captains essentially just look for the biggest WIP bubbles and goal those to get wafers moving, without giving much weight to what the downstream line segments look like.

While most captains would shade their methodology towards balancing the line if the inventory bubble were big enough, each captain has a different threshold for line imbalance before they would move away from "back-to-front" bias.

3.4 Need for consistency

The variability in how the factory is run from shift to shift is an impediment to continuous improvement. With different shifts applying different WIP management methodologies, there is not a common baseline process to improve upon. However, even if all OM's agreed on how the factory should be run, the fact that there is no systematic method for determining priorities and production goals at the operation level means that there is going to be significant variability in how the factory is goaled.

Creating a Decision Support Tool for the Captains should encourage more consistency across shifts and facilitate WIP analysis and improved priority setting. An improved goaling process should lead to a more balanced line, resulting in lower cycle times and less expediting (this hypothesis is discussed in more detail in Chapter 4). One of the main goals of my internship was to target specific improvement efforts and actually execute an experiment in the fab applying some intelligent WIP management and "pull" principles to the goaling process. Regardless of the results achieved, management is adamant about the need to learn through experimentation and to apply learnings in pursuit of continuous improvement.

4. Targeting Improvement Efforts

Observation of the current reality of goaling and priority setting in Fab 23 provided an impetus to focus efforts on investigating avenues for improving existing processes.

4.1 Fab Goals

The high-level goal of the fab is to ship the right wafers at the right time at the lowest cost. There are numerous potential approaches to meeting this goal. Fab 23 attempts to meet the challenge by working to maintain overall WIP velocity through the line and then expediting specific products towards the end of the line that fall behind schedule. OM's are held accountable for meeting velocity and ship goals, not the WIP management methodology used to achieve the goals.

High WIP velocity is promoted by measuring WIP turns on every shift. WIP turns are the number of activity steps performed during a shift divided by the average inventory[5]. Shifts are recognized favorably for achieving high WIP turns, and maintaining consistently high WIP turns is a source of pride for shifts. The thought is that promoting high WIP turns will result in lower cycle times through the fab.

While cycle time through the fab has always been an important metric, it was receiving increased attention around the time my internship began. Benchmark studies had shown that Intel had room for improvement with respect to cycle time per layer of circuitry when compared to competitors. This realization served to increase the urgency of cycle time reduction. Achieving the goal of reducing cycle time through the fab would translate to increased yields[6] and lower WIP levels. Additionally, less inventory tied up in production translates to lower costs.

---

[5] Activity steps are operations that permanently alter the physical or electrical characteristics of the wafer. For example, lithography steps are NOT classified as "activities" because a pattern is only etched into resist on the wafer. The physical wafer properties do not change.

[6] The amount of time wafers spend in production in the fab is inversely related to fab yield. The longer wafers sit in queue the more opportunities there are for contamination and damage.

4.2 mX WIP Management Team

An mX team focused on TpT (throughput time) reduction was kicked off around the time I was starting my internship. This team was investigating avenues for future efforts aimed at reducing cycle time through the fab. The team had fairly high level representation (functional area managers such as Manufacturing, Process Engineering, New Products, etc.). The major areas of investigation targeted for the team were:

- Reducing TpT for New Product Introductions
- Total Productive Maintenance
- TpT Benchmarking
- WIP Management

Because my project was focused on improving some of the goaling methods used, I naturally got plugged into the WIP Management effort. In September, we formed a cross-functional mX WIP Management Team as a spin-off from the main mX TpT team. This subteam had representation from Process Engineering, Manufacturing, Automation (IT), and Manufacturing (Industrial) Engineering.

The team's hypothesis was that an improved process for goaling the line and assigning priorities could help balance the line and lower cycle times. A balanced line (section 3.2.3 discussed evaluating line balance) would have less WIP bubbles and shorter queues. In an unbalanced line, WIP bubbles lead to wafers spending idle time in queue as capacity is not available to process the wafers. Many aspects of the current goaling process and tools used are obstacles to achieving line balance. It was decided that the major deliverable for the team that Quarter was to execute a WIP Management experiment to test improvements to the current goaling process.

4.3 Major Gaps in Existing WIP Management Processes & Tools

Before designing the experiment, it was necessary to evaluate the current processes and tools to determine what major gaps existed. This section identifies and discusses some of those key gaps.

4.3.1 Back-to-front not always advisable

While strictly following a "back-to-front" methodology may make sense in some situations, there are times when shared capacity would be better utilized by focusing on front-end operations. For instance in situations in which a weekly ship goal has already been hit early in the week, back-end operations should be less of a priority. There are times when the operations towards the end of the line may be far ahead of schedule, but operations at the front of the line have fallen behind schedule. It would make more sense to use shared tool capacity to smooth the WIP distribution towards the front of the line than continue to overship where there is no demand.

4.3.2 Determining which WIP bubble should get priority is subjective

Even when a more dynamic method of setting WIP priorities is used in an attempt to balance the distribution of WIP throughout the line, the process of prioritizing one WIP bubble over another is subjective. A captain who proactively pursues line balance will review detailed segment graphs that depict WIP distribution, and give priority to bubbles with excess inventory feeding into areas of the line with light inventory. However, there are typically multiple areas of the line with WIP bubbles and areas that are light on inventory at any point in time, and the decision making process for prioritizing one bubble over another is not systematic. The process is subjective.

4.3.3 Granularity of depicting WIP distribution not optimal

The two major sources of WIP information are weekly segment graphs (Figure 2), and detailed segment graphs that show WIP placement by day-long divisions. Looking at the WIP distribution by weekly segments does not give sufficient granularity to confidently determine whether or not the line is balanced.

Because the segment is one week long, there are several segments that have 3 or even 4 lithography layers contained in one segment. The weekly segment graph might show that actual inventory equals target inventory in each segment, however all of the WIP within the segment could be contained in just a few operations.

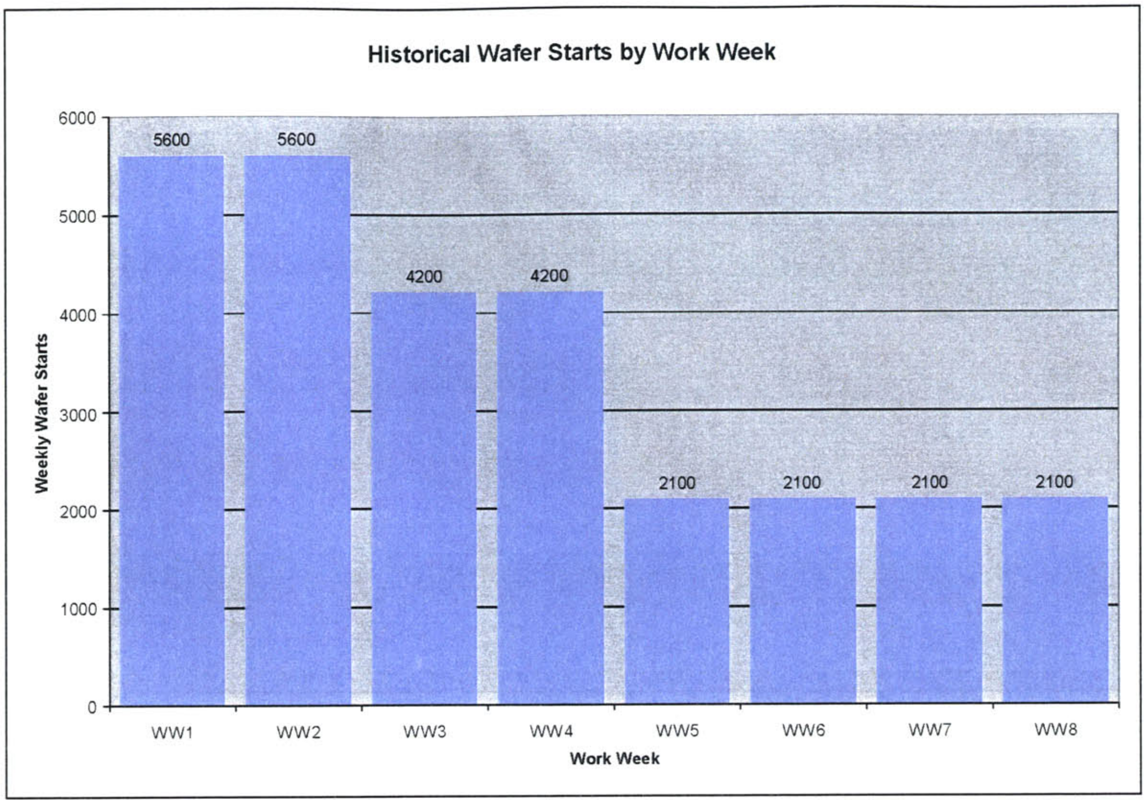
Conversely, the detailed daily segment graphs are often too granular to be of much use. Because some of the operations with longer target cycle times may tend to span the division from one day to another, decisions are made to assign an operation to a particular day. The result is that some of the "days" are longer than 24 hours, and other are shorter. There is also a certain amount of variability around these target cycle times, so the accuracy of determining the target cycle time of a small group of operations is less than when estimating the target cycle time of a larger group. For example, one can estimate the target cycle time of a wafer moving throughout the entire production process with much more confidence (~56 days) than one can predict the target cycle time of a wafer moving through one operation.

4.3.4 Target Inventory levels static throughout work week

As discussed previously, Captains evaluate segment inventory balance by comparing current inventory to what previous starts dictated should be in that segment. However, the starts number is only updated in the report on a weekly basis and stays static throughout the work week. This is not as much a problem when starts levels do not change from week to week. However, when the number of wafers introduced into the manufacturing process varies each week, this is an inaccurate gauge of target inventory.

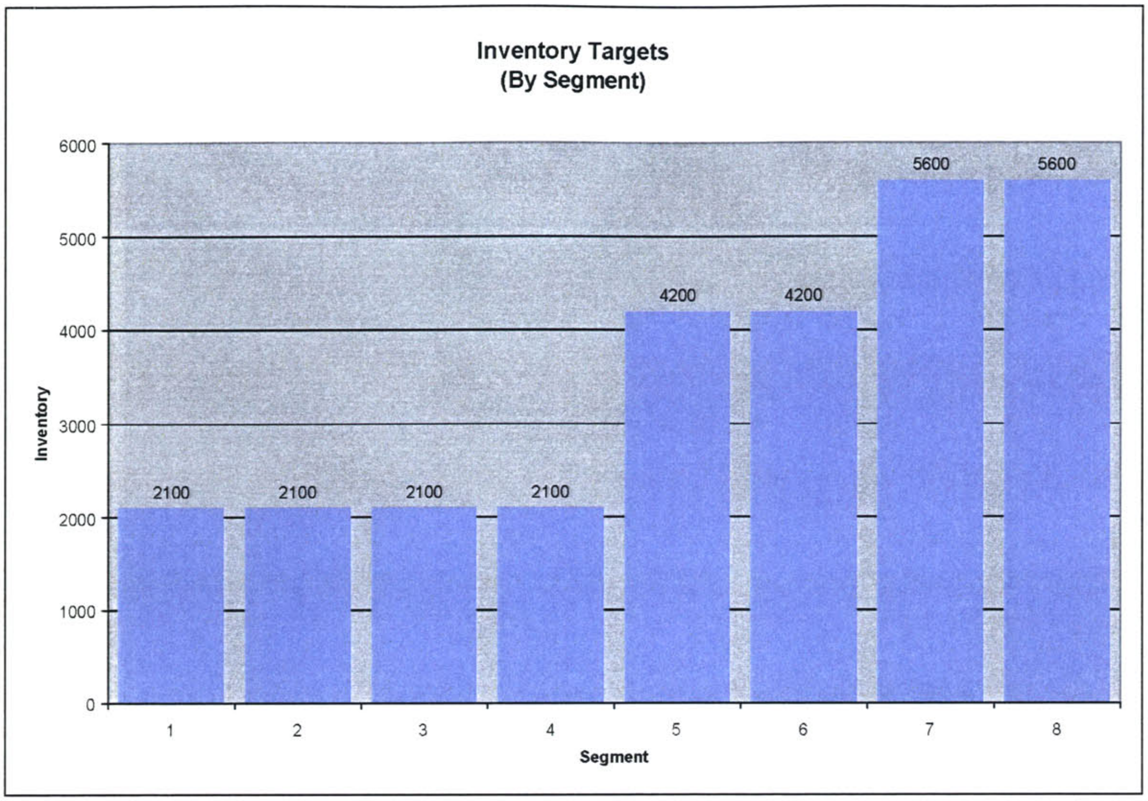
Let us consider the theoretical starts pattern discussed previously in Section 3.2.3.

Figure 6 – Historical Wafer Starts by Work Week



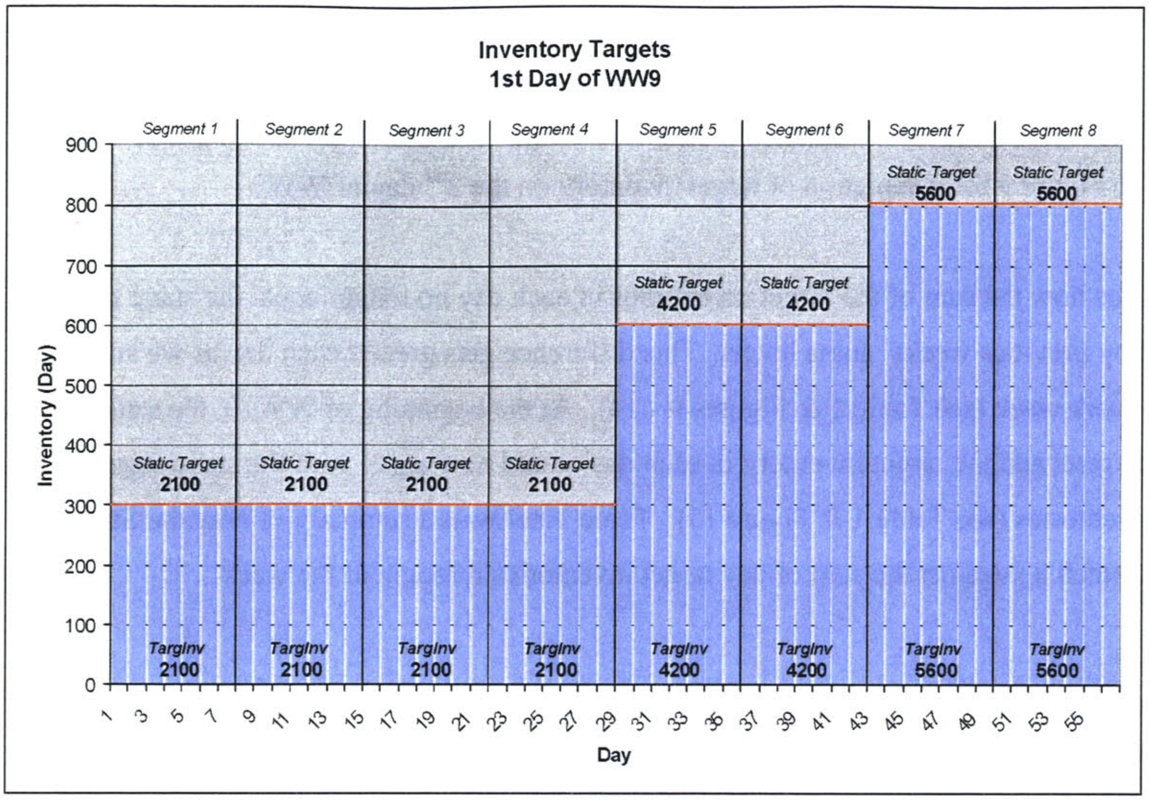
At the beginning of Work Week 9, the segment graph targets would dictate the following WIP distribution by Segment (mirror image of Figure 6):

Figure 7 – Inventory Targets by Segment



However, now let us look at the same target inventory profile divided by days (Figure 8). You can see that the target inventory as dictated by starts at the beginning of WW9 would be 2100 wafers in segment 1, with 300 wafers in each day (300 wafers/day * 7 days/segment = 2100 wafers/segment). The "Static Target" (labeled above the inventory bars) will remain the same throughout the week. The true target inventory (labeled "TargInv") can be calculated by summing up the target inventory levels for each day. At the beginning of the week, these numbers ("TargInv" and "Static Target") are equal.

Figure 8 – Inventory Targets (1st Day of WW9)



The concept of using previous wafer starts as target inventory levels implicitly assumes that WIP marches through the fab according to its target cycle time. If this assumption is valid, then the 2nd day of WW9 will have a slightly different WIP profile. All of the inventory bars have now been shifted by one day to the right (assume starts in WW9 will be 2800 wafers or 400 wafers / day). See Figure 9 for a depiction of target inventory on the 2nd day of WW9.

We can see how the sum of the target inventories in each day no longer equal the static target dictated by previous weeks' starts levels. This difference gets greater each day as we advance into the work week (see Table 1 & Figures 9 - 14). At the beginning of WW10, the static targets would be reset and the target inventory in each day would now agree with the static segment target inventories (see Table 1 & Figure 15). Table 1 below is a summary of how the Static Targets match up against the sum of day target inventories throughout the week.

Table 1 – Daily Inventory Targets throughout WW9

	Static	Target Inventory updated daily throughout WW9						
Segment	Target	1st Day	2nd Day	3rd Day	4th Day	5th Day	6th Day	7th Day
1	2100	2100	2200	2300	2400	2500	2600	2700
2	2100	2100	2100	2100	2100	2100	2100	2100
3	2100	2100	2100	2100	2100	2100	2100	2100
4	2100	2100	2100	2100	2100	2100	2100	2100
5	4200	4200	3900	3600	3300	3000	2700	2400
6	4200	4200	4200	4200	4200	4200	4200	4200
7	5600	5600	5400	5200	5000	4800	4600	4400
8	5600	5600	5600	5600	5600	5600	5600	5600

Figure 9 – Inventory Targets (2nd Day of WW9)

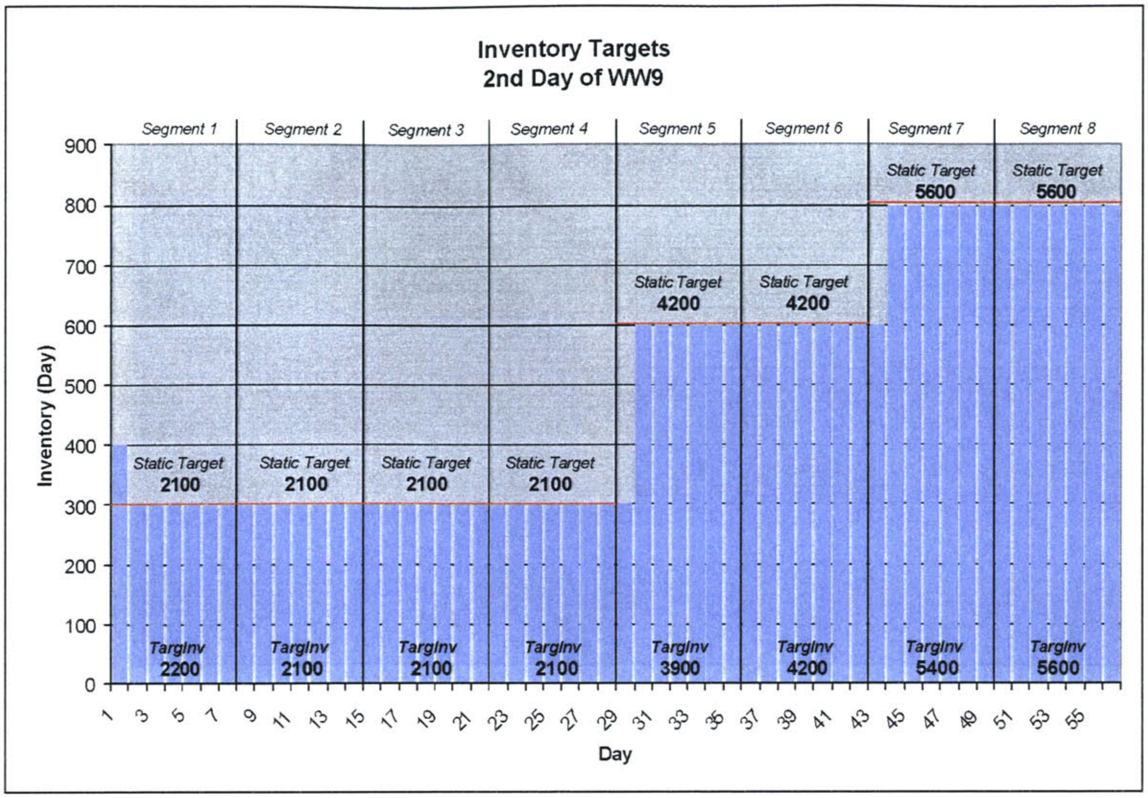


Figure 10 – Inventory Targets (3rd Day of WW9)

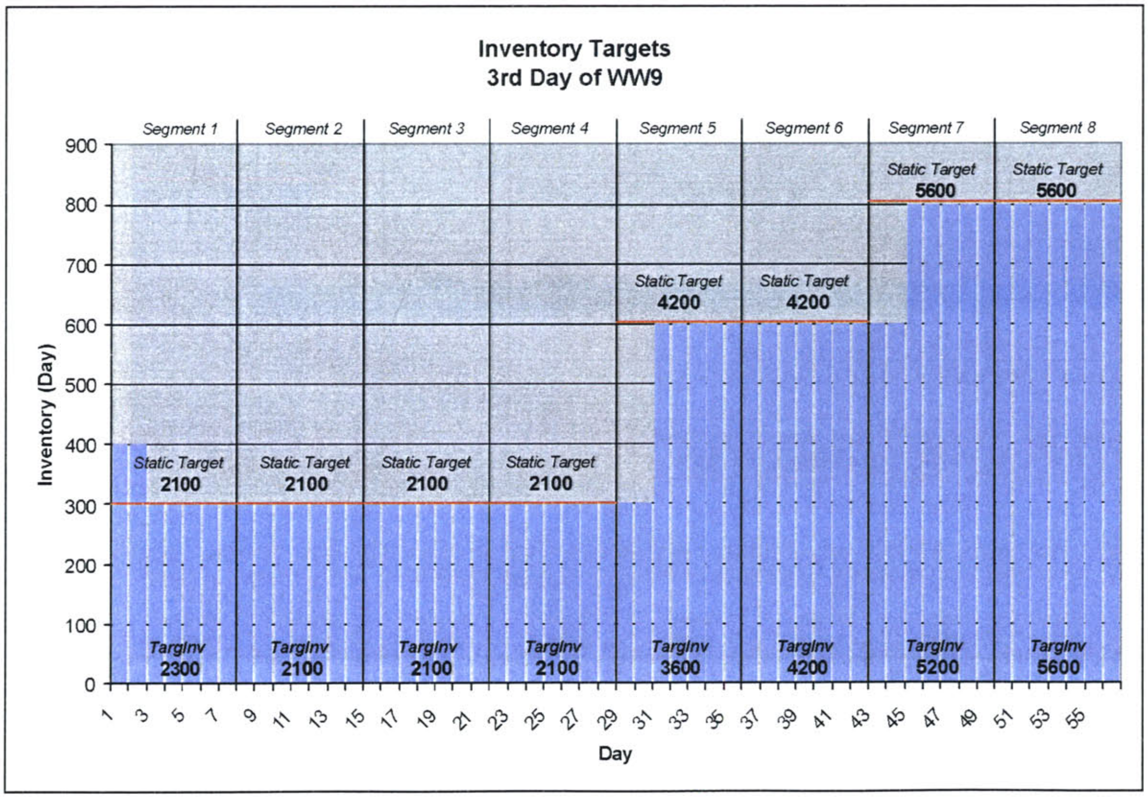


Figure 11 – Inventory Targets (4th Day of WW9)

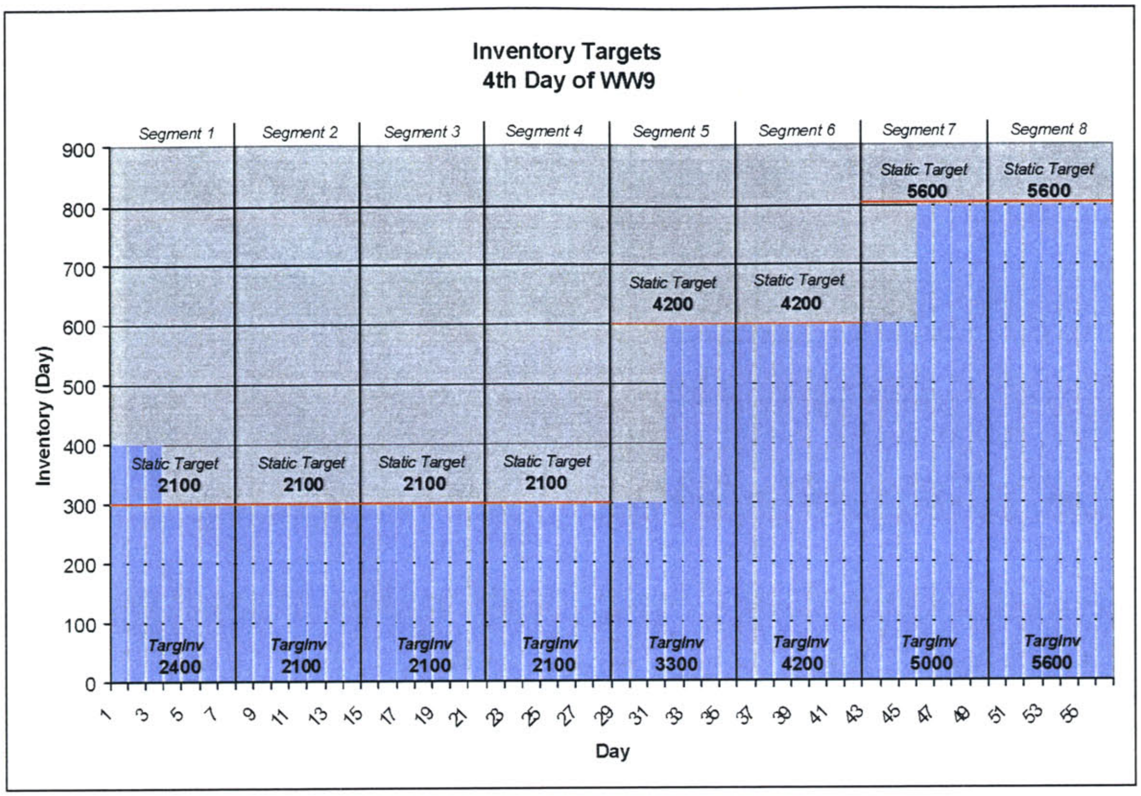


Figure 12 – Inventory Targets (5th Day of WW9)

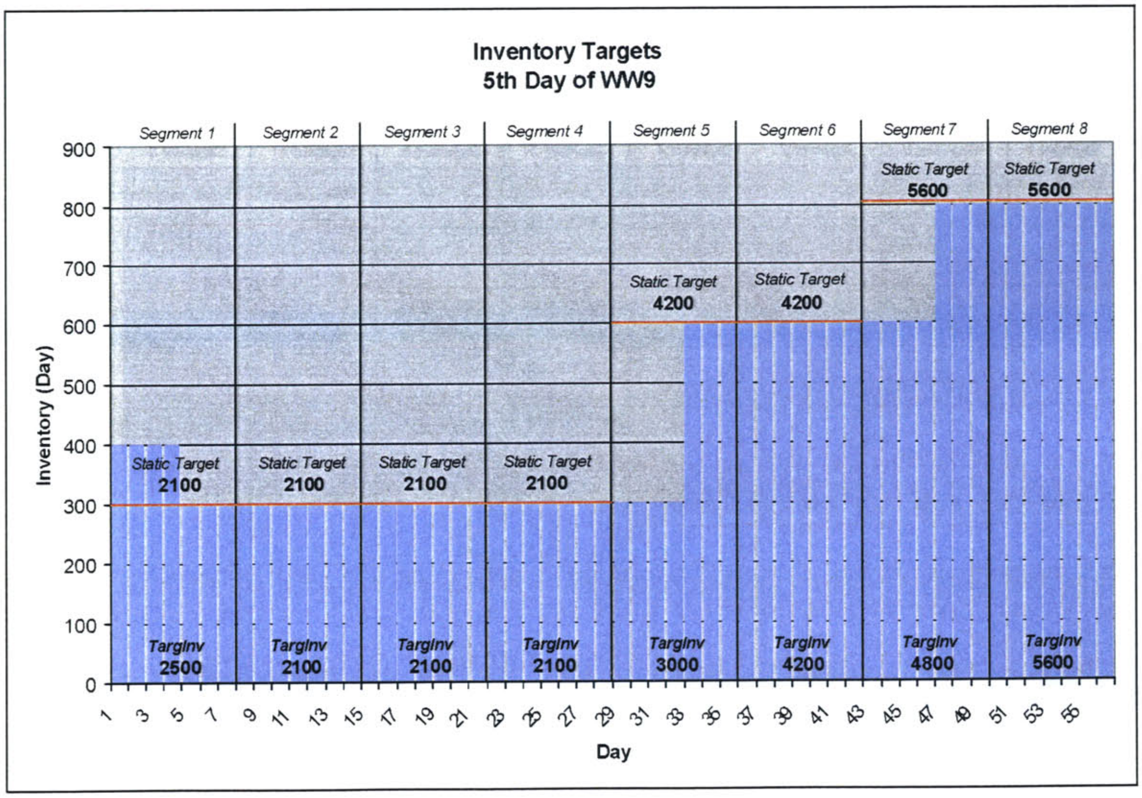


Figure 13 – Inventory Targets (6th Day of WW9)

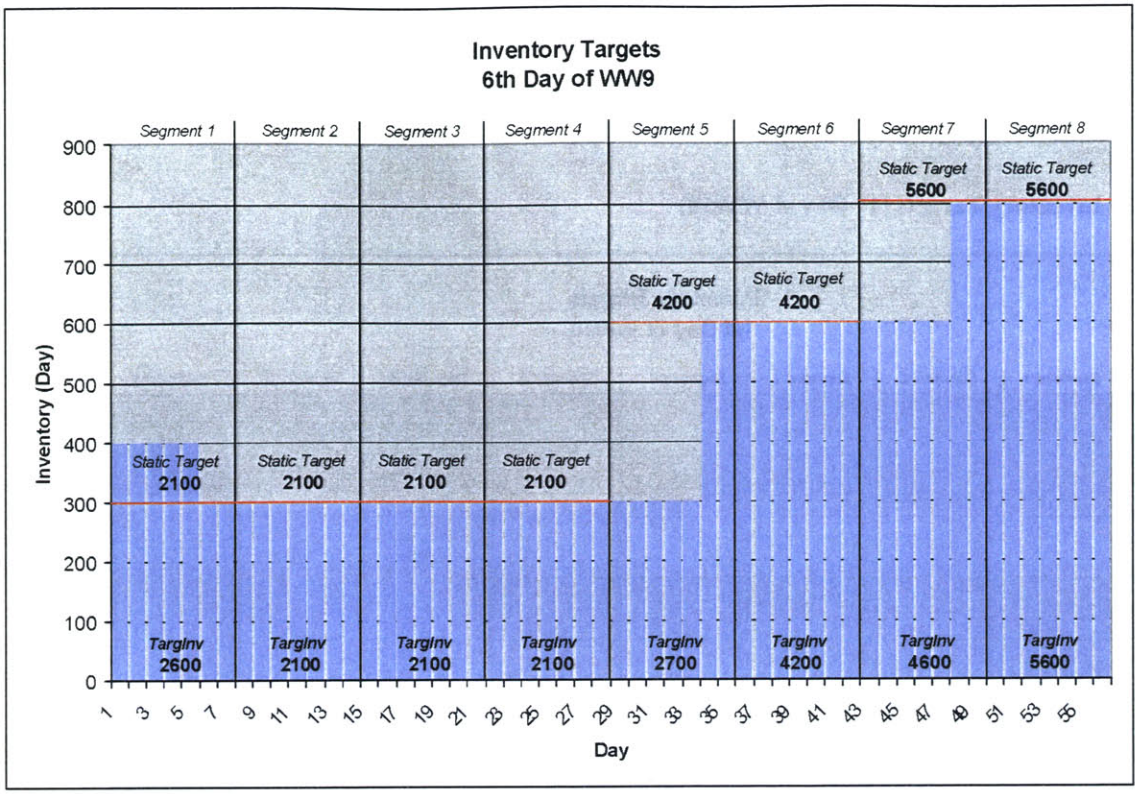
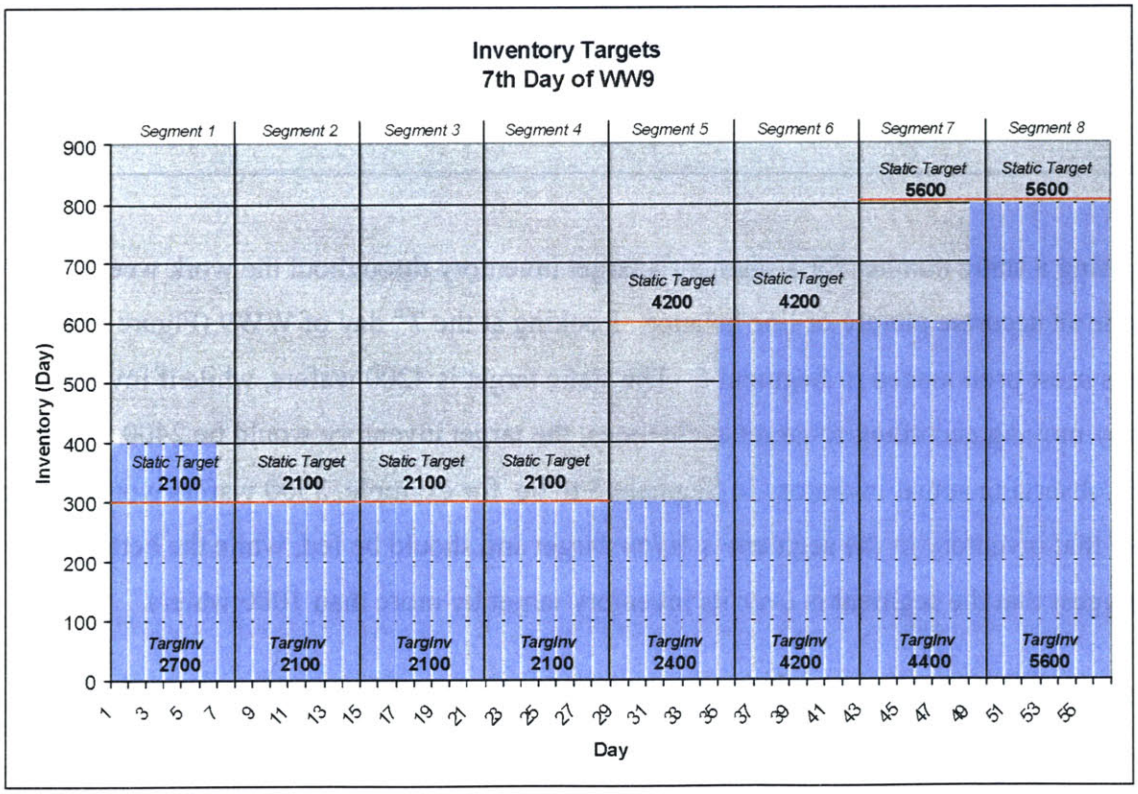
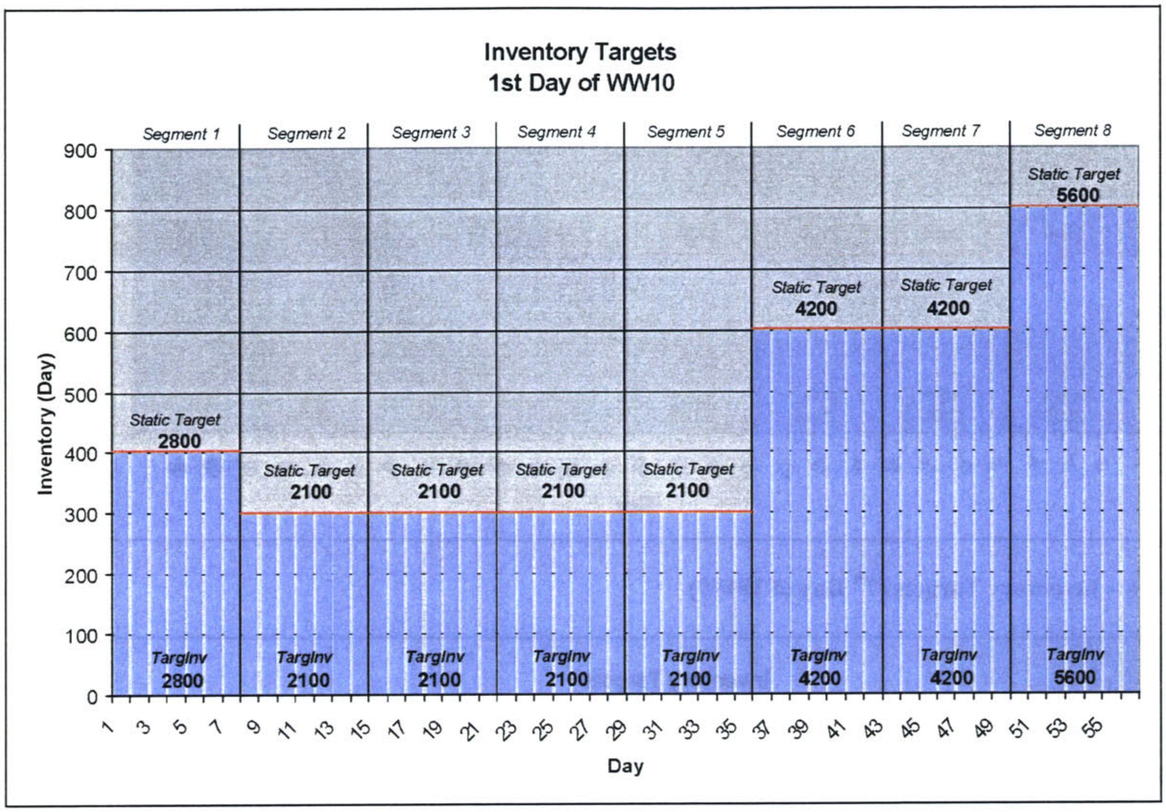


Figure 14 – Inventory Targets (7th Day of WW9)



At the beginning of WW10, the static target would again equal the sum of target inventories in each day:

Figure 15 – Inventory Targets (1st Day of WW10)



Clearly, using a static number for a segment's target inventory throughout the work week does not provide an accurate picture of line balance. Looking at the 7th day of WW9 (Figure 14), the problem is most pronounced in Segment 5. The static target is 4200 wafers, while if inventory were really moving according to target cycle times, the target inventory would be 2400 wafers. A captain observing actual inventory in Segment 5 to be, for example, 3500 wafers would be led to believe that inventory in the segment is *below* target and should be fed, while the better target would suggest that the segment is *over* its inventory target by more than 1000 wafers.

4.3.5 Segment target inventories should not be evaluated in isolation

As described earlier, operations managers who emphasize line balance will attempt to identify segments that are below target inventory levels and feed them, especially if the previous segment is over its target inventory level. However, just because Segment 5 is 500 wafers below its target inventory level and Segment 4 is 1000 wafers over its target inventory level, that does not necessarily mean that priority should be given to run Segment 4 inventory to feed Segment 5.

Consider the possibility that Segments 6 through 8 are 3000 wafers over their total target inventories. This would suggest that Segments 5 through 8 are 2500 wafers over their total target (500 wafer deficit in Segment 5 and 3000 wafer surplus in Segments 6 – 8). It may not make sense to give priority to Segment 4 to feed into a section of the line that is significantly over its target inventory levels when looked at as a group, even though the Segment immediately downstream may be under target inventory. In an article describing improvement work at Samsung, Robert Leachman et al. suggested this approach of examining downstream target inventory all the way to the end of the line when evaluating whether an inventory deficit or surplus exists [17].

4.3.6 Demand rather than starts should dictate target inventory levels

As discussed in section 4.3.4, using a static number throughout the week for inventory targets is a significant weakness in the current process. However, even if the target inventories were updated on a daily or shiftly basis, the problem remains that the targets are based on historical start levels. Historical start levels are not a good proxy for determining target inventory levels for several reasons.

Using historical start levels to determine target inventory levels implicitly assumes not only that WIP moves through the process according to target cycle times, but it also assumes that demand does not change within production lead time. For example, if 4000 wafers are started this week, the current methodology assumes that in a time period equal to the target cycle time through the fab, the demand will remain at 4000 wafers.

However, in reality, the start levels are dictated by what demand is *forecasted* to be in the future. The true demand needs in future weeks can differ from what was forecasted. Suppose we thought that demand in the current week was going to be 4000 wafers when starting wafers eight weeks ago, but now planning has determined that only 3000 wafers are needed to ship. This demand change dictates that we only need 3000 wafers in the last week-long segment of the line to meet demand, not 4000 wafers. Suppose next week's demand increased by 1500 wafers, from the original forecast of 2500 to 4000. Now the target inventory for the last 2 week-long segments of the line has increased by 500 wafers from 6500 wafers (4000 + 2500) to 7000 wafers (3000 + 4000).

Table 2 – Target Inventory using original starts forecast vs. updated forecast

	Demand Forecast		Target Inventory	
Scenario	**Current Week**	**Next Week**	**Last Segment**	**Last 2 Segments**
Forecast not updated (uses starts data)	4000	2500	4000	6500
Forecast updated to reflect expected demand	3000	4000	3000	7000

The table above illustrates that when using an updated forecast to determine target inventory levels rather than the original forecast that determined starts, the target inventory in the last week-long segment of the line went down by 1000 wafers, however the total target inventory in the last two segments of the line went up by 500 wafers. Suppose that WIP had consistently moved through the line according to target cycle times, and as a result the actual inventory in the last segment of the line is 4000 and the actual inventory in the last two segments of the line is 6500. The updated target inventory suggests that the fab is over target in the last segment, but under target in the last two segments. Therefore, it is more urgent to feed the 2nd to last segment of the line than running operations in the 2nd to last segment to feed the last segment of the line.

4.3.7 Maintenance of fab-out schedule inconsistent with forecasted demand

Shift managers are responsible for verifying the fab-out schedule on a weekly basis and making necessary adjustments. Sometimes there exist real demand-driven reasons to change the ship schedule. However, there are problems related with how this fab-out schedule is updated. If there are over-shipments in one week, the ship schedules for the following weeks will not be updated to reflect the differences in what was shipped versus what was needed. Also, shift managers will sometimes increase the ship target for the next week or two for no other reason than that there is a surplus of WIP towards the end of the line that can theoretically be shipped in that time frame.

4.3.7.1 Over- and under-shipments do not change following week's ship targets

If the fab exceeds its ship target one week, ship targets in following weeks are not reduced to account for the excess shipped. Suppose the fab needs to ship 5000 wafers / week from the fab during the next 2 weeks to meet demand. If actual shipments in the current week turned out to be 7000 wafers, the demand schedule should be updated to reflect 3000 wafers needed the next week. This would hit the 10,000 wafer goal over two weeks. However, in practice, the ship schedule for the next week would be left at 5,000 and not reduced to account for the overships in the previous week. This practice can likely be attributed to a back-to-front mindset that always seeks to keep the back-end a priority. Lowering ship goals would be looked upon as lowering the priority for the back of the line. If over-shipments were so great that there was an enormous difference between what was required over the next several weeks and the amount of WIP in the last half of the line, ship goals may be adjusted. Generally though, the practice of not updating future ship schedules to account for over-shipments serves to keep the backend urgency artificially inflated to require more shipments than forecasted demand suggests are needed.

4.3.7.2 Ship schedule modified to keep back-end urgency artificially inflated

Additionally, shift managers will sometimes raise the ship goal for the current week if it appears there is enough WIP in position to exceed the goal. An example of this might be that the shifts on the front end of the week have already hit the ship goal for the entire week. During the shift manager pass-down in the middle of the week, they would see that the goal has already been hit and that there are another 1000 wafers in the last three days of the line. They may actually raise the ship goal for the week by 1000 wafers because they can ship the extra wafers if priority is given to the end of the line.

This type of action serves to keep the back-end of the line as a priority and reinforces the "back-to-front" mentality of the line. If there is shared capacity at the end of the line, it may make more sense to not use it to ship the 1000 extra wafers, but rather use it towards the front of the line where downstream WIP deficits may exist.

4.3.8 Focus on WIP turns can be sub-optimal

As discussed in Sec 4.1, shifts are measured on their WIP turns performance. The numerator in the WIP turns calculation is *activity steps*. Focusing on maximizing activity steps performed can result in giving priority to areas of the line that have little downstream need for wafers. If an OM observes WIP staged in front of a cluster of activity steps in the line, he has incentive to goal these steps in order to raise their priority and increase his shift's WIP turns. This course of action would not be optimal if capacity could be allocated to other operations feeding areas of the line that are in greater need of WIP.

4.4 Opportunity to improve current process and tools

The gaps identified in the previous section provide a prime opportunity for experimenting with new ways to improve the process. In the spirit of mX experimentation, management and the mX WIP Management team were committed to planning and executing an experiment to do just that.

Sec 4.2 presented the hypothesis that line balance will lead to reduced cycle time. In order to achieve line balance, priorities must be given to the appropriate operations in the line. The gaps discussed in the previous section are obstacles to determining and assigning these priorities effectively:

- Back-to-front mentality leads to the front end of the line being neglected even if downstream operations are in more urgent need of WIP.
- Incentive to maximize WIP turns can lead to inefficient execution
- Even if an OM wants to balance the line and does not operate back-to-front or to strictly maximize WIP turns, the following obstacles exist:
    - Incorrect target inventories due to the use of static numbers throughout the week.
    - Incorrect target inventories due to reliance on old starts data for targets.

    - Target inventories evaluated in isolation rather than looking all the way to end of line when judging WIP surpluses and deficits.
- Maintenance of fab-out schedule serves to keep back-end urgency artificially inflated.

The methodology discussed in the next Chapter will attempt to address these gaps and provide a more systematic way to gauge priorities with the goal to improve line balance and reduce cycle time and WIP.

This page is intentionally left blank

5. Experimentation and Implementation

Experimentation is a highly valued aspect of the mX initative (see Section 2.3). One of the main goals of my internship was to actually execute an experiment in the fab applying some intelligent WIP management and "pull" principles to the goaling process. By creating a decision support tool for the captains to use when goaling the fab, we aimed to encourage more consistency across shifts and facilitate better decisions.

The proposal was framed as a 4-week experiment. This would take us up to the factory warm-down for Christmas break. We could then evaluate how the experiment went and incorporate learnings into a future experiment. Framing our effort as an experiment rather than a permanent change helped reduce resistance to our ideas and plans. The emphasis Fab23 places on the mX initiative and learning through experimentation was critical in allowing us to try these new methods. We sold the experiment by stressing that it was by no means a perfect solution, but it was a small step in the right direction. By painting a picture of major shortcomings of the current system (discussed in section 4.3) and then explaining how our experiment would help address them, it was difficult to argue that we should not proceed with our plan. It was actually refreshing how enthusiastic and supportive the captains, our key customers, were.

The framework for our approach was based on some of the principles described by Robert Leachman et al. in their paper "SLIM: Short Cycle Time and Low Inventory in Manufacturing at Samsung Electronics" [17]. This chapter will discuss the planning and technical details of the experiment, implementation tactics, and feedback received.

5.1 Technical Details of Experiment

Because captains primarily goaled the line on a gross wafer basis (irrespective of products), the experiment would focus on improving the goaling process on a gross wafer basis. It would be preferable to have a system to goal individual products specifically, but the resources required to develop the automation tools and user interfaces at the technician level for such a project put it out of our scope. However, the principles described here could readily be applied to take into account product specific priorities and goals in the future.

5.1.1 Dividing line into "micro-segments"

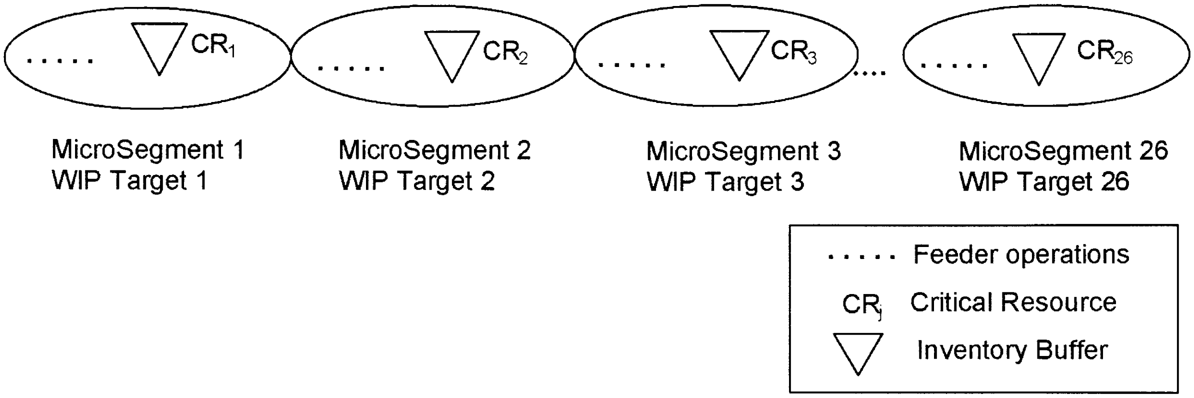
As discussed in chapter 4, the granularity of depicting WIP distribution left room for improvement. The week-long divisions of the traditional segment graphs were too long and could contain 3 or even 4 lithography layers. The day-long divisions used by the detailed segment graphs were too short. The cycle times of several operations are close to one day (considering process time and queue time), and operations often span what a strict 24-hour division would dictate.

For our experiment we chose to divide the line into 26 micro-segments. The division points were primarily lithography steps, but other near-constraint operations were also used. Lithography is commonly designed to be a factory constraint due to the cost of lithography tools. The operations at the division of these micro-segments are termed "Critical Resources" (CR). The CR is also the last step in a micro-segment.

WIP targets were also calculated for each micro-segment. The difference between actual inventory and target inventory throughout the line is the major driver when determining priorities and production goals [17].

Figure 16 shows a schematic of how the line is to be divided. Each micro-segment is comprised of one critical resource and the feeder operations upstream of that resource. Each micro-segment has a WIP target that is calculated at the beginning of every shift (see section 5.1.2 for discussion of this calculation). The goal is to keep the critical resources fed with the right amount of wafers, according to the priority of the critical resource. The priority of the critical resource represents how urgently downstream operations want to *pull* wafers out of that operation. Critical resource priorities will be discussed in more detail in Section 5.1.3.

Figure 16 – Line divided into micro-segments



The next section will discuss how target inventory levels were calculated for each micro-segment.

5.1.2 Calculating Target Inventory Levels

If the target production rate for an operation is known and the target cycle time for an operation is known, Little's Law can be applied to calculate what the target WIP level should be for that operation [18]

Little's Law tells us that WIP is equal to the product of throughput and cycle time [19]:

Equation 1 – Little's Law

WIP = Throughput * Cycle Time

Throughput is measured in wafers per unit of time so that when multiplied by Cycle Time, the WIP level in wafers is given. This equation can be applied to a single operation, line segment, or the entire line [19]. Whether or not they are aware of the *name* of the law, managers routinely apply Little's Law when evaluating inventory. When judging whether the fab is over or under target inventory, shift managers will compare actual inventory to what total starts have been over the target cycle time through the fab (see Section 3.2.3). As discussed previously in Chapter 4, when managers use starts as a proxy for target inventory, they are implicitly assuming that demand does not change within the production leadtime and thus starts can be translated into the *throughput* term in Little's Law. If 5,000 wafers were started every week over the past eight weeks (again picking eight week cycle time for illustrative purposes), it is assumed that throughput out of the fab should be 5,000 wafers per week over the next eight weeks.

Target WIP = Target Throughput * Target Cycle Time
Target WIP = 5,000 wafers / wk * 8 weeks
Target WIP = 40,000 wafers

So although managers are just adding up historical starts over the past eight weeks, they are essentially assuming demand / throughput goals do not change within target leadtime and applying Little's Law.

As discussed in Chapter 4, it would be better to use the forecasted demand and fab-out schedule to calculate target inventory levels. We will use the fab-out schedule to construct a **demand drumbeat profile** throughout the entire line (discussed in next section, 5.1.2.1). Knowing what the demand drumbeat is for every operation in the line along with the target cycle time of the operation will allow us to calculate a target inventory level for each operation. We can then aggregate operations into our micro-segments to get target inventory levels.

5.1.2.1 Constructing a demand drumbeat profile for the line

The demand drumbeat is independent of the actual inventory profile in the fab. It is the rate at which each operation should theoretically be outputting wafers in order to fulfill demand. If demand from the fab is forecasted to be constant at 10,000 wafers / wk, the demand drumbeat for every operation in the line would be 10,000 wafers / wk.

The situation is slightly more complicated when demand is not constant over the target production cycle time. Let us revisit the fab-out schedule described in section 3.2.3. Assume that this fab-out schedule is based on what demand is truly thought to be, not what historical starts levels were. Figure 17 depicts this demand schedule (same as Figure 3).

In order to be on track to ship 5600 wafers in the current week, operations in the last week of the line needs to be outputting wafers at the rate of 5600 wafers / week. Similarly if demand is forecasted to be 4200 wafers three weeks from now, operations that are three weeks from the end of the line should be outputting wafers at the rate of 4200 wafers / week. These target output rates are referred to as the demand drumbeat for the operations.

Using traditional one week segments makes it easier to illustrate the relationship between the weekly fab-out schedule and the demand drumbeat. Figure 18 shows the demand drumbeat for each segment of the line based on the fab-out schedule depicted in Figure 17. Every operation in Segments 7 and 8 would have a demand drumbeat of 5600 wafers / wk. Every operation in Segments 5 and 6 would have a demand drumbeat of 4200 wafers / wk. All operations in Segments 1 through 4 would have a demand drumbeat of 2100 wafers / wk.

Figure 17 – Fab-Out Schedule over next eight weeks

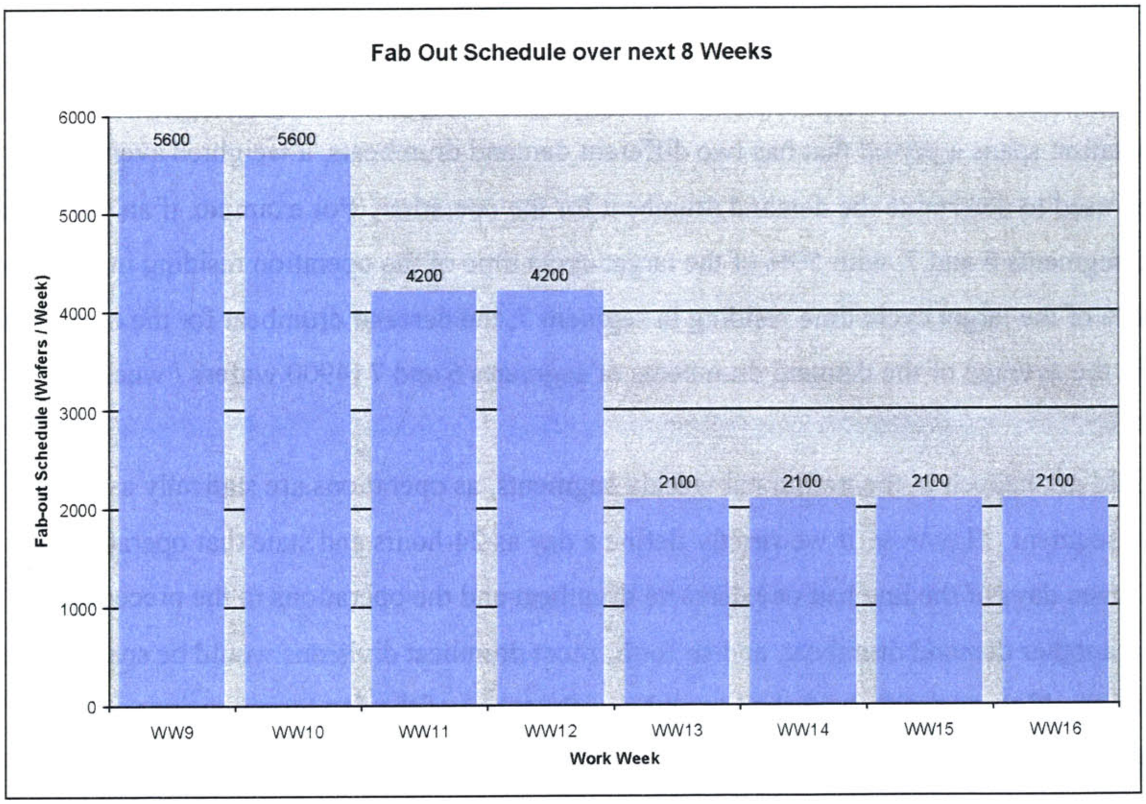
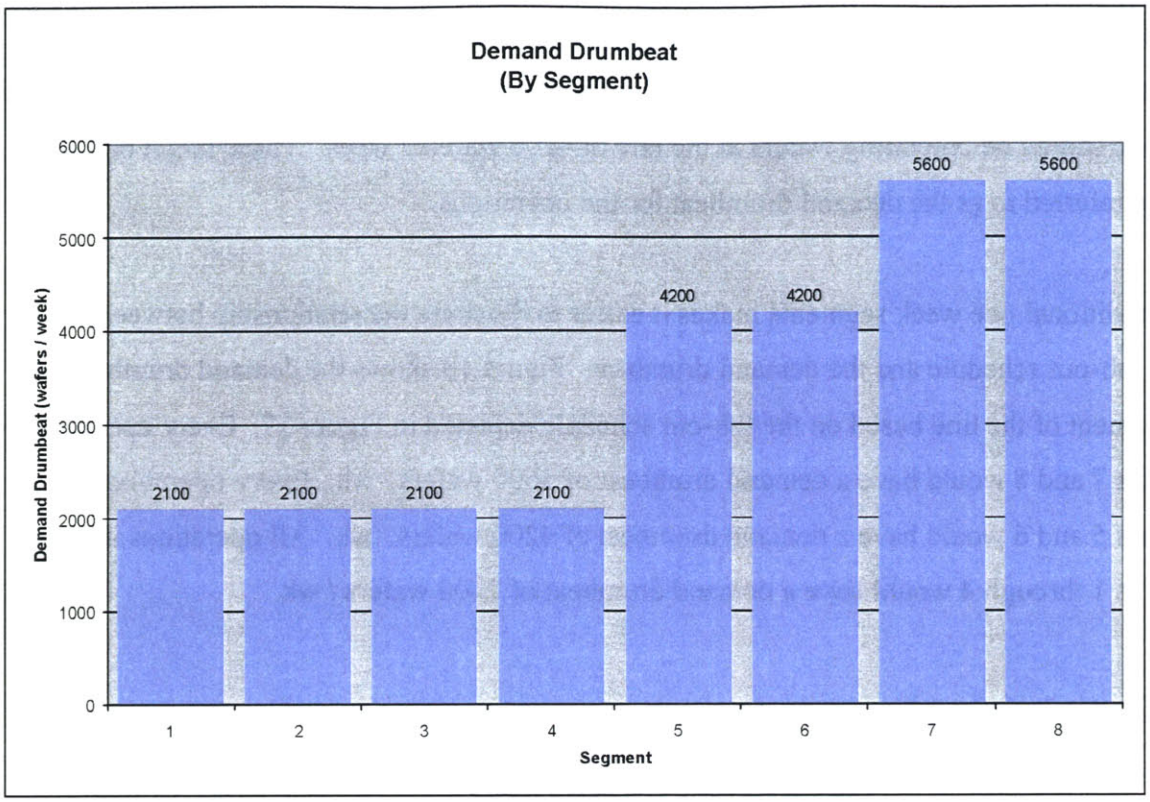


Figure 18 – Demand Drumbeat by Segment



If an operation spans a period that has two different demand drumbeats, a weighted average would be used to determine the demand drumbeat for the operation. For example, if an operation spanned segments 6 and 7, with 50% of the target cycle time of the operation residing in segment 6 and 50% of the target cycle time residing in segment 7, the demand drumbeat for the operation would be the average of the demand drumbeats of segments 6 and 7 (4900 wafers / week).

This would not happen in the traditional weekly segments, as operations are statically assigned to only one segment. However, if we strictly define a day as 24-hours and state that operations in the last seven days of the line had one demand drumbeat and the operations in the preceding 7 days had another demand drumbeat, and so forth, most drumbeat divisions would be spanned by an operation. If an operation happened to end exactly at one of the day-long divisions, it would be by chance.

5.1.2.2 Target cycle times for operations

The process engineering group at Fab23 maintains a cycle time model for all operations in the fab. First, a total target cycle time goal for the entire process is chosen. Let us stick with 56 days for the total target cycle time goal to be consistent with earlier examples. Next, the total process time for producing a finished wafer is subtracted from the total target cycle time, to get total queue time for a wafer. Process Engineering has data on run rates, batch sizes, and process time per wafer for all tools and processes from which total process time can be determined.

Suppose, the total process time for a wafer is 16 days. That would mean that a wafer spent 40 days in queue waiting to be processed. This 40 days of buffer time would then be allocated to all of the operations in the process. Historical queue time averages and variances are used in deciding how much buffer time is allocated to each operation[7].

5.1.2.3 Applying Little's Law to calculate target inventory

As discussed at the beginning of section 5.1.2, knowing the target throughput and target cycle time for an operation allows us to calculate the target WIP using Little's Law.

Target WIP = Target Throughput * Target Cycle Time

The demand drumbeat can be substituted for target throughput. The demand drumbeat is what we would want to process through each operation if had the WIP available in order to meet the fab-out schedule over the total target cycle time. Substituting into Little's Law, we get:

Target WIP = Demand Drumbeat * Target Cycle Time

Now that we have Target WIP for every operation, we can aggregate operations to calculate target WIP for all 26 micro-segments.

---

[7] Intel's cycle time calculations are confidential, however if the reader is interested in reviewing a similar methodology for calculating target cycle times, see "Stepper Scheduling in Semiconductor Wafer Fabrication Process" by Kim et al. [18]

It may be helpful to note that when looking downstream from a particular operation, all the way to the end of the line, the total target inventory from that operation to the end of the line is equal to what the demand is over the target cycle time to end of the line. For example, suppose we place ourselves at an operation that is 3 weeks from the end of the line and that the fab-out schedule over the next 3 weeks calls for 6000 wafers to be shipped. That will also be the sum of the target inventories for all the operations in that 3 week span of the line.

5.1.3 Key Metrics

Most of the key metrics used to facilitate the experiment were taken *or modified* from Robert Leachman et al.'s work at Samsung described in the paper "SLIM at Samsung". I will define the major metrics here and explain them in more detail later in this section:

- Ideal Production Quantity (IPQ) – represents how many wafers downstream operations want to pull out of CR in order to have WIP positioned to fulfill forecasted demand.
- Schedule Score (SS) - represents how far behind schedule an operation is.
- Balance Index (BI) - represents how full or empty downstream micro-segments are.
- Priority - represents how urgently downstream operations need WIP.

5.1.3.1 Ideal Production Quantity

The concept and calculation for an operation's Ideal Production Quantity (IPQ) was taken from Robert Leachman et al.'s work at Samsung [17]. For our experiment, we would calculate the IPQ for all identified Critical Resources. This value represents how many wafers downstream operations want to pull out of the CR in order to have WIP positioned to fulfill forecasted demand. Another way to phrase this, is that the IPQ is the number of wafers an operation needs to process in order to be on schedule.

The equation for calculating the IPQ over one shift for CR_j is:

Equation 2 – Ideal Production Quantity over one shift

$$\text{IPQ}_j = (\text{Demand Drumbeat}_j \cdot 1\,\text{shift}) + (\text{Target WIP}_{j\,\text{thru FabOut}} - \text{Actual WIP}_{j\,\text{thru FabOut}})$$

Examining this equation, we can see that we start off with the operation's demand drumbeat (described in Section 5.1.2.1). This is the rate at which wafers should be processed through an operation if the line was balanced and WIP were on schedule to meet forecasted demand.

The next part of the equation shows us that the difference between *Target WIP from the CR to the end of the line* and the *Actual WIP from the CR to the end of the line* is a major driver of the IPQ. Whenever downstream Actual WIP is less than downstream Target WIP, the IPQ will be scaled up. Whenever downstream Actual WIP is more than downstream Target WIP, the IPQ will be scaled down. It is possible that an operation could have a negative IPQ if actual downstream WIP were much greater than target downstream WIP.

5.1.3.2 Schedule Score

The "Schedule Score" (SS) represents how far ahead or behind schedule an operation is. The calculation of Schedule Score used in our experiment is a slight modification of Leachman's definition:

Equation 3 – Schedule Score

$$SS = -\frac{\text{IPQ}}{\text{Fab Drumbeat}}$$

The negative applied to the schedule score calculation so that the more negative the schedule score, the further behind schedule the operation is [17]. The fab drumbeat is the average demand drumbeat across the fab. If the targeted cycle time through the fab is eight weeks, the demand drumbeat would be the forecasted demand over the next eight weeks divided by 112 shifts (number of shifts in eight weeks).

Again, the units for schedule score are time units and the metric represents how far ahead or behind schedule an operation is.

5.1.3.3 Balance Index

A "Balance Index" (BI) is used to gauge how full or empty the line is downstream from an operation. The concept of a Balance Index or Balance Rate is described in both Leachman's Samsung paper and in a paper on bottleneck scheduling of a fab by Young Hoon Lee, Jongkwan Park, and Sooyoung Kim [17, 21]. While the spirit of gauging line balance is the same, we made modifications for the BI used in our experiment.

The "Balance Index" used in this experiment indicates how empty or full the next 3 downstream microsegments are. In order to introduce the concept of a Balance Index, let us initially assume that we are only concerned with how empty or full the next downstream microsegement is. The BI equation would look like:

$$\text{B.I.}_{\text{CR n}} = \frac{\text{Actual WIP}_{\text{MicroSegment n+1}} - \text{Target WIP}_{\text{MicroSegment n+1}}}{\text{Target WIP}_{\text{MicroSegment n+1}}}$$

It may be helpful to reference Figure 16 to visualize what the equation is actually telling us. We are looking at the difference between Target WIP and Actual WIP in the next microsegment as a ratio of the Target WIP. If we are on target and Actual WIP = Target WIP, the BI is zero. If we have no downstream WIP, the BI will be -1. If Actual WIP is greater than Target WIP, the BI will be positive. Thus, the more positive the BI the more full the next downstream layer is, and the more negative the BI the more empty the next downstream microsegment is (with a value of -1 indicating no WIP in the next downstream microsegment.

We wanted to look further downstream than just the next critical resource, but did not feel it necessary to always look downstream all the way to the end of the line (IPQ already takes into account downstream WIP surplus / deficit looking to the end of the line). We decided to look three micro-segments downstream when determining if the critical resource was feeding into an area of the line that was heavy or light with inventory.

Additionally, we are more concerned if the microsegment immediately downstream is empty than if the microsegment 3 downstream is empty. Therefore, we applied different weights to the downstream microsegments. Although the specific values chosen for the weightings are somewhat arbitrary, they served as a good starting point. The weighting given to downstream microsegments was 70% to the nearest microsegment, 20% to the next microsegment, and 10% to the third microsegment. Section 6.2.3 discusses recommendations for improving the weighting scheme.

The balance index equation used was:

Equation 4 – Balance Index

$$\boxed{\text{B.I.}_{\text{CR(j)}} = 70\% \cdot bi_{\text{CR}(j+1)} + 20\% \cdot bi_{\text{CR}(j+2)} + 10\% \cdot bi_{\text{CR}(j+3)}}$$

$$\text{Where} \quad bi_{\text{CR}(j)} = \frac{\text{Actual WIP}_{\text{MicroSegment(j)}} - \text{Target WIP}_{\text{MicroSegment(j)}}}{\text{Target WIP}_{\text{MicroSegment (j)}}}.$$

5.1.3.4 Critical Resource Priority

The "Schedule Score" and "Balance Index" combine to determine the priority. First, the Critical Resource is assigned to a priority bucket. A simplified version of the method we used is illustrated below, with only 5 buckets. This matrix was adapted from Leachman et al. [17].

Table 3 – Priority Bucket Matrix

	BI: empty	BI: okay	BI: full
SS: behind	1	2	3
SS: okay	2	3	4
SS: ahead	3	4	5

The most urgent critical resource would be found in category "1" where the BI indicates we are feeding into an empty portion of the line AND the SS indicates that we are behind schedule at that operation. The least urgent CR would be found in category "5" where the BI indicates that we are feeding into a WIP bubble AND the SS indicates that we are ahead of schedule.

Within each bucket, the operations are then ranked according their Schedule Score [17]. The operation furthest behind is the most urgent in that bucket. All operations in bucket 1 are higher priority than operations in bucket 2, and so forth. Using these 2 metrics (BI and SS) when determining priority allows us to de-prioritize an operation that is behind schedule if it is feeding into a WIP bubble. Or conversely, it allows us to increase the priority of an operation that is ahead of schedule when looking to the end of the line, but is feeding into an area of the line that is relatively empty.

Using this priority ranking criteria, we can now assign a priority rank 1-26 to each of the 26 Critical Resources. This priority represents how urgently downstream operations need WIP. Table 4 in section 5.2.1 shows an example of Critical Resource priorities. Section 5.2.3 will discuss how the priorities calculated using this methodology drive different behavior than what a "back-to-front" methodology would suggest.

5.2 Experiment Implementation

The primary goal of the WIP experiment was to provide a useful decision-support tool and methodology to the Captain of the Ship to enable more consistent and intelligent decisions when goaling the factory. The thought is that the proposed methodology will drive the fab to a more balanced line and facilitate more balanced flow through the line.

In order to get all shifts committed to executing the experiment, high level support was necessary. The plant manager and manufacturing manager were proactive supporters of the initiative and clearly stated the importance of the experiment. The shift managers committed to providing one of their senior Operations Managers as a primary point-of-contact (POC) for the experiment. We worked closely with the shift POC to coordinate training for Operations Managers, Area Coordinator, and Manufacturing Technicians.

5.2.1 Development of Excel based tool

An excel-based decision support tool was developed to automate calculation of the key metrics described in section 5.1.3 and provide other useful information for use in goaling decisions. The only manual input required from the operations manager is the fab-out schedule. Using the fab-out schedule and the target cycle times for operations, the tool develops a target inventory profile throughout the entire line as described in section 5.1.2. The tool also pulls in current actual inventory data for all operations.

The fab-out schedule allows the tool to build a demand drumbeat profile throughout the line as described in section 5.1.2.1. Having information for target inventory, actual inventory, and demand drumbeat enables the tool to calculate all of these key metrics: Ideal Production Quantities for critical resources, Schedule Score, Balance Index, and Priority.

Graphs were also created to give a visual representation of inventory status. Two key graphs are an inventory profile by micro-segment and total downstream inventory from each of the critical resources.

See Figure 19 for a depiction of target and actual inventory by micro-segment at a specific point in time. The red line represents the target inventory for each micro-segment as calculated from the fab-out schedule and target cycle times. The blue columns represent actual inventory in each micro-segment.

The balance index of CR3 (located at end of Micro-Segment 3) is positive, with a value of 1.96. Looking at the three downstream micro-segments in Figure 19, we can see that micro-segment 4 (MS4) has more than three times its target inventory, MS5 is slightly under its target inventory, and MS6 is more than twice its target inventory. The BI calculation is:

$$\text{B.I.}_{CR3} = 70\% \cdot bi_{CR4} + 20\% \cdot bi_{CR5} + 10\% \cdot bi_{CR6}$$

$$\text{B.I.}_{CR3} = 70\% \cdot (2.670) + 20\% \cdot (-0.306) + 10\% \cdot (1.559)$$

$$\text{B.I.}_{CR3} = 1.96$$

Figure 20 is a depiction of actual and target downstream inventory from the same point in time as the data in Figure 19. The red line represents target downstream inventory from each Critical Resource, and the blue columns represent actual downstream inventory from each Critical Resource. When observing the downstream WIP, we are looking all the way to the end of the line. When the blue column is higher than the red line, that means that actual downstream WIP is greater than our target downstream WIP and we have surplus WIP. Conversely, when the red line is above the blue column, we have a WIP deficit downstream. The difference between the blue column and the red line is the second part of the IPQ equation: Target $\text{WIP}_{\text{j thru FabOut}}$ – Actual $\text{WIP}_{\text{j thru FabOut}}$

$$\text{IPQ}_{\text{j}} = \left(\text{Demand Drumbeat}_{\text{j}} \cdot 1\,\text{shift}\right) + \left(\text{Target WIP}_{\text{j thru FabOut}} - \text{Actual WIP}_{\text{j thru FabOut}}\right)$$

Visually, it follows that the further below the red line the blue column is, the farther behind schedule the operation is, and the higher the IPQ will be. When the blue column is above the red line, the farther ahead of schedule the operation is, and the IPQ might often be negative. This would happen when the difference between target and actual downstream WIP is greater than the shiftly demand drumbeat.

Looking at Figure 20, again at CR 3, we can see that the actual downstream inventory is slightly above the target downstream inventory. Thus, the IPQ for the operation is slightly negative and the schedule score is 1.1. This indicates that CR 3 is ahead of schedule.

Evaluating CR 3 at the time period represented by Figure 19 and 20, we have seen that the operation is feeding into an area of the line that is heavy with inventory (BI = 1.96), and the operation is ahead of schedule (SS = 1.1). Using the priority matrix described in section 5.1.3.3, CR 3 would be designated as a low priority operation. In fact, the priority ranking of CR 3 was 22 out of the 26 critical resources. This suggests that downstream operations do not have an urgent need to pull wafers out of CR 3 when compared to other parts of the line.

Figure 19 – Micro-Segment Inventory: better line balance

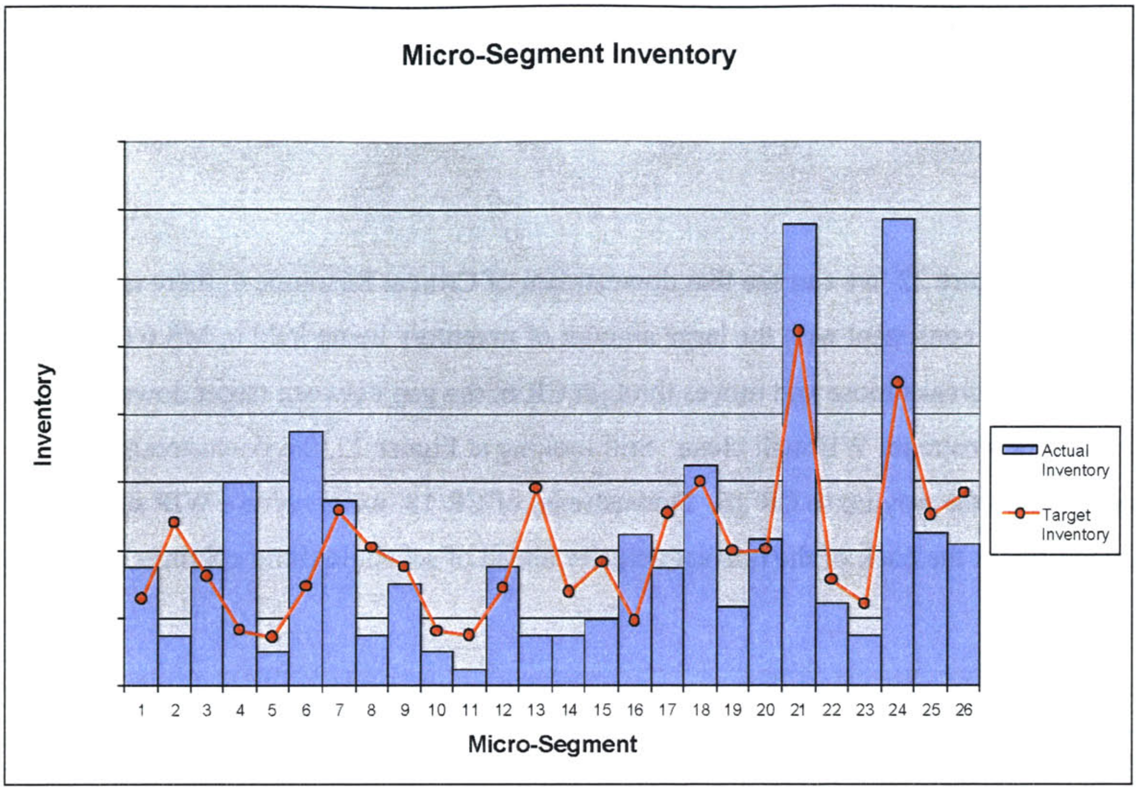
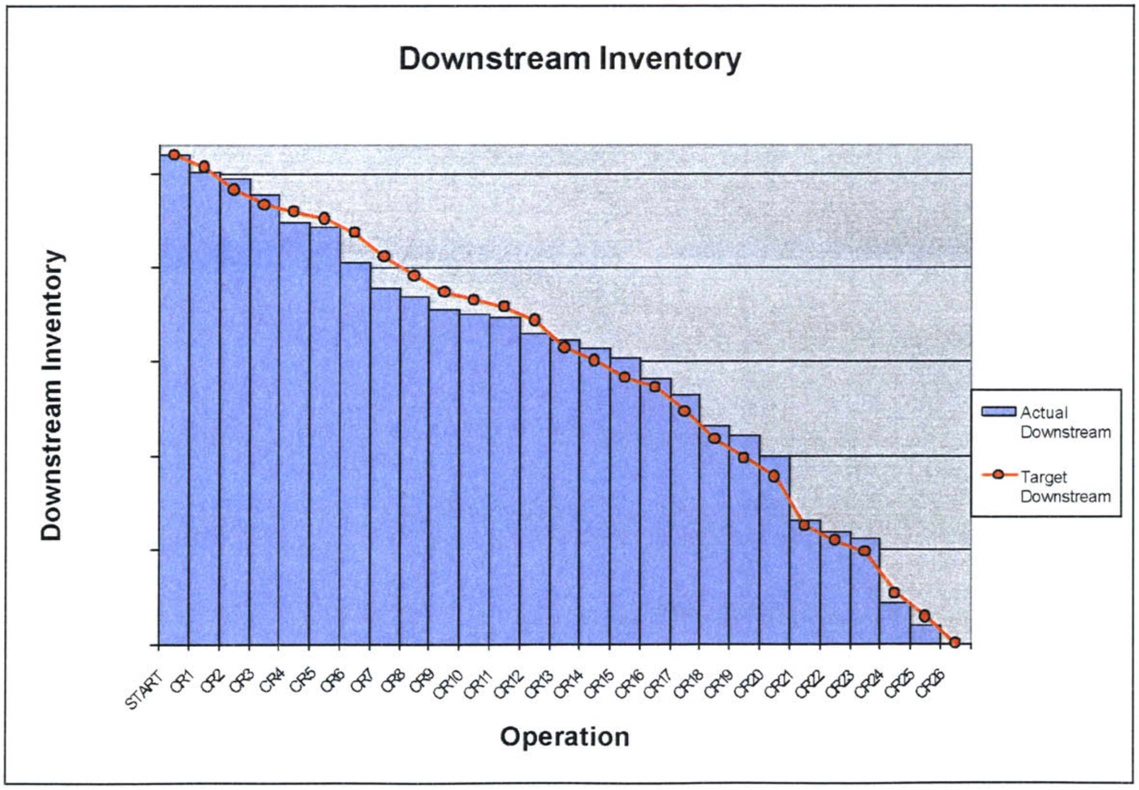


Figure 20 – Downstream Inventory: better line balance



The line was fairly balanced at the time represented by Figures 19 and 20. Now let us look at a time period when the line was not as balanced. Figures 21 and 22 are from such a time. It is immediately apparent from Figure 21 that the line is not well balanced as there are two extremely large WIP bubbles in MS 6 and MS 24. There are also long areas of the line that are relatively dry of WIP.

When looking at Figure 22, we can see that downstream of Critical Resource 6, there is large WIP deficit. This is consistent with the large amount of inventory being held in MS 6 (Figure 21). Once the WIP breaks loose and moves through CR 6, the gap between target downstream WIP and actual downstream WIP will close. Still looking at Figure 22, the downstream WIP deficit continues all the way up to CR 16. Downstream of CR 18, we observe a WIP surplus exists. Operations at the back of the line are actually ahead of schedule (blue columns are above red line).

Figure 21 – Micro-Segment Inventory: poor line balance

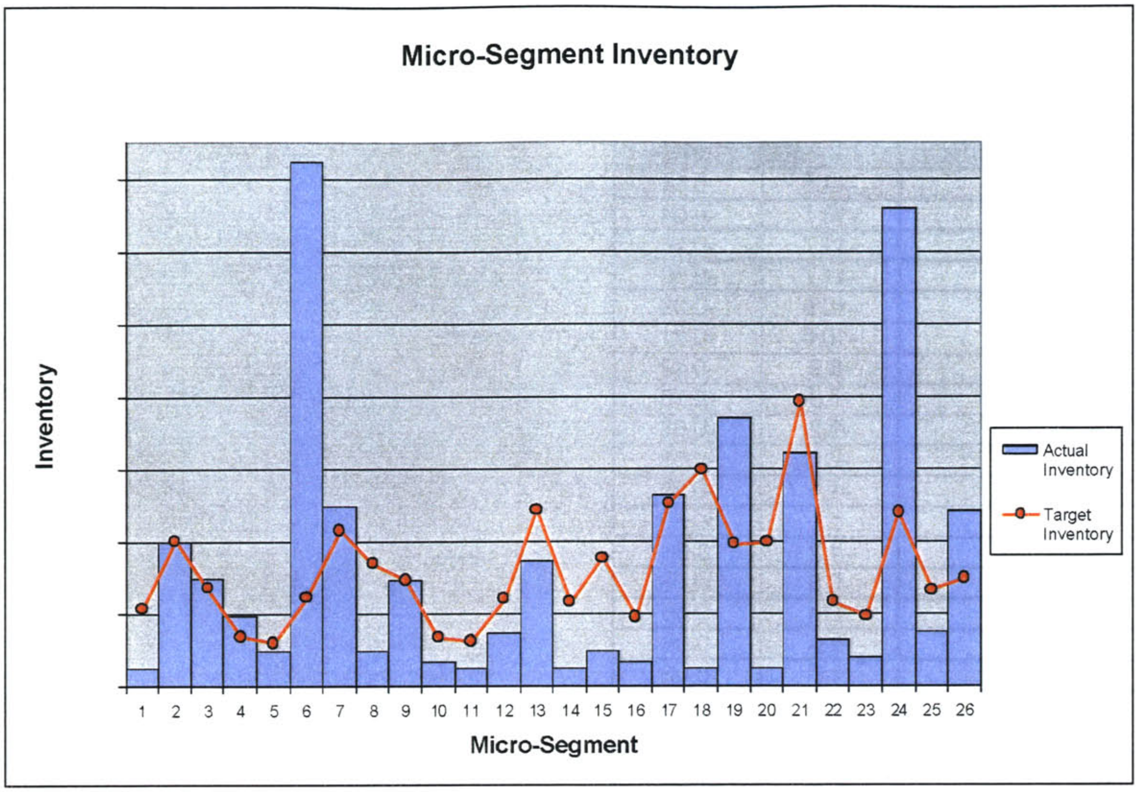
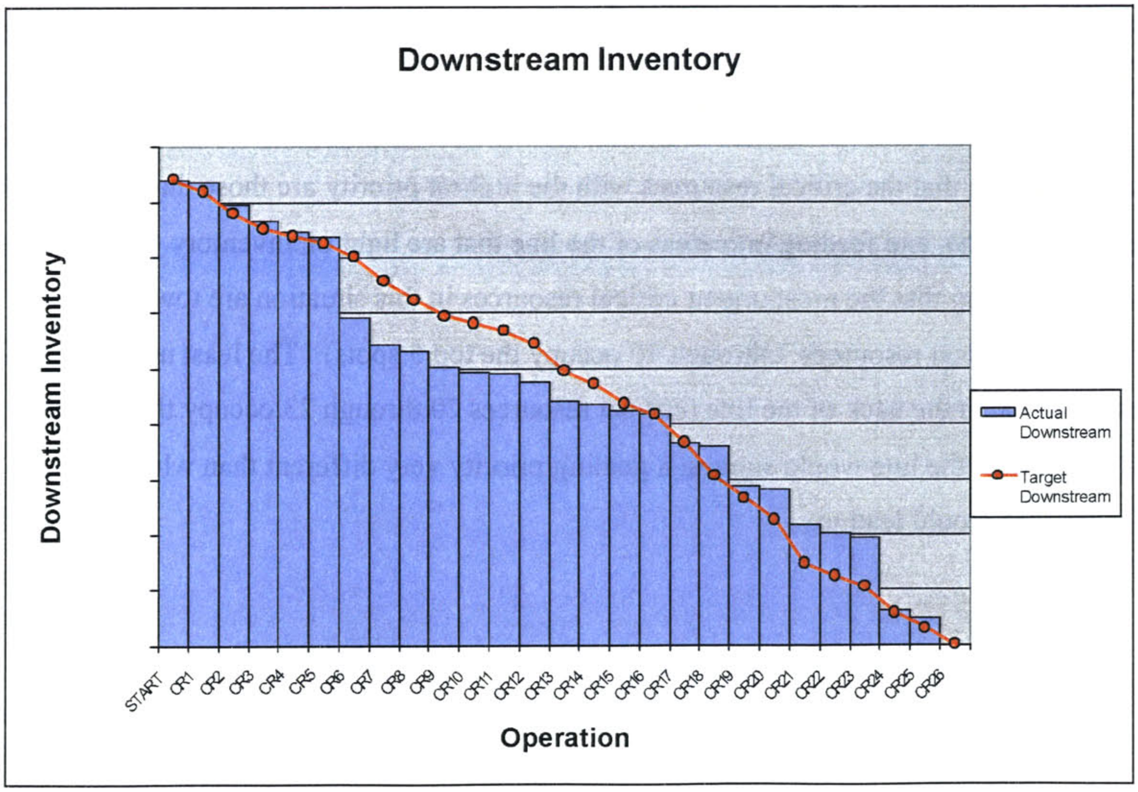


Figure 22 – Downstream Inventory: poor line balance



The rankings from the point in time represented by Figures 21 & 22 are shown below:

Table 4 – Priority Rankings consistent with Figures 21 & 22

Rank	Critical Resource	Sched Score	Balance Index
1	CR7	-14.5	-0.54
2	CR6	-13.7	-0.04
3	CR8	-11.7	-0.16
4	CR9	-11.7	-0.50
5	CR10	-10.9	-0.53
6	CR11	-10.0	-0.41
7	CR12	-8.9	-0.43
8	CR13	-7.3	-0.76
9	CR14	-5.2	-0.63
10	CR15	-2.2	-0.53
11	CR17	-1.1	-0.55
12	CR19	1.3	-0.69
13	CR26	-1.2	0.06
14	CR16	-0.8	-0.06
15	CR24	-0.4	0.06
16	CR25	0.9	0.06
17	CR1	1.0	0.06
18	CR2	1.0	0.14
19	CR4	0.0	0.85
20	CR5	0.3	3.37
21	CR3	0.7	0.75
22	CR18	5.3	0.42
23	CR20	5.4	-0.27
24	CR21	7.0	-0.25
25	CR22	8.2	-0.11
26	CR23	9.3	1.19

Table 4 above shows that the critical resources with the highest priority are those that are behind schedule (negative SS), and feeding into areas of the line that are light on inventory (negative BI). It is worth noting that the most urgent critical resources in this situation are towards the front of the line (critical resources 7 through 10 occupy the top 5 spots). The least urgent critical resources are towards the back of the line (critical resources 20 through 23 occupy the bottom 4 spots). This view of the line would suggest a goaling priority very different than what a "back-to-front" approach would lead to.

5.2.2 Experiment Execution

The captain updates the data in the excel-based decision-support tool at the beginning of every shift. This ensures that the target inventory profiles are up to date when comparing to actual inventory. This is a major improvement over using the static, wafer-starts based target inventories represented by the old detailed segment graphs.

To review, the spreadsheet calculates:

- Target Inventories
- Ideal Production Quantities
- Schedule Score
- Balance Index
- Priority Ranking

Fab23 automation actually provided a capability to upload the IPQ's and Priorities directly from the spreadsheet to the web-based goaling tool used by the Captains. This would prove very helpful when the captains are scrolling through the goaling interface and determining which areas of the line to goal.

The Priorities and IPQ's signal the CoS which feeder operations need to be given priority and goaled. Feeder operations are operations that are not designated as a Critical Resource that feed into a CR. The CoS will attempt to goal the appropriate feeders in an effort to put Critical Resources in a position to hit their IPQ, giving preference to the highest priority operations.

By manually goaling these operations in the web-based goaling tool, the CoS actually influences what the technicians on the floor see at their tool. The MT's have a graphical user interface at the tool level to see what lots are available for processing through their area/tool, and what the priorities are. When the CoS manually goals an operation, that operation goes to the top of the screen, is flagged as a priority, and the area's progress towards that goal is tracked in percentage terms.

The MT's will then focus on hitting their feeder goals. If all goals are hit, they will continue to run lots, giving preference to the highest priority operations.

Essentially, the critical resources can virtually ignore the priorities. The onus is put on the feeder operations to get the high priority material to the critical resource so it can be run. Because critical resources are constraint or near-constraint areas, they are going to run whatever WIP they have to run most efficiently [17].

Example

A lithography tool is responsible for 2 operations that are designated as CRs. One of the CR's is designated a Priority 2 with an IPQ of 500 and the other is a Priority 20 with an IPQ of 0. Suppose that only 25 wafers will be available for processing through the Priority 2 operation this shift, and 250 wafers will be available to be processed through the Priority 20 operation. Litho should run the 250 wafers of Priority 20 before touching the 25 wafers of priority 2. Because lithography is a constraint, high utilization needs to be maintained.

The correct way to interpret the priorities is this: the operations feeding the Priority 2 CR should be given high priority and specifically goaled. Feeders need to work to get the Priority 2 operation in a position to reduce its IPQ of 500. The operations feeding the Priority 20 CR are very low priority. That operation already has more than enough wafers to hit its IPQ of 0.

A situation like the above example might often go against the back-to-front philosophy manifested in the WIP management of the fab[8]. Using our decision-support tool, there are many times that preference will be given to operations further upstream because they are deemed to be higher priority according to the Schedule Score and Balance Index.

The lithography tool in the above example could still use priorities as a tie-breaker. Suppose they had 500 wafers of Priority 2 and 500 wafers of Priority 20 to run. Lithography could use priority ranking as a tie-breaker and choose to run the priority 2 operation. Additionally, if the

---

[8] The Priority 2 CR could be closer to the front of the line, while the Priority 20 CR is closer to the end of the line. "Back-to-front" would prioritize the back-end over the front.

lithography technicians understand the priorities, they can proactively contact upstream suppliers to request WIP for High Priority operations.

5.2.3 Example of Pull methodology driving different behavior

In the early stages of our 4-week experiment, I sat in on virtually all of the goaling meetings for every shift to observe how the new process was being implemented and get feedback from manufacturing. One of the shifts changed the OM acting as Captain mid-way through the week. Although I had spent time discussing the experiment with the new Captain, he was not quite as familiar with some of the details as were other Captains. He did goal some of the high priority feeders, although I noticed that he still gave a high priority to the end of the line although it was not flagged as a priority in the decision-support tool.

Table 5 shows some of the information depicted in the priority sheet from the decision support tool. In this particular situation, there were large WIP bubbles in the feeders to CR19, CR21, and CR24. The operations feeding these critical resources shared many of the same tools. The Captain used the vast majority of shared capacity in the feeders to feed into CR 24. The remaining capacity was used to feed CR 21. The captain left no shared capacity available to feed CR 19.

However, when we look at Table 5, we see that CR 21 and CR 24 are lower priorities than CR 19. Downstream operations want to pull wafers out of CR 19 more urgently than they need wafers out of CR 21 or CR 24. CR 24 is ahead of schedule with a schedule score of 2.8 and it is feeding into an area of the line that has an inventory surplus (BI = 0.4). CR 21 is also ahead of schedule with a schedule score of 2.1. CR 19 is behind schedule with a schedule score of -4.0 and is feeding into an area of the line that is light on inventory (BI = -0.7)! It does not make sense to give priority to operations feeding CR 21 and CR 24 over CR 19.

I met with the Captain after the goaling meeting and discussed this discrepancy. Once the priorities were explained, he completely agreed that it would have been better to have used the shared capacity to feed CR 19 over CR 21 and CR 24. He stated that he reverted to a "back-to-front" methodology as he was finishing up the goaling of the line. At the end of shift the line was even more unbalanced with CR 21 and CR 24 even further ahead of schedule, and CR 19 even further behind schedule.

This example was a good teaching tool for the rest of the shifts as well. It clearly shows how following a "back-to-front" methodology does not balance the line, and the advantages of using the decision support tool to make goaling decisions.

Table 5 – Priority Ranking to illustrate how pull drives different behavior

Rank	Critical Resource	Sched Score	Balance Index
1	CR 7	-20.9	-0.8
2	CR 8	-17.5	-0.6
3	CR 9	-15.2	-0.6
4	CR 11	-14.6	-0.6
5	CR 10	-13.6	0.3
6	CR 13	-12.8	-1.0
7	CR 6	-12.4	0.9
8	CR 12	-12.3	-0.2
9	CR 14	-10.0	-0.9
10	CR 15	-5.9	-0.7
11	CR 17	-5.9	-0.5
12	CR 19	-4.0	-0.7
13	CR 16	-3.5	0.2
14	CR 3	-3.5	-0.2
15	CR 1	-2.8	-0.2
16	CR 21	2.1	-0.7
17	CR 5	-3.5	2.4
18	CR 25	-1.7	0.4
19	CR 26	-0.9	0.4
20	CR 2	-0.5	0.6
21	CR 20	0.4	-0.4
22	CR 18	0.9	0.5
23	CR 4	-1.8	1.6
24	CR 24	2.8	0.4
25	CR 22	4.4	-0.3
26	CR 23	6.6	1.0

5.3 Feedback and changes to experiment

It was important to spend a lot of time observing the execution of the goaling and reflecting with the captains and area coordinators on the progress of the experiment. I had daily meetings with the operations managers to get feedback and gauge execution. These reflection sessions provided important feedback that allowed us to make changes to improve the experiment and served to help educate manufacturing on experiment principles and details.

5.3.1 Initial priority calculations not driving correct actions

The priority assignments were initially placing too much emphasis on the value for the Balance Index. If an operation was feeding into an empty part of the line, it was considered to be a high priority even if it had a very low IPQ. Conversely, if an operation were feeding into a full part of the line, it was considered to be low priority, even it had a very high IPQ. In response to this issue, we decided to adopt the concept of a priority matrix from Leachman et al. in which both Schedule Score (driven by IPQ) and Balance Index would jointly play a role in determining priority. The decision support tool was updated to reflect this change after the third day of the experiment.

5.3.2 Added granularity to schedule score when determining priority

When reviewing the priority output from one of the shifts, it was noted that some operations that were extremely behind schedule were less of a priority than operations that were only moderately behind schedule. Looking at the priority matrix described in section 5.1.3.4, we can see that if the Balance Index is "OK" (neither far behind schedule nor ahead of schedule), the highest priority bucket that the operation can be assigned to is "2". It was decided that additional granularity should be added to the degree of lateness communicated by the schedule score. If an operation met our threshold of being very far behind schedule, its priority would be bumped up more than if it was less behind schedule.

5.3.3 Searching for high priority resources in web-based goaling tool was cumbersome

We received feedback that scrolling through the web-based goaling tool to search for the high priority operations was time consuming. A user-friendly priority sheet was added to the decision support tool to facilitate easy identification of the top 5 feeder areas. Once the top 5 feeder areas were identified, the captains could go directly to those sections of the line in the web-based tool and goal the feeders.

5.3.4 Compliance difficult to monitor

Monitoring compliance was a very manual and time consuming process. I would perform spot checks comparing the output of the priority sheet to what the CoS actually goaled. When there were obvious differences, I would meet with the Captain to understand why. Sometimes it was due to tool availability in the fab, other times we got good feedback to improve the decision-support tool, and other times the CoS did not use the decision-support tool appropriately. Not using or interpreting the decision-support tool correctly was mainly an issue at the beginning of our experiment. The mX WIP management team had initially asked that a shift champion oversee the goaling of the line and coach the other operations managers. However, the role of captain rotated even more often than usual on some shifts. Much time was invested in meeting with all of the OM's across all of this shifts early on to ensure complete understanding. After the first week of experimentation, compliance was much better.

5.3.5 Improved maintenance of fab-out schedule

Section 4.3.7 discussed the fact that maintenance of the fab-out schedule was inconsistent with forecasted demand. Over-shipments in the current week did not reduce the fab-out schedule for following weeks. Additionally, the ship schedule for the current week would be adjusted to "move the carrot" if manufacturing was in position to hit the ship goal for the week early. Both of these issues served to keep the backend urgency artificially inflated and made it difficult to get away from the back-to-front mentality.

All of the operations managers I met with agreed that this did not make sense, but stated that is always the way it had been done at Fab23. I met with all of the Shift Managers to discuss this, and we agreed that this was driving the wrong behavior. About one week into the experiment the procedure for maintaining the fab-out schedule was updated to give credit for over-ships to following week's ship schedules. The following week the shift managers agreed that the fab-out schedule would not be modified upwards, even if enough WIP was in position to exceed the current goal. This did not mean that the fab would stop shipping wafers once a goal was hit, but it did mean that the backend would be less of a priority. The ship goal for any week would only be adjusted if there were truly a demand-driven reason to do so.

This page is intentionally left blank

6. Conclusions and Recommendations

6.1 Conclusions

6.1.1 Externalities affected experiment performance

Unfortunately, line balance deteriorated during the experiment and gains were not made in key metrics such as cycle time. Regardless of the fact that four weeks is not sufficient for observing the true trend in performance, there were several externalities that inhibited performance: tool availability issues, increased maintenance work in the fab, and the shutting down of tools to prepare for a holiday warm-down towards the end of the experiment.

6.1.1.1 Tool Availability Issues

At several times over the course of the experiment, tools either went hard down or their capacity was significantly limited due to quality issues. This resulted in over 30% of the total WIP in the fab being held at one operation several times. No matter how good a fab's WIP management system is, with poor tool availability it will be impossible to achieve a balanced flow through the line. The decision-support tool developed does suggest priorities to facilitate getting back to a balanced line, but without tools available to execute to the priorities, balance cannot be achieved.

6.1.1.2 Maintenance Work

During the time of planning and implementation, the fab was preparing to bring on a new process technology. This required that new tools be installed and others be qualified for the new process. The impacts of preparing the fab for the new process made it impossible at times for the Captain to follow the priorities suggested by the decision-support tool. Tool availability issues are compounded further by these impacts.

6.1.1.3 Preparation for Warm-down

The fab entered a warm-down period for the holidays, during which production was halted and maintenance crews performed major PM's. Preparation for this warm-down had a significant impact on constraining WIP movement towards the end of the experiment. Before beginning the warm-down, wafers had to be staged in "safe" zones of the line. Safe zones are essentially operations where there are less quality concerns associated with letting wafers sit for extended periods of time without being processed.

6.1.1.4 Other potential externalities in fab performance

The Competitive Semiconductor Manufacturing Program at the University of California at Berkeley studied about 30 fabs in the U.S., Japan, Korea, and Taiwan in the 1990's [20]. The charter of the study was to "measure manufacturing performance and identify underlying determinants of performance in the semiconductor industry." The variation in performance observed was independent of the technologies employed by each fab.

Some of the key metrics measured in the study were line yield, defect density, cycle time per layer, and labor productivity. The study found that fabs with the following characteristics performed worse in these key metrics than the other fabs in the study:

- Small fabs – metrics were poor in the areas of equipment and labor productivity for fabs with wafer starts less than 6,000 per week
- Fabs that are operated with "skeleton crews" or that often shut down production have poor performance.
- Fabs that are not "fully loaded" perform relatively worse.

The differences in fab performance were independent of technological differences. The criteria describing "bad fabs" was exaggerated in Fab 23 during the experiment [20]. Fab 23's wafer loading dropped significantly in the weeks preceding the experiment, and the fab continued to be under-loaded during the experiment. Also, production interruptions increased significantly due to maintenance work and tool qualifications for the new process technology. These are potential externalities that could have compromised fab performance.

6.1.2 Benefits of Decision-Support Tool

A decision-support tool that provides a baseline for priority recommendations is helpful in gaining consistency across shifts. As captains became familiar with the proposed methodology and execution of experiment, the difference in goaling styles decreased. Even captains who leaned heavily towards "back-to-front" priority setting prior to the experiment, now were presented with a different framework for looking at the line and saw the logic in setting priorities in a different way. General rules-of-thumb can be helpful, but they leave too much room for interpretation that can lead to very different results depending on who is making decisions.

Additionally, even if a tool provides more intelligent priorities, users will not be committed if the reasons for the suggested priorities are not clear. The visual representation of inventory distribution and line balance provided by the tool (illustrated by Figures 19 – 22) proved valuable in helping decision-makers understand what the methodology was attempting to accomplish. Captains could easily see where in the line production was ahead of schedule and behind schedule and why "back-to-front" is not always optimal.

6.1.3 Benefits of "Pull" methodology

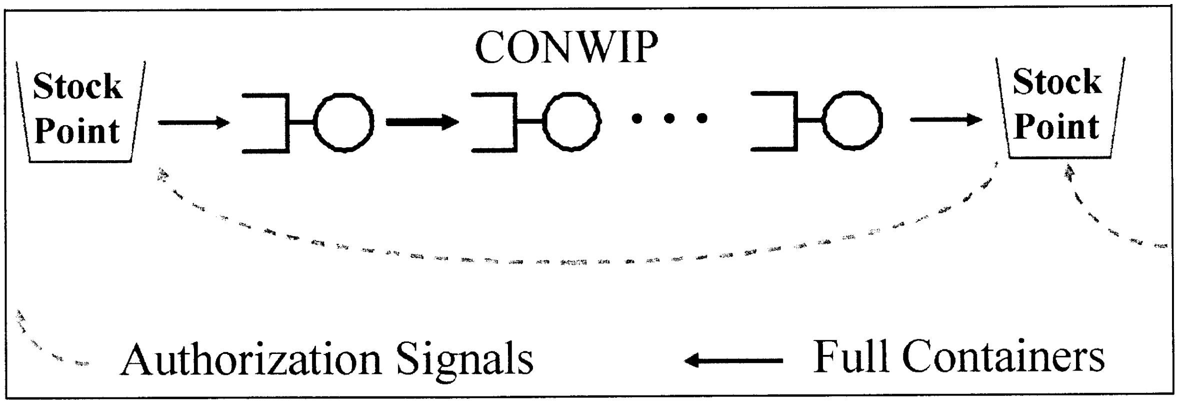
Despite the fact that external factors prevented achieving balance, anecdotal evidence suggests that decision-support tool drove better decisions. Sections 5.2.2 and 5.2.3 presented examples of decisions suggested by decision-support tool and how they were different from what many captains would have done independent of tool. It was agreed that the decision-support tool provided better direction, but without the tool the direction was not obvious.

Hopp and Spearman characterize a pull system as one that authorizes work release based on system status, and that by authorizing releases it helps put a cap on the level of WIP [16]. Reduced WIP then leads to lower cycle times. The methodology discussed in this thesis is not a strict pull system by any means, as wafer starts are continually pushed into the fab and the "authorization signals" communicated by the IPQ are suggestions rather than strict requirements. However, the proposed system does have some aspects of a pull methodology.

By considering downstream needs (IPQ and BI) at points throughout the line (CR's), the proposed methodology is attempting to authorize work through each CR based on system status. One can imagine that looking downstream from each CR to the end of the line, the target inventory represents a CONWIP inventory goal. The difference between actual inventory and this CONWIP goal results in upstream operations being "authorized" to release work via the CR's IPQ. See Figure 23 (adapted from "Factory Physics" textbook by Hopp & Spearman) for a depiction of a basic CONWIP pull system[19]. The IPQ is only a suggestion, as there are many potential issues that could make the IPQ infeasible such as lack of available WIP to process or machine availability.

However, by avoiding prioritizing production into areas of the line that have excess inventory (judged via the BI), wafers should spend less time in queue. And by prioritizing operations that have downstream WIP deficits (judged via IPQ), we reduce the chances of starving available machines. WIP is being sent where it is needed and can be processed. This should serve to increase throughput and decrease cycle time. Additionally, having lower queue times enables more rapid quality feedback [16].

Figure 23 – CONWIP schematic [19]



The system attempts to prioritize WIP so that the right number of wafers are at the right point in the line at the right time. If expanded to accommodate prioritization of different products rather than just gross wafers, this should lead to less expediting of behind-schedule products. It is also self-correcting in that if one shift overshoots its IPQ for a particular operation, the next shift will see that same operation as having a low (or negative) IPQ and low priority.

6.1.4 Keys to gaining support for change

Engaging the people affected by the process change in the development of the tools and methodology prior to implementation was imperative to gaining buy-in. Acknowledging that the proposed system is not a panacea for WIP management problems, but pointing out all of the areas in which it is an improvement over the current system provided a strong case for experimentation (see Chapter 3 for a discussion of the weaknesses in status quo that were addressed). An effort was made to engage the Captains as key players in the effort to fine tune the methodology and metrics used, as well as develop a user-friendly interface.

Building a network and identifying key stakeholders was important in getting traction. It is inevitable that a few of the affected individuals will continue to resist new methods, however if other key stakeholders can be turned into proponents, a tipping point can be reached where even the dissenters are willing to experiment.

Additionally, the active support of senior management and their advocacy of mX experimentation were key in overcoming inertia and enlisting support from the organization. The cross-functional WIP Management team was given significant visibility to senior management. This increased the attractiveness of being a team-member and was motivation to set ambitious goals.

6.1.5 Drive to experiment results in engagement and valuable feedback

The intricacies of the process are complex enough that the team could have easily taken 6 months to a year to plan. However, experimentation was the best way to get real engagement and valuable feedback. Rather than getting stuck in a seemingly endless "PLAN" phase of the PDCA loop (Plan, Do, Correct, Act), it is important at some point to "DO" something. The fact is that even if every aspect is planned to the nth detail, the reality of implementation will change these details.

Anyone executing such a change as our team did should know that the system will not work exactly as planned. It is important to be prepared to take feedback and spend significant time working with the people affected. This is especially important at the beginning of any implementation. Disciplined reflection with key stakeholders is imperative to creating a system people will use. The team made several important modifications to the experiment execution and interface early in implementation.

Because of the time invested in reflecting with the users, a tool was developed that people want to continue using. The best feedback received from the captains towards the end of the experiment was, "Why would we stop using this after 4 weeks?". In fact, management made the decision to continue using the tool even after bringing the fab back on-line following the warm-down.

6.2 Recommendations

6.2.1 Improve Tool Availability

Tool Availability should be a major initiative. Having a good WIP Management policy alone is not sufficient for improving fab performance. As stated previously, various tool issues throughout the experiment severely constrained WIP movement. Several times during execution, more than one fourth of total inventory in the fab was held up at single operations. An mX team should be initiated to focus on Total Productive Maintenance and tool availability.

6.2.2 Apply methodology to accommodate multiple product types

The principles espoused in this thesis can be applied to allow prioritization of different product types. Just as a fab-out schedule for gross wafers was applied to build a target inventory profile throughout the line for gross wafers, fab-out schedules by product could be used to build target inventory profiles for each specific product.

Comparing target inventory to actual inventory would allow for product-specific IPQ's and BI's. Although IPQ's for high-volume products would typically be higher than IPQ's for low-volume products, the Schedule Score would account for volume in its calculation (Fab drumbeat in denominator is representative of volume). Thus, a low-volume product with an IPQ of 500 would have a more negative schedule score than a high-volume product with an IPQ of 500. Assuming similar scores for BI, the low-volume product would be a higher priority.

However, the current web-based goaling interface and the tool-level interface do not allow for clear communication of production goals and priorities by product. IT resources would be needed to update existing tools or create new tools to facilitate extending the methodology to specific products.

Additionally, the concept of getting away from static inventory goals and recalculating goals every shift becomes even more important if the methodology is extended to discrete products (section 4.3.4 describes problems with static inventory goals). Even if ship goals for the foreseeable future remain constant in gross wafer terms, chances are that volumes for each product will vary from week to week.

6.2.3 Modify BI calculation to weight micro-segments according to CT

A potential weakness of the Balance Index calculation proposed (see Equation 4) is that the 70%, 20%, 10% weightings to downstream micro-segments are fixed. A more intelligent approach may be to modify the weightings of downstream micro-segments based on what the target cycle time of each micro-segment is.

Potentially, the longer the target cycle time of a micro-segment, the more concerned we would be about having it close to empty. For example, it would be more concerning to have an empty micro-segment that is 4 days long, than an empty micro-segment that is 1 day long. The weighting can be scaled up or down from the 70, 20, 10 baseline according to the relative differences in cycle time among the next 3 downstream layers. If the next 3 downstream layers were all the same cycle times, the weighting would remain 70, 20, 10. But if the next downstream layer was very short, and the 2nd and 3rd downstream layers were longer, the weighting could be closer to 50, 35, 15.

Example of new Balance Index calculation is below.

Equation 5 – Balance Index with scaled weightings

$$\boxed{\mathrm{B.I.}_{\mathrm{CR(j)}} = 70\% \cdot m_{j+1,j} \cdot bi_{\mathrm{CR}(j+1)} + 20\% \cdot m_{j+2,j} \cdot bi_{\mathrm{CR}(j+2)} + 10\% \cdot m_{j+3,j} \cdot bi_{\mathrm{CR}(j+3)}}$$

Where $bi_{\mathrm{CR}(j)} = \dfrac{\text{Actual WIP}_{\mathrm{MS(j)}} - \text{Target WIP}_{\mathrm{MS(j)}}}{\text{Target WIP}_{\mathrm{MS(j)}}}$ *and MS = Micro-segment*

And multiplier m: $m_{i,j} = \dfrac{CT_{MS(i)}}{70\% \cdot CT_{MS(j+1)} + 20\% \cdot CT_{MS(j+2)} + 10\% \cdot CT_{MS(j+3)}}$

A gain factor could even be applied to each micro-segment depending on the sensitivity desired to difference in downstream target cycle times. Control theory provides methods for calculating these gains. This is an area for the mX WIP Management team to explore as the groundwork provided is built upon.

6.2.4 Account for available inventory and capacity to provide feasible goals

As discussed earlier, the Ideal Production Quantities calculated are not necessarily feasible goals. The value represents the number of wafers that downstream operations want to "pull" out of the Critical Resource. There may not be enough inventory at the CR to satisfy the IPQ or there may not be enough available capacity to meet the IPQ request.

A potential improvement would be to take available inventory and capacity into account and provide feasible goals for each operation. Available inventory for processing in one shift would account for both inventory currently at an operation and upstream inventory that will advance to the tool in time to be processed that shift. Target cycle times can be used to make this determination.

Capacity considerations are much more complex. Most operations can be performed on several tools, and those tools can perform several operations. Additionally, some operations are subject to a lot-to-lens rule, which instructs that certain critical lithography operations have to be run on the same tool that processed a previous lithography operation. Taking all of the restrictions and qualifications into account is critical in setting operational goals and assigning capacity. Available capacity for an operation depends on how much capacity is allocated to other operations that share the same tool. Kim, Yea, and Kim propose several approaches (a Mixed Integer Programming Model, relaxed Linear Programming model, and a heuristic) for determining the optimal capacity allocation and goals with an objective to minimize the WIP imbalance [18].

Section 6.1.2 discusses how the decision-support tool developed served to reduce the variability across shifts in how the Captains assigned priorities and set goals. However, because the priorities calculated do not take into account if there is WIP or capacity available at operations throughout the line to execute to the priorities, the Captain still has discretion on how the suggested priorities are turned into feasible goals. Adding the functionality described in this section would further decrease the variability across this shifts by providing optimal, feasible goals from the outset.

6.2.5 Implement CONWIP system to modulate wafer starts

Finally, the full benefits of a "pull" system will not be realized until the introduction of WIP into the manufacturing process is controlled via "pull" rather than "push". Several derivatives of CONWIP exist for implementation in a fab. A standard CONWIP policy would serve to keep the total inventory in the fab at a certain limit. When a lot ships, a lot would be started. Little's Law shows that if inventory in the system is controlled, cycle time will be controlled as well.

However, several studies have shown that a standard CONWIP policy is less flexible and performs worse than some derivatives [21, 22]. A CONLOAD system would seek to maintain bottleneck utilization at a constant level [22]. Another variation of CONWIP would aim to keep a target WIP level in the first several layers of the process. Having an inventory limit on the beginning layers would prevent wafers from being started when there is a great excess of WIP at the beginning of the process. For example, a standard CONWIP system may want to start 500 wafers after 500 wafers shipped, however if there is a large WIP excess at the beginning of the line, a target inventory around the first few layers would prevent more starts. Modulating wafer starts based on fab status is another area that should be explored by the mX WIP management team.

References

[1] "Flash memory." *Wikipedia*. 19 Mar 2005. Wikipedia.org. <http://en.wikipedia.org/wiki/Flash_memory>

[2] LeClaire, Jennifer. "Intel Flashes New Memory Product Plans." *TechNewsWorld*. 03 Mar 2005. TechNewsWorld.com. <http://www.technewsworld.com/story/tech/intel-flash-memory-41068.html>

[3] Beekman, Scott, Brian Kumagai. "Choosing flash memory." *Electronic Products*. April 2004. ElectronicProducts.com. <http://www.electronicproducts.com/print.asp?ArticleUrl=toshiba.apr2004.html>

[4] Deffree, Suzanne. "Intel Loses Top Flash Position." *Electronic News*. 21 Nov 2003. ElectronicNews.com. <http://www.reed-electronics.com/electronicnews/article/CA337427.html>

[5] Kanellos, Michael. "Intel climbs flash rankings as market dips." *CNET News*. 16 Nov 2004. CNETnews.com <http://news.com.com/Intel+climbs+flash+rankings+as+market+dips/2100-1004_3-5455080.html>

[6] Dipert, Brian. "Flash forward to the future." *EDN*. 01 Nov 2004. EDN.com. <http://www.edn.com/article/CA474720.html?industryid=3170>

[7] Kanellos, Michael. "Samsung raises flash memory bar." *CNET News*. 10 Nov 2004. CNETnews.com. <http://news.com.com/Samsung+raises+flash+memory+bar/2110-1041_3-5446939.html>

[8] "NAND revenues to pass NOR in '05." *Electronics Weekly*. 02 Feb 2005. ElectronicsWeekly.com. <http://www.electronicsweekly.co.uk/articles/article.asp?liArticleID=38698>

[9] "Memory Reloaded." *Ferret*. 5 Mar 2004. Ferret.com.au. <http://www.ferret.com.au/articles/2a/0c01da2a.asp>

[10] 2000 Intel Environmental, Health, and Safety Performance Report. Intel website. Accessed 18 Sep 2004. Intel.com. <http://www.intel.com/intel/other/ehs/pdf/ehs_2000.pdf>

[11] "Intel in Your Community: Colorado." Intel website. 29 Mar 2005. Intel.com. <http://www.intel.com/community/colorado/aboutsite.htm>

[12] Spears, Steven and Kent H. Bowen. "Decoding the DNA of the Toyota Production System." *Harvard Business Review*. September-October, 1999. pp. 96-106.

[13] Flinchbaugh, Jamie. “Beyond Lean: Building Sustainable Business and People Success through New Ways of Thinking.” Lean Learning Center website. Accessed 04 Jun 2004. <http://www.leanlearningcenter.com/Beyond%20Lean.pdf>

[14] Wildeman, Roy. “An Application of Lean Principles within a Semiconductor Manufacturing Environment.” Thesis, Massachusetts Institute of Technology, 2005.

[15] “‘Copy Exactly’ Factory Strategy.” Intel website. 16 Mar 2005. Intel.com. <http://www.intel.com/pressroom/kits/manufacturing/copy_exactly_bkgrnd.htm>

[16] Hopp, Wallace J, Mark L. Spearman. “To Pull or Not to Pull: What is the Question?” *Manufacturing & Service Operations Management.* Vol. 6, No. 2, Spring 2004. pp. 133-148.

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

[18] Kim, Sooyoung, Seung-Hee Yea, Bokang Kim. “Stepper Scheduling in Semiconductor Wafer Fabrication Process.” *The Proceedings of the International Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM 2000)*. pp. 157-162

[19] Hopp, Wallace J, Mark L. Spearman. *Factory Physics: The Foundations of Manufacturing Management.* Chicago: Richard D. Irwin, 1996.

[20] Leachman, Robert C., David A. Hodges. “Benchmarking Semiconductor Manufacturing, A Research Program at the University of California at Berkeley.” *Competitive Semiconductor Manufacturing Program.* <http://esrc.berkeley.edu/csm/Csmlong497.pdf>

[21] Kim, Sooyoung, Young Hoon Lee, Jongkwan Park. “Experimental study on input and bottleneck scheduling for a semiconductor fabrication line.” *IIE Transactions.* Vol. 34, No. 2, February 2002, pp. 179-190.

[22] Rose, Oliver. “Conload – a new lot release rule for semiconductor wafer fabs.” *Proceedings of the 1999 Winter Simulation Conference.*