Master's Thesis

Operational Models for Evaluating the Impact of Process Changes on Cluster Tool Performance

by Niranjan Chandrasekaran
Advisor: Jeffery Herrmann

M.S. 99-4

ISR
INSTITUTE FOR SYSTEMS RESEARCH

ISR develops, applies and teaches advanced methodologies of design and analysis to solve complex, hierarchical, heterogeneous and dynamic problems of engineering technology and systems for industry and government.

ISR is a permanent institute of the University of Maryland, within the Glenn L. Martin Institute of Technology/A. James Clark School of Engineering. It is a National Science Foundation Engineering Research Center.

Web site http://www.isr.umd.edu

ABSTRACT

Title of Thesis: OPERATIONAL MODELS FOR EVALUATING THE IMPACT OF PROCESS CHANGES ON CLUSTER TOOL PERFORMANCE

Degree candidate: Niranjan Chandrasekaran

Degree and year: Master of Science, 1999

Thesis directed by: Assistant Professor Jeffrey W. Herrmann
Department of Mechanical Engineering

This thesis describes operational models that integrate process models to expedite process change decisions for cluster tool performance improvement. The process engineer attempting a process change needs to wait for the industrial engineer to approve the change after making sure it does not degrade cluster tool performance. Having a model that integrates process parameters into the operational model of the tool helps the process engineer quantify the impact of process changes on tool performance. This makes the process change decision faster.

Two integrated models for understanding cluster tool behavior have been developed here. One is a network model that evaluates the total time needed to process a lot of wafers for a given sequence of activities involved in the process. Including a manufacturing process model (in the form of a Response Surface Model) gives an integrated network model that relates the total lot processing time to process parameters like temperature and pressure and to process times.

The second model developed is an integrated simulation model that can be used when the sequence of wafer moves is not given but is determined by a scheduling rule. The model can be used to quantify the impact of changes to process parameters and product characteristics like deposition thickness on total lot processing time.

The thesis contains examples that illustrate the types of insights that one can gain into cluster tool behavior from using these integrated models.

OPERATIONAL MODELS FOR EVALUATING THE IMPACT OF PROCESS CHANGES ON CLUSTER TOOL PERFORMANCE

by

Niranjan Chandrasekaran

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland at College park in partial fulfillment
Of the requirements for the degree of
Master of Science
1999

Advisory Committee:

Professor Jeffrey W. Herrmann
Professor Shapour Azarm
Professor Gary W. Ruboff

DEDICATION

To my parents and sisters

ACKNOWLEDGMENTS

I am very grateful to my advisor Dr. Jeffrey Herrmann, for providing the impetus, the means and the support that enabled me to work on this project. I deeply appreciate his concern and the appreciation and encouragement he provided to enable me to become a professional. He has been an exemplary mentor and I thank him for his guidance and insight in my project. I especially appreciate the way he treats his group members as his peers.

I also thank Dr. Rubloff for his insight in my project at our weekly meetings. I would like to thank the members of my group Brian and Rock for the helpful discussions and comments. I would like to specially thank Priya and Sundar for always being there and my friends JR, Rman, Pappu, K2, Gogo, Guru, Sasi, Anita and Sangeeta who have made my life in UMCP rich, lively and interesting.

Finally I would like to thank my family for all their support. I thank my parents for always encouraging me to work hard and follow my dreams.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Motivation

This research is aimed at increasing the operational efficiency of semiconductor manufacturing processes by integrating process level models with operational methods. Previous research on operational methods in semiconductor manufacturing focused on improving production planning and scheduling by applying algorithms, with parameters like processing times as fixed inputs. But the motivation for this research is to also consider the process level interactions in the associated operational models.

This work addresses the gap that exists between the process engineer and the operations personnel in a semiconductor manufacturing fab. A process engineer attempting process changes for process improvement often does not know how the changes affect the manufacturing system performance. This requires many iterations of talking to the operations personnel before a balance is struck between the process change considered and its effect on system performance. So it would help to consider the process level interactions in the operational models of the system.

This work considers cluster tools as the manufacturing systems and provides ways to study the effect that process time and process parameter changes have on cluster tool performance. These methods help to eliminate the iterative process mentioned above by helping the process engineer analyze directly the effect of process changes on system performance.

Cluster tools are highly integrated machines that can perform a sequence of semiconductor manufacturing processes. Cluster tools reduce operator intervention and reduce queuing and cycle time.

Cluster tools are often used for the processes that create interconnects – the vertical plug structures that conduct signals between horizontal wiring planes. Interconnection technology is becoming the dominant technology challenge in advanced semiconductor manufacturing. The importance of specific sectors in the interconnection (or back-end-of-line) portion of the fab results first from the fact that products must traverse interconnect sectors several times, as the technology moves to six levels of wiring.

Also, the interconnection investment must be made in the fab toward the end of the process sequence, so that a very high price --- the time and effort invested in earlier process steps --- must be paid for any errors. Accordingly, a disproportionate share of the manufacturing cost for semiconductors is attributed to interconnection technology.

Tungsten is a popular metal for interconnect structures. The tungsten plug process consists of a contact-cleaning step, a sputtered Ti film, a TiN layer, a rapid thermal anneal (RTA), many layers of tungsten (W) deposited by the Chemical Vapour Deposition (CVD) process.

1.2 Research Objective

The objective of this research is to study the effect of process parameter changes on cluster tool performance. This work considers Applied Materials cluster tools like the AMAT Centura, and the processes of TiN barrier layer deposition and chemical vapor deposition of tungsten (CVD W). The performance metric of interest is the total lot processing time, which is the time the cluster tool requires to process an entire lot of wafers. This metric, called the makespan, is needed for throughput and cycle time calculation. The importance of this metric is explained in Chapter 3.

To model the processes, this thesis uses response surface models (RSMs), which are widely used in semiconductor manufacturing to relate process metrics like deposition rate to equipment and process variables like temperature and pressure. These RSMs can be continuously improved and adapted with process experience, process or equipment change, or expanded knowledge base.

1.3 Models and Insights

This thesis will describe two models that quantify the effect of process changes on makespan.

The first model is an integrated network model that evaluates the lot makespan for a given sequence of wafer moves. By integrating an RSM, the model expresses the lot makespan as a function of process parameters.

The second model is an integrated simulation model for the Centura cluster tool. This model integrates the RSM with existing cluster tool simulation software. This model is appropriate when the tool uses a dispatching rule to sequence wafer moves.

Both models allow one to quantify the sensitivity of lot makespan with respect to process times and process parameters. This helps process engineers understand how process changes affect the tool performance. Although each cluster tool configuration is different, our results provide some basic insights, as mentioned below:

1. The effect of process changes is not uniform. Sometimes the change in process results in a large change in makespan whereas sometimes the change is not as much as we expect. So knowing the sensitivity of makespan with process parameters gives an idea of the region in which the process change can be attempted.
2. From critical path analysis, we can find out which operations are more critical than others. Changing those operation times affect the makespan the most.
3. The rules that cluster tool controller uses are sensitive to operation times. It is possible that due to a process change, a rule generates a sequence that is detrimental to throughput performance. The results here give such insights.
4. The possibility of changing tool configurations without affecting performance in the event of technology shifts can be analyzed by methods outlined in this work.

The methodologies mentioned above can be applied to other cluster tools. With these results, process engineers can develop better processes, equipment purchasers can make better procurement decisions, and fab managers can improve their fab’s performance.

1.4 Outline of Thesis Report

Chapter 2 gives a brief explanation of the background literature concerning the tungsten plug process, cluster tool architecture, cluster tool performance evaluation metrics, and response surface methods.

The integrated network model for analyzing the makespan of a cluster tool operating on a fixed sequence of wafer moves is explained in Chapter 3. The RSM used in the network model and the simulation model is also given in Chapter 3. Chapter 4 details the integrated simulation model incorporating the RSM for deposition rate in the CTPS software.

Chapter 5 summarizes the work and lists the conclusions reached from this work. It also gives suggestions for future extension of this work.

CHAPTER 2

BACKGROUND LITERATURE REVIEW

2.1 Semiconductor Manufacturing Overview

The manufacture of integrated circuits (ICs) starts with the silicon wafer preparation. Then a series of steps like thermal oxidation, thin film deposition, lithography, etching, ion implantation and contact and interconnect development are carried out on the wafer [2]. Finally the wafer is diced into identical chips, which form the ICs, by drawing lines on the wafer surface.

Wafers are prepared by grinding the silicon ingot to the desired diameter and then cutting it into single slices of thickness 0.5-1 mm, depending on the wafer diameter. The wafers are then lapped, polished, and cleaned to remove the damage caused by slicing. Thermal oxidation of silicon to silicon dioxide by subjecting the wafer to oxygen or water vapor at elevated temperature is carried out at different stages of an integrated process technology. The grown film is used to properly terminate silicon bonds at the silicon surface, to isolate conductors and semiconductors, or to provide a high-quality dielectric for precision capacitors. The controlled deposition of thin organic and inorganic films is an important step in the manufacture of ICs. These films are deposited to remain an inherent part of the device structure, or to constitute intermediate layers that are used for particular processing steps and then removed. Two main methods for thin film deposition are chemical vapor deposition (CVD) and physical vapor deposition (PVD).

Another important step is the lithography step where a pattern is delineated in a layer of material sensitive to photons, electrons or ions. Lithography transforms complex circuit diagrams into patterns which are defined on the wafer in a succession of exposure and processing steps to form a number of superimposed layers of insulator, conductor and semiconductor materials. Typically 8-25 lithography steps and several hundred processing steps between exposures are required to fabricate a packaged semiconductor IC.

Clean and etch processes are used to selectively remove organic and inorganic materials from patterned and unpatterned substrate surfaces. When a process is used to remove particulates or unwanted films, it is called cleaning. When a film which was intentionally deposited or grown is removed, the process is called etching.

The electrical properties of semiconductor crystals can be modified predictably by introducing controllable amounts of dopant impurities into substitutional sites of the crystal. The most commonly used methods to introduce impurities into a semiconductor are doping during crystal growth, ion implantation and diffusion.

2.2 The Tungsten Plug Process

The tungsten plug provides the contact from the outside world to the underlying areas of silicon, in particular, from the metal layers defining digit lines and periphery circuits to highly doped underlying silicon nodes.

The manufacturing process for the plug uses a contact-cleaning step, a sputtered Ti film, a TiN layer, a rapid thermal anneal (RTA), and a CVD W layer [7]. Some important considerations in the formation of a tungsten-filled contact process are:

1. Dopant levels, which are required for low-resistance ohmic contacts, and the interaction of these dopant species with the formation of silicide metallurgical interface to silicon.
2. Cleans and their impact on contact quality
3. Contact layer as the metallurgical interface between tungsten and n^+ and p^+ silicon.
4. Barrier layer, which offers protection against the aggressive nature of the WF_6 contact fill reactant gas species.
5. Ti/TiN anneal as the thermal annealing step required to complete the metallurgical interface and barrier defence capabilities.
6. Tungsten fill, which covers the W fill process itself.
7. Tungsten etchback, which is the last step in the W plug formation.

The focus of this work is the tungsten fill step. Tungsten is generally deposited by chemical reactions involving WF_6, $SiH_{4,}$ H_2 and Ar and the process is called Chemical Vapor Deposition or CVD. It is usually a multistep process. The first deposit, called the seed layer, uses Silane (SiH_4) reduction of WF_6. Silane is preferred for seed layer because it protects the substrate Si or oxide layer from flourine attack in places where TiN barrier material is absent. The subsequent tungsten layer can be deposited with H_2 reduction because of the presence of the protective seed layer. Hydrogen is preferred over Silane, even though Silane reduction is faster, because hydrogen reduction is more conformal. This criterion becomes crucial when the aspect ratio (Depth/Width) of the plug is high.

Cluster tools are the most widely used manufacturing systems for carrying the tungsten plug step. Following sections discuss literature on cluster tools**.**

2.3 Cluster Tool Definition and Architecture

According to SEMI E10-96 standard, a cluster tool is defined as “A manufacturing system made up of integrated processing modules mechanically linked together (the modules may or may not come from the same supplier).”

Apart from the process chambers, the tool has one or more loadlocks. A loadlock is used mainly for loading and unloading a cassette of wafers. Sometimes a second loadlock may be used to hold work-in-process when there is waiting between chambers. Wafers are transported between two modules or between a module and a loadlock by a robot wafer handler. The robot wafer handler can be either be a single blade robot or a dual blade robot. The single blade robot can carry only one wafer at a time. The dual-blade robot consists of two blades pointing in opposite directions each capable of carrying a wafer. The two blades are tightly coupled by construction i.e., at any time only one of the two blades can pickup or drop off a wafer to or from a chamber. Rarely, we come across wafer handlers like the ones in Novellus Concept II , which index all wafers from their stations to the next simultaneously.

A cluster tool is in steady state when the flow of wafers in and out of the tool is the same. It is said to be in a transient state during the initiating and terminating of the processing of a batch of wafers. In cluster tools with multiple loadlocks, the cluster tool can be in steady state for long periods of time by having different loadlocks be alternatively ready for processing. This can be achieved by readying (i.e., pumping down to chamber pressures, etc.) one of the loadlocks to send wafers into the cluster tool, while wafers in the other loadlock is currently being processed.

Cluster tool processing may be differentiated by “sequential” vs. “parallel” processing. Examples of these are illustrated in Figure 1. Sequential processing requires that each module is functional for the tool to be functional. Usually the serial modules are all different, each representing a step in the process sequence. Conversely a parallel processing system can potentially be operated while one or more of the modules is not available. This is because all the modules are identical and perform the same function. Another variation on the cluster tool theme utilizes a “multi-visit” flow – closely related to standard sequential processing. In practice, a cluster tool is generally a combination of these extreme systems.

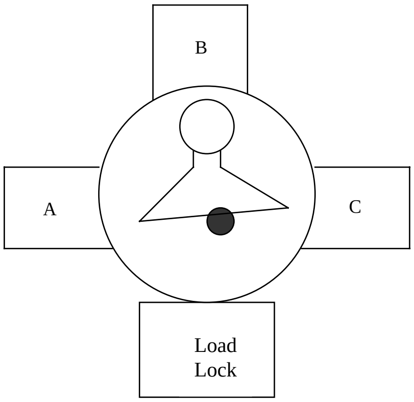


Figure 1(a). Serial Tool.

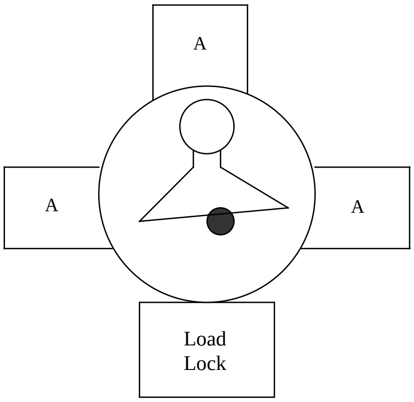


Figure 1(b). Parallel Tool.

Integration of single wafer modules to form a cluster tool leads to a dramatic increase in productivity. This is because operator intervention is kept minimal. Cycle times are thus reduced. Also there is lesser chance of contamination and this improves yield. Another benefit of cluster tool is that it allows for parallel processing of wafers at bottleneck process steps. Yet planning for such tools is still very local, still very much dependent on the specific processes at one point in the process sequence. The integration of such tools into an optimized processing line has not been realized because the dynamic nature of the tool's capacity and cycle time has not been appreciated.

2.4 Work on cluster tool performance metrics

Most of the research in the field of operations research applied to cluster tools has been aimed at defining metrics for cluster tool performance, and developing models and methods for tracking cluster tool performance.

Throughput has been one of the common measures that have been used to understand the behavior of cluster tools. Throughput is defined as the processing rate of the cluster tool, usually expressed in wafers per hour. A primary concern with cluster tools is the maximum throughput that they can achieve. Perkinson *et al.*(1994) have developed an analytical model that predicts the minimum theoretical time required to complete a series of processes in a cluster tool. Their deterministic model attempts to predict the throughput of a generic single-arm-N-chamber cluster tool. The throughput is expressed as the inverse of what they define as fundamental period (FP) – the time between subsequent completed wafers arriving at the loadlock. In their work, they mention three ways for throughput enhancement. Two of the techniques, lookahead algorithms and multispeed transporters increase the net transport speed by modifying the action of the transporter when it is not moving a wafer. The third technique is to incorporate dual load locks in the tool to minimize the length of the transient phase (beginning and end of processing a lot).

Following Perkinson's work, Srilakshmi *et al.*(1997) have derived analytical expressions for the throughput of a cluster tool with dual blade robot. They make the distinction between transport-bound and process-bound regions in the sequence of events. In the transport-bound region, an increase in throughput can be achieved only by decreasing the robot transfer times. Change in process times, so long as they are within some bounds, will not affect the overall throughput of the cluster tool. In the process-bound region, changes in process times and transport times will affect the throughput of cluster tool. In the transport bound region, by doubling the speed of transfer when the robot is not transferring a wafer, a single-blade robot can achieve better performance than a dual blade robot. In the process bound region, the dual blade robot performs better than a single blade robot.

Another line of research on cluster tool configuration looks into the effect of redundant chambers on tool performance. Redundant chambers are extra chambers than what would be required for the given processing times. They reduce the overall chamber utilization. Perkinson *et al.*(1996) continued their research on cluster tools by analyzing the effect of redundant chambers and revisitation sequences on throughput. The analytical models indicate that redundant chambers and revisiting schedules increase the performance-cost ratio of the tools. Redundant chambers are a good idea when one chamber’s processing time is significantly longer than the other clusters on the tool. Revisiting faster chambers can provide an economical method of maximizing tool utilization. But there is a trade-off to be made with the increased transporter movement.

Cycle time is another important performance index for a cluster tool. Lead times and service level are both affected by cycle time. Cycle time, the amount of time a product or its components spends as WIP, is especially crucial in the semiconductor industry since product life cycles are so short. Service level, the probability of on-time delivery, greatly impacts customer satisfaction and is significantly influenced by lead time and cycle time variability. Reduced cycle time also reduces WIP according to Little’s Law and hence there are savings on inventory.

Meyersdorf *et al*.(1997) discuss the importance of cycle time reduction and put down the characteristic curves between Cycle Time vs. Utilization and also Cycle Time vs. Capacity and WIP. They also prescribe a five-step cycle time reduction methodology:

1. Data collection
2. Quantitative analysis
3. Root cause analysis
4. Findings and Recommendations
5. Improvement Roadmap

Wood [25], in his paper, evaluates the impact of extensive integrated single-wafer processing on wafer production cost and cycle time in commercial fabs running contemporary process flows. He uses analytical models for minimum cycle time and waiting time. Formulae for calculating minimum wafer cost, fixed cost and total wafer cost are prescribed. Trade-off between Cost and Time is discussed.

While all the above mentioned work deal with a single cluster tool, Lopez and Wood [8] have studied the problem of configuration in a system of multiple cluster tools. Two systems are considered – serial (S) and parallel (P). In the S system, we have each tool performing the entire n-step process sequence and we have n such tools. In a P system, we have n tools and one tool is dedicated to perform one step in the n-step sequence. In both systems, each tool has n chambers. They derive analytical expressions for cycle time, throughput and WIP in terms of the various tool parameters like move times, process times, number of chambers etc. They conclude that for reliable tools, as process times of different steps become more similar and short compared to load/unload times, the serial configuration delivers lower cycle times than the parallel configuration, with the former able to attain throughput capacity at a smaller lot size than the latter.

2.5 Higher Level Performance Measures for wafer fabs – OEE and COO

Performance at the tool level can be measured by indices like cycle time, throughput etc. At the factory level, we need measures that address the entire system of tools. Two such measures are Cost of Ownership (COO) and Overall Equipment Effectiveness (OEE). Actually, OEE can be argued to be a subset of COO.

Semiconductor manufacturers are trying to sustain 25-30% annual gains in cost productivity; the industry is capital-intensive, and it wants to make existing equipment more efficient. For example, improving two workstations' OEEs from 40% to 60% is like getting a free workstation. A manufacturing tool's OEE reflects how well the production facility is using that piece of equipment. The OEE of any manufacturing tool is the product of four factors: OEE = Availability x Operating efficiency x Rate efficiency x Rate of Quality [13,17].

Availability is the percentage of time that a workstation is available for processing product. Breakdowns, setups, production qualifications, repairs and preventive maintenance comprise of unavailability time.

Operating efficiency is the percentage of available time during which the equipment is active. It reflects the equipment's idle time, which could result from no product, no operator, lunch breaks or production team meetings.

Rate efficiency is the ratio of the equipment's actual rate to its theoretical rate. Equipment wearouts and partial loads on the tool are reasons for a lesser actual rate.

Rate of quality is the percentage of end items that are good or acceptable. Some items have to be scrapped because of contamination during the manufacturing process and some may require to be reworked. The rate of quality accounts for these losses.

Even though it is desirable to have the OEE of every piece of machinery as high as possible, it would be a mistake to aim for 100% OEE for all tools. This would lead to wastage of resources and suboptimization. This is because the operation/tool with the least capacity – the bottleneck operation/tool - limits the manufacturing plant's output. Trying to achieve 100% OEE with every tool by running them at full load is equivalent to trying to exceed the bottleneck capacity and it only results in building up of piles of inventory at the bottleneck. So it is important that the limited resources are directed towards achieving the highest possible OEE at the bottleneck.

OEE deals with the productivity of the plant while COO includes cost factors in addition to all the factors that go into OEE. COO is the full cost of embedding, operating and decommissioning a processing system in a factory environment i.e., the life-cycle cost of owning a semiconductor tool [5,20]. An explanation of the various inputs to COO is as follows:

- Fixed costs: Cost of equipment, land etc. i.e., the capital costs
- Operating costs: Costs for using the equipment like labor, repair, utility, overheads
- Throughput: Number of wafers per hour the process system delivers to the factory
- Composite yield: The operating yield of the tool including breakage, defects
- Utilization: The ratio of production time compared to total available time
- Yield loss cost: A measure of the value of wafers and die scrapped through operational losses and defects

Equipment purchase decisions is one of the most important applications of COO for an IC manufacturer. The operating costs of installed equipment can also benefit from COO analysis. For tool suppliers, responding to a request for price quotation (RFQ) is often the first. The IC manufacturer either collects specific input to his own COO model, or the supplier provides a complete model. Equipment suppliers and end-users often cooperate on COO analyses.

Implementing COO for new purchases is an important application; however, using COO and OEE to evaluate the long-term benefit of manufacturing modifications may be the biggest value to the organization. This can be done through equipment benchmarking, comparative and bottleneck analysis, project prioritization and process optimization.

2.6 Response Surface Methodology

Response surface models are widely used in semiconductor manufacturing when process parameters have to be tuned for process improvement with respect to a specific response [3,14].

There exist methods for designing experiments and analysing data that enhance the exploration of a region of design variables in one or more responses. The underlying assumption for all the models is that a response Y is a function of a set of design variables X_1, X_2, $X_3 \ldots X_k$ and that the function can be approximated in some region of the X's by a polynomial model. Two commonly used polynomial models are:

- First-order model: $Y = A_0 + A_1X_1 + A\ _2X_2 + \ldots + A_kX_k$

- Second-order model: $Y = A_0 + \Sigma A_I X_I + \Sigma A_{ii} X_i^2 + \Sigma\Sigma A_{ij} X_i X_j$

The motivation for response surface analysis can be to solve the problem of planning and analyzing experiments in search for desirable conditions on a set of controllable (or design) variables that give rise to *optimal* response. But apart from finding the *optimal* response, RSMs can be used to identify regions where there is demonstrated improvement in response over that achieved by current operating conditions.

Four steps in a standard response surface analysis are:

1. Perform a statistically designed experiment
2. Estimate the coefficients in the response surface equation
3. Check on the adequacy of the equation
4. Study the response surface in the region of interest

There are commercially available software packages like ECHIP and SAS, which help in the design of experiments, given the range of design variables, by giving the combinations or levels of design variables to be used in the experiment and the total number of experiments in the form of a design table. The user can then go back and conduct real experiments. But nowadays dynamic simulation packages are available for many semiconductor processes since the real experiments are costly. Readymade experimental data can also be obtained from literature. Once the experimental data are ready, the software determines the coefficients and also plots the response surfaces.

The fourth step is to use the response surface obtained from the design of experiments for process improvement. Interesting regions in the response surface are identified to try out process changes. Thus RSMs help in narrowing down on the region where process changes need to be explored and hence save a lot of time and money spent on process improvement.

2.7 Summary

The manufacture of ICs is a complicated process involving more than a hundred steps. The quality of the tungsten plug step is crucial because the resistivity of the plug controls signal transmission between the horizontal wiring planes. And CVD process carried out in cluster tools is the universal choice for tungsten deposition.

There has been extensive research on the performance of cluster tools with respect to throughput and cycle time. Analytical expressions for throughput of different cluster tool configurations given the different operation times have been derived. The importance of cycle time has been understood and different tradeoffs between cycle time and utilization, capacity and WIP have been studied.

Higher level performance metrics like OEE and COO which deal with multiple cluster tools in a wafer fab, their cost and performance implications act as guidelines for improvements at fab level. They also aid equipment purchase decisions.

But little work has been done in understanding the impact of process parameter changes targeted at process improvement on operational performance of a cluster tool. And this work is aimed at understanding this interaction. Process models like RSMs have been integrated with operational models of a tool like discrete event simulations or network models explained in later chapters. These integrated models provide quantitative insights into the effect of process changes on cluster tool performance.

CHAPTER 3

INTEGRATED NETWORK MODEL

3.1 Need for this model

Most of the previous research on cluster tool performance has addressed throughput of the tool with fixed processing times for the chambers. And usually, the process times for all the chambers have been assumed to be equal. This work moves beyond the previous work by considering the effect of process time changes, due to process parameter changes, on cluster tool performance. The performance measure chosen here is the total lot processing time alternatively called as makespan (MS).

The makespan of the tool is important because the maximum throughput of the tool is inversely related to makespan. Also, if the tool is part of the routing through which a lot of wafers has to undergo processing, then the makespan of the tool directly contributes to the cycle time of the lot. And cycle time affects lead time, service level, WIP and yield.

The total lot processing time can be expressed as a function of the process times for the different steps in the manufacturing sequence. And the process times are a function of the process parameters like temperature and pressure. Thus modifying process parameters changes the lot processing time. Thus, process engineers need to be careful when tuning a process. However, in some cases, changing process times will not change the lot processing time much. The network model explained here helps to quantify the effect of process parameter changes or process time changes on lot makespan.

3.2 The Network representation

The analytical model developed here quantifies how process changes affect the lot makespan for a given sequence of activities. The advantage of this approach is that any kind of cluster tool with complex combinations of serial and parallel modules can be studied provided the sequence of wafer moves is known, which usually is the case. This model also integrates an RSM for the process steps so the process times for the steps are expressed as a function of process parameters. This makes it possible to study the effect of process parameter changes on makespan.

Processing a lot of wafers in a cluster tool involves a number of tasks. The operator loads the lot of wafers, in a cassette, into the loadlock. The wafer handler then moves the wafers from the loadlock to the chambers and also between chambers. The wafer handler can move a wafer into a chamber only when it is empty. The sequence of activities is represented by the sequence of wafer moves. A cluster tool controller controls the wafer handler movements. If the controller is intelligent, it can sequence the wafer moves dynamically, based on the state of the system and then the wafer moves are called anticipatory. Such controllers are not as widely used in the industry as the ones that follow a prescribed sequence.

Given the sequence, the cluster tool behavior can be modeled using a network. A network is a collection of nodes and directed arcs. Every node represents an activity. For example, the processing of the first wafer in the first chamber is an activity in the sequence that can be represented by one node in the network. The precedence constraints in the sequence are represented in the network by the directed arcs. The precedence constraints are obtained in two ways. First, for a given wafer, the order in which it should undergo the different processes for getting converted into the desired product lays down some constraints. Second, the order in which chambers get ready to accept new wafers and the availability of the wafer handler to move wafers also lays down additional constraints. For the purpose of drawing our network, the given sequence of activities contains the precedence constraints.

Figure 2 shows an example of a network for a tool processing a lot of two wafers. Each wafer has to undergo two processes – Orientation and degassing (OD), and CVD W Deposition. There is one chamber in the tool for OD and one for CVD W. Figure 2(a) shows the Gantt Chart or Timing diagram for the wafer handler and each chamber of the tool. Figure 2(b) shows the corresponding network.

In Figure 2, each node of the network carries the time that the corresponding activity requires. Definitions of some terms used in association with the network are:

1. Path: Any sequence of nodes connected by arcs starting from the first node and ending at the last node in the network is called a *path* through the network.

2. Length: The sum of the times of all the nodes in the path is called the *length* of the path.
3. Critical path: The longest path through the network is called the *critical path*. There can be more than one critical path for a network.

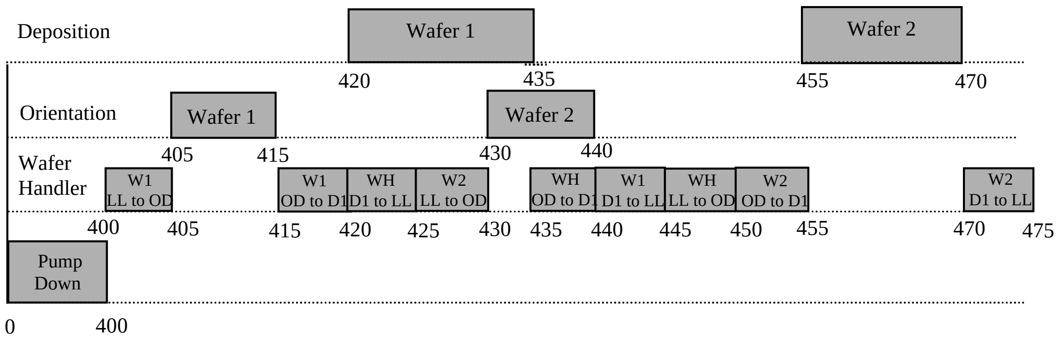


Figure 2(a). Gantt chart for two wafers.

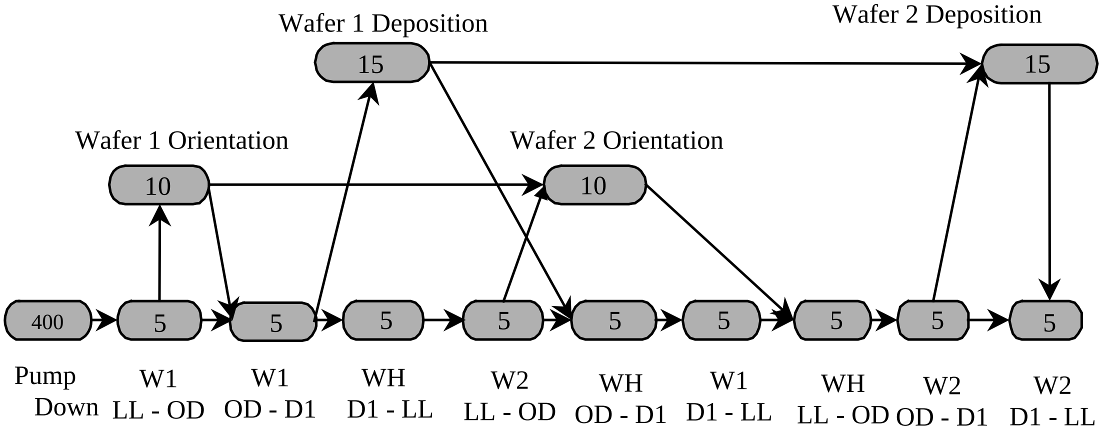


Figure 2(b). Corresponding Network.

Figure 3 shows the critical path for the network of Figure 2. Section 3.6 outlines an algorithm that can be followed to determine the length of the critical path of a network.

The length of the critical path equals the total time needed to process the lot of two wafers. This is the total lot processing time, which is referred to as *makespan.* The reciprocal of makespan is an upper bound on tool throughput.

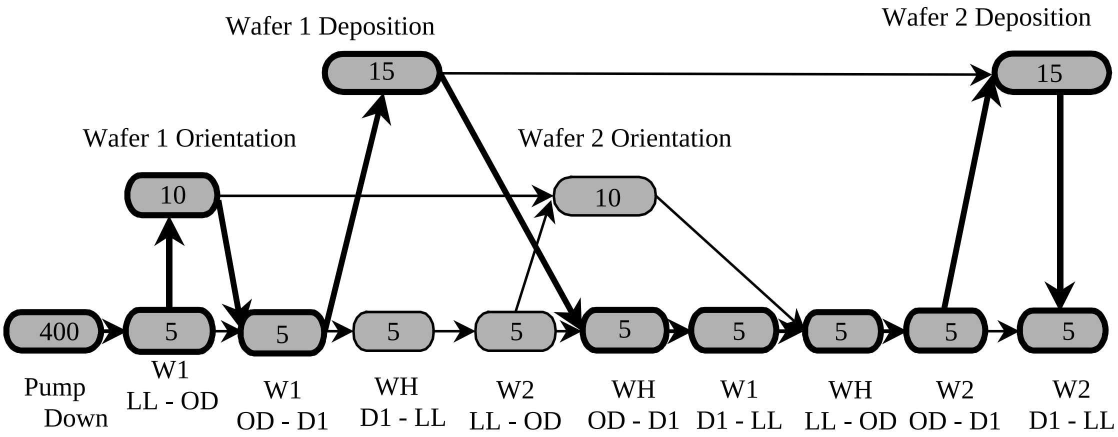


Figure 3. Critical Path for Network.

3.3 Process Time Changes

The network model can be used to study the effect of process time changes on lot makespan, thus providing some initial insight on process change considerations. Since the lot makespan is the length of the critical path, increasing the time of the activity (process) on the critical path will increase the length of the critical path and hence the makespan. In general, a small increase to an activity or node not on the critical path will not alter the critical path and hence the lot makespan is unaffected. Whereas a large change to time of the same node may alter the critical path to include that node.

The concepts mentioned above are clearly explained by the example network of Figure 4. Figure 4(a) presents the network and its critical path. Figure 4(b) illustrates an increase in the time of an activity that is on the critical path. Figure 4(c) illustrates a small increase in the time of an activity that is not on the critical path. Figure 4(d) illustrates a large increase in the time of that same activity, which creates a new critical path.

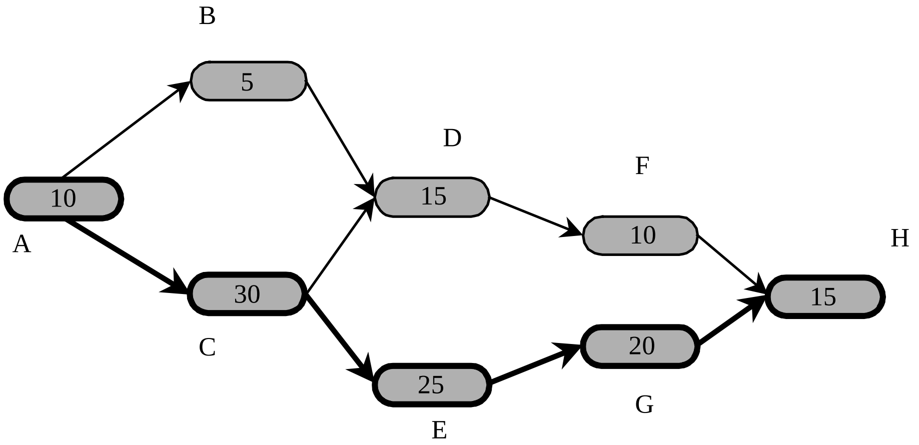


Figure 4(a). A Network with a Critical Path of Length 100.

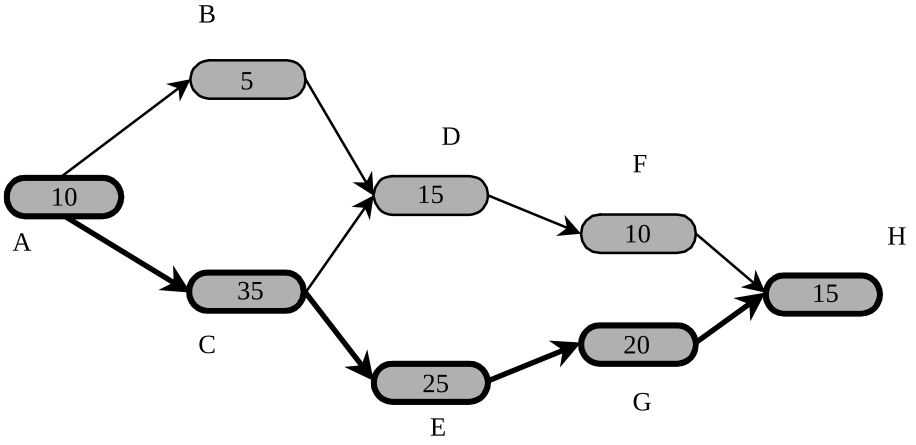


Figure 4(b). The Network after Increasing Activity C by 5.
The critical path is now 105.

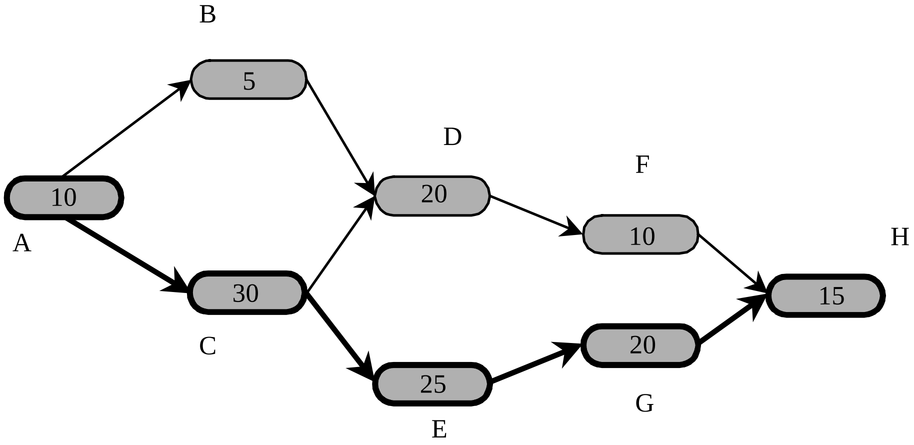


Figure 4(c). The Network after Increasing Activity D by 5.

The critical path remains 100.

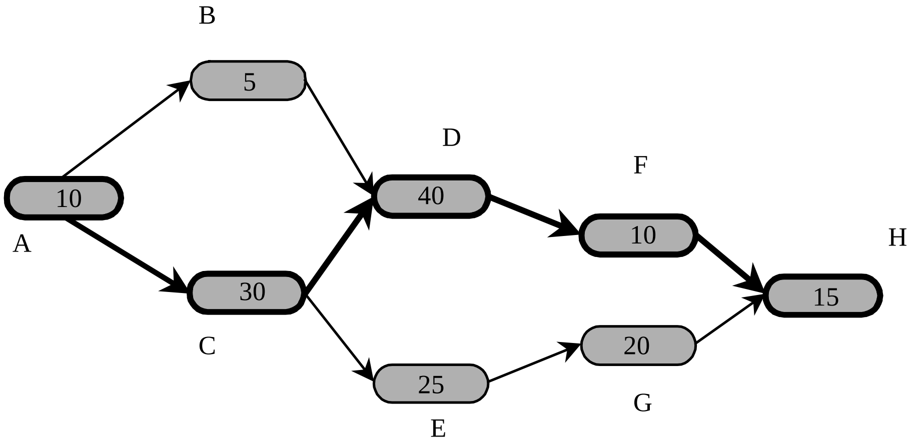


Figure 4(d). The Network after Increasing Activity D by 25.

The critical path is now 105.

We can apply the above concepts of critical path to a network representing cluster tool processing like the one shown in Figure 2(b). In the network, the nodes represent activities like deposition at the chamber, wafer handling etc. These activities keep repeating throughout the sequence for the entire lot because the same set of activities are carried out for each wafer. The term "operation" is used to describe such activities that occur multiple times. If there are m different operations, and operation k requires p_k time units, then $t_j = p_k$ if j is an operation of type k. Thus we can write makespan MS = $f(p_1,\ldots p_m)$. The lot makespan is a function of the operation times.

Consider a particular operation in the sequence. The more often this operation occurs on a critical path, the more it will affect the total critical path length and thus increase the lot makespan. If it occurs only once on the critical path, increasing its time by t time units will increase the lot makespan by t time units.

Consider, for example, the network in Figure 5(a). The deposition process time equals 7 seconds, and the critical path includes the deposition process for the second wafer. Increasing the process time by one second increases the lot makespan by one second. The other instance of this operation also requires more time, but it is not on the critical path. If the increase is large enough, the other activity may change the critical path, making it even longer (unless the sequence of activities changes). Consider the network in Figure 5(b). In this case, the deposition process time has increased to 15 seconds. This changes the critical path, which now includes both deposition processes.

Increasing the deposition process time by one second now increases the lot makespan by two seconds.

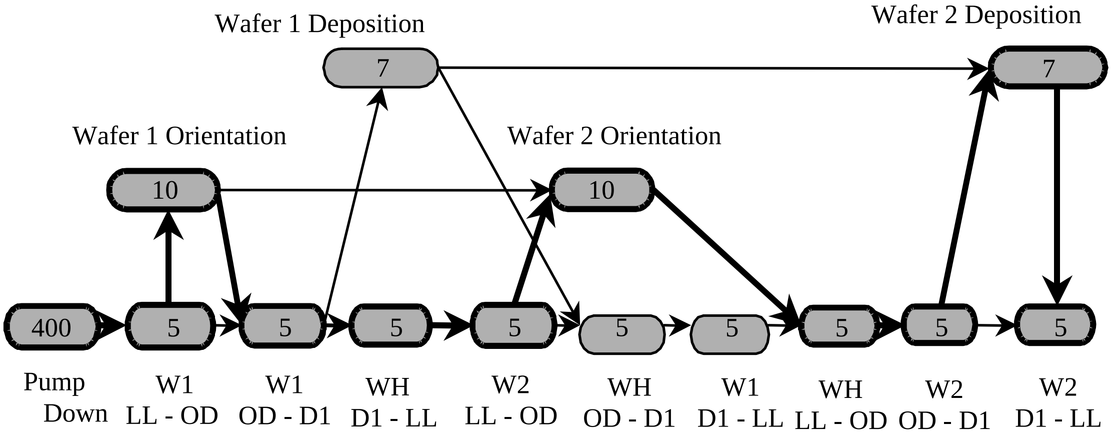


Figure 5(a). The critical path includes one deposition process.

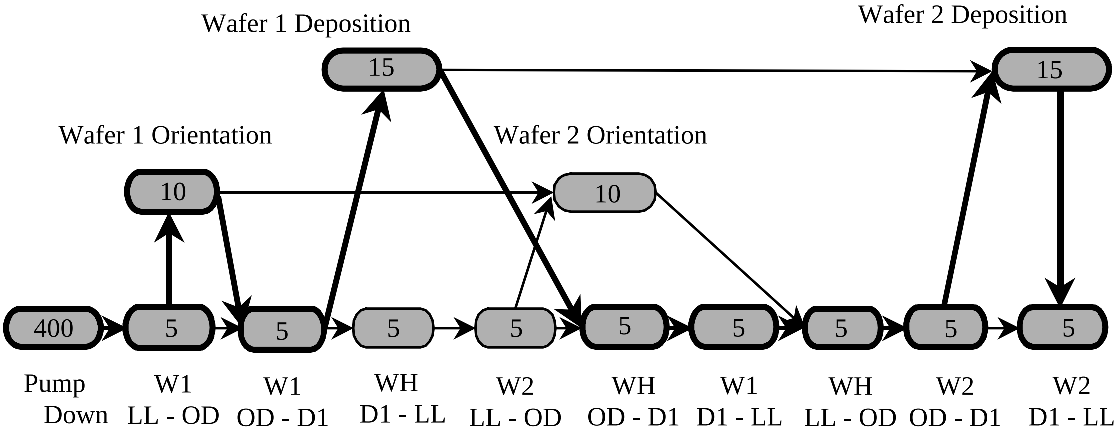


Figure 5(b). The critical path includes both deposition processes.

On increasing the time required for an operation, the operation features more often on the critical path. So a unit increase in operation time results in an amplified increase in makespan. Moreover, more than one operation occurs on the critical path. It is not possible to pinpoint a single operation as “bottleneck” because increasing the time of any operation will increase the total processing time. But we can say that one operation is “more critical” than another if the first occurs more often on the critical path than the second.

So the network model provides an efficient way of *visualizing* the sequence and identifying the processes which would make the largest impact on the lot makespan and hence throughput. It would be much more difficult to identify the same processes by just having the sequence written down.

3.4 Application of the methodology to a specific cluster tool example

The methodology of using the network model and critical path analysis to study lot makespan was applied to a specific cluster tool. The process time (time required) for an operation was varied and the impact on makespan was examined.

The cluster tool under consideration was the AMAT Centura cluster tool used for CVD W deposition. The processes involved in the sequence were Orientation/Degassing (OD) and CVD W. The lot size was fixed at 20 wafers and each wafer had to undergo both OD and CVD W. The tool had three chambers. The first was used for OD. The second and third were used as parallel chambers for CVD W and a wafer could go to any of the two depending on availability. The tool had a single loadlock and a single blade robot. Figure 6 illustrates the AMAT Centura tool.

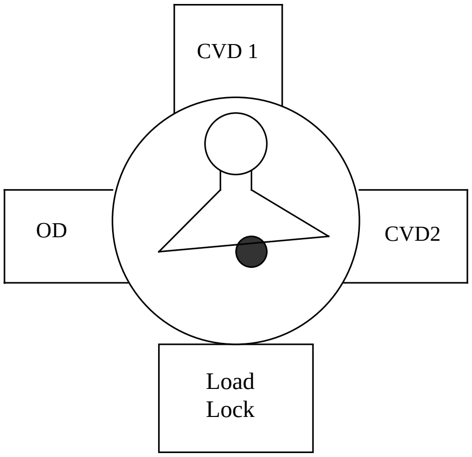


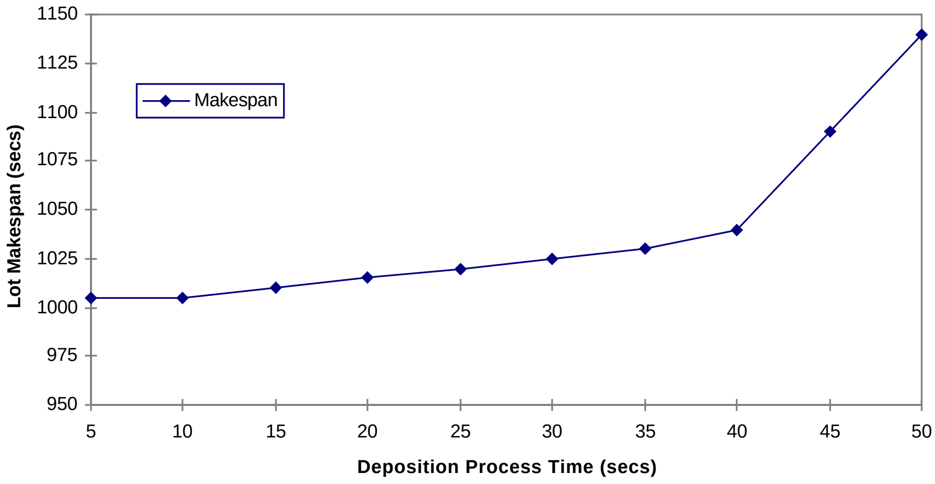
Figure 6. The AMAT Centura cluster tool.

In our example, the loadlock pump down time before the cassette of 20 wafers can be loaded into it is 400 seconds. The robot takes 5 seconds to move empty or loaded between the loadlock and any chamber or between any two chambers. And the time required for the OD operation is 10 seconds. All these times were kept fixed.

A network was constructed for the sequence of activities (given in Appendix) selected for processing the entire lot. To study the effect of deposition process time (D) on the lot makespan (MS), the network was analyzed for critical path and its length, which is MS.

Using the procedure outlined later in this chapter for finding critical path of a network, the critical path of this network was derived. The makespan is the sum of processing times of all the nodes along the critical path. The derivation of the critical path, along with the network, is included in the appendix. It gives us an analytical expression MS in terms of D i.e., MS = 640 + max{D,10} + max{D,35} + 8max{D,40}. Table 1 summarizes the observations in different ranges of D. Figure 7 illustrates the relationship. Figure 7. Lot makespan versus deposition process time.

(Note that the makespan axis does not start at zero.)



The variation of MS with D was validated by experimentation using the CTPS simulation software. The analysis was carried out by varying D starting from 1 second and increasing it upto 40 seconds and a few values above 40 seconds. It was observed that the makespan had different sensitivities to increase in D in different ranges and that the trend was constant after 40 seconds. So the variation of D was stopped after a few values above 40 seconds. The observation of variation of MS with D as seen from

CTPS was in perfect agreement with the derivation of MS as a function of D from the network.

Deposition Process Time D (seconds)	Lot Makespan (MS)	Sensitivity dMS/dD	Lot Makespan Range(seconds)
0 ≤ D ≤ 10	MS = 1005	0	MS = 1005
10 ≤ D ≤ 35	MS = 1005+ (D - 10)	1	1005 ≤ MS ≤ 1030
35 ≤ D ≤ 40	MS = 1030 +2 (D – 35)	2	1030 ≤ MS ≤ 1040
D > 40	MS = 1040 + 10 (D - 40)	10	MS > 1040

Table 1. Relationship between lot makespan and deposition process time.

The sensitivity of lot makespan to deposition process time is obtained by taking the derivative of MS to D, dMS/dD. The sensitivity shows that in the region of D between 1 and 10 seconds, the makespan is not affected by an increase in D. This is because the critical path does not include D at all. Similarly, the effect of D on makespan is maximum when D is 40 seconds or more. This is because as D is increased slowly, it keeps altering the critical path to include one deposition process, then two processes and finally ten deposition processes. Hence after 40 seconds, every one-second increase in D increases the makespan (MS) by 10 seconds.

Likewise, reducing an operation’s process time will have a more significant impact at first. As the process time continues to decrease, the operation occurs less often on the critical path, so the reductions don’t have the same benefit.

Also, note that these results depend upon the times chosen for the other operations and upon the sequence of wafer moves. Different process times or sequences yield different results.

3.5 Integrating Process Models

As explained in Chapter 2, a process engineer uses design of experiments and response surface methodology to relate process metric like process rate, yield etc. to process parameters like temperature and pressure. By doing so, the engineer can attempt process parameter changes to meet the process performance goals. However, if the process engineer is not careful, though the process performance may be improved, the effect on the manufacturing system performance may be undesirable. For example, one significant impact of changing process parameters is a change to the process time. If a cluster tool performs the process, the process parameter changes may affect the lot makespan, which affects the tool throughput. Although a higher rate should improve throughput by decreasing the nominal process time, evaluating this impact is not simple. As we saw in the previous section, a small change to the process time sometimes changes the lot makespan drastically. However, sometimes it does not. “Process improvements” that significantly lower a cluster tool’s throughput (especially if that tool is a bottleneck tool) can seriously degrade manufacturing system performance by increasing cycle time and decreasing maximum throughput.

Thus, it would be convenient for the process engineer to have a model that integrates process parameters into manufacturing system performance. In the cluster tool example presented earlier, an RSM for W CVD has been included in the network model for the deposition step to obtain an integrated network model. The RSM for the CVD W process was based on data collected by Stefani *et al*. (1996). The RSM has the following four process paramaters: reactor pressure, deposition temperature, the mole fraction of WF_6, and the mole fraction of H_2. The response is the average deposition rate in Angstroms per second (A/Sec). The CVD W deposition process is a H_2 reduction of WF_6, preceded by a short silane and WF_6 nucleation step that deposits a 400 A seed layer.

The RSM used here relates the actual deposition rate (DR) to the reactor pressure P in torr, and the deposition temperature T in Kelvin. The mole fractions were set to their median values because they did not affect the deposition rate as much as pressure and temperature. The RSM is as follows:

$$DR = 63.55 + 0.4248(P-80) - 364.6\left(\frac{1000}{T} - 1.346\right) - 2.079(P-80)\left(\frac{1000}{T} - 1.346\right)$$
$$+1.297x10^{-4}(P-80)^2 + 945.5\left(\frac{1000}{T} - 1.346\right)^2$$

Having the deposition rate DR (A/Sec), we can calculate the deposition process time D if we know the deposition thickness Th. Thus D = Th / DR(P,T). We can construct the integrated network model by using the above expression for D in our earlier network model for MS and thus expressing the makespan as a function of process parameters temperature and pressure.

We have MS = h(*Th*, *P*, *T*) = 640 + max {*Th*/*DR*(*P*, *T*), 10} + max {*Th*/*DR*(*P*, *T*), 35} + 8 max {*Th*/*DR*(*P*, *T*), 40}. If we increase temperature and pressure, then the deposition rate would increase thus reducing deposition time. So the makespan reduces. The integrated model has been used to quantify an example using a deposition thickness of 3000A. Figure 8 illustrates the results.

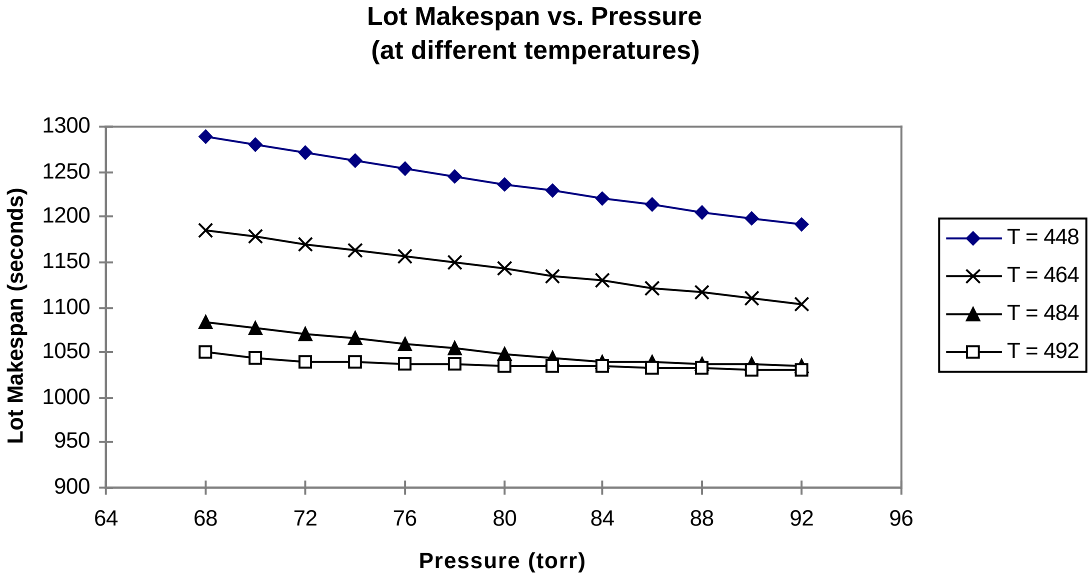


Figure 8. Lot makespan versus pressure at different temperatures
(note that the makespan axis does not start at zero).

At lower temperatures, the deposition rate is low and deposition time is high. So the operation occurs more often on the critical path. So the impact of temperature and pressure change is large. At higher temperatures, the deposition time is low and the operation occurs less often on the critical path. So the decrease in makespan with increase in pressure and temperature is low.

Apart from knowing the absolute effect of process parameter changes on lot makespan, we can also use the integrated model to evaluate the sensitivity of makespan change with process parameters. This would help the process engineer to know which region of process parameter values he should target for maximum change in throughput.

This is done by finding the partial derivatives of makespan with respect to temperature and pressure. We first differentiate the deposition process time D with respect to pressure and temperature as follows:

$$\frac{\partial D}{\partial T} \quad = \quad \frac{-3x10^{6}}{(DRxT)^{2}}\left[2743 + 2.079P - \frac{1.9x10^{6}}{T}\right]$$

$$\frac{\partial D}{\partial P} \quad = \quad \frac{-3000}{DR^{2}}\left[3.2 + 2.6x10^{-4}P - \frac{2080}{T}\right]$$

We know the sensitivity of MS with respect to D as the derivative dMS/dD shown in Table 1. If we multiply the above terms by this sensitivity, we get the partial derivatives of MS with respect to P and T.

Figure 9 graphs the derivative of lot makespan with respect to pressure (dMS/dP) as pressure and temperature change. The more negative derivative shows that at lower temperatures and pressures, the lot makespan is more sensitive to changes in pressure, which is what we observed previously.

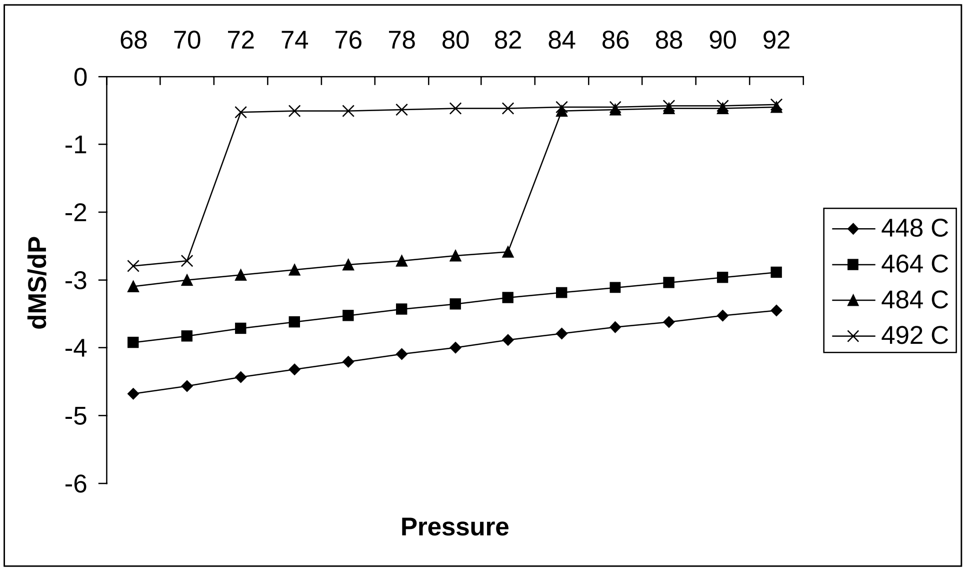


Figure 9. Derivative of lot makespan with respect to pressure.

We can also study the impact of technology shifts in the manufacturing processes on manufacturing system performance using the integrated network model. The National Technology Roadmap for Semiconductors gives semiconductor manufacturers guidelines on where the technology is heading in different processes. The roadmap gives the thickness of tungsten to be deposited in the via and contact layers. As the technology shifts to nodes with lower dimensions, the gate widths decrease and the interconnect diameter decreases, and hence the required tungsten deposition thickness also decreases. For example, in the 250 nm technology node, the tungsten thickness in via 3 and via 4 are 3750 A whereas at the 130 nm technology node, the thickness reduces to 1896 A, a straight reduction of around 50% in the required deposition thickness. The immediate question that comes to mind is whether the drastic reduction in deposition thickness will lead to drastic savings in makespan. Probably we may even be able to do away with some deposition chambers because of this makespan reduction. The network model can be used to quantitatively analyze such situations.

In the network model, the lot makespan is a function of the deposition thickness. If the process parameters remain the same, the reduction in deposition thickness decreases deposition time and hence makespan. Table 2 shows the results of the analysis using the model and the different deposition thicknessses as technology shifts. The most interesting result is that even if deposition thickness reduces by 50%, the reduction in makespan is only around 16%. This is because in this region, specifically under the selected process conditions (median values), the sensitivity dMS/dD =1 and hence the lot makespan is not very sensitive to deposition thickness.

Table 2. The impact of deposition thickness on lot makespan.
(Under a fixed sequence)

Technology Node (nm), Layer	Dep. thickness Th (A)	Deposition process time D (seconds)	Percent reduction in process time	Lot makespan MS (seconds)	Percent Reduction in lot makespan
250, Contact	2100	33	-	1028	-
180, Contact	1500	24	27	1019	0.9
150, Contact	1275	20	39	1015	1.3
130, Contact	1050	17	48	1012	1.6
250, Via 1-2	2700	42	-	1060	-
180, Via 1-2	1950	31	35	1026	3.2
150, Via 1-2	1575	25	40	1020	3.8
130, Via 1-2	1350	21	50	1016	4.2
250, Via 3-4	3750	59	-	1230	-
180, Via 3-4	2625	41	31	1050	15
150, Via 3-4	2188	34	42	1029	16
130, Via 3-4	1896	30	49	1025	17
250, Via 5	7500	118	-	1820	-
180, Via 5	5250	83	30	1470	19
150, Via 5	4375	69	42	1330	27
130, Via 5	3792	60	49	1240	32
150, Via 6	5250	83	-	1470	-
130, Via 6	4550	72	13	1360	7.5

3.6 Algorithm for critical path evaluation

We have discussed the importance of the network model and critical path analysis in the evaluation of cluster tool performance. It is appropriate to outline an algorithm for finding the critical path of a network. This is required when the network gets large and complicated. For simple networks, the critical path can be obtained by inspection.

The following algorithm will identify the critical path and calculate its length. Given a network N = (V, A), where V = {v_1, ..., v_n} is the set of nodes and A is the set of directed arcs {(j, k)}. Each node v_k represents a distinct activity in processing the lot. Arc (j, k) is in A if activity j must precede activity k. Associated with each node v_k is the time t_k that the activity requires. Let P(v_k) be the set of immediate predecessors of node v_k. That is, v_j is in P(v_k) if and only if (j, k) is in A. Similarly, let S(v_k) be the set of immediate successors of node v_k. That is, v_j is in S(v_k) if and only if (k, j) is in A.

Then, calculate the earliest completion time C_k of each node v_k as follows:

C_k = t_k if P(v_k) is empty. Otherwise C_k = max{C_j: j is in P(v_k)} + t_k. Repeat for all v_k in V.

Let T = max{C_k: v_k is in V}. Calculate the latest completion time D_k of each node v_k as follows:

D_k = T if S (v_k) is empty. Otherwise D_k = min {D_j - t_j: j is in S (v_k)}. Repeat for all v_k in V.

All nodes v_k such that $C_k = D_k$ are on a critical path. The length of any critical path is T.

This algorithm can be implemented in a computer program for the purpose of automation of critical path evaluation.

3.7 Summary

This chapter explains the use of integrated network model to study the effect of process changes on cluster tool performance. Previous research on cluster tools deals with fixed processing times. But the network model developed here can help when process times change because of change in process parameters like temperature and pressure.

The cluster tool controller follows either a prescribed sequence or follows a rule to sequence wafer moves for lot processing. When a prescribed sequence is followed, it can be represented as a network of nodes connected by arcs. The makespan is the total time taken to process a lot of wafers, which is the total time taken to complete all the steps in the sequence. In the network representation, the makespan is the critical path of the network. The process of critical path length determination is explained in the appendix using specific examples. This process gives analytical expressions for makespan in terms of process times.

When process times change, the critical path of the network changes and the new length determines the changed makespan. So studying the network model can help in quantifying the change in makespan due to a change in process time. RSMs have been integrated into the network model by using them to calculate the process times on the nodes. This directly brings in the effect of process parameters on makespan.

The use of the network model has been demonstrated by studying two specific cluster tool examples. The sensitivity of makespan to process parameters like temperature and pressure has also been evaluated. The sensitivity study helps to identify regions of process parameters where maximum improvement on makespan can be attempted. Alternatively, it can also give regions where makespan is not sensitive to process parameters. This is useful because the process engineer can attempt process improvement in those regions without worrying about degrading operational performance. A concrete example of such a situation is seen in the effect of technology shift on tool performance. The decreased deposition thickness due to technology improvement results in a deposition process time which falls in the less sensitive region because in that region, the deposition process occurs less often on the critical path.

CHAPTER 4

INTEGRATED SIMULATION MODELS

4.1 Need for Simulation models

Sequencing the wafer moves involved in processing a lot of wafers can become complicated. The cluster tool controller can follow a prescribed sequence of activities if it is known that this would be the best (or at least reasonable) sequence in the range of process times encountered. But if the process times undergo a large change, then the sequence may not be a good one to follow since it may lead to reduced throughput. In such cases, it is better for the cluster tool controller to follow some rules. This changes the sequence of activities according to the range of operation times.

Analytical models like the one explained in the previous chapter may not be suited for such situations. Discrete event simulations are natural tools in such situations, since the simulation can be programmed to use the same rule. The simulation can then be used to determine the sequence of activities and the total lot processing time. Once the sequence is known, one could use the network model to quantify the impact of small process changes that do not change the sequence.

This work has addressed the use of simulation models in studying the impact of process changes on cluster tool performance. A simulation software called CTPS (Cluster Tool Performance Simulation) developed by Lee Schruben at Cornell University has been used to study the AMAT Centura cluster tool example of the previous chapter by integrating with it the same RSM and analyzing lot makespan for different sequences and different cluster tool configurations.

4.2 Integrated Simulation Model of Centura Tool using CTPS software

4.2.1 The CTPS Software

The CTPS software uses an Excel spreadsheet as the front end for obtaining user input. A printout of the input screen is attached in the Appendix. The input includes the following:

1. Tool Configuration: The number of process modules in the tool to be simulated, the number of loadlocks (Single/Dual), pump down time needed to prepare loadlock for loading, and vent time needed to vent loadlock.
2. Chamber descriptions: The chambers can be of three types – single, batch or index. A single chamber processes one wafer at a time. A batch chamber starts processing after a set number of wafers have been loaded. In an index chamber, several processes occur within the same chamber on the same wafer.
3. Robot: The transfer times from loadlock to chambers and between chambers. The input is in the form of a matrix.
4. Product: The wafers in a single cassette can have different processing requirements (different routes through the system). So the number of wafers for each particular routing in a cassette has to be specified.

5. Routes: The sequence of process stations that a wafer has to follow and the processing times are specified in the route.
6. Run control: The lotsize and the simulation rule to be used are specified here. There are two rules available for simulation – Push and Pull. In the Push rule, wafers enter modules as soon as possible and in the Pull rule, wafers leave the modules as soon as possible.

This information is sufficient for evaluating the cluster tool performance in a given scenario. If the process time changes, one can change the input parameters and recalculate the tool performance. However, in practice, process times themselves are a function of the product characteristics and the process parameter values. Thus, it would be desirable to include these attributes as the input to the simulation model. Then, one can use the simulation to evaluate the cluster tool performance when the product characteristics or process parameter values change.

We have created this type of integrated simulation model. We started with the CTPS software mentioned above. To model the cluster tool described in Section 3, we added the W CVD RSM to the software. The user enters the product's deposition thickness and the process parameter values (reactor pressure, deposition temperature, the mole fraction of WF_6, and the mole fraction of H_2) in addition to the other model inputs. The software then uses the RSM to calculate the deposition process time and continues by simulating the cluster tool behavior.

We shall demonstrate the use of the integrated simulation model in the section on technology shift example. Now let us examine the CTPS software and the rules in detail.

4.2.2 Push vs. Pull rules

The CTPS simulation allows the user to choose between two rules that generate the sequence of moves for the wafer handler – the PUSH and the PULL rules. In either of the rules, the wafer handler moves only after it receives a signal from the chamber that the processing is completed. That is, after a chamber ends processing, the wafer handler moves empty to that chamber, picks up the wafer and moves it to the next processing chamber in the routing. But the rules differ in the situation when the robot is busy and more than one chamber have finished processing and are waiting for the robot.

In the Push rule, the robot attends the chamber nearer to the start of the routing first. In the Pull rule, the chamber nearer to the end of the routing is served first. That is, in the Push rule, wafers enter modules as soon as possible, whereas in the Pull rule, wafers leave the modules as soon as possible. Figure 10 shows the timing diagrams for the two rules when we simulate the Centura tool with and OD and two CVD chambers using CTPS. It illustrates the difference in the wafer move when the two chambers are waiting for the robot. In the Push rule, the wafer handler moves a wafer from loadlock into OD first before serving a CVD chamber. Whereas in the Pull rule, it first moves a wafer from CVD chamber back to loadlock before serving OD. Note that both wafer 1 and

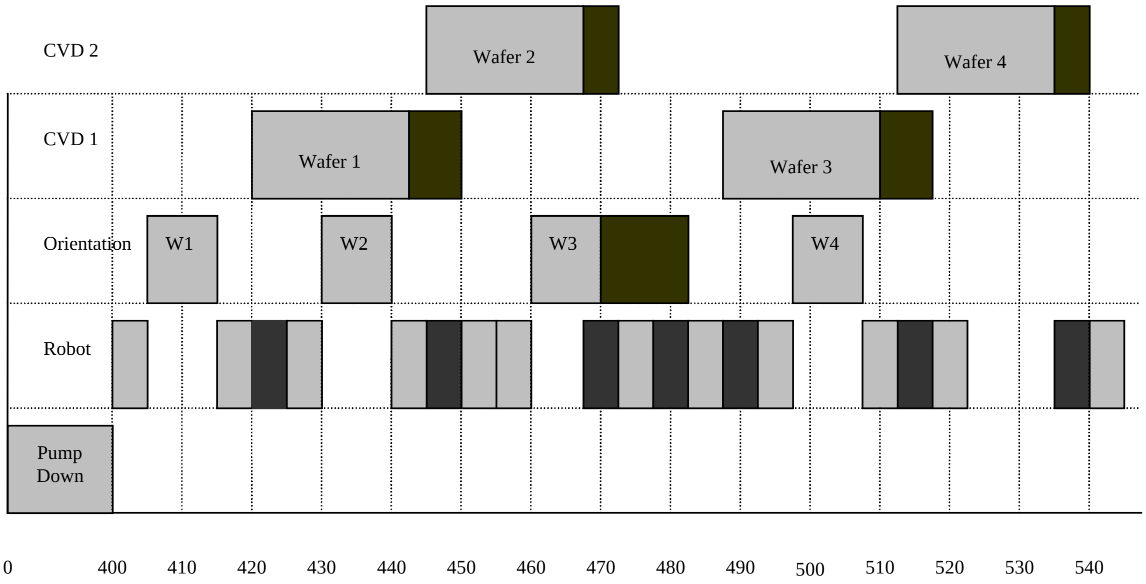


Figure 10(a). Pull Rule – Gantt Chart for 4 wafer lot with OD = 10 secs, D = 23 secs.

■ Waiting for robot □ Processing

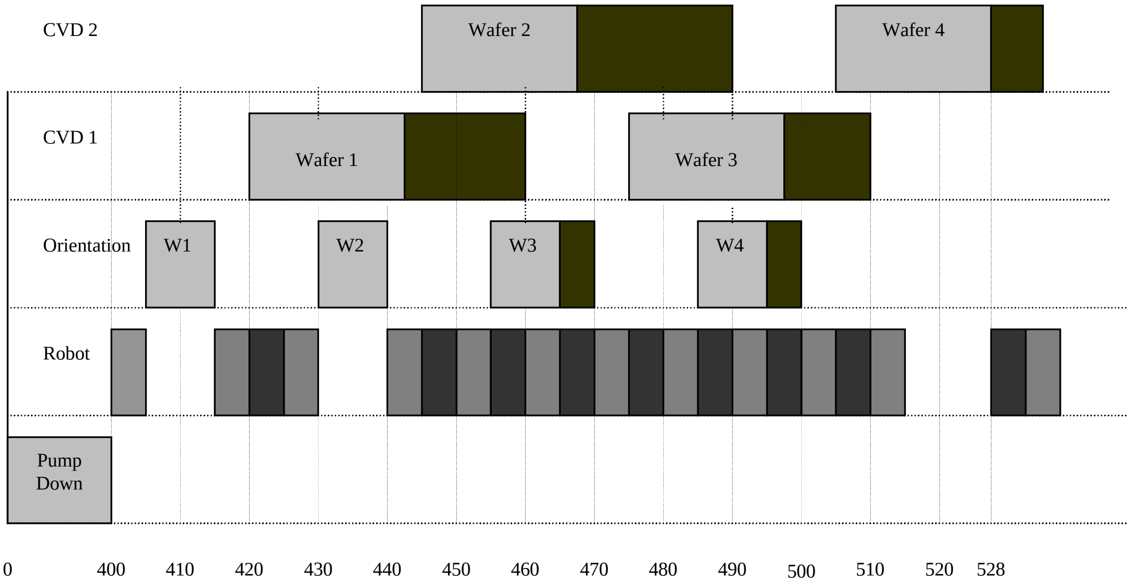


Figure 10(b). Push Rule – Gantt Chart for 4 wafer lot with OD = 10 secs, D = 23 secs.

■ Waiting for robot □ Processing

wafer 3 have finished processing and are waiting for the robot. Under the Pull rule, wafer 1 is unloaded from CVD before wafer 3 is moved into OD. Whereas under the Pull rule, wafer 3 is first pushed into OD before wafer 1 is removed from CVD 1.

4.2.3 Effect of process time changes on sequence of activities

It was said earlier that if the process time changes are large, then the cluster tool controller changes the sequence of activities according to a rule to make sure the cluster tool performance is acceptable.

Such sequence changes have been demonstrated under the Push and Pull rules by using the CTPS software to simulate our cluster tool example from previous chapter. The deposition process time D is a part of the routing information provided to CTPS. We varied D from 1 to 40 seconds and ran the simulation to study the effect of the change in D on the lot makespan. Figures 11 and 12 illustrate the results for the Pull and Push rule respectively.

In Figure 11 and 12, the different slopes for the different ranges of D can be explained in a fashion similar to the example in Chapter 3. The network model for the particular range can be drawn because in that range, the sequence does not change and the slope can be explained by studying how often the deposition operation occurs on the critical path.

It is more interesting to note that between the ranges of D from1-20 seconds and 20-40 seconds and above, the controller changes sequence. For example, in the 1-20 seconds

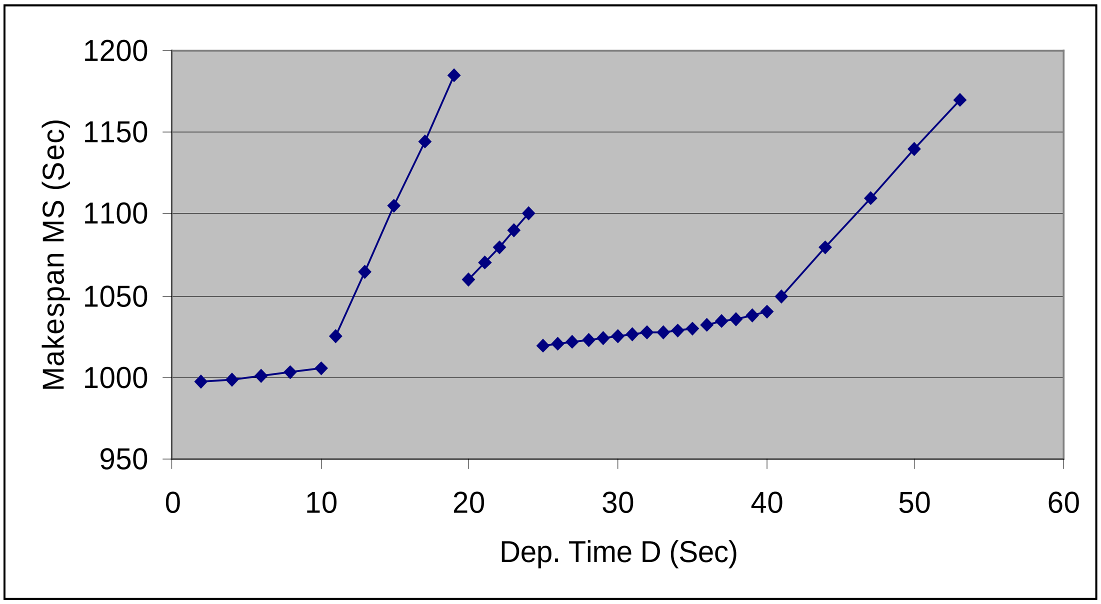


Figure 11. MS vs D. under Pull Rule

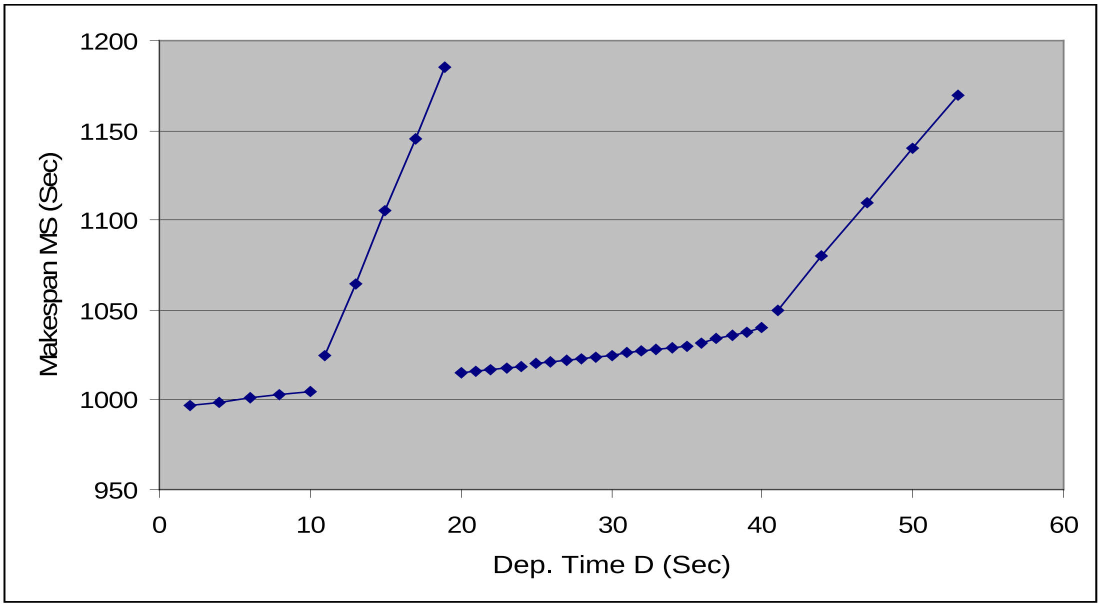


Figure 12. MS vs D under Push Rule

range, though there are two chambers available for W CVD, the controller uses only one chamber. This is because in that range, the deposition time is small and hence the wafer always finishes processing and waits for the robot to transfer it. Whereas in the 20-40 seconds range, the robot is relatively more free and the CVD chambers take more time for processing. So both the chambers are used.

Also from Figure 11 and 12, we see that there is a difference between the sequence that the Push and Pull rules follow in the range of D from 20-25 seconds. There is a difference between the two rules only in this range but this specific to the combination of processing times used here. If we used a different set of numbers for OD and CVD W, then there might be more places where the rules give different sequences. The rules differ only when two wafers are waiting for the wafer handler. In our example, suppose there are unprocessed wafers in the loadlock, an empty OD chamber, and a deposition chamber is holding a processed wafer. The pull rule will give priority to the wafer that has finished deposition. The push rule will give priority to the next unprocessed wafer that needs to visit OD. In the range of D from 1-20 seconds, the situation explained above never occurs because deposition chambers always finish faster than OD. Similarly, in the range of D above 25, the deposition chamber takes a lot of time to finish, but the wafer handler and OD chamber are relatively free. So again the above situation does not occur. Only in the range from 20-25 seconds, the situation of an empty OD and a processed wafer waiting at CVD W chamber occurs.

We can see that in such a case, the Push algorithm results in lesser makespan and hence higher throughput than the Pull rule. This is because in the Push rule, a new wafer is loaded into the empty OD chamber and the next activity is to move the same wafer from OD to CVD W. So the robot saves on an unloaded move. Whereas in the Pull rule the robot has to return a processed wafer from CVD W to loadlock and then move unloaded to the OD to move a wafer from OD to CVD. It is this extra unloaded move of the handler that leads to higher makespan. So it is important to make sure that we use the right rule because the rule affects the sequence which in turn affects the makespan and throughput.

4.2.4 Technology shrink example – The use of integrating product and process information in CTPS

As explained in section 3.5, with improvements in process technology, gate widths decrease and the interconnect diameter decreases. So the required deposition thickness decreases and the deposition process time decreases. If the fab manager is satisfied with a certain level of throughput but is interested in reducing capital equipment costs, then such a technology shift may inspire us to consider reducing the number of process modules to achieve the same level of throughput because the deposition process time has reduced. For instance, if the deposition thickness decreases by 50%, perhaps one module can do the work of two. We can use the integrated simulation model to analyze such a situation because the product characteristic like deposition thickness is now a direct input of the simulation.

Table 3 shows the example under consideration with decreased thickness at each layer due to technology shift. In the simulation model, we set the process parameters to their median values, vary the thickness, and run the simulation under both the rules to get makespan. By running the simulation with two routings – one with two chambers and one with one chamber, the impact of removing the second deposition chamber can be evaluated.

Table 3. The impact of dispatching rules and chambers on lot makespan.

Note: when Th = 1050, the tool uses only one deposition chamber.

Technology Node (nm), Layer	Deposition Thickness Th (A)	Deposition process time D (secs)	Lot Makespan MS (seconds)		
			Two chambers, push	Two chambers, pull	One chamber
250, Contact	2100	33	1028	1028	1465
180, Contact	1500	24	1019	1100	1285
150, Contact	1275	20	1015	1060	1205
130, Contact	1050	17	1145	1145	1145
250, Via 1-2	2700	42	1060	1060	1645
180, Via 1-2	1950	31	1026	1026	1425
150, Via 1-2	1575	25	1020	1020	1305
130, Via 1-2	1350	21	1016	1070	1225
250, Via 3-4	3750	59	1230	1230	1985
180, Via 3-4	2625	41	1050	1050	1625
150, Via 3-4	2188	34	1029	1029	1485
130, Via 3-4	1896	30	1025	1025	1405
250, Via 5	7500	118	1820	1820	3165
180, Via 5	5250	83	1470	1470	2445
150, Via 5	4375	69	1330	1330	2185
130, Via 5	3792	60	1240	1240	2005
150, Via 6	5250	83	1470	1470	2445
130, Via 6	4550	72	1360	1360	2245

The most important conclusion is that great reductions to deposition thickness do not necessarily cause great reductions in lot makespan (or great increases in throughput). Also, sometimes the result of technology shift is counter-intuitive because of a poor dispatching rule. That is, for the 2-chamber configuration under pull rule, the makespan for via 1 and 2 in 250 nm node is 1060 (when thickness is 2700 A). Whereas when the thickness goes down by 50% in the 130 nm node, the makespan actually goes up by 10 seconds which is against our expectations. This is because the robot move time is now comparable to the deposition time and it becomes more critical. So the dispatching rule shifts to a different sequence which is less optimal. The simulation model can prepare us for such situations where poor control results from the dispatching rule.

Moreover, The impact of change in lot makespan with change in deposition thickness is greater for the one chamber case than the two-chamber case. But the performance with one chamber never reaches the two-chamber performance.

Also, the one-chamber configuration is worse than the existing scenario even after technology shifts and the deposition thickness decreases. Moreover, the decreasing thickness can lead to worse performance, so a new control scheme may be necessary. These results illustrate the potential of the integrated simulation model. And they illustrate how the dispatching rule and tool configuration influence the impact that a process change has.

4.2.5 Effect of process parameter changes on makespan

The previous section explained how the integrated simulation model could be used to experiment with different tool configurations under changing process times. Usually, the process time change is a result of the process engineer tuning the process parameters for process improvement. So it is important for the engineer to identify the region of process parameter to target for maximum effect on makespan. Also, it is necessary to know quantitatively the sensitivity of makespan to process parameters. The integrated model takes care of this requirement by providing the engineer with an input sheet containing process parameters and gives the makespan as direct output.

For our Centura cluster tool example, a sensitivity analysis of makespan with temperature (448-492 C) and pressure (68-92 torr) was carried out. Table 4 summarizes the results of the analysis. The results are in agreement with our expectations from the integrated network model for the tool. At low temperature and pressure, the deposition rate is low and so the deposition process time is high. In this range, the deposition process occurs more often on the critical path and hence the sensitivity of makespan to temperature and pressure is high. At high temperature and pressure, the deposition process time is low and in this range, the deposition process occurs less often on the critical path. So the makespan does not change much with temperature and pressure.

						Pressure							
Temp.						(Torr)							
(C)	68	70	72	74	76	78	80	82	84	86	88	90	92
448	1290	1280	1271	1262	1253	1245	1237	1229	1221	1213	1206	1199	1192
452	1261	1252	1244	1235	1227	1219	1211	1203	1196	1189	1182	1175	1168
456	1235	1226	1218	1210	1202	1194	1187	1179	1172	1165	1159	1152	1145
460	1210	1202	1194	1186	1178	1171	1164	1157	1150	1143	1137	1130	1124
464	1186	1178	1170	1163	1156	1149	1142	1135	1129	1122	1116	1110	1104
468	1163	1156	1148	1141	1134	1128	1121	1115	1109	1102	1097	1091	1085
472	1141	1134	1127	1121	1114	1108	1102	1095	1090	1084	1078	1073	1067
476	1121	1114	1108	1101	1095	1089	1083	1077	1072	1066	1061	1055	1050
480	1102	1095	1089	1083	1077	1071	1066	1060	1055	1049	1044	1039.8	1038.8
484	1084	1077	1071	1066	1060	1054	1049	1044	1039.8	1038.8	1037.8	1036.8	1035.8
488	1066	1060	1055	1049	1044	1039.8	1038.6	1037.6	1036.8	1035.8	1034.8	1034	1033
492	1050	1044	1039.8	1038.8	1037.8	1036.8	1035.8	1034.8	1033.8	1033	1032	1031.2	1030.4

Table 4. Makespan (in Secs.) vs. Temperature and Pressure

4.3 Another Example – OD/PVD/CVD Tool.

The effect of process time changes on sequence of activities and hence makespan of another cluster tool example has been studied using the simulation. This cluster tool has six chambers and can carry out TiN barrier layer deposition, which precedes tungsten deposition, and tungsten deposition as well. The first chamber is used for orienting and degassing the wafers, the second one for contact clean of the wafers. The next two are parallel chambers for PVD TiN deposition. The last two chambers are parallel chambers for CVD tungsten deposition. Figure 13 illustrates the tool.

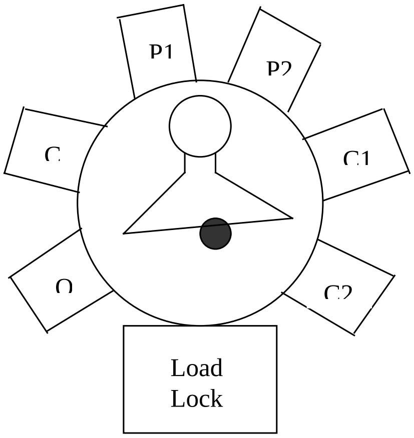


Figure 13. OD/PVD/CVD Tool

The process times for OD, Clean and CVD W were fixed at 10, 30 and 60 seconds respectively and the effect of varying the TiN deposition time from 40 to 80 seconds on makespan of a 10-wafer lot was studied. TiN deposition times lower or higher than this range were not studied because in those ranges, either the CVD or the PVD process would be more critical than the other.

Figure 14 shows the change in makespan with change in TiN deposition time under the Push rule and Pull rule. In the Push rule, the makespan remained at 840 seconds for PVD times from 40-50 seconds. This insight is very useful because the process engineer is free to adjust the process parameters for process improvement without affecting the tool performance. And in the range of PVD times from 50-80 seconds, every one-second increase in the time increases the makespan by 1 second. For PVD deposition times over 80 seconds, every one second increase in the time increases makespan by 9 seconds (also seen from the network model for this sequence, given in the appendix). Having this knowledge, the process engineer can now try to limit the deposition time to less than 80 seconds when he attempts process change. Otherwise, it might prove detrimental to the operational performance of the tool.

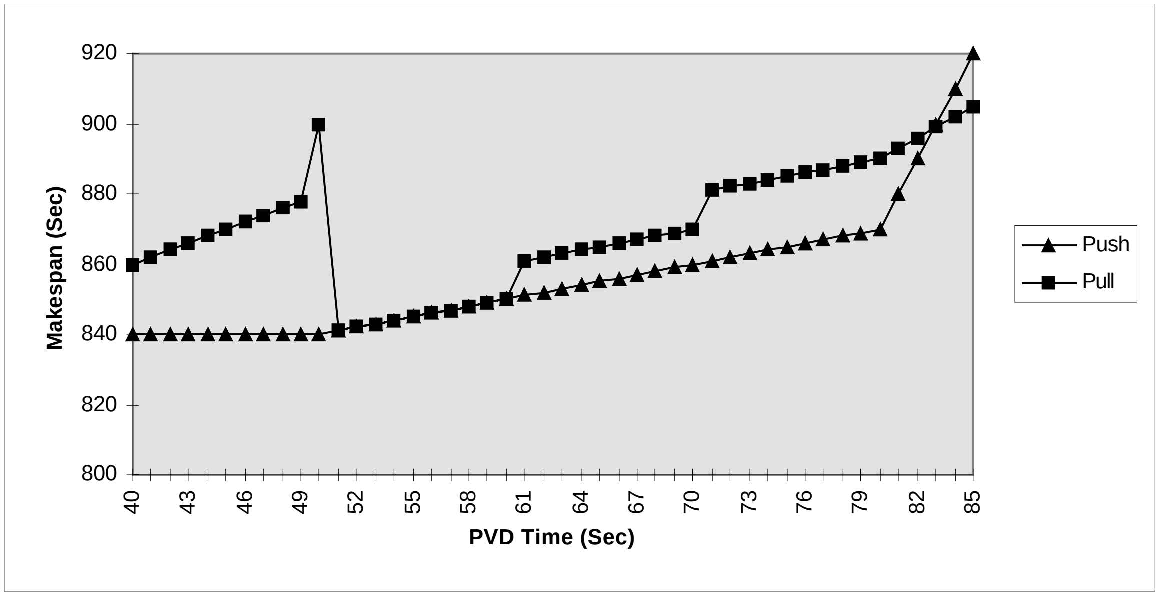


Figure 14. Lot Makespan versus PVD process time for Push and Pull Rule.

Under the Pull rule, the controller shifts to a more efficient sequence when the PVD deposition time changes from the 40-50 seconds range to the above 50 seconds range. So there is a drop in makespan from 900 seconds to 841 seconds when the deposition time changes from 50 to 51 seconds. So, if the tool is operating in this range of deposition times, then it would be worthwhile for the process engineer to change the process parameters so that the controller shifts to the more efficient sequence.

4.4 Summary

When the cluster tool controller follows a rule to dynamically schedule the wafer moves, we cannot use the network model explained in the previous chapter. So we use discrete event simulators like the one explained in this chapter. Integration of RSMs into the simulation is done by using them to calculate the process times specified in the routing information given to the simulation as input. So the effect of process parameters on tool performance can now be quantified by specifying the process parameters as direct simulation inputs.

When a process change is made, the process time changes and this may affect the sequence of wafer moves because of the control rule. Different rules affect the tool performance differently. So the use of the integrated simulation model is in the quantification of the effect of process change on sequence and hence makespan. A specific example of the effect of two rules – Pull and Push, on a Centura tool has been analyzed using its integrated CTPS simulation model.

The integrated model can also be used to analyze different cluster tool configurations. For example, the reduction in deposition thickness due to technology shift leads to reduced deposition times. So we may be able to process wafers in fewer chambers. An analysis of different configurations of the Centura tool example under technology shift has been done. It shows that sometimes the sequencing rule can affect the tool performance in a counter-intuitive fashion. So the integrated model can be helpful for the process engineer as a predictive tool in such situations.

CHAPTER 5

SUMMARY AND CONCLUSIONS

This work addresses the gap that exists between the process engineer and the operations personnel in a semiconductor manufacturing fab. A process engineer attempting process changes for process improvement often does not know how the changes affect the manufacturing system performance. This requires many iterations of talking to the operations personnel before a balance is struck between the process change considered and its effect on system performance. This work provides ways to study the effect of process time and process parameter changes on cluster tool performance. These methods help to eliminate the iterative process mentioned above by helping the process engineer analyze directly the effect of process changes on system performance.

Lot makespan is used as the performance metric for our analysis of the cluster tool. The inverse of makespan is an upper bound on the tool throughput. The importance of cycle time for the entire IC manufacturing process has been explained in Chapter 2. The lot makespan for a particular tool used for one step in the process is a part of the cycle time. So it becomes an important performance measure for the tool in the analysis of the entire system.

Two models for analysis of cluster tools is presented here – an integrated network model and an integrated simulation model. The integrated network model can be used for cluster tool controllers using fixed sequences for wafer moves. It models the makepsan as the critical path of the network. Process time and process parameter changes can be considered as changes in the nodes of the network. So their effect on makespan can be studied by analyzing the change in length of the critical path or the change in path itself. The AMAT Centura cluster tool used for W CVD deposition is analyzed using the integrated network model. Analytical expression for Makespan (MS) as a function of Deposition process time (D) is derived from the network. Sensitivity of MS to D is also evaluated using the network. Integration of the RSM with the network model has been used to directly analyze the effect of process parameter changes and deposition thickness changes on MS.

The integrated simulations are used for controllers that use rules to sequence wafer moves. The CTPS software has been used to demonstrate the change in sequences with change in deposition times for the same AMAT cluster tool example. The rules used by CTPS have been examined in detail. In addition, the RSM has been integrated into CTPS and the effect of technology shifts and different tool configurations on makespan have been quantified. The integrated simulation model has also been used to study the effect of change in process times on change in sequence of wafer moves in a more complicated cluster tool with six chambers used for TiN barrier layer and W deposition.

The examples discussed here clearly show the kind of insights that can be gained by the use of the above methods. For example, the sensitivity analysis of MS vs. D helps the process engineer to select the range of operating process times for maximum effect on throughput and stable performance. Similarly, the integration of process models into the network or simulation models help quantifying the effect of change in process parameters on makespan directly. Since the change in makespan with process parameters is large under some conditions and small sometimes, this quantification of the effect is very useful in process tuning. Additionally, since the cluster tool's maximum throughput depends upon the process parameters, the range of possible process recipes should be considered when evaluating a tool's potential performance and cost-effectiveness.

The method of critical path shows that more than one operation fall in the critical path and changing any of them affects makespan and hence throughput. So no specific operation can be called as the bottleneck. We can say that one operation is more critical than another if it occurs more often on the critical path.

The study of the effect of different rules on performance using the integrated simulation model on the technology shift example shows how important it is to use the right rule. An inappropriate rule may lead to degraded system performance.

These models can be used for analyzing different tool configurations to see if the same performance can be achieved with fewer process chambers in the light of technology shrink which may lead to reduction in deposition thickness.

Future work can focus on applying the methodology of RSM integration into operational models for more process steps in the manufacture of ICs. Research can be done to identify rules that controllers can use to generate reasonable sequences for wafer moves, which lead to stable performance, for a wide range of operation times. An attempt could be made to develop computer programs for analyzing complicated networks using the algorithm mentioned in this work.

APPENDIX

SEQUENCE OF WAFER MOVES AND CRITICAL PATH EVALUATION

A1. Centura Tool

We have discussed in Chapter 3 the integrated network model and its use when the wafer handler follows a prespecified sequence of wafer moves. And we have discussed in detail the Centura cluster tool example and its network model. To derive the network model, the sequence given below was used:

Tool configuration: One loadlock (LL) with 20 wafers. One chamber (OD) for orientation and degassing. Two chambers (CVDA and CVDB) for W CVD. Pump down time = 400 seconds. Each wafer handler move requires 5 seconds. Each OD requires 10 seconds.

Pump down tool
Move wafer 1 from LL to OD. Begin wafer 1 OD.
When wafer 1 OD ends, move wafer 1 from OD to CVDA. Begin wafer 1 CVD.
Move wafer handler to LL.
Move wafer 2 from LL to OD. Begin wafer 2 OD.
When wafer 2 OD ends, move wafer 2 from OD to CVDB. Begin wafer 2 CVD.

REPEAT NEXT 12 MOVES UNTIL WAFER 20 BEGINS CVD.

Move wafer handler to LL.
Move wafer 3 from LL to OD. Begin wafer 3 OD.
When wafer 1 CVD ends, move wafer handler to CVDA.
Move wafer 1 from CVDA to LL.
When wafer 3 OD ends, move wafer handler to OD.
Move wafer 3 from OD to CVDA. Begin wafer 3 CVD.
Move wafer handler to LL.
Move wafer 4 from LL to OD. Begin wafer 4 OD.
When wafer 2 CVD ends, move wafer handler to CVDB.

Move wafer 2 from CVDB to LL.
When wafer 4 OD ends, move wafer handler to OD.
Move wafer 4 from OD to CVDB. Begin wafer 4 CVD.

When wafer 19 CVD ends, move wafer handler to CVDA.
Move wafer 19 from CVDA to LL.
When wafer 20 CVD ends, move wafer handler to CVDB.
Move wafer 20 from CVDB to LL.
End.

The network for this sequence can be drawn by representing each activity by a node and connecting the nodes using arcs in the order given by the sequence. The value of each node is the time of the activity represented by the node. The network for the first six steps in the sequence is shown in Figure A1.

To evaluate the critical path of the network, we first define 'Earliest Starting Time (EST)' and 'Earliest Finishing Time (EFT)' of a node. Taking the earliest starting time of the first node as zero, its earliest finishing time is the processing time needed for the activity it represents. The earliest starting time of the second node is equal to the earliest finishing time of the first node and so on. In general, when a node is preceded by more than one node, its *earliest starting time* is defined as the maximum of the earliest finishing times of all the preceding nodes. The *earliest finishing time* of any node is defined as the sum of its earliest starting time and the processing time of the activity it represents.

The critical path length is defined as the *minimum* time in which the entire sequence of operations can be completed and is equal to the *earliest finishing time* of the last node in the network. So to evaluate the length of the critical path, we start from the first node

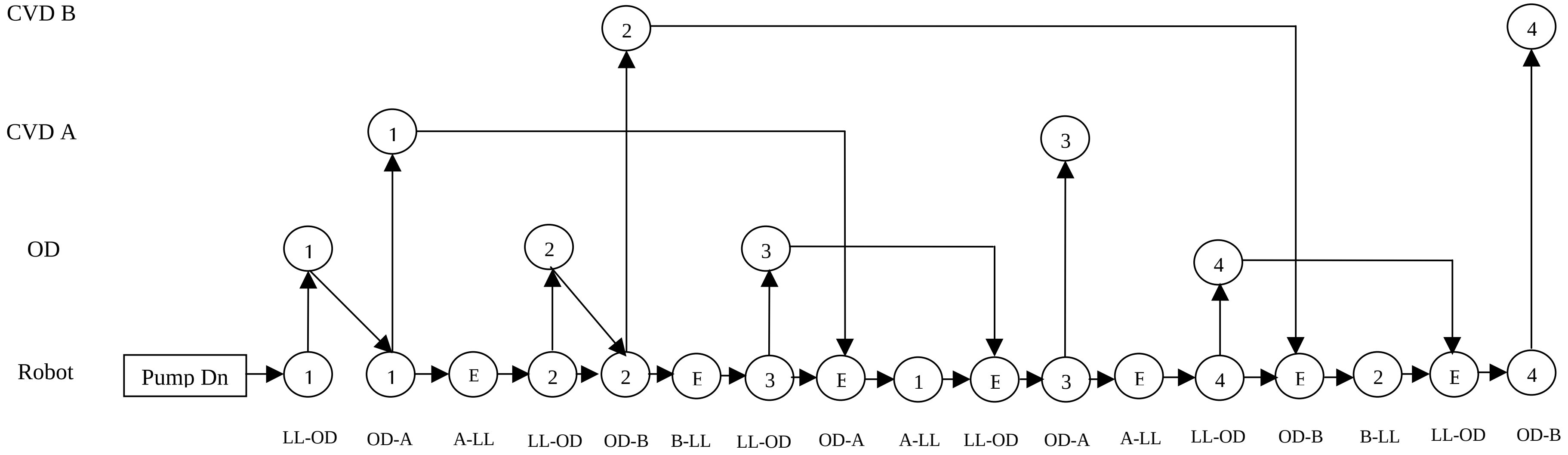


Figure A1. Network for the frst six steps in the sequence

and proceed systematically evaluating the EST of all the nodes of the network till the last node. The critical path length is the sum of EST of the last node and its processing time.

The following set of calculations illustrates the evaluation of EST for the nodes of the network in Figure A1.

ES_1(OD-A) = MAX [400+5, 400+5+10] = 415
ES_2(OD-B) = 440
ES_E(B-LL) = 445
ES_2(PM2) = 445
ES_E(OD-A) = MAX [455, 420+D] = 420 + MAX(D, 35)
ES_E(LL-OD) = MAX [430 + MAX(D, 35) , 465] = 430 + MAX(D, 35)
ES_E(OD-B) = MAX [450 + MAX(D, 35), 445 + D] = 450 + MAX(D, 35)
ES_E(LL-OD) = MAX [460 + MAX(D, 35), 460 + MAX(D, 35)] = 460 + MAX(D, 35)
ES_4(OD-B) = 465 + MAX(D, 35)

Figure A2 shows the network representation of the next 12 moves that are repeated in the sequence. It begins with the moving of wafer 4 from OD to CVD B and ends with the moving of wafer 6 from OD to CVD B. We repeat this until wafer 20 begins processing in CVD B. The calculations that follow gives the EST of the corresponding node.

ES_4(OD-B) = 465 + MAX(D, 35)
ES_E(OD-A) = MAX [480 + MAX(D, 35) , 440 + MAX(D, 35) + D]
= 440 + MAX(D, 35) + MAX(D, 40)
ES_E(LL-OD) = MAX [450 + MAX(D, 35) + MAX(D, 40), 490 + MAX(D, 35)]
= 450 + MAX(D, 35) + MAX(D, 40)
ES_E(OD-B) = MAX [470 + MAX(D, 35) + D, 470 + MAX(D, 35) + MAX(D, 40)]
= 470 + MAX(D, 35) + MAX(D, 40)
ES_6(OD-B) = 485 + MAX(D, 35) + MAX(D, 40)
.
.
.
ES_{20}(OD-B) = 625 + MAX(D, 35) + 8MAX(D, 40)

Figure A3 shows the network for the last four steps of the sequence and the calculations that follow evaluates the EST of the last node and hence its EFT which is the critical path length.

ES_{20}(OD-B) = 625 + MAX(D, 35) + 8MAX(D, 40)

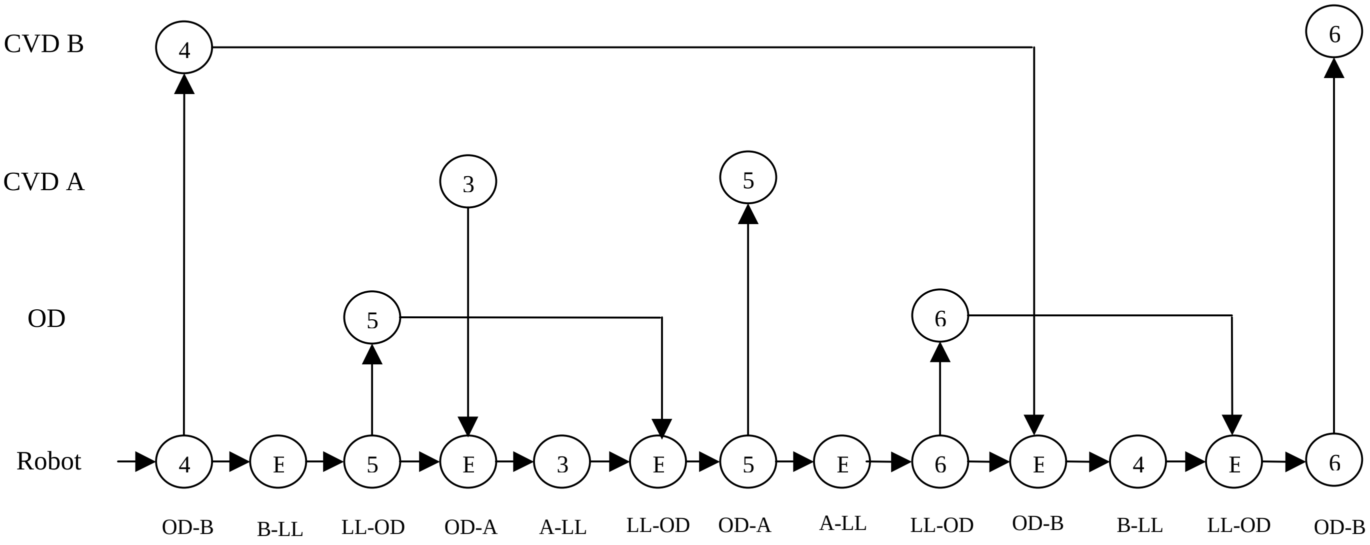


Figure A2. Network for the twelve repeated steps in the sequence

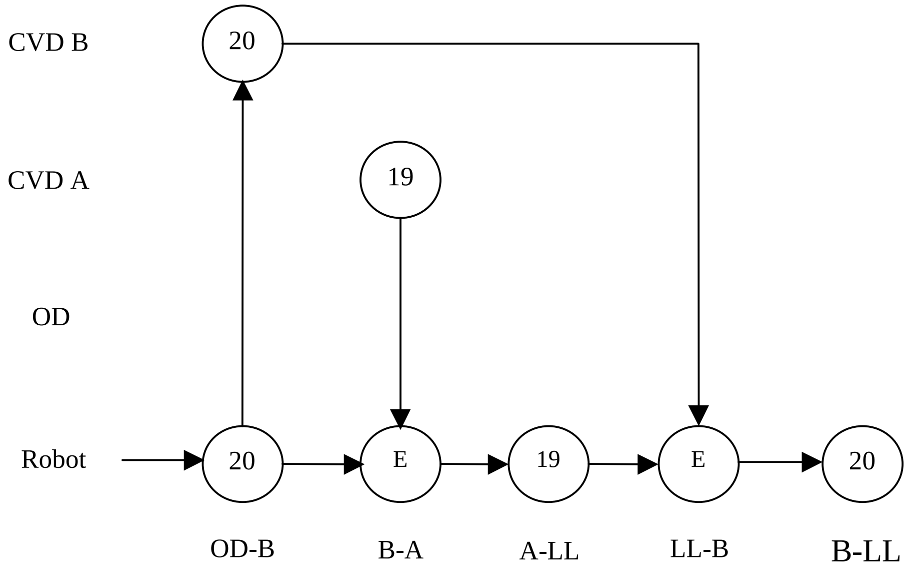


Figure A3. Network for last four steps in sequence

ES_E(LL-B) = MAX [630 + MAX(D, 35) + 8MAX(D, 40) + D ,
640 + MAX(D, 35) + 8MAX(D, 40)]
= MAX(D, 10) + MAX(D, 35) + 8MAX(D, 40) + 630
ES_{20}(B-LL) = 635 + MAX(D, 10) + MAX(D, 35) + 8MAX(D, 40)

Therefore, **EF_{20}(B-LL) = 640 + MAX(D, 10) + MAX(D, 35) + 8MAX(D, 40)** which is the analytical expression given in Chapter 3 for the critical path of the Centura tool sequence.

From the above expression, it is evident that makespan has sensitivities of 1, 2 and 10 with respect to deposition time in regions of D from 0-10 seconds, 10-35 seconds and above 40 seconds respectively.

A2. OD-PVD-CVD Tool

Tool configuration: One loadlock (LL) with 20 wafers. One chamber (OD) for orientation and degassing. One chamber for Contact clean. Two chambers (PVDA and PVDB) for TiN PVD. Two chambers (CVDA and CVDB) for W CVD. Pump down time = 400 seconds. Each wafer handler move requires 5 seconds. Each OD requires 10 seconds and each clean requires 30 seconds.

Pump down tool
Move wafer 1 from LL to OD. Begin wafer 1 OD.
When wafer 1 OD ends, move wafer 1 from OD to Clean. Begin wafer 1 Clean.
Move wafer handler to LL.
Move wafer 2 from LL to OD. Begin wafer 2 OD.
When wafer 1 Clean ends, move wafer handler to Clean.
Move wafer 1 from Clean to PVD A. Begin wafer 1 PVD.
When wafer 2 OD ends, move wafer handler to OD.
Move wafer 2 from OD to Clean. Begin wafer 2 Clean.
Move wafer handler to LL.
Move wafer 3 from LL to OD. Begin wafer 3 OD.
When wafer 2 Clean ends, move wafer handler to Clean.
Move wafer 2 from Clean to PVD B. Begin wafer 2 PVD.
When wafer 3 OD ends, move wafer handler to OD.
Move wafer 3 from OD to Clean. Begin wafer 3 Clean.
Move wafer handler to LL.
Move wafer 4 from LL to OD. Begin wafer 4 OD.

When wafer 1 PVD ends, move wafer handler to PVD A.
Move wafer 1 from PVD A to CVD A. Begin wafer 1 CVD.
When wafer 3 Clean ends, move wafer handler to Clean.
Move wafer 3 from Clean to PVD A. Begin wafer 3 PVD.
When wafer 4 OD ends, move wafer handler to OD.
Move wafer 4 from OD to Clean. Begin wafer 4 Clean.
Move wafer handler to LL.
Move wafer 5 from LL to OD. Begin wafer 5 OD.
When wafer 2 PVD ends, move wafer handler to PVD B.
Move wafer 2 from PVD B to CVD B. Begin wafer 2 CVD.

REPEAT NEXT 22 MOVES UNTIL WAFER 18 BEGINS PVD.

When wafer 4 Clean ends, move wafer handler to Clean.
Move wafer 4 from Clean to PVD B. Begin wafer 4 PVD.
When wafer 5 OD ends, move wafer handler to OD.
Move wafer 5 from OD to Clean. Begin wafer 5 Clean.
Move wafer handler to LL.
Move wafer 6 from LL to OD. Begin wafer 6 OD.
When wafer 1 CVD ends, move wafer handler to CVD A.
Move wafer 1 from CVD A to LL.
When wafer 3 PVD ends, move wafer handler to PVD A.
Move wafer 3 from PVD A to CVD A. Begin wafer 3 CVD.
When wafer 5 Clean ends, move wafer handler to Clean.
Move wafer 5 from Clean to PVD A. Begin wafer 5 PVD.
When wafer 6 OD ends, move wafer handler to OD.
Move wafer 6 from OD to Clean.
Move wafer handler to LL.
Move wafer 7 from LL to OD. Begin wafer 7 OD.
When wafer 2 CVD ends, move wafer handler to CVD B.
Move wafer 2 from CVD B to LL.
When wafer 4 PVD ends, move wafer handler to PVD B.
Move wafer 4 from PVD B to CVD B. Begin wafer 4 CVD.
When wafer 6 finishes Clean, mover wafer handler to Clean.
Move wafer 6 from Clean to PVD B. Begin wafer 6 PVD.

When wafer 19 OD ends, move wafer handler to OD.
Move wafer 19 from OD to Clean.
Move wafer handler to LL.
Move wafer 20 from LL to OD. Begin wafer 20 OD.
When wafer 15 CVD ends, move wafer handler to CVD A.
Move wafer 15 from CVD A to LL.
When wafer 17 PVD ends, move wafer handler to PVD A.
Move wafer 17 from PVD A to CVD A. Begin 17 CVD.

When wafer 19 Clean ends, move wafer handler to Clean.
Move wafer 19 from Clean to PVD A. Begin wafer 19 PVD.
When wafer 20 OD ends, move wafer handler to OD.
Move wafer 20 from OD to Clean. Begin wafer 20 Clean.
When wafer 16 CVD ends, move wafer handler to CVD B.
Move wafer 16 from CVD B to LL.
When wafer 18 PVD ends, move wafer handler to PVD B.
Move wafer 18 from PVD B to CVD B. Begin wafer 18 CVD.
When wafer 20 Clean ends, move wafer handler to Clean.
Move wafer 20 from Clean to PVD B. Begin wafer 20 PVD.
When wafer 17 CVD ends, move wafer handler to CVD A.
Move wafer 17 from CVD A to LL.
When wafer 19 PVD ends, move wafer handler to PVD A.
Move wafer 19 from PVD A to CVD A. Begin wafer 19 CVD.
When wafer 18 CVD ends, move wafer handler to CVD B.
Move wafer 18 from CVD B to LL.
When wafer 20 PVD ends, move wafer handler to PVD B.
Move wafer 20 from PVD B to CVD B. Begin wafer 20 CVD.
When wafer 19 CVD ends, move wafer handler to CVD A.
Move wafer 19 from CVD A to LL.
When wafer 20 CVD ends, move wafer handler to CVD B.
Move wafer 20 from CVD B to LL.
End.

As explained in the previous section, we draw the networks for the steps in the sequence above. The makespan for the 20-wafer lot is the length of the critical path for the network representing this sequence. Figures A4, A5 and A6 illustrate the networks and the calculations below explain the EST calculation for the nodes in the networks. The EFT of the last node in the network of Figure A6 is the makespan of the lot processed with this sequence.

For the network in Figure A4,

$ES_1(\text{OD-CL}) = 415$
$ES_{1E}(\text{OD-CL}) = 450$
$ES_{2E}(\text{P1-OD}) = 460$
$ES_{2E}(\text{OD-CL}) = 500$
$ES_{3E}(\text{P2-OD}) = 510$
$ES_{1E}(\text{OD-P1}) = \text{MAX}\ [\ 460 + P, 530\ ] = 460 + \text{MAX}(P, 70)$
$ES_{3E}(\text{C1-CL}) = \text{MAX}\ [\ 550, 470 + \text{MAX}(P, 70)\ ] = 470 + \text{MAX}(P, 80)$

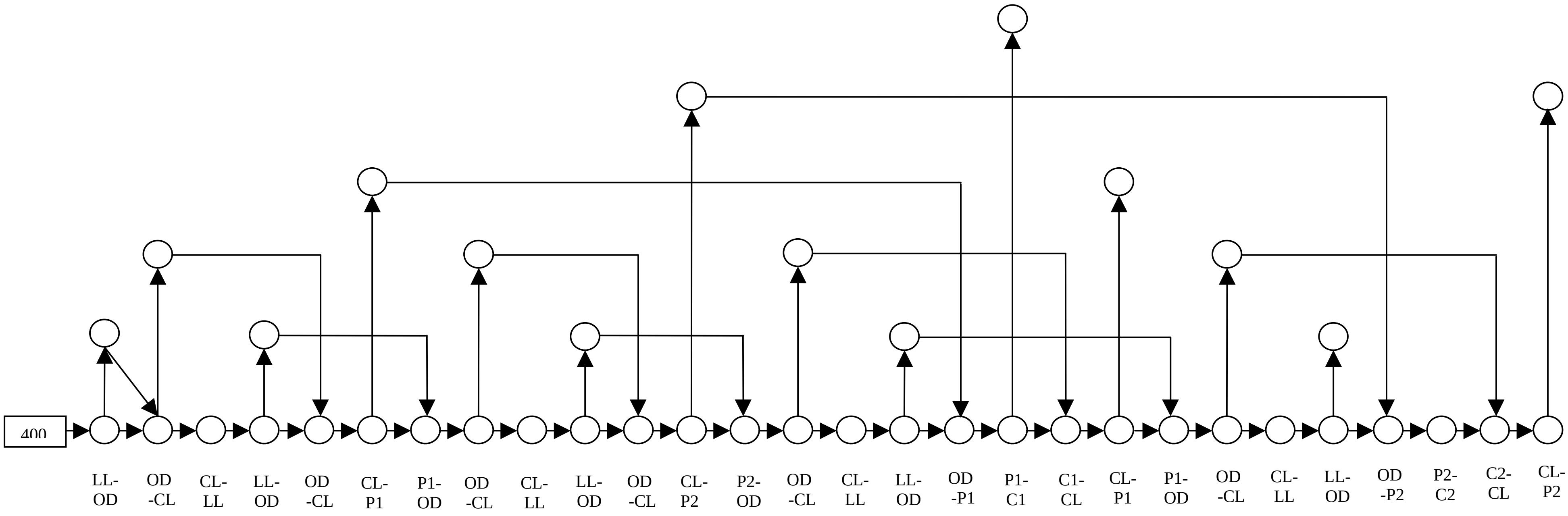


Figure A4. Network representing first 27 moves in the sequence

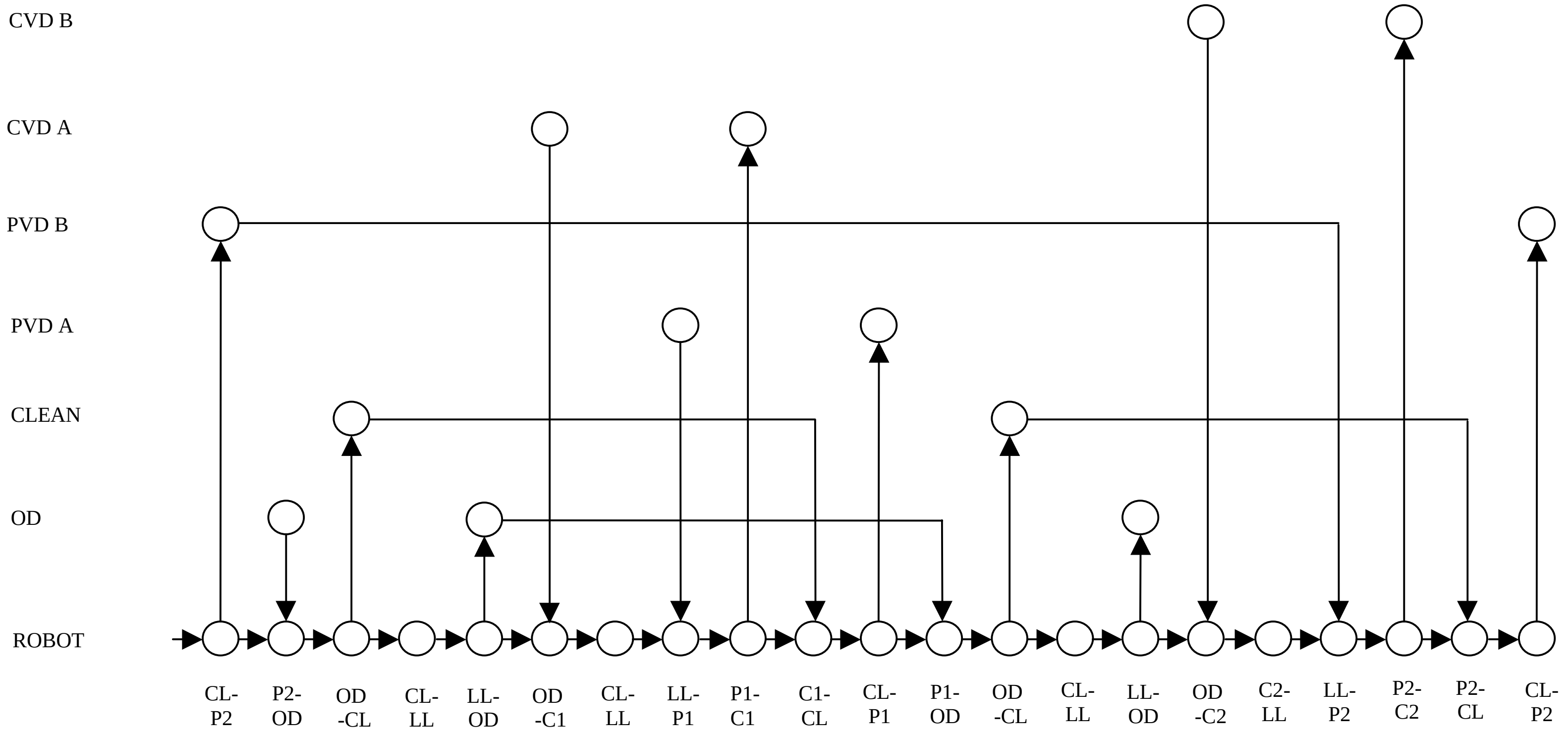


Figure A5. Network for the 22 repeated moves in the sequence

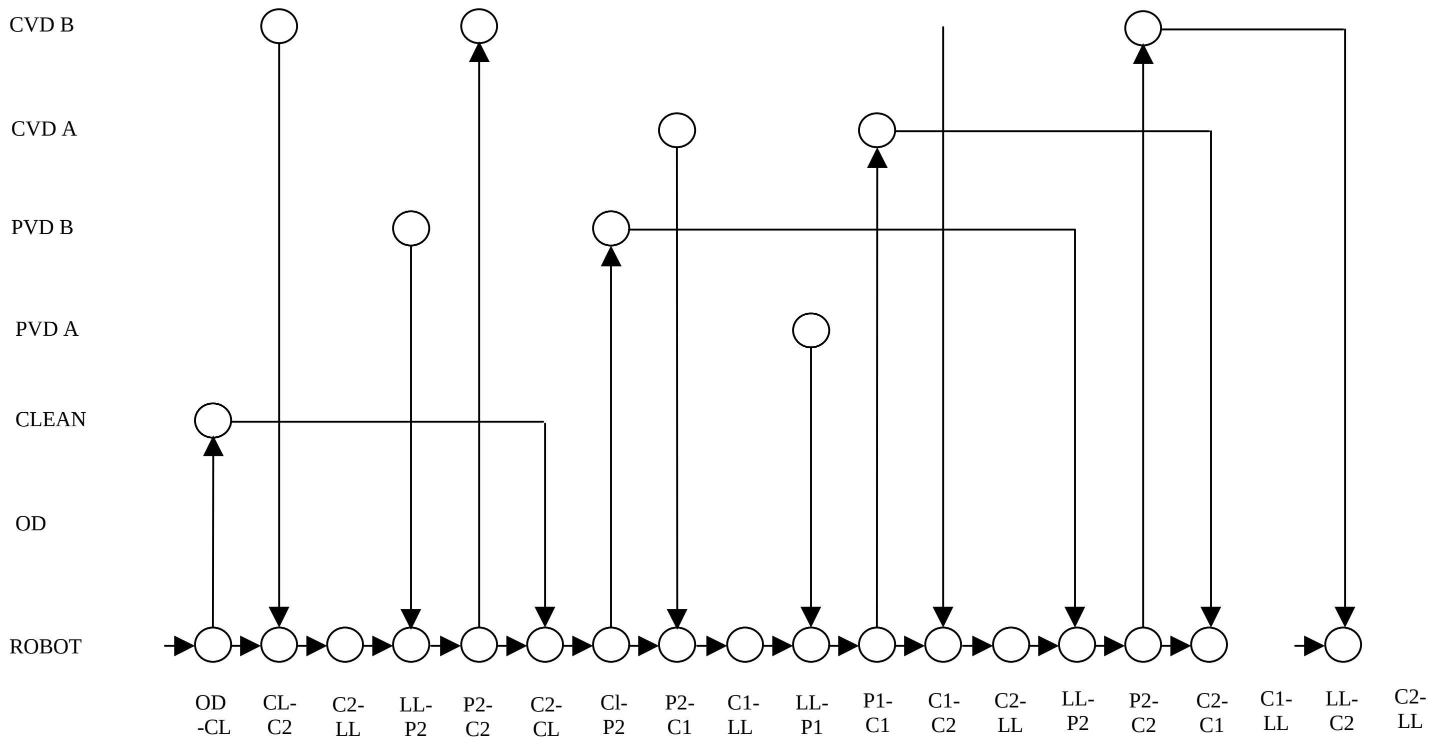


Figure A6. Network for the last 31 moves in the sequence

ES_{4E}(P1-OD) = 480 + MAX(P, 80)
ES_{2E}(OD-P2) = MAX [510 + P, 500 + MAX(P, 80)] = 510 + MAX(P, 70)
ES_{4E}(C2-CL) = MAX [520 + MAX(P, 70), 520 + MAX(P, 80)] = 520 + MAX(P, 80)
ES_4(CL-P2) = 525 + MAX(P, 80)

For the network in Figure A5,

ES_4(CL-P2) = 525 + MAX(P, 80)
ES_{5E}(P2-OD) = 530 + MAX(P, 80)
ES_{1E}(OD-C1) = 550 + MAX(P, 80)
ES_{3E}(LL-P1) = MAX[480 + MAX(P, 80) + P, 560 + MAX(P, 80)]
= 480 + 2MAX(P,80)
ES_{5E}(C1-CL) = 490 + 2MAX(P, 80)
ES_{6E}(P1-OD) = 500 + 2MAX(P, 80)
ES_{2E}(OD-C2) = 520 + 2MAX(P, 80)
ES_{4E}(LL-P2) = MAX[530 + MAX(P, 80) + P, 530 + 2MAX(P, 80)]
= 530 +2MAX(P,80)
ES_{6E}(C2-CL) = 540 + 2MAX(P, 80)
ES_6(CL-P2) = 545 + 2MAX(P, 80)

Having derived the analytical expressions for ES for Clean to PVD B move of wafer 4 and wafer 6, we can extrapolate the relationship for repeated cycles of the network until wafer 18 begins PVD. Then,

ES_{20}(OD-CL) = 645 + 9MAX(P, 80)
ES_{16E}(CL-C2) = 650 + 9MAX(P, 80)
ES_{18E}(LL-P2) = MAX[670 + 8MAX(P, 80) + P, 660 + 9MAX(P, 80)]
= 670 + 8MAX(P, 80) + MAX(P, 70)
ES_{20E}(C2-CL) = MAX [680 + 9MAX(P, 80), 680 + 8MAX(P, 80) + MAX(P, 70)]
= 680 + 9MAX(P, 80)
ES_{17E}(P2-C1) = 690 + 9MAX(P, 80)
ES_{19E}(LL-P1) = MAX[700 + PMAX9P, 80), 640 + 9MAX(P, 80) + P]
= 640 + 9MAX(P, 80) + MAX(P, 60)
ES_{18E}(C1-C2) = MAX[740+8MAX(P,80)+MAX(P,70),650+9MAX(P,80)+MAX(P,60)]
= 650 + 8MAX(P, 80) + MAX(P, 70) + MAX(P, 90)
ES_{20E}(LL-2)=MAX[690+9MAX(P,80)+P,660+8MAX(P,80)+MAX(P,70)+MAX(P,90)]
= 690 + 9MAX(P, 80) + MAX(P, 50)
ES_{19E}(C2-C1) = MAX[710+9MAX(P,80)+MAX(P,60),700+9MAX(P,80)+MAX(P,50)]
= 710 + 9MAX(P, 80) + MAX(P, 60)

ES_{20E}(LL-C2) =MAX[760+9MAX(P,80)+MAX(P,50),720+9MAX(P,80)+MAX(P,60)]
= 760 + 9MAX(P, 80) + MAX(P, 50)
ES_{20}(C2-LL) = 765 + 9MAX(P, 80) + MAX(P, 50)
EF_{20}(C2-LL) = 770 + 9MAX(P, 80) + MAX(P, 50)

Therefore, the makespan for the 20-wafer lot following the above sequence for processing in the OD/PVD/CVD tool is

Makespan = EF_{20}(C2-LL) = 770 + 9MAX(P, 80) + MAX(P, 50)

From the above expression, we can clearly see that the makespan does not change from 770 for PVD deposition times between 40 and 50 seconds. Makespan has sensitivities of 1 and 10 for deposition times between 50 and 80 seconds and above 80 seconds respectively. So the network and critical path evaluation explains the variation of makespan with PVD deposition time observed with the CTPS simulation.

REFERENCES

1. Atherton, R.W., F.T. Turner, L.F. Atherton, and M.A. Pool, "Performance analysis of multi-process semiconductor manufacturing equipment," Proceedings, pages 131-136, IEEE/SEMI Advanced Semiconductor Manufacturing Conference, 1990.

2. Badih El-Kareh, "*Fundamentals of Semiconductor Processing Technologies*," Kluwer Academic Publishers, 1995.

3. Box, G.E.P., and N.R. Draper, "*Empirical Model-Building and Response Surfaces,*" Wiley, New York, 1987.

4. Connors, Daniel P., Gerald E. Feigin, and David D. Yao, "A queueing network model for semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, Volume 9, Number 3, pages 412-427, 1996.

5. Dance, Daren L., Devid W. Jimenez, and Alan L. Levine, "Understanding equipment cost-of-ownership," *Semiconductor International*, pages 117-122, July, 1998.

6. Dhudshia, Vallabh H., and Clyde Hepner, "Cluster tool performance tracking," *Future Fab International*, Volume 1, pages 173-175, Technology Publishing Ltd., London, 1996

7. Ireland, P.J., "High aspect ratio contacts: A review of the current tungsten plug process," *Thin Solid Films*, Volume 304, pages 1-12, 1997.

8. LeBaron, H.T., and M. Pool, "The simulation of cluster tools: a new semiconductor manufacturing technology," Proceedings, pages 907-912, Winter Simulation Conference, Buena Vista, Florida, December, 1994.

9. Lopez, Marcel J., and Samuel C. Wood, "Performance models of systems of multiple cluster tools," 1996 *IEEE/CPMT International Electronics Manufacturing Technology Symposium.*

10. Lopez, Marcel J., and Samuel C. Wood, "Systems of multiple cluster tools: configuration and performance under perfect reliability," *IEEE Transactions on Semiconductor Manufacturing*, Volume 11, Number 3, pages 465-474, 1998.

11. Mauer, J., and R. Schelasin, "Using simulation to analyze integrated tool performance in semiconductor manufacturing," *Microelectronics Engineering*, Volume 25, pages 239-246, 1994.

12. Meyersdorf, Doron, and Taho Yang, "Cycle Time Reduction for Semiconductor Wafer Fabrication Facilities," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 418-423, 1997.

13. Murphy, Robert, Puneet Saxena, William Levinson, "Use OEE; don't let OEE use you," *Semiconductor International*, pages 125-132, September, 1996.

14. Myers, Raymond H., Andre I. Khuri, and Walter H. Carter, Jr., “Response Surface Methodology: 1966-1988,” Technometrics, Volume 31, Number 2, pages 137-157, 1989.

15. Perkinson, Terry L., Peter K. McLary, Ronald S. Gyurcsik, and Ralph K. Cavin, “Single-wafer cluster tool performance: an analysis of throughput,” *IEEE Transactions on Semiconductor Manufacturing*, Volume 7, Number 3, pages 369-373, 1994.

16. Perkinson, Terry L., Ronald S. Gyurcsik, and Peter K. McLary, “Single-wafer cluster tool performance: an analysis of the effects of redundant chambers and revisitation sequences on throughput,” *IEEE Transactions on Semiconductor Manufacturing*, Volume 9, Number 3, pages 384-400, 1996.

17. *Semiconductor Business News*, “Applied Materials, Novellus, LAM Research lead cluster tool market,” CMP Media Inc., 1998.

18. Singer, Peter, “1996: A New Focus on Equipment Effectiveness,” *Semiconductor International*, pages 70-74, January 1996.

19. Stefani, Jerry A., Scott Poarch, Sharad Saxena, Purnendu K. Mozumder, “Advanced process control of a CVD tungsten reactor,” *IEEE Transactions on Semiconductor Manufacturing*, Volume 9, Number 3, pages 366-383, 1996.

20. Srinivasan, R.S., “Modeling and performance analysis of cluster tools using petri nets,” *IEEE Transactions on Semiconductor Manufacturing*, Volume 11, Number 3, pages 394-403, 1998.

21. Thompson, Alan G., W.Kroll, M.A.KcKee, R.A.Stall, and P. Zawadzski, “A Cost of Ownership Model for CVD Processes,” *III-Vs Review*, Volume 8, Number 3, pages 14-20, 1995.

22. Venkatesh, Srilakshmi, Rob Davenport, Pattie Foxhoven, and Jaim Nulman, “A steady-state throughput analysis of cluster tools: dual-blade versus single-blade robots,” *IEEE Transactions on Semiconductor Manufacturing*, Volume 10, Number 4, pages 418-424, 1997.

23. Wolf, H., R. Streiter, S.E. Schulz, and T. Gessner, “Growth rate modeling for selective tungsten LPCVD,” *Applied Surface Science*, Volume 91, pages 332-338, 1995.

24. Wood, S.C., “Adaptable manufacturing systems for integrated circuit production,” Technical Report ICL 94-032, Integrated Circuits Laboratory, Stanford University, 1994.

25. Wood, Samuel C., “Simple performance models for integrated processing tools,” *IEEE Transactions on Semiconductor Manufacturing*, Volume 9, Number 3, pages 320-328, 1996.

26. Wood, Samuel C., “Cost and Cycle Time Performance of Fabs Based on Integrated Single-Wafer Processing,” *IEEE Transactions on Semiconductor Manufacturing*, Volume 10, Number 1, 1997.