Reducing Cycle Time at an IBM Wafer Fabrication Facility

LIEVEN DEMEESTER
Anderson Graduate School of Management
University of California, Los Angeles
Los Angeles, California 90024

CHRISTOPHER S. TANG
Anderson Graduate School of Management
University of California, Los Angeles

In 1991, IBM San Jose decided to produce and sell magnetic heads for computer disk drives on the open market to original equipment manufacturers. However, as IBM's wafer fabrication facility increased the number of products it manufactured, its manufacturing cycle time lengthened. Since cycle time is important in competing in the open market, IBM San Jose formed a study team (in cooperation with UCLA) to examine the wafer fab and to develop ways to reduce cycle time. The team designed a new production control system and proposed new performance measures for operators and engineers. IBM implemented the new production control system and established the performance measures in June 1992, and the cycle time decreased by 50 percent by the end of 1992.

IBM San Jose is a division of IBM that manufactures data storage and retrieval systems (disk drives) for IBM mainframe computers and workstations. It produces magnetic heads for the disk in a wafer fabrication facility (wafer fab)—a capital intensive facility whose fixed costs are high and whose variable costs are relatively low. To lower the average fixed cost (per unit), IBM San Jose decided to produce and sell magnetic heads on the open market in 1991. However, because different types of magnetic heads require different types of operations, increasing the number of product types complicates production. Consequently, the manufacturing cycle

Copyright © 1996, Institute for Operations Research and the Management Sciences
0092-2102/96/2602/0034$01.25
This paper was refereed.

INDUSTRIES—ELECTRIC/ELECTRONIC
PRODUCTION/SCHEDULING—FLOW SHOP

INTERFACES **26:** 2 March–April 1996 (pp. 34–49)

time (or manufacturing lead time) lengthened. Long manufacturing cycle time decreases IBM's competitiveness, especially when the competitors (such as Seagate, Hitachi, and Fujitsu) not only compete on price and quality but also on speed [Blackburn 1991, Hill 1988]. To examine the wafer fab and to develop ways to reduce the cycle time, IBM San Jose and UCLA formed a study team at the beginning of 1992.

After an initial investigation of the operations at the wafer fab, we identified two major causes for long and unstable (or unpredictable) cycle time. The first cause is the complexity of the manufacturing process and its inherent uncertainties, such as rework and yield loss. These characteristics tended to cause high work-in-process inventories (WIP) and long cycle time. The second cause was the existing production control system. It did not fully exploit the system information on rework, yield loss, and work-in-process inventories to stabilize (or control) the cycle time. Because cycle time was unpredictable, the managers released work orders early to prevent late shipments. They also maintained a fairly large inventory of finished wafers to meet the demand. These two practices increased the total inventory in the system as well as the cycle time. In addition to observing the operation, we interviewed the managers and various personnel to understand the incentives for reducing cycle time. We learned that the existing performance measures did not include cycle time [Demeester and Tang 1993]. For this reason, we recommended that IBM modify its performance measures to give the managers and personnel incentives to reduce cycle time.

Based on our observations and input from management, we believed that a new production system could reduce cycle time. However, it was made clear to us that a new production control system (a) should be based on the existing system, and (b) should be simple.

An optimal production control system that enabled management to meet the demand with the minimum amount of WIP and the shortest cycle time would have been ideal. However, because of the system's dynamics (such as the process yield, rework, and demand forecasts) and its

The performance measures did not include cycle time.

complexity (the number of products, the number of circuit layers, and the number of work centers), it is extremely difficult to evaluate the performance of a production control system analytically. In fact, there is no known optimal production control system for a reentrant flow shop with rework, uncertain yields, and uncertain demand. Several researchers have developed various production control systems that utilize some basic ideas. For example, Glassey and Resende [1988b] have developed a production control system for maximizing the throughput rate of the wafer fabs. The key idea of their system is to release the wafers to the fab so that the bottleneck work center does not become idle.

While the existing production control systems described in the literature have focused on maximizing throughput, our production control system is intended to stabilize and to reduce cycle time. It is based on

the existing system and it divides production control decisions into *production release decisions* and *production dispatch decisions*. The production release decisions are made by the line manager, who specifies the number of wafers to be released to the system. The production dispatch decisions are made by operators, who schedule the wafers to be processed at the work centers. The production release decisions affect the production dispatch decisions as follows. The production dispatch decisions depend on the type and the number of wafers waiting to be processed at each work center; however, the total number of wafers in the system is controlled by the production release decisions. The line manager bases the production release decisions on an aggregate model of the wafer fab. Despite its simplicity, this aggregate model still captures the stochastic nature of the wafer fab and the dynamics of the material flows in the wafer fab. At the same time, it retains a level of aggregation that makes it easy to understand. The production dispatch decisions are based on a simple myopic policy that gives highest priority to the wafer that is furthest behind the planned completion schedule. The combined effect of the production release and production dispatch decisions is that the cycle time will be stabilized and that a simple mechanism can be applied to reduce the cycle time gradually.

The Manufacturing Process

The wafer fab at IBM San Jose manufactures the circuits for magnetic heads of different computer disk drives as follows: a number of circuits for many identical magnetic heads are built on top of a single ceramic wafer (Figure 1).

To build the circuits, the wafer fab uses thin film processing technology that adds or removes very thin patterns of materials to or from the surface of the wafer. The circuits generally consist of multiple layers of different materials (Figure 2).

The wafer fab uses similar processes to form each layer of the circuit. These processes can be grouped into the following six categories: vacuum deposition, photolithography, inspection, plating, etching,

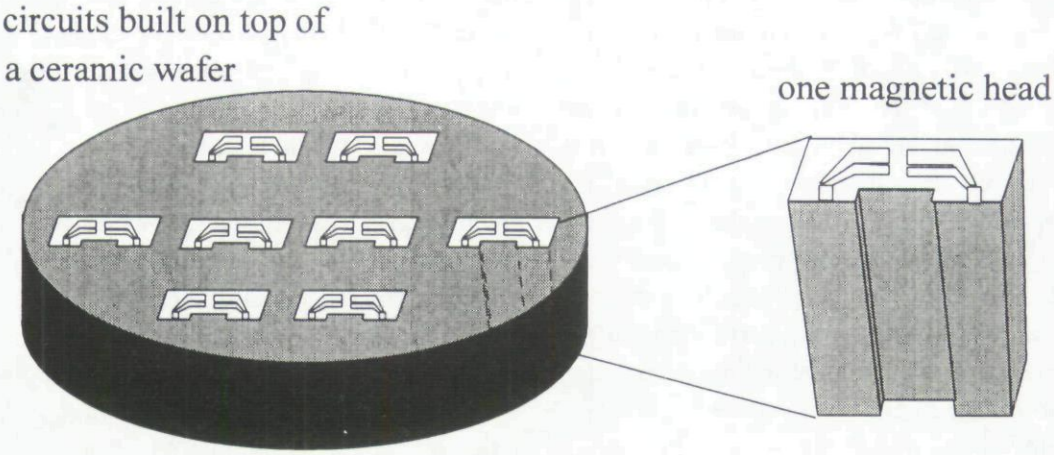

Figure 1: This ceramic wafer (left) can produce eight identical magnetic heads. One is shown enlarged on the right. On its top surface are the circuits built during wafer fabrication.

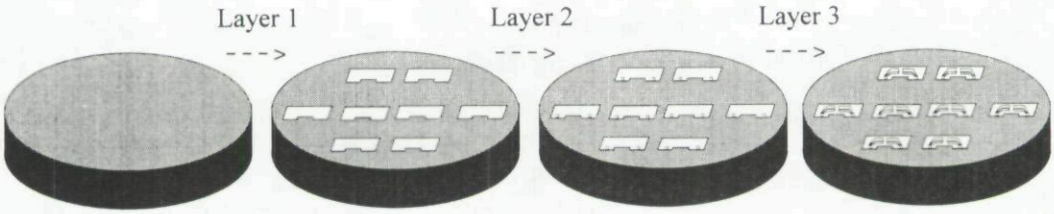

Figure 2: Using thin film processing technology, the wafer fab builds the necessary circuit layers, one on top of the other. The circuits on this wafer require only three layers.

and ion milling. They are similar to the processes used in fabricating integrated circuits that are described by Burman et al. [1986], Chen et al. [1988], Mead and Conway [1980], and Reinhard [1987].

Six types of processes are used for fabricating wafers:

(1) Vacuum deposition is a batch process that deposits a thin layer of a certain type of metal on a batch of wafers in a vacuum chamber.

(2) Photolithography consists of three steps: (a) photoresist; (b) masking; and (c) developing. In the photoresist step, a light-sensitive coating (photoresist) is placed on top of the wafer. Then, in the masking step, the coated wafer is exposed to ultraviolet light through a mask that contains the pattern of the photoresist to be removed. Finally, in the developing step, special chemicals are used to remove the unexposed photoresist on the wafer [Burman et al. 1986, Chen et al. 1988].

(3) Inspection consists of various procedures used to check whether the wafers have been processed properly and if the circuits on the wafers satisfy certain specifications.

(4) Plating is an electrolytic batch process in which various types of metallic materials are deposited on a batch of wafers immersed in a tank that contains the right types of metallic ions.

(5) Etching is a batch process in which special chemicals are used to remove certain types of materials from the wafer according to the pattern determined in the photolithography process.

(6) Ion milling is a process in which ion bombardment is used to remove material from the wafer with very high precision.

In the wafer fab, machine tools that perform the same process, for example, ion milling, are grouped into a work center.

During wafer fabrication, multiple layers are formed on top of a wafer and each layer has its own requisite sequence of operations. Therefore, each type of wafer has its own process flow or required sequence of operations (Figure 3). The process flow of IBM's system is similar to those of other wafer fabs [Cory 1986, Martin-Vega et al. 1989]. Specifically, each wafer makes multiple visits to the same work center at different points in the fabrication process. Such a system is known as a *reentrant flow shop* (Figure 3).

Cycle Time

Reentrant flow shops are known to have long and unstable cycle times. Planning and controlling the flow of materials through such a shop is difficult because of the reentrant nature of the process flow and the uncertain demand [Glassey and

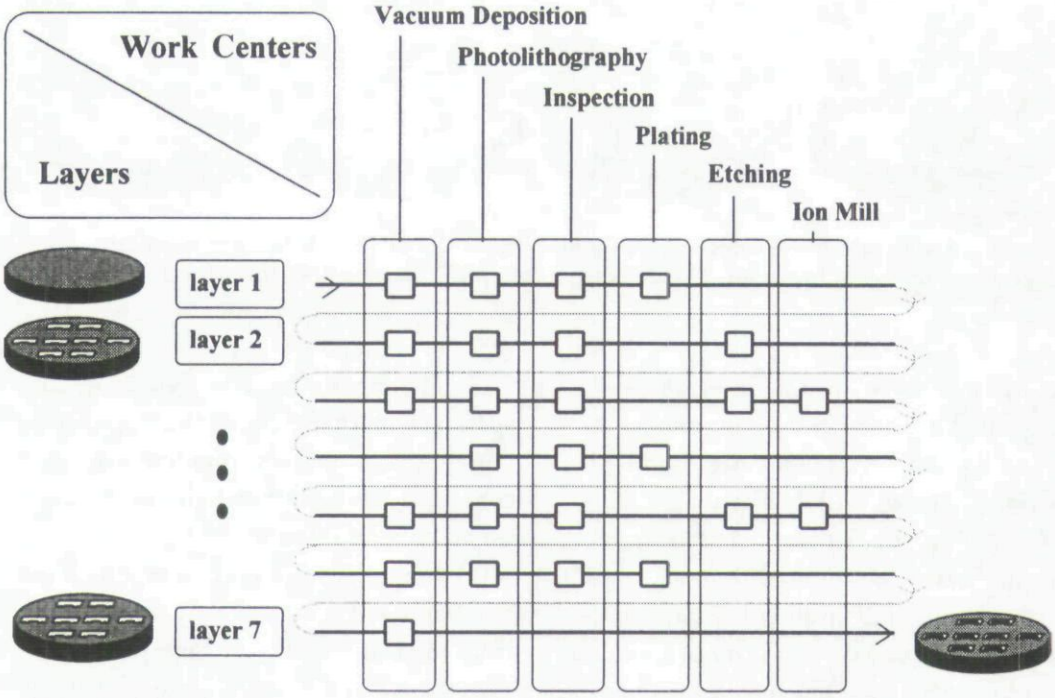

Figure 3: In this process flow of a wafer with seven layers, all operations required for the same circuit layer are aligned horizontally, and all operations performed by the same work center are aligned vertically. For simplicity, we did not include the process flow due to rework.

Resende 1988a, 1988b]. At the IBM wafer fab, other factors also contribute to the long and unstable cycle time: the long net processing time, the uncertain yields, the rework, the system uncertainties, and the delays caused by batch processing and machine setups.

The total processing time each wafer takes is fairly long because each wafer requires over 200 operations at a number of work centers.

The system uncertainties include machine failures, uncertain process yield, and rework. Machine tools sometimes fail to perform within specifications; this disrupts the flow of materials in the fab and causes the cycle time to increase and to fluctuate. Wafers are tested at different points in the process. Those that pass continue the process, those that fail are either scrapped or sent back to an earlier operation for rework. This increases the cycle time and causes it to fluctuate even more.

Since the fab uses different fixtures or different material-handling systems for different processes, batch sizes differ by wafer type and process. Therefore, it is not uncommon for one batch of wafers to wait for another one to form the right batch size for the next operation. In addition, setup times incurred when a machine tool switches from one type of wafers to another or from one layer to another increase the cycle time.

We could reduce the cycle time by improving the fabrication process: reducing

the net processing time, increasing the machine reliability, increasing the processing yield, reducing the batch sizes, and reducing the setup times. We could also reduce it by reducing the size of work-in-process inventories in the system [Graves 1988], so we designed a new production control system that stabilizes the WIP and reduces the WIP if IBM makes certain improvements in the fabrication process.

A New Production Control System

The primary objective of our production control system is to control the flow of materials in the fab so that it will complete the wafers according to the planned schedule (and meet customer demands). The secondary objective is to control the work-in-process inventories so that the cycle time will be reduced and stabilized. To accomplish these objectives in a reentrant flow shop, we based our system on the existing system in which the production control decisions are divided into two types: production release decisions and production dispatch decisions. The production release decisions are based on global information (such as WIP, yields, and planned cycle time) and specify the number of wafers of each type to release to the fab at the beginning of each time period (or shift). Within each time period, the production dispatch decisions are based on local system information (such as machine availability, yields, and due dates) and set priorities for the different types of wafers at different stages of the process that are waiting at each work center. These system-generated decisions are then used by people to make real decisions. The line manager, who manages the WIP in the system, makes the release decisions. Operators, responsible for processing the wafers according to schedule, make the dispatch decisions. The release decisions affect the dispatch decisions. The dispatch decisions depend on the distribution of WIP (the type and the amount of WIP waiting to be processed at different work centers); however, the production release decisions control the total amount of WIP.

Because the process flow is very complex, any production release decisions that depended directly on the detailed information from each work center would be too complex to be practical. We therefore developed an aggregate model of the wafer fab to generate simple production release decisions (appendix). To monitor the progress of the wafers through each layer (instead of at each step of the process), we aggregate the operations that belong to a single layer into a single production stage. In this way we modeled the entire wafer fab as a serial production system that faces uncertain yield at each stage and uncertain demand. We then adapted the production rule developed by Tang [1990] to generate release decisions for each stage (or layer)

The line managers were measured by how well they met demand.

(appendix). The production release decisions do not deal with the reentrant flow process directly; however, we considered the reentrant nature of the process flow explicitly in the design of the production dispatch decisions.

In our production control system, the goal of the production release decisions

and production dispatch decisions is to complete the wafers according to a planned production schedule. The schedule is based on a planned cycle time, which serves as a benchmark for the actual cycle time. For instance, if all the wafers are completed according to the schedule, then the actual cycle time will be equivalent to the planned cycle time. At IBM San Jose, we noticed that the production schedule was based on too short a planned cycle time. Since the existing production control system was based on an unrealistic planned cycle time, the actual average cycle time exceeded the planned cycle time by a wide margin, and managers could not rely on the operational decisions it specified. To gain the confidence of management, we used a more realistic planned cycle time that was based on historical performance and that included a safety time at the end of each stage (or layer) (appendix).

Average cycle time was reduced by up to 50 percent.

Within each time period, the actual yield for each operation is uncertain. For this reason, the WIP fluctuates from period to period. To maintain the right amount of WIP in the system (so that the actual average cycle time will be close to the planned cycle time), we developed production release decisions (using the release rule developed by Tang [1990]) that consist of a push component and a pull component (appendix). The push component specifies the number of wafers to be released to the fab at each stage; these quantities are based on the planned cycle time and the updated demand forecasts. However, to compensate for fluctuations in WIP due to yield loss and rework the pull component adjusts the quantity specified by the push component. This adjustment is positive or negative depending on whether the actual WIP level is below or above the target level. It follows the pull philosophy in which production decisions depend on the WIP level of the system.

Because of the reentrant nature of the wafer fabrication process, at any time different types of wafers at different stages of the process are waiting to be processed at each work center. We used information about the distribution of the WIP in the system, the planned yield of each operation, and the planned cycle time to estimate the due date and the completion time for each of the wafers. Our production dispatch policy assigns a higher priority to those wafers that are behind schedule (appendix). Although other dispatch policies may outperform ours, we think that ours is simple and practical.

The production release decisions and the production dispatch decisions are aimed to complete the wafers according to the planned schedule so that the actual average cycle time will be close to the planned cycle time. Based on the theoretical results presented by Tang [1990], the actual average cycle time will equal the planned cycle time if the following assumptions are satisfied: (a) no defective wafers are cycled back for rework to an earlier operation that belongs to an earlier layer; (b) process yields at different layers are independent; (c) processes have small batch sizes and low setup times; (d) each work center has

multiple machines with enough capacity to perform the different operations required at different layers, which minimizes the interactions among different layers that are caused by the reentrant flow; and (e) the demand process and the yields are stationary. Assumptions (a) through (d) are satisfied to a large extent at the IBM wafer fab; however, it is clear that assumption (e) might be violated. Our experience at the IBM wafer fab showed that the bias created by this violation is minimal because the pull component of the production release decisions makes the system self adjusting [Denardo and Tang 1994]. In any event, those applying our production control system should take the limitations created by the above assumptions into consideration.

Reducing Cycle Time Continuously

To reduce the actual cycle time, we need to reduce the planned cycle time. We developed a simple mechanism to reset the planned cycle time when certain system improvements are observed.

We can reduce the planned cycle time when we reduce the mean or the variance of actual cycle time of the work centers. We can do this by making system improvements (such as reducing setup time and improving process yield) or by making operational improvements (such as improving the coordination of activities among work centers). In addition, we can reduce the variance of the actual cycle time of the work center by using a production control system that stabilizes the cycle time.

Once we have reduced the mean or the variance of the actual cycle time of the work center, we can reset the planned cycle time. Suppose that the managers and the engineers have developed a way to improve the process yield at a work center. This will make its actual average cycle time less than its planned cycle time. (Improving the process yield at a work center reduces the amount of rework going from this work center to other work centers in the same layer. Hence, it reduces the actual average cycle time for those work centers as well.) We can then reset the planned cycle time for this layer (appendix) by using the updated value of the average actual cycle time of the work centers. Hence, we can reduce the total planned cycle time accordingly.

In addition, suppose that we have reduced the variance of the actual cycle time of a work center. Then we can reduce the planned safety time and the planned cycle time for this layer accordingly (appendix). The cycle time reduction for this layer and its corresponding reduction in WIP will not affect the improvements made in succeeding layers because the production release decisions control exactly which WIP gets released from one layer into the next. So, we can carefully reset the planned cycle time for each layer. In this way, we can gradually reduce the planned cycle time as management and employees continue to improve the system and its operation.

To stabilize and reduce the cycle time at each work center, one must have the full cooperation of operators, engineers, and managers. Before we implemented the production control system, we noticed that inconsistent performance measures at the IBM wafer fab could hinder implementing any program to reduce the cycle time. The line managers were measured by how well

they met demand, the process engineers were measured by the process yield, and the operators were measured by the production volume. With such inconsistent performance measures, it would be difficult to coordinate the activities needed to improve the system and its operation and to reduce cycle time. For example, operators would favor reducing setup times, but process engineers have no incentive to develop procedures to reduce setup times. To help coordinate the activities needed to improve the system and its operation, we proposed forming a cross-functional team for each work center that is composed of a manager, engineers, and operators. We also proposed that the performance measures for the members of the team include the actual average cycle time of the work center. The facility adopted both recommendations in April 1992. The reorganization has empowered the cross-functional teams to develop ways to improve operations at each work center.

Implementation Results

In March 1992, we presented our plans for the new production control system to top management and discussed the plan for continuous cycle time reduction (Figure 4). The support from top management and the commitment from other personnel led the information systems department at IBM to develop computer codes for the new production control system. IBM implemented the new production control system in early June 1992. From May 1992 to August 1992, the total demand and the product mix remained fairly stable but average cycle time changed. The actual average cycle time was reduced by up to 50

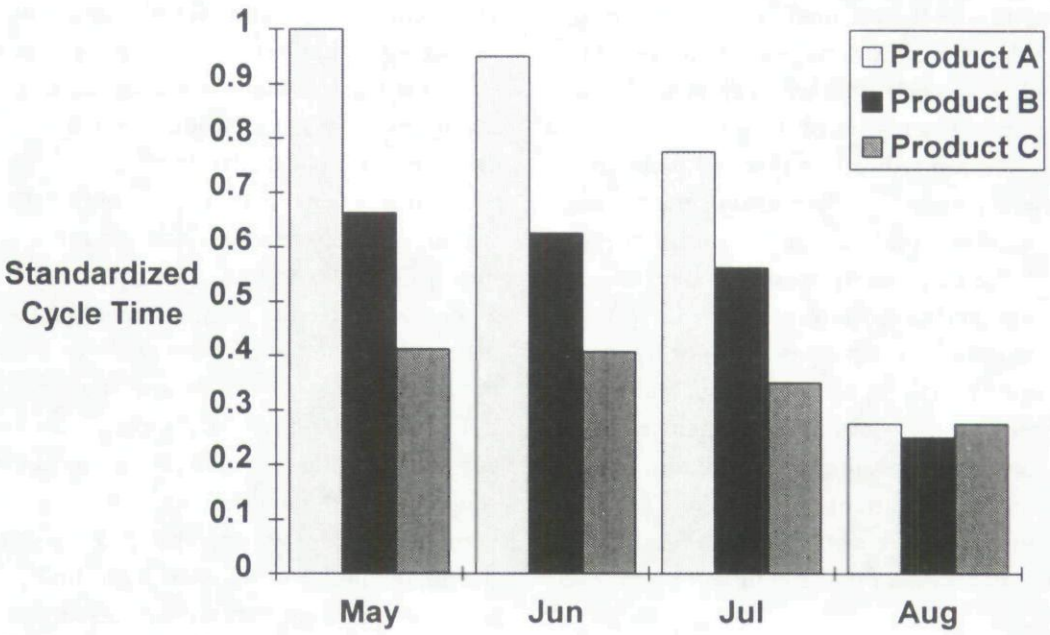

Figure 4: The actual average cycle time of three major products (A, B, and C) decreased from May 1992 to August 1992 at the IBM wafer fab. The vertical axis was standardized with respect to the actual average cycle time of product A in May 1992.

percent. In addition, actual average cycle times of different products were initially very different but became very similar. (When average cycle times of different products are similar, managers find it easier to quote lead times for their customers.) Moreover, the work-in-process in the system fell by as much as 50 percent between June and August 1992. (This is not surprising, given the direct relationship between WIP and cycle time.) These major improvements were partly due to the new production control system; however, we believe that the cross-functional approach and the new performance measures are important contributing factors in reducing cycle times.

Acknowledgments

We acknowledge Dennis Bettencourt, Dave Gill, and Bob Larocca of IBM San Jose for participating in the IBM-UCLA study team. In addition, we are grateful to Bob Larocca and the information systems department for developing the computer codes for the production control system. We are indebted to Glenn Larnerd (vice-president of IBM), Barbara Grant, Sofia Laskowski, Gela Russell, and Ralph Ahlgren of IBM San Jose for supporting this joint IBM-UCLA project. Finally, we thank Rick So of the University of California at Irvine and the departmental editor, Candace Yano, for their constructive comments.

APPENDIX

Layer Aggregation

We used a simple way to aggregate the operations belonging to a single layer into a single production stage (Figure 5).

In our simple example, there are three work centers A, B, and C. Work center A performs operations 1 and 3, B performs operation 2, and C performs operation 4. This fab processes a single type of wafer with two circuit layers. Layer 1 requires operations 1 and 2, while layer 2 requires operations 3 and 4. This wafer fab is a reentrant flow shop with each wafer visiting work center A twice. Based on the test result after each operation, each wafer is scrapped, cycled back to previous operations for rework, or proceeds to the next operation. Here we assume that a defective wafer is either scrapped or cycled back for rework to an earlier operation (or the same operation) that belongs to the same layer (that is, it is not sent for rework on an earlier layer). In this case, there is no interaction between different layers in terms of rework. This assumption turns out to be very reasonable for the IBM wafer fab in San Jose.

Since the production release decision is made at the beginning of each time period, the planned yield for each operation is updated at the beginning of each time period. Within each time period, we assume that the actual yield equals the planned yield. Clearly, this assumption would be more reasonable if the time period were reasonably short. (The actual data we obtained from the engineers shows that the actual yield does not vary much within a time period.) In our example, the planned yields are as follows. After operation 4, a wafer has a 0.05 chance of being scrapped, a 0.25 chance of being sent back to operation 3 for rework, a 0.25 chance of being sent back to operation 4 for rework, and a 0.45 chance of entering the finished wafer buffer. IBM estimates these probabilities using historical data and updates them regularly.

To keep track of the work-in-process inventories in the system, our new production control system specifies two types of buffers that store the work-in-process inventories: an *intralayer buffer* located immediately before each operation and an *interlayer buffer* located immediately after

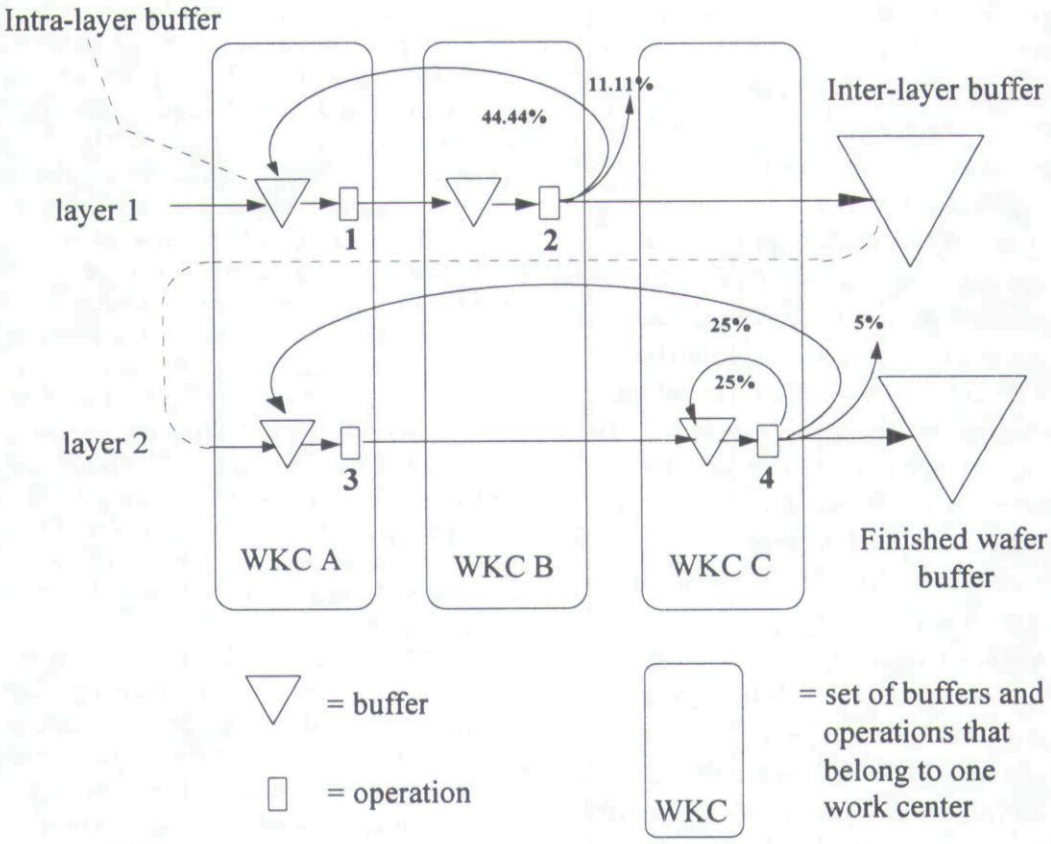

Figure 5: In the material flow in this wafer fab with three work centers, two layers, and four operations, operations in the same layer are aligned horizontally and operations in the same work center are aligned vertically. We indicated material flows, rework, and scrap, along with the corresponding probabilities.

each layer. The distinction between these buffers is conceptual. For instance, the physical location of the intralayer buffers for operations 1 and 3 and the interlayer buffer of layer 1 could be a single storage space that is physically located near work center A.

To aggregate the operations that belong to a single layer into a single production stage, we aggregate the planned yields of the operations within a layer into the *aggregated layer yield*. In addition, we aggregate the WIP in the intralayer buffers into the *aggregated layer inventory*. We now show how to aggregate the planned yields and the WIP within a layer.

Aggregated Layer Yield

For any operation i that belongs to layer k, where $i = 1, 2, 3, 4$, and $k = 1, 2$, let b_i^k be the probability that a wafer, starting from operation i (that belongs to layer k), will reach the interlayer buffer located immediately after layer k. By assuming that the planned yields are independent, we can compute these probabilities by modeling the flow of a wafer within layer k as an absorbing Markov chain. (Here we assume that the yields at different operations are independent. At the IBM wafer fab, the actual yield information provided by the engineers shows that the correlations of the yields at different operations are low.

Hence, this assumption is quite reasonable.) Specifically, we consider each operation as a state, scrap as an absorbing state, and exit to the interlayer buffer k as another absorbing state. In this case, we can compute the probability b_i^k for each operation at each layer. (The details of this simple computation are given in Theorem 3.3.7 on page 52 by Kemeny and Snell [1960]. Details are omitted.) In our example, it can be easily shown that $b_1^1 = 0.8$, $b_2^1 = 0.8$, $b_3^2 = 0.9$, and $b_4^2 = 0.9$. These probabilities enable us to aggregate the planned yields of the operations within a layer into the aggregated layer yield. Let y^k be the aggregated yield of layer k (that is, the probability that a wafer, starting from the beginning of layer k, will reach the interlayer buffer k without being scrapped). Then $y^k = b_j^k$, where operation j is the first operation in layer k. In our example, the yield of layer 1 is given by $y^1 = b_1^1 = 0.8$, and the yield of layer 2 is given by $y^2 = b_3^2 = 0.9$. Since the aggregated layer yields depend on the planned yields, they are updated at the beginning of each time period.

Aggregated Layer Inventory

At the beginning of each time period, we aggregate the WIP in the intralayer buffers (that belong to layer k) into aggregated inventory for the interlayer buffer located immediately after layer k. Let $w^k(t)$ be the inventory level at the interlayer buffer located immediately after layer k at the beginning of period t. Let $I_i^k(t)$ be the work-in-process inventory in the intralayer buffer located in front of operation i that belongs to layer k at the beginning of period t. It follows from the definition of b_i^k that out of these $I_i^k(t)$ wafers, $b_i^k I_i^k(t)$ wafers are expected to complete all operations in layer k successfully. In our example, the expected aggregated inventory level for layer 1, denoted by $B^1(t)$, is equal to

$$B^1(t) = b_1^1 I_1^1(t) + b_2^1 I_2^1(t) + w^1(t). \quad (1)$$

Similarly, the expected aggregated inventory level for layer 2, denoted by $B^2(t)$, can be written as

$$B^2(t) = b_3^2 I_3^2(t) + b_4^2 I_4^2(t) + w^2(t). \quad (2)$$

Using the aggregated yield and the expected aggregated inventory level of each layer, we can model the operations for processing a layer as a production stage. For example, the two-layer wafer fab in Figure 5 can be modeled as an aggregated system with two production stages. A buffer located immediately after each stage stores the aggregated inventory of each layer (Figure 6). In addition, the yield of each stage is the aggregated yield of the corresponding layer.

Planned Cycle Time for Each Stage

The aggregate model describes a serial production system that faces uncertain

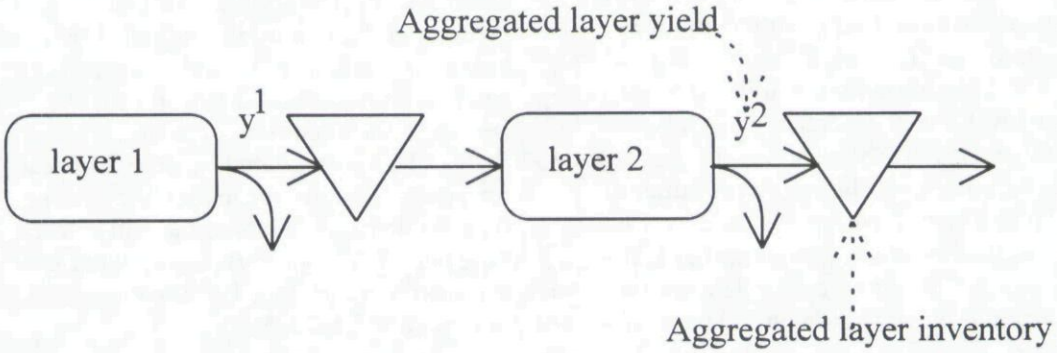

Figure 6: This represents the aggregate model for the wafer fab in Figure 5, along with the aggregated layer yields and the aggregate layer inventories.

yield and uncertain demand. We can adapt the model developed by Tang [1990] for serial production systems to develop production release decisions for this aggregated system. To use Tang's production release rule for each stage (or layer), we need to specify the time to release the wafers and the target work-in-process inventory needed to meet the future demand. In Tang's model, it is assumed that the cycle time of each stage is equal to one period. However, the cycle time of each stage in our aggregated system might not be equal to one period. In our system, the planned cycle time of a stage (or a layer) is equal to the sum of the planned cycle time of the work centers within the layer and the planned safety time. The planned cycle time of a work center is equal to the estimated amount of time that a wafer spends at that work center in that layer, which includes both the time that a wafer waits in the intralayer buffer (including the time due to rework, machine failure, setup time, and waiting time to form a complete batch for the next operation) and the actual processing time. The planned safety time is equal to the estimated run-out time of the safety stock in the interlayer buffer.

To set the safety stock and the planned safety time for layer 2 (Figure 5), let τ_A^2 be the actual cycle time of work center A in layer 2, where τ_A^2 is a random variable. Define τ_C^2 in a similar way. Let D_i be the demand in period i. For simplicity, we assume that D_i, $i = 1, \ldots,$ are *i.i.d.* random variables. Suppose we utilize the order-up-to level policy described by Peterson and Silver [1979]. Then we can set the planned safety stock for the interlayer buffer after layer 2 to $n\sigma_2$, where n is the safety level specified by management and σ_2 is the standard deviation of the demand over the actual cycle time of layer 2. In this case, $\sigma_2 = \sqrt{\text{Var}\,(\sum_{i=1}^{\tau_A^2+\tau_C^2} D_i)}$. The planned safety stock, $n\sigma_2$, is expected to be depleted after $E(\min\{\pi: \sum_{i=1}^{\pi} D_i > n\sigma_2\})$ periods. For simplicity, we approximate this quantity by setting the planned safety time S^2 to $n\sigma_2/E(D_i)$. It is easy to check that σ_2 decreases as the variance of the actual cycle time of the work center decreases. Hence, we can reduce the planned safety stock and the planned safety time of a layer when the variance of the actual cycle time of the work center decreases [Graves 1988].

So, at the beginning of a time period, let t_A^k, t_B^k, t_C^k be the planned cycle time of work center A, B, and C within layer k and let S^k be the planned safety time for the interlayer buffer located after layer k, where $k = 1, 2$. These times are based on actual historical data and are updated at the beginning of each time period. Since layer 1 requires operations performed in work centers A and B, the planned cycle time for layer 1, denoted by c^1, is

$$c^1 = t_A^1 + t_B^1 + S^1. \tag{3}$$

Similarly, the planned cycle time for layer 2, c^2, can be written as

$$c^2 = t_A^2 + t_C^2 + S^2. \tag{4}$$

Target Inventory for Each Stage

To adapt Tang's production release rule, we need to specify the target inventory for each stage (or layer). This target inventory serves as the benchmark for the WIP at each layer. (Since each stage in the aggregated system corresponds to a layer of operations in the wafer fab, the target inventory for a stage corresponds to the target level for the aggregated WIP for that layer.) At the beginning of period t, let $D(t + j)$ be the "updated" demand forecast for period $t + j$. In our model, we assume that the demand occurs at the end of each time period. We set $T^k(t)$, the updated target inventory for stage k at the beginning of period t, as follows:

$$T^1(t) = \sum_{j=c^2}^{c^2+c^1-1} D(t + j)/y^2 \tag{5}$$

$$T^2(t) = \sum_{j=0}^{c^2-1} D(t + j). \quad (6)$$

Essentially, the target inventory for each layer is equal to the cumulative demand forecasts (adjusted for the aggregated yields of the succeeding layers) in the periods that correspond to the planned cycle time of that layer. The planned cycle time of each layer includes the planned safety time, and the planned safety time accounts for the demand uncertainty. Thus, the target inventory for each layer includes the necessary safety stock to cope with demand uncertainty.

The Production Release Decision for Each Layer

Given the planned cycle time c^k and the target inventory $T^k(t)$ for layer k, we can obtain the production release decision for the aggregated system at the beginning of period t. Let $X^k(t)$ be the number of wafers to be released from the interlayer buffer to layer k during period t, where $k = 1, 2$. Our production release decisions can be expressed as follows:

$$X^1(t) = D(t + c^1 + c^2)/(y^1y^2) + (T^1(t) - B^1(t))/y^1 + (T^2(t) - B^2(t))/(y^1y^2). \quad (7)$$

$$X^2(t) = D(t + c^2)/y^2 + (T^2(t) - B^2(t))/y^2. \quad (8)$$

The production release decision consists of a push component and a pull component. The first term in (7) depends on the total planned cycle time for both layers, $c^1 + c^2$, and on the effective aggregated yield for both layers, y^1y^2. It $(D(t + c^1 + c^2)/(y^1y^2))$ represents the number of wafers to be released in period t for layer 1 so that the expected number of wafers to be completed in period $t + c^1 + c^2$ is equal to $D(t + c^1 + c^2)$. Hence, this term corresponds to the push component because it is based on the demand forecasts and the planned cycle time.

The second and the third terms are intended to restore the interlayer buffer inventory $B^k(t)$ back to its target $T^k(t)$. They compensate for the deviations of the aggregated layer inventory from the respective target layer inventory at each stage downstream from layer 1. (These deviations are adjusted by the aggregated layer yields.) For instance, if the aggregated layer inventory at layer 2, $B^2(t)$, is above (below) its target $T^2(t)$, the second term will decrease (increase) the number of wafers to be released at layer 1 (that is, upstream from layer 2). The third term can be interpreted in a similar way. Therefore, the second and the third terms correspond to the pull component of the production release decisions. (Denardo and Tang [1992] and Tang [1990] discuss this class of linear production control policies in detail.)

The production release decisions specified in (7) and (8) enable the managers to specify the release quantities ($X^1(t)$ and $X^2(t)$) as quotas for the operators to meet in period t. (We do not encourage the operators to exceed these quotas because early completion would make it difficult to identify the work center at which we can reduce the planned cycle time in the future.) If the quotas are met in each period and the aggregated layer yields are accurate, the layer inventories will be restored to their targets, the demands will be met, and the average actual cycle time will be equal to the planned cycle time. This observation suggests that our production release decisions can be used to control the WIP in the system so that the actual cycle time is close to the planned cycle time.

The Production Dispatch Decision for Each Work Center

The production release decision is made only at the beginning of each time period, while production dispatch decisions are made at various points within each time period. Because of the reentrant nature of the process, there are wafers from different layers waiting to be processed at a work center at any time. Our production dispatch decisions prioritize the wafers to be processed at a work center so that the

wafers will be completed according to schedule. The priority is based on an estimated tardiness of the wafers waiting to be processed in front of the work center. Specifically, the estimated tardiness of a wafer is the difference between its estimated completion time and its estimated due date (that is, the time when it is expected to be needed to satisfy demand). Both estimates are updated at the beginning of each period.

For example, in Figure 5, wafers are waiting in the intralayer buffers in front of operation 1 and operation 3 to be processed by work center A at time z within time period t. By using the updated information on the process yields at the beginning of period t, we can check that the total expected number of completed wafers that can be generated from the work-in-process inventories that are downstream from operation 1 is equal to $I_2^1(z)b_2^1y^2 + w^1(z)y^2 + B^2(z)$, where $I_2^1(z)$ is the work-in-process inventory in the intralayer buffer located in front of operation 2 at time z, $w^1(z)$ is the interlayer buffer inventory for layer 1 at time z, and $B^2(z)$ is the aggregated layer inventory for layer 2 at time z. We can compute $B^2(z)$ in a way similar to that in (2). Since the demand occurs only at the end of each period, we can determine d_1 (the estimated due date for the wafers that are waiting in the intralayer buffer for operation 1) by comparing this amount with the demand forecasts in future periods. In this case, the estimated due date d_1 (measured in terms of time periods) is given as follows:

$$d_1(z) = Min\ \{\tau: \sum_{s=t}^{\tau} D(s) > I_2^1(z) \cdot b_2^1 \cdot y^2 + w^1(z) \cdot y^2 + B^2(z)\}. \quad (9)$$

This estimated due date d_1 can be interpreted as the expected depletion time (measured in time periods) for the inventories that are downstream from operation 1. By using the same approach, we can determine d_3, the estimated due date for the wafers that are waiting in the intralayer buffer for operation 3, as

$$d_3(z) = Min\ \{\tau: \sum_{s=t}^{\tau} D(s) > I_4^2(z) \cdot b_4^2 + w^2(z)\}. \quad (10)$$

For each of the intralayer buffers, we compare the estimated due date with the estimated completion time of those wafers waiting in the intralayer buffer. For instance, the estimated completion time of those wafers waiting for operation 1 is equal to $[z + (c^1 + c^2)]$. The estimated tardiness for those wafers that are waiting to be processed at work center A for operation 1, denoted by $u_1(z)$, is given by

$$u_1(z) = [z + (c^1 + c^2)] - d_1(z). \quad (11)$$

Similarly, we can express $u_3(z)$, the estimated tardiness for those wafers that are waiting to be processed at work center A for operation 3, as

$$u_3(z) = [z + c^2] - d_3(z). \quad (12)$$

In this case, our production dispatch decision follows a myopic policy in which a higher priority will be assigned to an operation that has a higher value of estimated tardiness. For example, operation 3 will be given a higher priority if $u_3 > u_1$. Hence, our production dispatch decision aims to complete the wafers according to the planned schedule by assigning a higher priority to tardy wafers. The production dispatch decisions provide guidelines only for determining the order in which different types of wafers should be processed. The operator decides whether to process partial batches or to wait (in order to process full batches).

References

Blackburn, J. 1991, *Time Based Competition*, Richard D. Irwin, Homewood, Illinois.

Burman, D. Y.; Gurrola-Gal, F. J.; Nozari, A.; Sathaye, S.; and Sitarik, J. P. 1986, "Performance analysis techniques for IC manufactur-

ing lines," *AT&T Technical Journal*, Vol. 65, No. 4, pp. 46–57.

Chen, H.; Harrison, J. M.; Mandelbaum, A.; Van Ackere, A.; and Wein, L. M. 1988, "Empirical evaluation of a queueing network model for semiconductor wafer fabrication," *Operations Research*, Vol. 36, No. 2, pp. 202–215.

Cory, L. 1986, "Just-in-time approach to IC fabrication," *Solid State Technology*, Vol. 29, No. 5, pp. 177–179.

Demeester, L. L. and Tang, C. S. 1993, "Reducing cycle time at an IBM wafer fabrication facility," unpublished manuscript, Anderson Graduate School of Management, UCLA.

Denardo, E. V. and Tang, C. S. 1992, "Linear control of a Markov production system," *Operations Research*, Vol. 40, No. 2, pp. 259–278.

Denardo, E. V. and Tang, C. S. 1994, "Effective control of a stochastic production system: Stability and inferability of proportional restoration rules," working paper, Anderson School of Management, UCLA.

Glassey, R. and Resende, M. 1988a, "A scheduling rule for job release in semiconductor fabrication," *Operations Research Letters*, Vol. 7, No. 5, pp. 213–217.

Glassey, R. and Resende, M. 1988b, "Closed loop job release control for VLSI circuit manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 1, No. 1, pp. 36–46.

Graves, S. 1988, "Safety stocks in manufacturing systems," *Journal of Manufacturing and Operations Management*, Vol. 1, No. 1, pp. 69–84.

Hill, T. 1988, *Manufacturing Strategy*, Macmillan, London, England.

Kemeny, J. and Snell, J. 1960, *Finite Markov Chains*, Van Nostrand, Princeton, New Jersey.

Martin-Vega, L.; Pippin, M.; Gerdon, E.; and Burcham, R. 1989, "Applying just-in-time in a wafer fab: A case study," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 2, No. 1, pp. 16–22.

Mead, C. and Conway, L. 1980, *Introduction to VLSI Systems*, Addison-Wesley, Reading, Massachusetts.

Peterson, R. and Silver, E. 1979, *Decision Systems for Inventory Management and Production Planning*, John Wiley and Sons, New York.

Reinhard, D. K. 1987, *Introduction to Integrated Circuit Engineering*, Houghton Mifflin Company, Boston, Massachusetts.

Tang, C. S. 1990, "The impact of uncertainty on a production line," *Management Science*, Vol. 36, No. 12, pp. 1518–1537.

Ralph Ahlgren, Superintendent of Manufacturing, IBM, 5600 Cottle Road, San Jose, California 95193, writes "The modifications made to the manufacturing control system have enabled us to extend our system's ability to control our shop floor product movement and yielded a net reduction in cycle time. Dr. Tang and Mr. Demeester's paper is an accurate portrayal of a comprehensive application of theory to real manufacturing problems."

Copyright 1996, by INFORMS, all rights reserved. Copyright of Interfaces is the property of INFORMS: Institute for Operations Research and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.