1. Introduction

The primary purpose of factory floor scheduling is to ensure the factory production plan is fulfilled. As described in the previous chapter, the factory production plan specifies a target output schedule to be achieved. If this schedule is achieved, then factory management is doing its part to ensure customer deliveries will be on-time and that marketing will have the supply of build-to-plan products they requested.

As a practical matter, the overall approach to planning and execution of a complex production process such as semiconductor fabrication must be hierarchical. That is, the production plan is not expressed at a level of detail sufficient to explicitly guide process execution. Instead, the plan was based on a model of the capacity consumption and the elapsed cycle time associated with execution. The production planning function developed a plan believed to be feasible with respect to capacity and cycle time. It is the job of factory floor scheduling to break the production plan down into detailed directions that guide the operation of individual machines, transport carriers, robots, workers, etc.

Factory floor scheduling starts with the production plan’s targeted output schedule as a primary input and strives to achieve that output schedule. Other aspects of the production plan may help guide the development of the factory floor schedule, e.g., the planned quantity and timing of releases of new lots (and the planned allocation of raw materials or intermediate product inputs to constitute those lots), and the projected utilizations of equipment or other resources.

As will be discussed, the factory floor scheduling methodology has an impact on the cycle times that are achievable. Thus a secondary purpose of factory floor scheduling is to enable the factory to achieve product cycle times and equipment utilizations of which it is capable.

2. Cycle Time Metrics

Before delving into scheduling logic, it is useful to delineate various metrics and definitions of cycle time as well as review industry performance. At the outset, it must be noted that cycle time is viewed differently by a customer placing product orders and by a manager of a manufacturing facility. From the customer’s viewpoint, cycle time is the elapsed time from when an order is placed until the ordered product is received. This perception of cycle time is perhaps best termed lead time. Lead time has a number of components: There is time consumed by data entry and paperwork for the customer order, by design effort if the product is custom, and there is time required to communicate the production order (and the production specifications if the product is custom) to manufacturing facilities. If the ordered product is not already available in finished goods inventory, there is the time to manufacture the product. If multiple manufacturing facilities are involved, there is the time consumed by interplant shipping of intermediate products. There may be a certain amount of safety time built into the quoted lead time to protect against
unforeseen design, manufacturing or shipping problems. And then there is the time required to ship the completed product to the customer.

Cycle time from the manufacturing manager’s point of view is the elapsed time from the launch of a manufacturing lot into a process flow until the lot completes the process flow. In some companies this time is termed the turn-around time or TAT for short. In some industrial engineering textbooks, this perception of cycle time is termed the manufacturing flow time.

Manufacturing cycle time includes all elapsed time until the manufacturing lot is completed: time to pass through each of the process steps, transport time between the steps, wait time at each step, and time the lot was placed “on hold” because of quality control alarms or issues.

Semiconductor fabrication facilities routinely tabulate statistics on manufacturing cycle times. It is common for managers to view statistics concerning the average cycle time for each process flow as well as an overall weighted-average cycle time for the manufacturing facility. Cycle time statistics are typically collected for each process step in each process flow. Typically, the cycle time statistic for a process step is computed as the average time from lot “track-out” (i.e., completion) of the preceding process step until lot track-out of the step in question.

We define the standard cycle time (\(SCT\)) of a process step as the minimum time required to pass one lot through the step without waiting or interference. That is, \(SCT\) includes transport time and process time, but not wait time or hold time. \(SCT\) is thus the minimum feasible lot cycle time. For this reason, in some companies it is termed the theoretical cycle time or the intrinsic cycle time. Standard cycle time of a process step has a fixed value that remains fixed until the process specification, the equipment or the lot size is changed. One can speak of the total standard cycle time of a process flow; it is defined as the sum of the standard cycle times of all the steps in a process flow.

Recall that we defined theoretical process time \((ThPT)\) of a process step as the minimum time between completions of consecutive lots for a machine in perfect working order. We also defined the average or effective process time \((EPT)\) considering the actual rate efficiency losses. In some companies, process time is termed tact time. (‘Tact’ is the German word for heartbeat, i.e., tact time expresses the pulse rate of the machine.) At this point, the reader might wonder: What is the difference between standard cycle time and theoretical process time? Indeed, for some types of equipment, \(ThPT\) and \(SCT\) are the same. But for many and perhaps most types of semiconductor manufacturing equipment, a lot is not done on a machine at the time processing can be initiated on a subsequent lot. That is, there is some parallel activity on the machine, i.e., the processing cycles of consecutive lots overlap. This is sometimes referred to as pipelining. For process steps performed by such machines, \(SCT\) is longer than \(ThPT\), i.e., the lot is resident on the machine for a time that is longer than the time the lot consumes capacity. There are even some machine types where \(SCT\) is shorter than \(ThPT\). For example, some deposition machines undergo an automated self-cleaning procedure after every lot is processed and before another lot can be processed. While the self-cleaning procedure is underway, the completed lot can depart for its next process step. Thus the self-cleaning procedure contributes to process time but not to standard cycle time.
There are various statistical approaches to tabulating cycle times. The different approaches provide different values for cycle time, so it is important to understand the differences and avoid the pitfalls of apples-to-oranges comparisons.

*Static cycle time* is the elapsed time from launch of a production lot into its process flow until the completion of the last step in its process flow. A common statistic viewed by managers is the average static cycle time for each process flow, developed by averaging the static cycle times for all lots departing the process flow in a recent observation period such as a week or a month.

The *dynamic cycle time* metric is developed as follows. First, dynamic cycle time for an individual process step is computed by averaging the cycle time over all lots that tracked out of the step in a given observation period such as a day (a 24-hour observation period). Here, “cycle time” is the elapsed time from track-out of the previous step to track-out of the step in question. The dynamic cycle times so calculated for all steps in a process flow are then summed to obtain the dynamic cycle time metric for the process flow.

Because the dynamic cycle time is computed by summing cycle times computed for the same observation period for every step, dynamic cycle time is not achieved by any particular lot. It does provide a current snapshot of cycle time performance over the entire process flow, and evidently for that reason it is termed “dynamic” cycle time.

The *throughput time* metric is developed as follows. For an individual process step in a given observation period such as a day (a 24-hour period) or a production shift (typically a 12-hour period), information is obtained concerning the WIP level at the start of the period (“BOH” – short for beginning-on-hand) and the WIP level at the end of the period (“EOH” – short for ending-on-hand). A crude estimate of the average WIP level during the period is a simple average, i.e., \((\text{BOH} + \text{EOH})/2\). The number of units tracked out of the process step (termed the “moves” during the observation period) also is obtained. The throughput time is an estimate of the time required to clear the average WIP at the step considering the moves rate achieved at that step, i.e.,

\[
\text{Throughput time} = \frac{\text{Average WIP level}}{\text{Moves}} = \frac{\text{BOH} + \text{EOH}}{2 \times \text{Moves}}.
\]

The reciprocal of the throughput time is termed the *WIP turns* achieved at the process step, i.e., the number of times the average WIP was cleared at the step per period, considering the achieved moves rate. Like the case of dynamic cycle time, throughput times for individual steps may be summed to obtain a metric of throughput time for the whole process flow.

These three metrics of cycle time for a process flow are *not* equivalent, and they will usually have different values. Most commonly, but by no means always, the metrics will compare as follows:

*Static cycle time > Dynamic cycle time > Throughput time*

The reasons for this: First, semiconductor fabrication is characterized by events that are infrequent but devastating to cycle time, e.g., a process step is shut down for 16 hours because of
an out-of-control alarm and difficulty encountered in restoring process stability; a critical photolithography scanner is down for 48 hours for a lens change-out, etc. Suppose such events have an overall frequency of, say, one event every 30 days. Suppose too that the total cycle time for the process flow is long, e.g., 40-50 days. Then for any given lot it is likely that some devastating event transpired while it was somewhere in the process flow. On the other hand, on most days, no such events transpire or have transpired in recent days, so cycle time performance looks good. Hence, on most days, the dynamic cycle time is good. But static cycle time is not as good. Second, by simply looking at BOH and EOH WIP levels, the throughput time calculation could be underestimating the average WIP level. Third, the throughput time metric implicitly downplays the impact of lot holds. Suppose there are a number of lots on long-term hold while other lots pass right through the step. The throughput time metric may look favorable, yet once the lots on hold are released they will contribute to a serious degradation of cycle time. That will show up in the dynamic cycle time metric but not show up in the throughput time metric.

Whether utilizing static cycle time, dynamic cycle time or throughput time metrics, most fab managers look at a normalization of the raw metrics. The two most common normalizations are (1) Cycle time per mask layer, whereby the cycle time for a process flow is divided by the number of layers of circuitry in the flow, and (2) The theoretical cycle time multiplier, whereby the measured cycle time for a process flow is divided by the total standard cycle time for the flow. The latter normalized metric is often expressed as an “X” factor. For example, a process flow running at “1.5X theoretical” means the cycle time multiplier is 1.5.

Cycle time per mask layer and the cycle time multiplier make more meaningful the comparison of cycle times of different process flows, different fabrication facilities, etc. Typically, a cycle time metric for an entire facility is a weighted average of the cycle times for the process flows in the facility, whereby the weights are the wafer starts in each flow or the wafer output in each flow in some observation period such as a week or a month.

Figure 1 displays monthly cycle time per layer scores for ten semiconductor fabrication facilities participating in the Competitive Semiconductor Manufacturing benchmarking survey conducted by U. C. Berkeley. The plotted points are weighted-average fab-wide scores for fabrication facilities built in the mid-1990s. Included are fabs located in Japan, Korea, Taiwan, Europe and the United States. Note that, generally, cycle time improves as a fab gets older. This reflects engineering efforts to reduce disruptions associated with process or equipment trouble as well as improvements in planning and scheduling methodology. But note too that for every fab there are some months where cycle time performance worsened. This could happen for a variety of reasons: First, it must be remembered that the scores are weighted averages across all process flows in the fabs. At an arbitrary point in time, volume could have been shifting out of a mature process flow whose problems were mostly worked out into a new process flow experiencing much more trouble. Second, management might have deliberately increased utilization levels in the fab. Or third, there might have been a serious disruption such as a major process quality issue in a certain month. After the fabs had been in operation four years or so, the cycle times per layer were all in the range of 1.3 – 2.6 days per mask layer, with the average somewhere between 2 and 2.3 days per layer. Note in particular the performance of fab M1; in its first two months of operation, it had the worst cycle time among the 10 fabs, about 3.5 – 3.7 days per mask layer.
Figure 1. Cycle Time Performance in CSM Survey Fabs
But three years later, it had the best cycle time performance. This fab made major improvements in its factory floor scheduling methodology that shall be discussed later.

Managers at some fabs producing make-to-order chips consider a different metric of cycle time: They look at a high percentile of the distribution of static cycle time. For example, the 90th percentile of static cycle time in a month is the cycle time such that 90% of the manufacturing lots output during that month achieved that cycle time or achieved a shorter cycle time. Figure 2 provides a histogram of the static cycle times for all lots output from a custom-chip wafer fab during the month of December, 1992. This cycle time is measured from a wafer inventory point just before the first metallization layer until fab output. (The nature of this fab’s business was that generic wafers were fabricated on a build-to-plan basis up to this inventory point, and then production was make-to-order beyond this inventory point. As may be seen, ten lots achieved a cycle time of 14 days, 20 achieved a cycle time of 15 days, and so on. While the average cycle time during this month was between 16 and 17 days, the 95th percentile of the cycle time distribution was at 23 days. For a 95% probability of on-time delivery, this company would need to quote a fabrication lead time of at least 23 days.

3. A Queuing System Perspective on Manufacturing

To better understand the trade-off described in the second paragraph of the introduction, it is helpful to have a little background about queuing systems. First, we note a simple but famous conservation law known as Little’s Law. Little’s Law applies to any production or service system with a stationary distribution for the arrival rate of work and a stable queue of work (i.e., the average completion rate of work equals the average arrival rate of work in the system). If we think in terms of a manufacturing context, the work arriving at a bank of machines consists of production lots. In words, Little’s Law, states

$$(\text{Avg. arrival rate of lots})(\text{Avg. time a lot is in the work station}) = (\text{Avg. number of lots resident at the work station})$$

If we denote the arrival rate by $\lambda$, the average time in the system by $w$ and the average number of customers (or lots) in the system by $L$, the classic notational form of Little’s Law is

$$L = \lambda w.$$  \hspace{1cm} (1)

Note, that, given a stable production rate (arrival rate), Little’s Law indicates that cycle time and WIP are proportional.

In manufacturing terms, we can restate Little’s Law as

$$(\text{Avg. production rate})(\text{Avg. cycle time}) = (\text{Average WIP level})$$

This conservation law applies to a work station, to an individual step performed at the work station, to a series of production steps performed at several work stations, to the entire process flow for a family of products, even to the entire factory.
Figure 2. Histogram of Lot Cycle Times

95th percentile = 23 days
Now consider Figure 3. The curve in Figure 3 plots the level of work-in-process (WIP) present at an arbitrary work station (i.e., a bank of parallel processing machines) vs. the utilization of availability of the machines in the work station. In their experience, nearly all factory managers have observed this trade-off between WIP and utilization. The same sort of trade-off exists between cycle time and utilization; we simply would need to re-scale the vertical axis by the reciprocal of the production rate, and then we would have the graph of cycle time vs. utilization.

The particular shape of the curve of WIP vs. utilization (or cycle time vs. utilization) depends on the amount of variability in the arrival and processing rates. To understand this, imagine customers walking into a bank staffed with a single teller. Suppose customers arrive once every two minutes – exactly. Suppose the teller finishes serving every customer in two minutes – exactly. And suppose the teller never needs to take a break. This is a system with zero variability. Every customer arrives to find the teller is free and ready to process his/her transaction. Such a system can achieve 100% utilization with no customer waiting and with a cycle time of two minutes. The case of zero variability in a factory is portrayed by the dashed right-angle curve in Figure 4.

Now let’s introduce some processing variability to our hypothetical bank example. Suppose that for some customers, the teller requires longer than two minutes to process the transaction. Or worse, suppose sometimes the teller needs to find the bank manager to get an approval, and sometimes the teller takes a bathroom break. In this case, some customers are going to wait. And as we increase the arrival rate of customers, more and more customers are going to wait, and they are going to wait longer. So cycle time rises super-linearly as utilization is raised. For the case of a factory with relatively low variability, this trade-off is illustrated by the “low-variability” curve in Figure 4.

Now let’s add some arrival variability to our bank example. Suppose the arrival distribution is roughly as follows: Many customers come into the bank during the noon hour, then it is pretty slow during midafternoon, and then there is a final big surge of customers after 4pm. If we wanted to ensure our teller was at 98% utilization all afternoon, we would need to have so many customers arrive during the noon hour that some of them would still be waiting at 4pm! And for that scenario, cycle time is explosive. This case is suggested by the “high-variability” curve in Figure 4.

One more feature of the curves in Figures 3 and 4 is worth observing. These hyperbolas describing cycle time vs. utilization of availability do not obey a law of averages. That is, suppose the average utilization of availability of a particular resource is 80%. Imagine operating half the time in long observation period at 90% and half the time at 70%. Given the shape of the curves, the cycle time when operating at 70% will be somewhat faster than the cycle time when operating at 80%, but the cycle time when operating at 90% will be much longer than when operating at 80%. Thus operating half the time at 90% and half the time at 70% will result in an average cycle time significantly longer than if operation was at 80% utilization of availability throughout the period. On the other hand, suppose the average utilization of availability was
Additional WIP cannot increase throughput \((U = A)\)
Figure 4. Several Cases of the Trade-Off of WIP vs. Utilization
40%, and we compare that to the case of operating half the time at 30% utilization and half the time at 50% utilization. The case of time-varying utilization will have a somewhat longer average cycle time than the case of constant utilization, but the difference will not be nearly as great as it was for the 80% vs. 70%/90% case. This suggests that controlling variability in the utilization of the high-utilization resources is more important than controlling variability of low-utilization resources.

The cycle time – utilization trade-off in our hypothetical bank may make the reader wonder: What are the sources of variability in a factory? A partial list is as follows:

- Fluctuating production workloads, particularly if the arrival rate of lots sometimes approaches or even exceeds the capacity of the available machines. As noted above, this is most concerning for high-utilization resources.
- Machine down time, process out-of-control stoppages, operator absence
- Inflexible or disqualified machines, i.e., WIP available and machines available but the available machines are not qualified to process the available WIP
- Large-batch production runs, whereby lots wait for setups or qualification runs
- Changing production priorities, “hot” lots

Now let’s return to the manufacturing-context version of Little’s Law. If we expand terms, we can isolate the contributions of waiting time and standard cycle time to the WIP level:

\[
WIP = (\text{Production rate})(\text{Cycle time})
\]

\[
WIP = (\text{Production rate})(\text{Waiting time} + \text{Standard cycle time})
\]

\[
WIP = (\text{Production rate})(\text{Waiting time}) + (\text{Production rate})(\text{Standard cycle time})
\]

We make the following identifications with the right-hand-side terms: The first term is called the buffer WIP, i.e., the portion of WIP not undergoing processing or transport activity. The second term is called the active WIP, i.e, the portion of WIP being processed or transported. That is,

Total WIP = Active WIP + Buffer WIP
= (Production rate)(Standard cycle time) + (Production rate)(Waiting time). \hspace{1cm} (2)

The buffer WIP is a consequence of the variability in the system; if there were no variability, all WIP would be active, just as in our hypothetical bank example.

In general, Little’s conservation law also applies to just the queuing portion of the customer time in the system, i.e.,

\[
(\text{Avg. arrival rate})(\text{Avg. time in queue}) = (\text{Avg. number of customers in queue}),
\]

and it equally applies to just the service portion of the time in system, i.e.,

\[
(\text{Avg. arrival rate})(\text{Avg. standard cycle time}) = (\text{Avg. number of lots undergoing service}).
\]
With thoughtful design of the scheduling logic, one may be able to mitigate the impact of some of the variability described above, thereby reducing the buffer WIP and thereby making the factory capable of a shorter cycle time. Imagine a scheduling logic that was able to maintain lot arrivals to a work station at a pace the available and qualified machines could handle. Imagine a different logic that gave no heed to how many machines were available and qualified. We might expect the former logic to provide better cycle time performance than the latter. We will return to this point later.

4. The Theory of Constraints

Equation (2) implies a certain WIP level – specifically, the Active WIP – must be present in order to maintain the targeted production rate. If the WIP level at a step or in a series of steps drops below the Active WIP level for that step or series of steps, the target production rate cannot be maintained.

The most concerning steps in this regard are steps performed by resources that in the production plan were targeted to operate at their maximum allowed utilizations of availability. We refer to such resources as bottlenecks. In contrast, WIP levels at steps performed by resources with slack capacity are less concerning, because such resources are capable of operating faster than the target production rate. If steps performed by non-bottleneck resources fall behind the targeted lot completions, such resources still have some ability to catch up. Even if operated above the targeted utilization (while catching up), the impacts on cycle time are relatively minor. On the other hand, resources planned to run at their allowed capacity or close to allowed capacity have little or no ability to catch up. Worse, trying to run such resources beyond their allowed capacity is likely to sharply extend cycle times. Thus there should be concern with controlling the variability of lot arrivals to steps performed by bottleneck equipment.

The Theory of Constraints espoused by Goldratt et al stresses an identification of bottlenecks inhibiting the generation of more salable output. The theory emphasizes efforts to increase bottleneck throughput and to increase the capability of the bottleneck for more throughput. In particular, the theory emphasizes regulation of workloads for the bottlenecks. If there is variability in arrivals to steps performed by a bottleneck resource, then maintaining a buffer WIP among such steps can be helpful for ensuring the WIP level at such steps does not drop below the Active WIP level. Moreover, the scheduling logic applied at steps performed by non-bottleneck resources should be attuned to maintaining sufficient WIP in front of the bottlenecks.

5. Impact of the Flexibility of Machines

A common situation in semiconductor manufacturing is one in which a work station consists of many parallel machines, and they work station performs many different process steps, but for any give process step only a few machines are qualified to perform that step. Scheduling is complicated by the need to intelligently allocate available WIP to available machines. It is easy to make an allocation which has WIP left over and machines left over, but the unassigned machines are not qualified to process the unallocated WIP.
Machine down time or machine disqualifications have an exacerbated impact when machines are inflexible. To illustrate, suppose two machines are qualified to perform a certain step. Suppose the overall production rate of the work station is 240 lots per 12-hour shift, and there are 10 machines working in parallel in the work station. Suppose the production rate of the step in question is 48 lots per shift, and suppose the process time is 0.4 hours per lot for all steps. Suppose the machine availability averages 90%.

Now suppose one of the two machines qualified to perform the step in question has been down for a long time, and the WIP of that step has built up to 16 lots. Suppose the normal WIP level for that step is 2 lots. Suppose the WIP level is normal for all other steps performed by these machines. How long will it take to reduce the WIP of that step to normal?

Note that 240*0.4 = 96 hours of workload completed per shift is required to keep up with the overall production rate. The machines have 85% availability, so, on average, the 10 machines can perform 10*0.85*12 = 102 hours of work per shift. If allocated equally, the workload per machine per shift is 9.6 hours and the surplus capacity per machine is (102 – 96)/10 = 0.6 hours per shift. Note also that 48*0.4 = 19.2 hours of workload is required to keep up with the production rate of the step in question. This is exactly 20% of the total required workload to keep up with overall production rate, so if the two machines just process lots at the step in question, they will be doing their “fair share.” For that step, the two qualified machines can perform 2*0.85*12 = 20.4 hours of work per shift. This leaves 20.4 – 19.2 = 1.2 hours of machine time per shift to work off surplus WIP of the step in question. The surplus WIP constitutes (10 – 2)*0.4 = 3.2 hours of work. Thus it will take 3.2/1.2 = 2.67 shifts to bring the WIP level back to normal.

Now suppose one of the other 8 machines becomes qualified to perform the step in question, i.e., there are now three machines qualified to perform the step in question. In that case, there would be 1.8 hours of surplus capacity per shift that could be devoted to the step in question. It would then take 3.2/1.8 = 1.78 shifts to bring the WIP level back to normal, i.e., 33% shorter. Thus cycle time is strongly influenced by the number of qualified machines when machine or process availability is variable.

6. Scheduling Paradigms

There are at least six paradigms underlying scheduling systems that have been implemented in industrial practice. The gist of each of these paradigms is explained as follows. In a typical work station to be scheduled, the station performs multiple manufacturing steps from multiple process flows. Multiple manufacturing lots are in queue, awaiting processing. Workers (or robots) face choices about which lot to select for processing next from among the lots waiting to be worked on. There may be multiple machines or workers available for performing processing and multiple lots waiting, so the selection problem can be more of an allocation problem (i.e., which lots assigned to which machines).

1. Negative Scheduling. The Negative Scheduling paradigm provides information to workers about what lots not to work on. When an excessive level of WIP has accumulated at a certain manufacturing stage, then lots at the preceding stage are de-prioritized or even prohibited from
being worked on. In effect, an absolute cap is placed the WIP level between stages. A “stage” in this context may be an individual manufacturing step or a series of steps. The application of negative scheduling is sometimes referred to as a pull system or a “just-in time” system because lots are not processed unless the workers or machines in the subsequent stage are “ready” for the lots in the sense that the WIP level at the subsequent stage is not too high.

The Negative Scheduling paradigm was pioneered in the Japanese automobile industry, in particular, at Toyota Motors. In an automobile assembly line, machines are utilized in a serial fashion, i.e., each machine type is visited once. In an assembly line the primary resource is labor, a relatively flexible resource in the sense that labor can be shifted between stages. Under this paradigm, when the WIP at a certain point in the line reaches a limit, a signal is sent to the upstream stage to cease production. Workers at that stage stop work and can move to assist the downstream stage experiencing trouble.

The original mechanism for signaling downstream WIP levels was through the use of tickets attached to each work piece. A fixed total number of tickets was allowed to circulate between consecutive stages. When the downstream stage began work on a piece of WIP, the ticket was detached and delivered to the upstream stage. The upstream stage was required to possess a detached ticket in order to commence work on another piece of WIP. If the downstream stage stopped completely, note that the upstream stage would eventually run out of detached tickets and therefore also stop. This stoppage encouraged an investigation of the line blockage, a reallocation of workers and a rebalancing of the production line. The Japanese word for ticket is kanban. Thus a Negative Scheduling system is also referred to as a Kanban control system. The general managerial strategy is to gradually reduce the number of kanbans in circulation, forcing the line to learn to maintain the production rate with a lower WIP level. To maintain that rate with a lower level of WIP, manufacturing staff and management will have to innovate and implement changes that reduce variability, thereby enabling more linear workflow and hence shortened cycle times.

Note that the Negative Scheduling paradigm is most meaningful in the case that manufacturing is repetitive, i.e., in the case a stream of output is being generated.

2. Lot Dispatching. The Lot Dispatching paradigm provides information to workers about which lots are most urgent to work on. The underlying idea is that, as machines finish what they were working on or return from maintenance, the operators can reference a Dispatch Priority List to find out which lots are most urgent. Choosing which lot to process next when a machine becomes available is termed the dispatching decision.

There are two variants of the Lot Dispatching paradigm. Under the Project Scheduling paradigm, a calculation is made that translates a target date for project completion into target schedules for every required project task. This calculation is made in advance of performing the project. Applying this paradigm to a manufacturing context, each new manufacturing lot is assigned a target output date (the lot’s due date), and the output due date is translated into due dates for each manufacturing step to be performed on the lot. The step due dates are based on planned or allowed cycle times at each step. In effect, each manufacturing lot is viewed as a separate project.
A *slack score* is calculated for each waiting lot, equal to the due date for the lot less the estimated remaining cycle time to reach the end of the process flow less the current date. Considering all the lots waiting at a work station, priority is given to the lot furthest behind its schedule (or least ahead if none are behind). This is known as the *least slack* dispatching rule.

Note that the least slack rule reduces variability in lot cycle times. Lots running ahead of the target cycle time are de-prioritized, while those running behind the target cycle time are given priority. The variance in total cycle time is thus reduced compared to other priority rules.

We see that, under the Project Scheduling paradigm, lots are prioritized based on planned schedules for the lots and comparison of current lot progress to its schedule.

The other variant of the Lot Dispatching paradigm is the *Job-Shop Scheduling paradigm*. Job-Shop Scheduling also is a methodology for dispatching lots awaiting processing at a work station. A “job shop” is a custom manufacturing environment in which every manufacturing lot is a unique product or job. Job Shop Scheduling methodology includes the due-date prioritization of project scheduling, but it often utilizes a different metric for measuring lot lateness: the so-called *critical ratio* (defined as the ratio of the time until due date divided by the estimated time to reach the end of the process flow). The lot with the smallest critical ratio is prioritized first. Job-Shop Scheduling also embraces other kinds of priority rules, e.g., least work in next queue (reflecting the spirit of the Negative Scheduling paradigm), shortest processing time first (reflecting the notion that if productivity is measured in terms of jobs completed, then performing the short jobs first provides the best report card), same-setup compatibility (minimizing setups and thus maximizing productivity), and same-batch compatibility (striving for large-as-possible batches on batch equipment and thus maximizing productivity). There is little theory available about which rule to use in which situation.

3. *WIP Management*. The WIP Management paradigm works to achieve a target output schedule for the factory, expressed as quantities of wafers of each product due in periods such as days or weeks. Using target cycle times for each step, the target output schedule is translated into target output schedules for each product/step. Subtracting downstream WIP from the target schedule for a product/step determines the remaining production still required for that product/step. Collectively, the targets so calculated for all product/steps performed in a work station become the production goals for that work station. Suppose, for example, the goal for a particular product/step is three lots completed by the end of the day, and suppose five lots of that product/step are available in the on-hand WIP. Under the WIP Management paradigm, there is little concern for which three of the five lots are selected for processing, as long as three are indeed processed (thereby fulfilling the goal).

The portions of the production goals for a work station that are supported by on-hand WIP can serve as the input data for *Execution Planning*, described next.

One variant of the WIP Management paradigm is so-called *Cutoff Scheduling*, commonly practiced at Silicon Valley fabrication lines in the early decades of the industry. To illustrate, suppose there is a monthly target for the production output of a certain product. As the end of the
month draws near, line supervisors will assess the output to date and the current WIP in the line
to identify the lot of that product most distant from fab out that needs to make it to fab out before
the end of the month. Supervisors will then manage the operators and technical staff to make
sure all of the WIP back to that lot makes it out before the end of the month.

When applied periodically such as in the case described above, Cutoff Scheduling tends to favor
dispatching lots near the end of the line late in a period, perhaps slowing the movement of WIP
near the front of the line.

4. Execution Planning. Instead of generating priority lists or announcing inhibitions of certain
product/steps, the Execution Planning paradigm entails the preparation of explicit execution
plans for each work station. These plans specify which lots will run on which machines in what
order. The execution plans may include not only production tasks but also non-production tasks
such as equipment maintenance and re-qualification, engineering work, etc. Execution plans are
conveniently summarized in terms of Gantt Charts prepared for each machine or other resource
of interest. The Gantt Chart specifies the timing and sequence of every tasks to be undertaken by
the resource. Note that the Execution Planning paradigm goes well beyond the Negative
Scheduling, the Project Scheduling and the Job-Shop Scheduling paradigms because it specifies
a complete execution procedure, i.e., a true schedule. However, because the supply of WIP to
work on may be uncertain, and because unforeseen events may arise that are not anticipated
when the execution plan was prepared, the execution plan may become stale. Either it must be
frequently re-computed, or else the information in the plan must be treated as goals to generally
guide production and non-production activity. The logic utilized to generate execution plans
could involve formal optimization, heuristics, or constraint-satisfaction techniques. The priorities
underlying the development of the execution plans could be lot due dates or target output
schedules.

5. Agent-Based Scheduling. A relatively new paradigm is so-called Agent-Based Scheduling.
This paradigm involves the simulation of market mechanisms applied in the factory. Under this
paradigm, equipment or work stations “sell” appointments to lots and non-production tasks. Each
lot seeks to make a “profit” based on its timeliness of completion and what it spent to be
processed. Each lot and each resource has an “agent” that handles the booking of appointments.
Agent-Based Scheduling falls under the Execution Planning paradigm because Gantt Charts are
prepared for all resources, i.e., a complete, explicit schedule is prepared. Its departure point is the
simulation of internal, artificial markets to resolve conflicts and make scheduling decisions. It is
somewhat reminiscent of Job-Shop Scheduling, except lots select machines rather than the other
way around. While a few implementations have been reported in other industries, implemented
examples of Agent-Based Scheduling in the semiconductor industry are unknown. Perhaps this
reflects the challenge of setting up an effective market mechanism. (The reader may recall
reading about the problems encountered in setting up deregulated markets for energy.)

6. Release Scheduling. Release scheduling concerns the determination of the timing of release
into the factory of new production lots. It can incorporate elements of any of the preceding
paradigms. For example, it can reflect the Negative Scheduling paradigm (i.e., releases are
allowed when the current WIP level, or the workload associated with the current WIP, is not too
high), the Project Scheduling paradigm (i.e., the target date for release depends on the target
output schedule and the target cycle time), or the Execution Planning paradigm (i.e., a model of factory workloads associated with current WIP and with new lot releases is prepared, and a determination is made of the best time to start each lot). The production planning methodology discussed in the previous chapter is a form of release scheduling, although typically developed on a coarse time scale such as weeks. A finer time scale such as production shifts may be effective for refined scheduling the releases of lots in the production plan.

One might imagine that a single paradigm must be selected for developing a scheduling system. That is not the case at all. For example, it is quite feasible to have Negative Scheduling elements (such as WIP limits) and at the same time have Project Scheduling or Job-Shop Scheduling elements (such as dispatch lists), WIP management elements (such as production targets), and Execution Planning elements (goals for lot completions by each work station during a short horizon such as production shift). Indeed, many fabrication facilities utilize scheduling methodology reflecting multiple paradigms. Nonetheless, often one paradigm is more strongly embraced than the others. Some examples follow.

7. Case Studies in Cycle Time Reduction

Company A – A Focus on Reducing Variability

Back in 1993-1994, a fabrication facility belonging to company A was fabricating 4MB and 16MB DRAMs on 150mm wafers. A corporate initiative was launched to reduce TAT. At the fabrication facility a task force consisting of the regular engineering and management staff was set up to work on TAT reduction. Over a two-year period, average fab TAT was reduced from 77 days to 41 days (the latter figure equivalent to about 2 days per mask layer). The reductions in cycle time included a reduction in wait time of 30 days and a reduction in standard cycle time of 6 days.

The TAT Task Force adopted the following strategy:

- Establish TAT goals for every process step
- Monitor actual TAT vs. goal TAT, and monitor the variation in production rate at each process step
- In each month, identify the four worst-performing steps and try to fix them within one month

The Task Force found that the biggest single issue was batching performed by the operators and their leaders (i.e., their supervisors). Isolated or small-volume lots (meaning the other WIP waiting at the work station was largely not WIP waiting for the same process step or same process recipe) tended to wait huge amounts of time because the operators would select other types of WIP which had many more lots sharing the same recipe. This was being done by the operators because they wished to maximize their productivity, so they did not want to go through the effort of making a setup for a small run on that setup.

The Task Force made efforts in two directions to overcome this problem. First, they changed the scheduling policy to require that all recipes with WIP on hand must be set up and run at least
once every shift. Second, they worked on making setups less time-consuming. This was done by converting some of the setup time into “external setup”, meaning a portion of the setup procedure could be performed in parallel with other production work and therefore not take up machine time.

The Task Force found the worst-performing steps were often in the photolithography work station. Moreover, they noted the strategic importance of photolithography. Every mask layer has a photo step at the start; thus the photo area has more leverage on the distribution of fab WIP than do other work stations. “If we control photo, we control the fab.”

Over a two-year period, the photo area average WIP declined from 700 lots to 388 lots while maintaining an overall production rate of 600 lots per day. (Note that, applying Little’s Law, we can deduce that their average photo cycle time dropped from 1.17 days to 0.65 days.)

Investigating the photo area, they found that move-ins (i.e., WIP arrivals) and move-outs (i.e., WIP completions) by step were very disparate, and they were not meeting the photo step TAT goals. Worse, the queue of waiting lots would grow very fast any time there was a major breakdown of one or more of the photo machines.

The following actions were taken:

- Reticle changes (i.e., photo setups) were simplified and expedited. Reticle changes became planned. Reticle storage and tracking became more organized and rigorous. Reticles were inserted in the steppers while the previous reticle was still running.
- Photo engineering relaxed the rules on when test wafer “send-aheads” were required. This was facilitated by requiring every photo step to set up every shift, meaning the last demonstration of process in-control became more recent.
- Photo machine maintenance was coordinated with the photo WIP level. When WIP was below average, maintenance was accelerated, and when it was above average, maintenance was delayed.

As a result of these changes, move-ins and move-outs at each step became closer. They were still disparate, but TAT goals were being met.

**Company B – Effective Negative Scheduling and Innovative Communication**

Back in 1991 a fabrication facility operated by company B produced Bipolar logic products on 125mm wafers. The fab operated four process flows with many device types within each flow. Company B executives established very aggressive goals for cycle time reduction, demanding it be reduced by 25% per year for several years. The regular production management at this fab directed the cycle time reduction project, assisted by industrial engineering and other engineering staff at the fab. During one six-month period within the project, fab cycle time dropped from 2.1 days per mask layer to 1.4 days per mask layer. This was achieved with a 30% reduction in theoretical cycle time plus an even larger reduction in lot waiting time. When operating at 1.4 days per mask layer, the fab was achieving a 1.5X multiplier on theoretical cycle time.
The tactics pursued by the Cycle Time Reduction Team at this fab was as follows:

- Establish goals for TAT at each step in each of the four process flows
- Track actual production rate vs. target production rate by hour at each step
- Track actual TAT vs. target TAT at each step
- Focus attention on the most problematic steps
- Develop and implement an elaborate Kanban system throughout the fab using large display screens. The Kanban system guides both production and maintenance activity
- Standardize processes so that similar steps in the four different flows utilize exactly the same process recipe, thereby eliminating recipe changeovers
- Cross-train operators for maximum flexibility. All operators in each equipment area became qualified to perform all tasks in the area.
- Using robotics, the photoresist coating, photolithography exposure and photolithography development steps were linked into a single, pipelined operation. The linked photo cell they developed was termed a “photocluster.” The robotics was made flexible so that lots undergoing different photo steps could be pipelined through the robotic photocluster.
- Over a period of years, the layout of machines in the fab evolved from a pure Farm Layout to what was termed a “Cell Layout.” For example, the photo department was split into Implant Layers Photo, Metal Layers Photo, Oxide/Poly Layers Photo. Photoclusters assigned to process implant layers were positioned close to ion implant machines. Photoclusters assigned to process metallization layers were positioned close to metal etch machines, and so on. This reduced transport time and facilitated coordination of photo steps with implant steps and with etching steps.
- Hot lots were eliminated.
- On the corporate level, company B implemented automated company-wide production planning system generating a revised plan for all factories every weekend. The planning calculation comprehended fab equipment capacity and reticle supply, ensuring that planned fab starts were capacity-feasible and reticle-feasible. The planning logic ensured that changes in fab production mix were only gradually phased in. Moreover, the planning logic did not reschedule fab WIP, it only scheduled new fab starts.

The kanban system developed by the company B fab is remarkable. The process steps in each of the four process flows were aggregated into process blocks generally corresponding to the WIP tracking logging points. For each block, a WIP limit was established. The WIP limit was based on the Active WIP for the block (in turn based on the theoretical cycle time for the block and the target output rate for the block) plus an allowed Buffer WIP. The allowed Buffer WIP was based on an allowed time for how long it would take to bring the WIP level back down to the Active WIP level during typical breakdowns and during preventative maintenance to machines in the block. For this purpose, statistics were collected on MTTPMs, MTTRs, UPHs and machine availabilities. The limit was calculated in wafers and then rounded up to accommodate an integer number of lots. The calculation of WIP limits was computerized and automated, so that WIP limits were automatically refreshed as target production rates changed, machine statistics changed, or other data changed.

Rather than employ some sort of ticket system, the company B fab chose to communicate WIP status vs. limits using large display screens. This fabrication facility was a “ballroom fab”
without interior walls separating different equipment areas. The display screens were suspended from the ceiling throughout the production area, grouped into threes. In a three-screen group, each screen would display a matrix with four columns, one for each process flow, and with rows corresponding to process blocks. It took three screens to cover all blocks.

Figure 5 illustrates the information conveyed on one of these screens. The actual screens are in color, but the figure is black and white, so some notation has been added. The cells of the matrix on the display screen were actually colored green, yellow or red. In the figure, “R” denotes a cell colored red and “Y” denotes a cell colored yellow. (No letter means the cell was colored green.) The colors convey the status of WIP vs. the WIP limits. A red cell means the WIP in that block of that process has reached or exceeded the WIP limit. This conveys the information that operators manning equipment performing steps in the preceding block should not process lots heading for the red block. A yellow cell means the WIP limit will be reached once one more lot enters the block. A green cell means the WIP is below the WIP limit by more than one lot. This suggests to the operators that it is fine to send more lots to this block. Not shown in the figure is that the display in some cells maybe flashing.

A flashing block means the block is not keeping up with its target production rate for the shift. This conveys the information that operators manning equipment performing steps in that block should give priority to lots in that block in order to catch up to the target production rate. Where the word “Low” is illuminated, it means that the WIP level in that block has fallen below the Active WIP level, rendering it impossible for that block to maintain the target production rate. This is a signal to operators manning equipment in the preceding block that it would be helpful to process and send more WIP.

There also is useful information in the display for engineering staff. For example, an “M2” symbol in a block means that two process machines serving that block are currently down. A down arrow appearing in a block means that there currently is an SPC out-of-control alarm in that block that has inhibited the process.

Regardless of where one is within the fab, a trio of screens displaying the complete fab WIP and trouble status was close at hand. And a quick look at the screens provided a tremendous amount of information. All production and engineering staff could know all the time what the situation was in the fab, where there were problems, and which problems were most urgent. Consider an operator manning equipment in block 095. Block 102 in flow 2 is red, so we should not work on lots in flow 2. We have green blocks in flows 1, 3 and 4 at the 102 block, but in flow 1 the “Low” light is illuminated, indicating the WIP there is inadequate to keep up with its target production rate. If we have WIP available in flow 1, we should work on it first. Consider an operator performing new lot releases. There are green cells in flows 1, 3 and 4. But in the cell for flow 4 there also is “Low” light illuminated, indicating the WIP is inadequate to keep up with the target production rate. Another lot release is needed in that flow; it would be wisest to prepare that release first.

Now consider a machine repairman. There is a down machine serving block 095 in flow 4, and there are two machines serving block 075 in flows 1 and 2 that are down. Block 095 in flow 4 is still green, indicating that not a lot of WIP has backed up yet. On the other hand, in flow 2 block
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>065</td>
<td>R 065</td>
<td>065</td>
<td>Low 065</td>
</tr>
<tr>
<td>R 075 M2</td>
<td>R 075 M2</td>
<td>075</td>
<td>Low 075</td>
</tr>
<tr>
<td>085</td>
<td>085</td>
<td>Low 085</td>
<td>085</td>
</tr>
<tr>
<td>Y 095</td>
<td>Low 095</td>
<td>095</td>
<td>095 M1</td>
</tr>
<tr>
<td>Low 102</td>
<td>R 102</td>
<td>102</td>
<td>102</td>
</tr>
<tr>
<td>Low 104</td>
<td>104</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>115</td>
<td>R 115</td>
<td>Low 115</td>
<td>115</td>
</tr>
</tbody>
</table>

Note: “R” means red, “Y” means yellow, “↓” means process hold "M2" means 2 machines are down, “Low” means low WIP

Figure 5. Production Control Display Screen at Company B
075 has filled up, and the WIP is so backed up that block 065 is filled up too. From the point of view of minimizing cycle time impacts, attending to the machines serving block 065 has the highest priority.

Finally, consider a process engineer responding to out-of-control alarms. There are alarms in block 102 of flow 4 and in block 075 of flow 2. As noted above, there is not a WIP back-up in block 102 yet, but the WIP back-up in block 075 is severe. This suggests that it is more urgent to investigate and fix the trouble in block 075.

Not only were these screens displayed throughout the fab, but the screens could be viewed on the computer at every engineer’s and every manager’s desk. Production operators and supervisors received clear guidance concerning which type of lots should be worked on, and they could see which areas were having trouble. Engineers and technicians could see which process or equipment trouble should be worked on first. And fab management could see trouble as soon as it developed, rather than learning about it in the next start-of-shift review meeting.

The screen displays and associated logic were iteratively developed at the company B fab by the Production Team with support and programming by the Information Systems Team. At first, displays showed only the red, yellow and green color codes. Later they added the down machine and down-process indicators. Still later, they added the blinking displays and low-WIP indicators.

Looking back at the end of the project, when cycle time had been dropped to 1.2 days per mask layer, they estimated that 15-20% of the cycle time reduction stemmed from the implementation of the photoclusters and layout changes in the fab, while 80-85% stemmed from WIP reduction and changes made to accommodate WIP reduction. It was remarked that equipment efficiency improved as well. At the same time cycle time dropped from well over 2 days per layer to 1.2 days per layer, the throughput of the G Line steppers increased from 550 to 750 wafers per stepper per day.

Company C – Project Scheduling to Manage Custom Production

Back in 1992, the subject fab at company C produced application-specific integrated circuits (i.e. custom chips) using gate-array technology. Before a mid-fab option point (just before the first metal layer), generic base wafers were built to plan in four process technologies with design rules ranging from 1.2 microns down to 0.6 microns using G Line and I Line steppers. An inventory of completed base wafers was maintained for each process technology at the option point. Beyond the option point, many low-volume devices were fabricated to order; the typical count of devices in fab WIP after the option point was on the order of 600! The most common lot size after the option bank was only one wafer; the median lot size was 8 wafers. On rare occasions a customer would order enough custom chips to fill out a 25-wafer lot.

The very broad mix, with little volume in any particular device, meant that every lot processed at photo required a reticle change and possibly reticle qualification by performing a “send-ahead” involving expose, develop and inspect steps performed on the first wafer of the lot (the send-
ahead wafer) before exposing the rest of the wafers in the lot. Reticle storage and retrieval was a major challenge; the reticle library serving the fab contained more than 18,000 reticles!

Despite this complexity, the company C fab managed to reduce cycle times in all four process technologies to two days per mask layer while maintaining output rates in each technology. Significantly, this fab reduced the variance of lot cycle times dramatically. In each month the fab measured the 100th percentile of cycle time from option bank to fab out, i.e., its metric was the cycle time of the worst lot. This fab managed to reduce the difference between mean cycle time and the 100th percentile of cycle time to only 2 days. Remarkably, company C consistently achieved 95% or better LIPAS (line item performance against schedule) for its custom production.

The tactics employed by company C to achieve this performance are summarized as follows:

- Lead time for fabrication quoted to customers was based on the 95th percentile of fab cycle time
- Fab management was evaluated by company executives in terms of the 100th percentile of fab cycle time, i.e., cycle time performance was judged based on the worst lot each month
- Fab management established cycle time goals for each step. UPH goals for bottleneck equipment also were established. Actual cycle times and actual UPHs were closely monitored by fab management. Senior management met with line supervisors three times daily to monitor performance and review any issues that arose.
- Computerized lot dispatching based on the Project Scheduling paradigm was implemented. Priority lists of lots, listed in least slack order, were maintained in real time at all work stations. Operators were required to select the lot with the least slack score when a machine became free. This resulted in many setups (i.e., many recipe changes), almost every lot. This meant that lot sequence was negatively correlated with elapsed lot cycle time, thereby dramatically reducing the variance in lot cycle times.
- The photolithography area worked to stage reticles for the high-priority lots. This required extra manpower to fetch reticles from the library and insert the reticles in the steppers before the lots were ready to run.

Monthly cycle time performance of the company C fab is depicted in Figures 6, 7 and 8 for the months of April, August and December, 1992, respectively. These are histograms of lot cycle times from option bank to fab out. During the period April – December, 1992, the fab implemented the least slack dispatching system and other improvements. As may be seen, the average fab cycle time did not change much during this period, falling from 16.7 days in April to 15.6 days in December. But look at the 95% percentile: It declined by 5 days, from 23 days to 18 days. Thus the lead time quoted to customers also declined by 5 days. From the viewpoint of the customers, the company C fab sped up by 5 days. As a result, company C became the preferred vendor of gate-array custom chips and became the first custom chip company to reach a billion dollars in annual sales.

Finding that fab speed was something the company could profitably sell, company C decided to offer different cycle times with different prices. Out of 450 total lots in the post-option-bank
Figure 6. Cycle Time Performance at the Company C Fab for the Month of April, 1992
Figure 7. Cycle Time Performance at the Company C Fab for the Month of August, 1992
Cycle Time Distribution - 12/92

95th percentile = 18 days

Figure 8. Cycle Time Performance at the Company C Fab for the Month of December, 1992
WIP, up to 15 were allowed to be “prototype” lots with the fastest cycle time; 20 – 30 were allowed to be “express” lots with the next fastest cycle time; and up to another 15 were allowed to be “semi-express” lots with the next fastest cycle time. To accommodate this marketing initiative, the dispatching logic was modified to prioritize prototype lots ahead of express lots, express lots ahead of semi-express lots, and semi-express lots ahead of regular lots. Lots within the same class continued to be prioritized by the least slack rule. The semi-express category was not a service sold to customers but was used by the marketing department to provide favorable treatment of targeted customers or to make up for a lot scrap (if and when that happened).

Company C also worked hard on automated incremental production planning to support their custom chip business. A production plan by process technology was prepared, allowing a certain number of lot releases per day from the option bank in each technology. As customer orders were received, the planned starts would be consumed and actually scheduled. The scheduled releases collectively were known in the company as the “fab starts queue.” A “commit date,” i.e., a delivery quotation, was given to a customer by the Order Management Department only after finding a slot in the fab starts queue for the inquiry. If the customer accepted the quote, the slot became a scheduled release. The fab starts queue was visible on-line to fab management. In effect, updates to the fab starts schedule were made in real time.

Given this success in manufacturing planning and execution, the focus at company C turned to reduction of the non-manufacturing time required to fulfill a customer order. The company undertook the development of an automated “order configuration system” that allowed the customer to enter its desired specifications which were then converted into manufacturing specifications and transmitted to both the company’s fabrication facilities and to its back-end-plant subcontractors and its reticle supplier. This eliminated paperwork and collapsed the administrative time needed to prepare manufacturing to process the customer’s order.

The IMPReSS Planning System at Harris Semiconductor

The primary intent of the IMPReSS Planning System described in the previous chapter was to improve on-time delivery and provide a real-time delivery quotation capability. As a by-product of this effort, lead times quoted to customers declined, and manufacturing cycle times also declined. Lead times were reduced because manufacturing performance became more predictable; safety time added to average cycle times could be reduced. Manufacturing performance became more predictable because (1) WIP between corporate inventory points was never rescheduled, and (2) new releases at fab start or from corporate inventory points were scheduled only if capacity-feasible, raw-materials-supply-feasible and WIP-feasible. Capacity available to new releases was net of capacity required to flush WIP at targeted cycle times. With this control of factory workloads, factory cycle times were well-controlled.

The implementation of IMPReSS revealed a number of organizational weaknesses that had to be corrected in order for the system to be successful:

- Data quality was poor, and data ownership did not exist. An “owner” was appointed for each piece of data, whereby the owner had the responsibility for ensuring the data was correct and complete in the official planning database.
- Factory managements did not want to flush out “dead WIP,” i.e., WIP for which demand had evaporated. The discipline of IMPReSS required all WIP to move according to target cycle times up to the next corporate inventory point, where, if not needed, it would be scheduled to stop.
- Non-bottleneck factories did not want to slow down production even though their output could not be handled by a follow-on bottleneck plant. Again, the discipline of IMPReSS, coupled with a change in management philosophy to embrace the Theory of Constraints, forced factories to follow schedules consistent with the plans for the overall production network.
- Within factories, non-bottleneck work centers did not want to make setups for low-volume products. To fulfill the IMPReSS target output schedules, it was necessary to enforce the discipline to set up every operation with WIP on hand at least once every shift. Again, the Theory of Constraints was embraced to change organizational behavior.

There is one more important case-study, an example of scheduling improvement following the Execution Planning paradigm. This is a much more involved story to tell, and one with greater impact, so it deserves a section to itself.

8. The SLIM Project at Samsung

Introduction

During the late 1990s and continuing through the 2000s, Samsung Electronics Corp., Ltd. (SEC) was a leading merchant of dynamic random access memory (DRAM) devices, static random access memory (SRAM) devices, and other advanced digital integrated circuits. During the late 1990s, SEC has sustained about a twenty-percent market share of the vast DRAM market. In terms of unit volume, SEC was the largest manufacturer of digital integrated circuits in the world at that time. SEC’s advanced memory chip products required more than 400 fabrication steps in multi-billion-dollar factories that include several hundred processing machines of various types. At the Kiheung, South Korea, site, probably the largest semiconductor fabrication site in the world during the late 1990s, SEC fabricated more than 300,000 silicon wafers per month (about half six-inch wafers and half eight-inch); the site headcount exceeded 10,000. The headcount at SEC’s chip assembly and test site in Onyang, South Korea, exceeded 4,000. SEC also operated a large semiconductor fabrication plant in Austin, Texas, under the auspices of its US semiconductor subsidiary, Samsung Austin Semiconductors (SAS).

The Competitive Semiconductor Manufacturing (CSM) Program at the University of California at Berkeley benchmarked the performance of semiconductor fabrication plants (“fabs”) around the world during the period 1992 - 2001. The CSM Program documented manufacturing metrics including yields, equipment and labor productivity, and manufacturing cycle time from participating fabs and analyzes management practices to identify those practices that underlie top performance (Leachman and Hodges, 1996). In December, 1995, the CSM program visited SEC’s Line 3 fab in Kiheung, South Korea. The author was part of the CSM Site Visit Team. The CSM Program found that SEC achieved excellent yields and excellent productivity of equipment and labor, but the manufacturing cycle time was the worst of 29 fabs in the CSM
survey. Line 3’s cycle time was about 35% longer than average in the survey and more than double the leading-edge performance of 2.0 days per layer of circuitry. It was difficult for the CSM researchers to even obtain cycle time data from SEC; at that time, SEC did not routinely compute the cycle time metric. A special effort had to be made just for the purposes of the CSM survey to tabulate cycle time data. The Berkeley team presented to SEC manufacturing management their evaluation, indicating that while the company had many strong points relative to the rest of the industry, cycle time was a glaring weak point. Unbeknownst to the Berkeley research team, at almost the same time, the SEC manufacturing managers were hearing from SEC executives that cycle time improvement had become a priority in the company.

Nineteen ninety-five was a memorable year for manufacturers of dynamic random access memory (DRAM) devices. Soaring world-wide demand, coupled with tight capacity, pushed prices of four megabit DRAMs upwards in a sellers’ market the likes of which has not been seen before or since in the memory chip business. DRAM producers, including SEC, enjoyed record profits. As a consequence, the year saw heavy investments in new fabrication capacity by both existing DRAM manufacturers and new entrants, and at the end of 1995, the market turned. Prices collapsed in early 1996, and the market quickly transformed into a buyers’ market.

Recognizing the impending shift, in December, 1995, Mr. Y. W. Lee, President of SEC’s Semiconductor Business, informed the SEC Semiconductor Manufacturing Dept. of the urgent need to reduce cycle time. Prices were starting to fall, and he correctly foresaw that they would fall rapidly for an extended period of time. The company’s financial results were sure to take an abrupt turn for the worse, but if the company could reduce cycle times, the drop in selling prices might not be as severe as otherwise, and the company’s financial results would be better than otherwise. Equally important, President Lee recognized that customers would become much more selective about their DRAM suppliers. The CSM survey’s indication that SEC’s cycle times were non-competitive implied SEC’s customers might be enticed to switch to other DRAM vendors able to offer shorter lead times. The company was becoming vulnerable to loss of market share.

In January, 1996, SEC contacted Leachman and Associates LLC, requesting consulting assistance for cycle time reduction. A proposal for a one-year project was submitted to SEC in February, 1996, and accepted. A project team consisting of staff from both Leachman and Associates and SEC was formed. The project was given the acronym “SLIM” (Short cycle time and Low Inventory in Manufacturing) by SEC and was initiated in March, 1996. The project was extended for four more one-year periods, ultimately terminating in June, 2001.

The Problem

Wafer fabrication is one of the most complex manufacturing processes extant. Lots of twenty-five silicon wafers pass through a hundreds of manufacturing steps in which a series of 18-25 layers of integrated circuitry is fabricated on the surface of the wafers. Each layer involves visits to specialized processing equipment for performing photolithography, diffusion, etching, ion implantation, chemical vapor deposition, cleaning, ashing, measurement and other processes, in a specified sequence. Process batch sizes at various equipment types range from one wafer to 150 wafers.
Process equipment is subject to periods of non-availability for performance of preventive maintenance, engineering work, or unplanned repairs and re-qualification. The manufacturing process is very delicate and somewhat unstable, especially for new devices and new process technologies. Individual lots or process steps may be placed in a “hold” status while engineering effort is made to restore process stability. Out of a group of seemingly identical process machines, only a few may be successfully qualified by the process engineers to perform a particular process step on a particular device, and the list of qualified machines may be quite dynamic as the engineers struggle to achieve process controllability.

These factors make the flow of work-in-process (WIP) through the factory quite turbulent relative to the manufacture of other kinds of products. To attain reasonable productivity, a significant amount of WIP within the manufacturing process is necessary. For example, in the 1995 CSM survey, the best-performing memory device fab achieved an average cycle time of about 2.0 days per mask layer (Leachman, 1994, Hodges and Leachman, 1996). The standard cycle time, i.e., the sum of times required for machine processing and material handling, was about 0.7 days per layer. Thus, at the time, even in the best-performing fab, the amount of inactive WIP exceeded the amount of active WIP.

Since a particular equipment type may perform as many as thirty different manufacturing steps to fabricate a given device, the variety of process steps each equipment type is called upon to perform is generally large. For most machine types, there is some loss of productivity when the equipment must be changed over to perform a different kind of process step. Worse, frequent changeovers of the equipment increase the difficulty of sustaining process control. Imbalances in the distribution of WIP through the manufacturing process arising from the turbulence described above can be mitigated or they can be exacerbated, depending on the choices made for which process steps to perform. Thus improved production scheduling can play an important role in cycle time reduction.

Generally, in SEC fabs, the machines performing photolithography are the most heavily utilized and are thus the bottleneck equipment types. Machines performing the photolithography steps are heterogeneous. There are different makes and models that have varying capabilities as to which photolithography steps they can perform accurately. Moreover, to achieve process controllability, only a small subset of a particular make and model may be qualified by the process engineers to perform a certain step on a certain device. Adding even more complexity, the engineers sometimes make the list of suitable machines for a given lot at certain photo steps a function of which photo machine was utilized to process the lot at some preceding photo step performed days or even weeks earlier.

In this situation, capacity of a fab line is quite complex and dependent on the product mix. It is difficult to plan changes in the mix produced of the various devices in a way that keeps the factory running at maximum output yet does not cause a WIP build-up at any photo machines nor a loss of process stability. For the same reason, it is also difficult to plan how many photo machines need to be qualified at each step. Especially difficult to manage are transitions between generations of DRAMs, whereby the volume of a new generation device must be ramped up while ramping down the volume of the older-generation device. In the most unfortunate case, an
overload of certain photo machines is not recognized until a nasty WIP back-up occurs in the fab. Thus improved production planning can play an important role in cycle time reduction as well.

In terms of planning and scheduling practices, the starting point for SEC was not too different from many other semiconductor manufacturers at the time. Production planning was based on an aggregate capacity expressed for each fab line and did not delve into analysis of the device/step qualifications of each photo machine. A monthly target fab out quantity was established for each device in each fab line. The scheduled input of blank wafer lots roughly corresponded to the fab out schedule shifted backwards in time by a target cycle time. But some batching of fab input lots into groups corresponding to the largest batch size of all equipment types was performed, and early input of lots was made when believed necessary to sustain maximum productivity.

A lot dispatching system known as “MSS” (Manufacturing Scheduling System) was in use in the fab lines. MSS prioritized production lots waiting for processing at each equipment type based on the age of the lot vs. a target for the elapsed cycle time up to its current process step. (This is equivalent to the familiar “least slack” dispatching rule.) But this priority list could not be followed too closely since that would result in an excessive number of changeovers of the equipment. More influential on manufacturing activity was the so-called “cutoff schedule” method of manufacturing management. A monthly target fab out quantity was prepared for each device in each fab line. Manufacturing supervisors would identify what WIP needed to make it out of the factory in the current month in order to fulfill the target output, and then they identified the device/steps they must operate in order to flush this WIP. (The needed process steps for each device were said to fall within the cut-off for the current month’s fab outs; hence the name.) Operators were instructed to set up the equipment to run the identified device/steps. Device/steps that corresponded to the subsequent month’s output were set up only if there was no WIP for the current month target remaining at the equipment.

In essence, SEC had become a victim of its own success. The emphasis on productivity in the seller’s market of 1994-1995 led to very high WIP levels in the SEC fabs. All the Kiheung fab lines had cycle times in excess of four days per circuitry layer (also termed mask layer) when the SLIM project was approved in early 1996.

**Project Strategy**

SEC management imposed certain constraints on the project direction. There could be no diminution of the company’s leadership yields and wafer throughput as cycle time was reduced. It would have been easy to achieve major reductions in cycle time with a wholesale reduction in production levels or a relaxation of the constraints imposed by process control on machine allocation. Instead, we were charged with devising methods by which the fabs would work smarter. Notwithstanding our efforts, management viewed major reductions in WIP levels as risky, so we were instructed to initiate the SLIM project working with only two of the six fab lines extant at Kiheung at that time. If the project with these two fabs was successful, the methods subsequently could be propagated to the other fab lines.
Policies for managing production at SEC are very much consensus-driven. Manufacturing managers, supervisors, equipment leaders and equipment operators all must understand and agree with the logic by which schedules are set and work decisions are made. Schedules proposed by systems were and are subject to override by the humans actually controlling the factory. This meant the basic methodology proposed for floor scheduling must be understandable and convincing to factory staff, and that a substantial education and training program would have to accompany any changes in the methodology for production management.

Finally, the urgency for cycle time reduction dictated a certain approach to the SLIM project. The sophistication with which cycle time was managed would have to be improved incrementally, enabling realization of the benefits of cycle time reduction as early as possible. Waiting to implement a sophisticated scheduling system that would take months or even years to develop was out of the question.

**SLIM Principles**

At the time of the SLIM project, most semiconductor companies operated fabrication lines under the Lot Dispatching paradigm for production management. As described earlier, the focus in this paradigm is to manage the cycle times of production lots. A newly released lot is assigned a due date equal to its release date plus a target cycle time. Scheduling on the factory floor is driven by a prioritization of the lots waiting at each equipment bay. Most commonly, this prioritization is based on a comparison of the estimated remaining cycle time of the lot to the time remaining until its due date, expressed either as a ratio (“the critical ratio” rule) or as a difference (“the least slack” rule).

SLIM reflects the WIP Management paradigm. Instead of lot due dates, factory floor scheduling is driven by a target fab out schedule for each device. This output schedule is viewed as a continuous-time schedule; for example, if the fab out schedule is expressed in terms of output quantities each day, then is assumed that one quarter of the quantity scheduled on a particular day of a particular device is due six hours into that day. SLIM includes methodology to translate this continuous-time target output schedule into a target profile of WIP through the sequence of process steps for each device.

The primary scheduling concern in SLIM is cumulative production of each device/step, not the completion of particular lots. Based on a review of actual output to date vs. the target fab out schedule and of actual downstream WIP vs. target WIP, production targets and priorities and quantitative goals are established for each device/step.

There are several advantages of the WIP Management paradigm compared to the more common Lot Dispatching paradigm:

(a) It is easier to control to a practical level the number of recipe changes that are scheduled. Typically, using SLIM, each device/step is scheduled to be set up on one machine once per shift (or on more if parallel machines are needed to produce higher volumes) to meet production targets computed for the shift. This substantially reduces the number of recipe changes compared to lot-based dispatching. The SLIM schedules are thus much more palatable to the line staff and
to process engineers than the Lot Dispatch schedules.

(b) In wafer fabrication, lots of the same device will get out of order, relative to the order in which they were released. There are many time-consuming inspections and tests that are performed on only a fraction of the lots passing through the fabrication process. Moreover, individual lots may be placed on hold for extended periods of time when process control issues arise. Following the Lot Dispatching paradigm, the operators would have to work to put lots of the same device back in order. This is unnecessary to meet the fab out schedule for the device, and it does not help product cycle time. Worse, since the lot-based logic views due-date performance of each lot as equally important, it sometimes thinks it is more important to re-order lots of a device whose WIP is ahead of schedule rather than meet the fab out schedules for other devices whose WIP is in deficit relative to target. Thus the Lot Dispatching paradigm can compromise the fab out schedule. SLIM never compromises the fab out schedule.

(c) If a lot of one product passes another lot containing the same product, the due dates for these lots become incorrect. Worse, if a lot is scrapped, then the lot due dates for all upstream lots become incorrect in the dispatching system. Even worse, if the fab out schedule is changed, all lot due dates become invalid. Because SLIM incorporates on-line analysis of the downstream WIP and the fab-out schedule, its targets and priorities never become stale.

**SLIM Scheduling Logic**

*Determining target cycle time for individual process steps (SLIM-M).*

The effectiveness of the WIP Management paradigm rests on the establishment of an efficient WIP profile. If we consider a fab with a constant output rate, then Little’s Law indicates that determination of target WIP levels and determination of target cycle times are equivalent. Thus great care is taken in SLIM for establishing target cycle times for the process steps to fabricate each device.

The fragile nature of advanced semiconductor process technology renders the actual distribution of WIP through the fabrication process quite dynamic. Consider first the case when all equipment is operating and all process steps are in control. See Figure 9. The pipe depicted in Figure 9 represents the production line, its width represents the maximum flow rate or capacity at various process steps. SEC fabs are designed so that the photo machines are the bottleneck, and other machines have some amount of surplus capacity. Thus the pipe in the figure is sketched to have a narrow width at each photo step. Normally, WIP in SEC fabs is concentrated in the photo area; typically, thirty to forty percent of total fab WIP is resident in the photo area, as suggested by the large piles of WIP at photo steps in the figure.

Now consider the case where some serious process or equipment trouble has arisen at some other process area besides photo. See Figure 10. This kind of trouble is depicted in the figure by a constriction of the pipe at a non-photo step. While this condition persists, the supply of WIP is reduced at downstream steps, in particular, at the next downstream photo step. The photo machines are somewhat inflexible, since only a limited number of the entire stock of photo machines are qualified to perform a particular step on a particular device. Thus at some point, the
When all the equipment are up and the process is in control, the steppers are the bottleneck, and the largest concentration of WIP is at photo.

Figure 9. WIP distribution in the case of no process or equipment trouble
When there is process or equipment trouble, the WIP at photo can run down. A buffer is needed, suitable for the risk of trouble in that layer.

**Figure 10. WIP distribution in the case of process trouble**
disruption will cause utilization of the photo machines to decline, and fab throughput will be diminished. While other machines have excess capacity and can be run faster than the fab output rate, the photo machines cannot. Losses of photo throughput are unrecoverable losses of fab throughput. It is clear that some amount of buffer WIP at photo steps is necessary to insulate the photo machines from upstream disruptions.

If one were to observe an accelerated animation of the distribution of WIP in the fab over time, one would observe a series of transients. At first glance the equipment and process may be running well, and WIP will be concentrated in the photo area. Then a disruption happens, and the WIP population shifts upstream from one or more photo steps. Next, the process or equipment problem is resolved, and the WIP population migrates back downstream to the photo steps. Then another disruption occurs, and so on. The philosophy underlying SLIM is that the most effective target distribution for WIP is one that puts the fab in the best position to cope with the next disruption. That is, while the equipment and process are working well, the fab should strive to move as much as WIP as possible to the photo bottleneck, to be as prepared as possible to cope with the next disruption.

Suppose the target cycle time from fab-in to fab-out for a particular device is given. We designate this time as the total target cycle time $TTCT_i$ for device $i$. From industrial engineering studies, the intrinsic cycle time $ICT_{ij}$ for each process step $j$ on device $i$ is known, i.e., $ICT_{ij}$ is the irreducible time required for material handling and processing of one lot of device $i$ through step $j$. If we compare the total target cycle time for the device to the total intrinsic cycle time, the difference is what may be termed the total buffer time for the device, i.e.,

$$TBT_i = TTCT_i - \sum_{j=1}^{N_i} ICT_{ij}$$

is the total budget for waiting time for lots of device $i$, where $N_i$ denotes the total number of process steps to complete device $i$.

For the purposes of establishing target cycle times, all of the budgeted buffer time is allocated to bottleneck steps by SLIM. That is, the target cycle times for non-bottleneck steps are set equal to the intrinsic cycle time. The target cycle times for bottleneck steps are set equal to the intrinsic cycle time for that step plus an allocation of the total buffer time. This reflects the philosophy described above that strives for maximum insulation of the bottleneck steps when process and equipment so allow.

The allocation is made with the idea of providing buffers in proportion to the amount of disruption to which each bottleneck step is subjected. That is, bottleneck steps subject to more trouble occurring in the immediate upstream stretch of the fabrication process since the previous bottleneck step need a larger buffer to assure the same utilization of its assigned photo equipment, compared to bottleneck steps subject to less upstream trouble. For example, in Figure 2, if Layer 2 of the process has more trouble than Layer 3, then the bottleneck step immediately following Layer 2 should be awarded a larger buffer than the bottleneck step following Layer 3.
A metric for the amount of trouble in a stretch of fabrication process is not obvious. As a proxy, in SLIM we measured the discrepancy between the intrinsic cycle time and the actual average cycle time for each process stretch between consecutive bottleneck steps. In SEC’s experience, the larger this discrepancy, the more the trouble was occurring in this stretch. The buffer time allocated to the downstream bottleneck step is made proportional to this difference. Let $DCT_{ij}$ denote the difference between the actual average cycle time and the total intrinsic cycle time for the portion of fabrication process for device $i$ that ends with the step immediately preceding bottleneck step $j$ and begins immediately after the bottleneck step preceding bottleneck step $j$ (or begins at fab start if there is no preceding bottleneck step). Then for bottleneck steps $j=1, 2, \ldots, NB_i$ performed on device $i$, the target cycle time is expressed as

$$TCT_{ij} = ICT_{ij} + BT_{ij}, \quad (4)$$

where

$$BT_{ij} = \frac{DCT_{ij}}{\sum_{k=1}^{NB_i} DCT_{ik}} TBT_i. \quad (5)$$

For all non-bottleneck steps $j$ performed on device $i$, we set

$$TCT_{ij} = ICT_{ij}. \quad (6)$$

It is of course unrealistic to expect non-bottleneck steps to perform at their intrinsic cycle times. But that is not the point of the SLIM strategy. The point of the target cycle times is to provide incentive to push as much WIP as possible through an unreliable stretch of the process between bottleneck steps while the equipment and process in that stretch is working. Sooner or later, the equipment or process will go down, and the WIP pushed downstream while they were up will serve to keep the bottleneck step productive despite the upstream disruption.

The SLIM methodology automatically adjusts the target cycle times as the process technology evolves. If problems in a particular stretch of the process are mitigated, then the difference between intrinsic and actual cycle time in that stretch will decline, and some of the buffer time will be reallocated to bottleneck steps other than the next downstream bottleneck step.

It is revealing to contrast the SLIM methodology for setting target cycle times with methodology proposed by other authors. Many authors have proposed using average cycle times from fab history or from discrete-event simulations of the fab to establish target cycle times, e.g., Lu et al (1994). Because these methods average over the transient WIP distributions as well as the distributions experienced when all process and equipment are working well, these targets express much less concentration of WIP at the bottleneck than is targeted by SLIM. One therefore might expect less bottleneck utilization for the same total fab WIP in that case. Other authors have proposed a constant ratio of target cycle time to intrinsic cycle time for all process steps, e.g.,
Fordyce et al (1992). This method tends to award the greatest buffer WIP to steps with very long process times, diffusion steps in the case of wafer fabrication. Unless diffusion furnaces experience high utilization of their available time, this also does not seem to promote the maximum bottleneck utilization afforded by the available WIP.

**Determining target WIP for individual process steps (SLIM-M).**

An integral generalization of Little’s Law is used to translate target cycle times into target WIP levels. If the fab out rate were constant, and no yield loss was anticipated, then the target WIP level for a step would simply be the product of the target cycle time for the step and the target fab out rate for the device. In the case the fab-out schedule is time-varying, note that if the target cycle times and the fab out schedule are exactly achieved, the WIP population at step \( j \) should equal the target fab outs in the interval \([TCTFO_{ij}, TCTFO_{ij-1}]\), where \( TCTFO_{ij} \) denotes the sum of target cycle times for all steps subsequent to the completion of step \( j \) on device \( i \). That is, the target WIP level for step \( j \) on device \( i \) is

\[
TW_{ij} = \left( \frac{1}{FY_{ij}} \right) \int_{TCTFO_{ij-1}}^{TCTFO_{ij}} FO_i(t) \, dt
\]

(7)

where \( FO_i(t) \) is the continuous-time function defining the target fab-out rate for device \( i \) at time \( t \), and \( FY_{ij} \) is the planned yield loss from completion of step \( j \) to fab out of device \( i \). Note that, in the case of a constant fab-outs rate \( FO_i \) for product \( i \) and line yields are 100%, (7) reduces to the familiar form of Little’s Law, i.e.,

\[
TW_{ij} = (FO_i)(TCTFO_{i,j-1} - TCTFO_{i,j}).
\]

**Short-term production targets and priorities (SLIM-M).**

Analysis of the actual downstream WIP vs. the target WIP and the fab-out schedule is performed in SLIM to determine production targets for a short horizon, specifically, an 8-hour production shift. To illustrate, suppose actual fab-outs to date are exactly on time with respect to the target fab-out schedule. Suppose also there is no planned yield loss from completion of step \( j \) to fab out of device \( i \). Now consider the progress of step \( j \) relative to a time horizon of 8 hours (i.e., 0.33 days). If the total WIP downstream from step \( j \) is less than the target fab-outs of device \( i \) over the interval \([0, TCTFO_{ij} + 0.33]\), then there is a downstream deficit of device \( i \). An amount of WIP equal to this deficit should be processed through step \( j \) during the current shift in order to render step \( j \) exactly on time with respect to the fab out schedule and the target cycle time. This amount is termed the *ideal production quantity* (IPQ) for step \( j \). The amount is an “ideal” since it may be infeasible to process this amount for a variety of reasons. There might not be enough WIP supplied to step \( j \) during the shift to complete the IPQ; there might not be enough qualified machines available to complete the IPQ in one shift; or step \( j \) on device \( i \) might be running so far ahead of schedule that the IPQ could be negative.
In words, the IPQ computed at the start of a production shift for a given device/step is the total fab outs due until the target cycle time from the step to fab out plus one production shift, less actual fab outs to date, less the downstream WIP of that device. Mathematically, the ideal production quantity for step \( j \) on device \( i \) is computed as

\[
IPQ_{ij} = \frac{1}{FY_{ij}} \left[ \int_{-\infty}^{0} [FO_i(t) - AFO_i(t)] dt + \sum_{k=j+1}^{N} \left( \frac{FY_{ik}}{FY_{ij}} \right) (TW_{ik} - AW_{ik}) + \frac{1}{FY_{ij}} \int_{TCTFO_{ij}+0.33}^{TCTFO_{ij}} FO_i(t) dt \right]
\]

(8)

Here, \( AW_{ij} \) denotes the actual WIP at step \( j \) on device \( i \), \( AFO_i(t) \) denotes the actual fab out rate at time \( t \), and other notation is as before. The first term expresses the discrepancy in fab outs to date, the second term expresses the discrepancy between actual downstream WIP and target downstream WIP, and the third term expresses one more shift of fab outs due after the target cycle time from step to fab out. All terms are adjusted by the planned yields from step to fab out, so that all terms are expressed in units of production of step \( j \). In view of (7), (8) can be rewritten as

\[
IPQ_{ij} = \frac{1}{FY_{ij}} \left[ \int_{-\infty}^{0} [FO_i(t) - AFO_i(t)] dt + \sum_{k=j+1}^{N} \left( \frac{FY_{ik}}{FY_{ij}} \right) (TW_{ik} - AW_{ik}) \right] + \int_{0}^{TCTFO_{ij}+0.33} FO_i(t) dt - \sum_{k=j+1}^{N} \left( FY_{ik} \right) AW_{ik}
\]

(9)

That is, the ideal production quantity for step \( j \) on product \( i \) is given by the discrepancy in target fab-outs to date vs. the actual fab outs to date, plus the fab outs due from now until the target cycle time to fab out plus one shift, less the yielded downstream WIP, with the result normalized for the projected line yield from step \( j \) to fab out. Note also that, in the special case that fab-outs to date are on time, line yields are 100%, and there is a constant target fab-out rate \( FO_i \) for product \( i \), then (8) reduces to

\[
IPQ_{ij} = \sum_{k=j+1}^{N} TW_{ik} - \sum_{k=j+1}^{N} AW_{ik} + 0.33 FO_i,
\]

i.e., the ideal production quantity for a shift is the difference between the total downstream target WIP and the total downstream actual WIP, plus one shift of target fab-outs.

A more complicated formula for the IPQ arises in the case there is one or more split points in the product structure midway through the fabrication process. At a split point, WIP for a generic device is allocated to become WIP of two or more specific devices. We omit details here, but the basic idea is that one must properly account for downstream surplus WIP. That is, surplus WIP
downstream from the split point must be compared to target fab outs for the specific device in order to determine the consumption schedule for that WIP. Even though the total WIP for a device family may equal or exceed the target, shortages for one or more specific devices result in positive IPQs at upstream steps. We omit details here.

Dispatching priorities for the various device/steps processed by a given equipment type are prepared by SLIM based on these targets. The IPQ is translated into a dispatching score by dividing it by the average fab out rate over the interval \([0, TCTFO_i + 0.33]\) and changing the sign. We call the result the "schedule score" (SS), reflecting how many shifts early or late step \(j\) will be if no WIP for that step is completed this shift. For example, SS = 1.0 means the step is running ahead of schedule by an amount equal to one shift’s normal production. A negative schedule score indicates the step is behind schedule; the IPQ quantity for the step indicates how many wafers need to be processed this shift to restore the step to on-schedule status by the end of the shift.

The schedule score and the ideal production quantity comprise the basic foundation of the machine scheduling algorithms of SLIM. Device/steps are prioritized by SS. Once a device/step is assigned to a machine, processing of that device/step is not interrupted until either the IPQ is completed or the available WIP is exhausted. Since the IPQ is a target quantity for a production shift, device/steps tend to be scheduled for setup once per machine per shift by SLIM.

We note that in the extreme case that every lot is a different product, prioritization of device/steps by schedule score becomes equivalent to prioritization of lots using the least slack rule on the SLIM-M target cycle times. But when the WIP includes multiple devices and many lots of each device, the SS/IPQ logic generates schedules quite different from those generated by lot dispatching. For the reasons cited earlier, the SS/IPQ schedules are demonstrably much better, both in terms of maintaining the target WIP profiles as well as in terms of the number of device/step setups.

Non-bottleneck machine scheduling (SLIM-L).

We first consider on-line scheduling of non-bottleneck machines. Here, three concerns are taken into account. The first concern is the aforementioned importance of maintaining the target WIP profile for each device, as reflected in the SS and IPQ scores. The second concern is maintaining an adequate supply of WIP to facilitate full utilization of the machines utilized at the next downstream bottleneck step. Consider the case of two device/steps with comparable schedule scores. Referring to Figure 11, suppose the supply of WIP at the next downstream bottleneck step for device/step A is relatively high, and the amount in the pipeline en route to the bottleneck step is also relatively high. On the other hand, the supply of WIP at the next downstream bottleneck step for device/step B is low, and the amount in the pipeline en route to the bottleneck step is also low. Since the machines qualified to perform each bottleneck step are somewhat inflexible, there is a serious risk that the machines qualified to perform the next photo step on device/step B may suffer a loss of utilization. Thus we prefer to dispatch device/step B.

A metric included in SLIM for assessing the supply of WIP to the next downstream bottleneck step is termed the *balance index* (BI). The balance index \(BI_i\) for step \(j\) on device \(i\) is computed as
where \( j_{B} \) denotes the next bottleneck step performed on device \( i \). The balance index for a given device/step expresses the difference between downstream actual WIP and target WIP up to and including the next bottleneck step, divided by the target WIP up to the next bottleneck step. A balance index of 0 indicates actual WIP exactly matches target WIP; a score of minus one indicates no downstream WIP at all. The lower the value of BI, the more urgent the dispatch of the device/step, from the point of view of sustaining utilization of the bottleneck machines.

To integrate both concerns, priority levels are defined in SLIM-L corresponding to ranges of schedule score and balance index, as displayed in Table 1. Device/steps in level 5 are prioritized ahead of device/steps in level 4 and so on. Within each priority level, device/steps are further prioritized by least schedule score first.

### Table 1. Priority Levels in SLIM-L

<table>
<thead>
<tr>
<th>BI &lt; -0.5</th>
<th>-0.5 &lt; BI &lt; 0.5</th>
<th>BI &gt; 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS &lt; -16</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>-16 &lt; SS &lt; 16</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>SS &gt; 16</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

The third concern addressed in SLIM-L is that the number of device/step setups scheduled should be controlled to a level acceptable to line staff. This is done two ways: First, a machine that is working on a device/step with a positive IPQ will not be changed to run a different device/step until either the WIP of that device/step is exhausted or the IPQ becomes negative. Second, the number of parallel machines that are set up to process a particular device/step is limited to the minimum number needed to complete the IPQ by the end of the shift. The minimum number of machines for step \( j \) of device \( i \), \( MNM_{ij} \), is computed as

\[
MN_{ij} = \left[ \frac{(IPQ_{ij})(PT_{ij})}{(60)(TR)} \right]^+ \tag{11}
\]

where \( PT_{ij} \) is the process time per wafer (expressed in minutes) of step \( j \) on device \( i \), \( TR \) is time remaining in the shift, expressed in hours, and \( \lceil x \rceil^+ \) denotes the smallest positive integer greater than or equal to \( x \).
Downstream WIP situation:

Recipe A

Recipe B

According to BI, recipe B should be dispatched first.

Figure 11. Illustration of WIP balance index
For example, suppose there are ten lots of the highest priority device/step in available WIP and there are eight machines. Suppose processing times are short, so that the machines become ready for dispatching frequently. Suppose MNM for this device/step is 3. Lot dispatching systems might assign all eight machines to be set up to run this device/step. SLIM will only set up three (unless other machines would otherwise be idle). From the point of view of the line staff, the SLIM schedule is much more rational.

The SLIM-L algorithm (“L” for lot scheduling) is outlined as follows. The algorithm is triggered whenever a machine is about to complete a lot (or complete a machine load, if a load is larger than one lot) or when there is a new WIP arrival and there are one or more idle machines.

1. Consider an idle machine, or if none, consider the next machine to complete processing of its current load.

2. If there exists unassigned WIP of the device/step for which the machine is currently running, and if the IPQ of that device/step is positive, then dispatch another lot (or another machine load) of that device/step to run on the machine. Reduce the IPQ by the number of wafers assigned.

3. If not, consider the highest priority device/step with available WIP that this machine is qualified to run, say, step \( j \) on device \( i \). If \( MNM_{ij} \) other machines are not already scheduled to run that device/step, then dispatch a lot (or machine load) of this device/step to run on the machine. Reduce \( IPQ_{ij} \) by the number of wafers assigned.

If after step 3 there are still one or more idle machines and there is still unscheduled WIP, dispatch a lot (or machine load) of the highest priority device/step available for which the machine is qualified. Reduce the IPQ by the number of wafers assigned.

**Bottleneck machine scheduling (SLIM-S).**

For the photo area, it is not wise to structure the scheduling problem as one of selecting device/steps to assign to machines as they become idle, as is done in SLIM-L. This is because of the tight capacity juxtaposed against the restrictive lists of machines qualified to perform the various device/steps. A larger perspective, in which one develops an intelligent allocation of the device/steps among the entire set of machines, is necessary to avoid situations with idle machines and unassigned WIP that are not compatible.

In SLIM, a scheduling algorithm known as SLIM-S (“S” for stepper) is applied periodically to develop Gantt chart schedules for every stepper (i.e., for each photo machine). This algorithm assigns as much as possible of the currently available WIP. The concerns addressed in this algorithm are as follows. The first concern is to achieve maximum utilization and minimum setups of the steppers, since these are the bottleneck machines. The second concern is to maintain the target WIP profile for each device, as reflected in the SS and IPQ scores.

SLIM-S incorporates a three-pass algorithm. In the first pass, current setups are examined to see if the assigned device/step has fulfilled its IPQ; if not, more WIP is assigned as available. In the
second pass, new setups are planned so as to complete the IPQs of as many device/steps as possible. In the third pass, remaining WIP is assigned to achieve full utilization. In each pass, steppers are selected for WIP assignment so as to minimize setups and also with regard to the restrictions imposed by limited stepper qualifications. A “least candidate WIP” rule is utilized for cases where steppers are qualified to process a limited number of device/steps. By making the assignment to the stepper that has the least total WIP to choose from, we mitigate the likelihood that that stepper will become idle for lack of suitable WIP. For cases where the qualifications are quite flexible, the “most available time” rule is used in order to accelerate schedule recovery.

In more detail, the SLIM-S (“S” for steppers) algorithm is summarized as follows.

1. Determine available time this shift for each stepper after accounting for time to complete currently assigned lots, time to perform planned maintenance and an estimate of time to return to service if currently down.

2. For all steppers currently running, if completion of the lot now running is not sufficient to reach the IPQ for that device/step, assign more WIP until the IPQ is reached or until available WIP is exhausted or the end of the shift is reached, whichever occurs first. Update available stepper time, available WIP and the IPQs.

3. Consider highest priority device/step. Find qualified steppers with minimal setup time. (For a stepper that already has this device/step assigned to it, the setup time is set to zero.) Among those, find the stepper with the most available time (least candidate WIP). Assign WIP until IPQ is reached, WIP is exhausted or the end of shift is reached, whichever occurs first. If IPQ is not reached and more WIP is available, repeat this assignment with next most preferred stepper. Update available stepper time, available WIP and the IPQs.

4. Repeat step 2 for next highest priority device/step. Update available stepper time, available WIP and the IPQs.

5. After all device/steps with positive IPQs have been considered, re-compute schedule scores and re-sort device/steps.

6. Consider highest priority device/step. Find qualified steppers with minimal setup time. (For a stepper that already has this device/step assigned to it, the setup time is 0.) Among those, find the stepper with the most available time (or alternatively, the stepper with the least candidate WIP). Assign WIP until WIP is exhausted or end of shift is reached, whichever comes first. If more WIP is available, repeat this assignment for the next most preferred stepper. Update available stepper time and available WIP.

7. Repeat step 5 for the next highest priority device/step. Update available stepper time and available WIP.

8. After all WIP has been considered and/or there is no remaining stepper time, obtain the final schedule by re-sorting the order of lot assignments to each stepper so as to eliminate any unnecessary setups.
During the course of the shift, more WIP will arrive at the photo area. This WIP is considered for insertion into the Gantt chart schedules of the steppers by an algorithm known as “sub-SLIM-S”. In principle, this algorithm could be run on a transaction basis every time new WIP arrives, but in practice at SEC, it is run every ten minutes to schedule the WIP arrivals that occurred over the last ten minutes. Sub-SLIM-S inserts urgent WIP arrivals into the Gantt charts of the current SLIM-S schedule, where “urgent” is understood to mean those device/steps whose IPQs are not yet scheduled to be fulfilled. Less urgent WIP is assigned to the tail ends of the Gantt charts where time is available. The sub-SLIM-S algorithm is summarized as follows.

1. Order the device/steps with available WIP by SS.

2. We first consider only the new WIP arrivals for which the IPQ of its device/step is positive. Consider these device/steps in priority order. Given a device/step, if there exists one or more steppers already scheduled to run the device/step, choose the stepper among them whose available time is largest. Insert new lots of that device/step into the Gantt chart for that stepper, displacing subsequently scheduled device/steps that are pushed beyond the end of the shift, until the WIP is exhausted or the IPQ is reached, whichever occurs first. If the end of the shift is reached before the just-assigned WIP of this device/step is completed, remove the excess and consider it for assignment to another stepper scheduled to run this device/step. Assign WIP until the IPQ is reached or the WIP is exhausted to the stepper among them with the next largest available time, again displacing other assignments that are pushed beyond the end of the shift.

If there does not exist any stepper scheduled to process this device/step, search among the qualified steppers with minimum setup time for this device/step and identify the earliest time a stepper is changed over to run a device/step of lower priority. Insert a setup and new lots of the considered device/step into the Gantt chart for that stepper at this point, displacing subsequently scheduled device/steps that are pushed beyond the end of the shift, until the WIP is exhausted or the IPQ is reached, whichever occurs first. If the end of the shift is reached before the just-assigned WIP of this device/step is completed, remove the excess and consider it for assignment to another qualified stepper. Assign WIP until the IPQ is reached or the WIP is exhausted to the stepper among them with the next earliest time a lower-priority device/step is scheduled, again displacing assignments that are pushed beyond the end of the shift.

Update available WIP, IPQs and schedule scores. Repeat step 2 for the next such device/step, including consideration of device/steps that may have been displaced by the new assignments.

Now consider all available WIP in priority order, including assignments displaced in steps 2 and 3. Given a device/step, choose the qualified stepper with minimum setup time first and the largest available time second, and assign WIP lots to the end of the Gantt chart for that stepper until the WIP is exhausted or the end of the shift is reached, whichever occurs first. If WIP of this device/step is not yet exhausted and other steppers are qualified to run this device/step, assign WIP to the stepper among them with the next largest available time (if
Another issue in stepper scheduling concerns the masks utilized to print the patterns on the wafer. There are sometimes fewer masks than there are qualified steppers for a given device/step, and in some cases, a different mask is required depending on which stepper is to be used. SLIM-S and sub-SLIM-S schedule the use of masks as well as steppers. We omit details here.

**Diffusion batch scheduling (SLIM-D).**

Diffusion furnaces can accommodate as few as one or as many as six lots for a fixed-duration machine cycle. Typically, several device/steps are identical furnace process steps. Thus there are additional scheduling decisions concerning what device/steps to include in a run of a particular furnace process step, and concerning whether to initiate a partially full machine cycle of a particular furnace step or to wait for more WIP to arrive.

Like SLIM-L, SLIM-D is an on-line, transaction-based algorithm. The same priority ordering of device/steps as is used in SLIM-L is used in SLIM-D. For each device/step, SLIM-D formulates tentative furnace batches considering the on-hand WIP. Lots of the particular device/step under consideration are selected first; if insufficient to fill the furnace, lots of the next highest priority device/step compatible in the same furnace run are considered, and so on, until a maximum sized furnace batch is formed from the available WIP. A minimum batch size (MBS) is specified for each furnace step. If the batch so formed exceeds the MBS, the batch is dispatched using the first available machine. If not, dispatching of the device/step is postponed until arrival of additional compatible WIP, and a tentative batch is constructed for the next highest priority device/step.

**Scheduling release of new lots (SLIM-I).**

Release of new lots into the fab is decided periodically such as once per shift or once per day. For the chosen time horizon (shift or day), an IPQ is calculated for the starting step on each device. That is, we first calculate for each device what release quantity is necessary to replenish up to the target WIP level. This value is rounded up to an integer multiple of the standard release batch size for the device. Next, an estimate is computed of the workload on the steppers qualified to perform the first photo step on the device. This estimate is of the workload over the estimated lead time from release to arrival at the photo step plus the length of the horizon. If there is no remaining stepper capacity, then release is blocked. Otherwise, release is scheduled as calculated. As may be seen, SLIM-I incorporates elements of constant WIP control, e.g., Hopp and Spearman (1996), as well as bottleneck workload regulation, e.g., Wein (1988).

**Line Simulation (SLIM-F).**

All of the SLIM scheduling modules are incorporated in a discrete-event simulation model known as SLIM-F. The SLIM-F simulation is fed with initial conditions corresponding to the actual factory state (e.g., WIP status, equipment status, machine arrangement tables, target out schedule, etc.) Factory operation is then simulated up to a user-specified horizon.
SLIM-F is used by systems technology engineers to test changes in the SLIM logic and to monitor factory usage of SLIM through comparisons of actual WIP movement to simulated movement. If an observable discrepancy between simulated movement and actual movement arises, it is an indication that usage of SLIM has faltered. This could result from a failure in data maintenance, a software error, or a lack of acceptance of SLIM schedules by factory floor staff. Such a discrepancy triggers a review to find root cause and take corrective action.

SLIM-F also is used to support maintenance and engineering activities. For example, there often is some flexibility in the scheduling of preventive maintenance procedures. By studying the results of SLIM-F simulations of the next several shifts or the next several days of factory operation, maintenance engineers can identify periods of low WIP movement through particular machines, i.e., periods when preventive maintenance would be the least disruptive of WIP movement. As another example, when process controllability issues arise in photolithography, a particular stepper may be temporarily disqualified from performing a particular device/step. By reviewing future WIP photo levels simulated by SLIM-F, process engineers can identify the time by which the stepper needs to be re-qualified in order to avoid a serious impact on cycle times.

**SLIM Planning Logic**

The effectiveness of the SLIM floor scheduling logic strongly depends on a target fab out schedule that is consistent with the bottleneck machine capacities. Thus there is a need for production planning that performs accurate analysis of fab out demands with respect to machine capacities.

Capacity analysis at the SEC fabs is challenging because of many factors. We highlight three important factors here. First, the product mix is highly dynamic. For even a single type of device, such as a 64M synchronous DRAM, normally two generations of the device are in production at the same time in the same fab. Each generation has distinct but overlapping sets of machines qualified to perform the various process steps, different process times and cycle times, and perhaps different numbers of steps as well. At a given point in time, production of the oldest generation could be at high volume but ramping down; and production of the new generation could be at low volume but rapidly ramping upwards. Product mix is thus continually evolving. For the re-entrant semiconductor manufacturing process flow, a static capacity analysis tool is undesirable. Workloads on individual machines over time need to be accurately assessed.

Second, the fab out schedule often is changed inside the target cycle time, sometimes dramatically, as revisions are made to customer demand forecasts. This second factor means that the fab WIP of each device should be explicitly scheduled in the planning model; to be accurate, one cannot assumed that initial WIP should simply continue to move through its process flow according to pre-specified cycle times.

The third factor is that the machines of a general type (such as photo exposure machines) are very non-homogeneous, owing to the variety of makes and models and the limited numbers of machines that are qualified to perform each device/step. In the most challenging case, the machines qualified to perform a certain device/step are a function of what machines were utilized at a previous step on that device. This factor means that an accurate planning model cannot treat
the machines of a particular type as a homogeneous group with an aggregate capacity. Instead, feasible allocations of WIP to individual machines must be calculated in the model.

A linear programming (LP) model was formulated to cope with these challenges. The model includes variables for the release of new lots into the fab as well as the release of initial WIP from every major manufacturing step in discrete periods such as work days out to a horizon defined by the user. Additional variables are defined to route new releases and initial WIP through alternative machines. Constraints are formulated expressing the capacity of individual process machines in each period and the consistency of WIP movement through different process steps in different time frames. Constraints also are formulated expressing the backorders or finished goods inventory relative to different classes of demands. Objective functions minimize backorders and finished goods inventory levels. An application developed by the authors generates this formulation from the SLIM database, and after optimization is complete, prepares user reports. When interfaced with a commercial optimization software and the SLIM databases, the result is what is known as the “SLIM-O” (“O” for output) capacity analysis tool. Mathematical details are provided in Lin (1999) and Leachman (2001).

SLIM-O incorporates a much more precise capacity analysis than is typical in LP models. Lead time parameters are time-varying and non-integer, and constraints relating production variables to workloads and demands are formulated at non-integer points of time to ensure mass conservation through continuous time. In application to SEC fab lines, the SLIM-O models have tens of thousands of constraints and variables, but are readily solved in minutes on UNIX workstation computers.

An important use of SLIM-O is in what SEC terms Transition Planning, which concerns the period when volume of a new-generation DRAM device is being ramped up while volume of an older-generation DRAM device is being ramped down. The new device will go through a significant learning curve of cycle time in the first months of its existence. Initially, its process flow is loaded up with many tests and inspections, and equipment choices are few, as the process engineers struggle to achieve a robust and controllable process. As time goes on, tests and inspections are reduced, and more pieces of equipment are qualified. SLIM-O is designed to help schedule the engineering work needed to qualify various machines for production of the new device so that SEC can achieve a fast production ramp and a fast cycle time learning curve for the new device. By making sure production volumes are consistent with machine qualifications, cycle time is efficiently managed.

Typically, it is desired to achieve constant total wafer output during the transition. Considering the learning curve, there must be a planned build-up of WIP as the slower-cycle-time new device displaces the fast-cycle-time end-of-life device. SLIM-O helps SEC Production Control engineers plan an appropriate build-up.

Another version of SLIM-O was implemented for capacity analysis in device assembly and test manufacturing areas. The primary technical issue here is the considerable tooling that is required to perform certain manufacturing steps, and limited compatibility of different types of tools. In addition to other features, the SLIM-O model for this case includes variables assigning production workloads to alternative sets of compatible tools. We omit details here.
SLIM Implementation

The implementation of SLIM scheduling applications at SEC was made in phases. First, an interdisciplinary SLIM project team was formed, focusing on SLIM systems development, training and implementation in two or three large fab lines. Ultimately, seven such SLIM Teams were formed. Members of each SLIM Team included staff from SEC Manufacturing, Photo Engineering, Total Productivity (an IE department), Production Control and Systems Technology (MIS), as well as staff from Leachman and Associates (L&A). Team members from the SEC Total Productivity, Production Control and Systems Technology as well as L&A included individuals with graduate-level OR/MS academic training.

At the outset, each Team set up an on-line database supporting SLIM, and any gaps in electronically stored data (intrinsic cycle times, process times, machine qualification tables, etc.) were identified so that required industrial engineering efforts and data interface efforts could be initiated immediately. Fortunately for the SLIM project, SEC was relatively data-rich compared to other semiconductor companies. At project start, the Total Productivity Team already maintained a fairly complete database of intrinsic cycle times and process times for most devices as well as efficiency and availability metrics for the major equipment types in all fab lines.

Software development of the various SLIM algorithms depended on the algorithmic complexity involved and required linkages to other systems. SEC Systems Technology coded all of the user interfaces, database queries and interfaces to factory floor execution systems. The SLIM-M, SLIM-L, SLIM-S and SLIM-I formulas and algorithms developed by the authors were coded and maintained by SEC Systems Technology staff, while SLIM-O and SLIM-F were coded and maintained by L&A. Various commercial software are employed in the SLIM scheduling modules. The CPLEX\textsuperscript{tm} optimization software from ILOG is used to solve linear programs in SLIM-O; the ASAP\textsuperscript{tm} software from AutoSimulations is used to implement SLIM-F; and the RTD\textsuperscript{tm} software from AutoSimulations is used to convey SLIM-L and SLIM-S scheduling decisions to the factory floor execution system as well as to update the SLIM database with factory floor status.

Training of management, manufacturing staff and engineering staff in the SLIM principles and logic is essential to the acceptance and use of SLIM. Classes were organized and delivered by the SLIM Teams to practically every manufacturing department employee in the subject fabs (including all supervisors and operators staffing every production shift) as well as to all of the management and to many of the process and equipment engineers. One Point Lesson sheets, tutorial reports and technical reports were prepared and widely distributed.

Each Team made a phased implementation of the SLIM scheduling modules, in an effort to reduce cycle times as early as possible. SLIM-M was implemented first, so that production managers and supervisors could begin directing shift production activity towards maintaining the target WIP profile as well as begin taking down the fab WIP to levels consistent with the target cycle times.

SLIM-I, SLIM-L, SLIM-S and SLIM-D were implemented next. Typically, schedules were first displayed in an advisory mode, e.g., the SLIM-S living Gantt chart schedule for all the steppers is displayed in the photo area, and the SLIM-L on-line device/step priorities (SS, IPQ, how many
machines to set up) are displayed in other areas. Later, in Kiheung Lines 6, 7, and 8 and in the SAS fab in Texas, SLIM-L and SLIM-S become fully integrated with the factory execution system, reducing the human involvement in scheduling to an on-exception basis. After such integration, it was possible for SEC to prevent human over-ride of the SLIM schedules. On occasion this has been done, but normally manufacturing operators and supervisors have the capability to reject the SLIM schedules if they feel they can make a better decision.

SLIM-O was implemented next. This application is used for a variety of purposes: refinement of fab-out demands into the maximum capacity-feasible target fab out schedule; making detailed projections of fab WIP-outs for company-wide supply chain management systems; and planning the qualifications of steppers and other equipment needed to accommodate new devices (i.e., transition planning).

Finally, SLIM-F was implemented. By simulating fab operation in previous periods, usage of SLIM is monitored. By simulating fab operations is future periods, scheduling of maintenance and engineering can be done taking impacts on cycle time into account.

As each module was implemented, some additional amount of cycle time reduction was realized. As actual cycle times were reduced, the target cycle times were reduced further, generally set about 0.2 days less than the actual cycle times per mask layer. Generally, each SLIM Team achieved significant cycle time results three months after team formation. Full benefits took about two years to achieve.

In March, 1996, the first SLIM Project Team was formed. Cycle time reduction in Kiheung Lines 4 and 5 was the subject of a one-year project. One-week joint working sessions were held each month between consultants’ staff and internal staff. For the period June through August 1996, some of the L&A staff spent full-time on site in Korea. In March, 1997, the SLIM Project was renewed, and a second SLIM Team was set up to reduce cycle time in Lines 6 and 7. In March 1998, the SLIM Project was renewed again, and the project scope was expanded to embrace all SEC fab lines. L&A posted staff to full-time duty in Korea, and additional SLIM Teams were set up for Lines 2 and 3, for the Samsung Austin Semiconductors (SAS) fab line in Texas, and for three SEC fab lines in Bucheon, South Korea. Kiheung Line 8 was added to the responsibility of the Line 6/7 SLIM Team. In March, 1999, the SLIM Project was renewed for a fourth year, and the project scope was broadened to embrace cycle time management in electrical die sort (EDS) and device assembly/test manufacturing areas. SLIM Teams were set up for Kiheung EDS and for SEC’s large assembly/test facility in Onyang, South Korea. In March, 2000, the SLIM Project was renewed for a fifth year, with the scope of SLIM project efforts extended to treat the scheduling of maintenance and engineering and work, including the qualification of machines to process new devices.

Throughout this period, the SLIM project teams enjoyed strong SEC executive support. The importance of the cycle time reduction project was impressed upon every manufacturing and engineering employee, and the SLIM Teams generally received excellent cooperation. The President of the Semiconductor Business (Mr. Y. W. Lee) approved each year of the SLIM project. The project director in the first year was the Senior Director in charge of manufacturing and engineering in Lines 4 and 5 (Mr. J. H. Lim); in subsequent years, the project director was
the Senior Manager of the Total Productivity Team (Mr. Y. I. Kim). An executive Steering Committee including all Senior Manufacturing Managers and chaired by Senior Vice President J. W. Kim began meeting monthly in the second half of 1998 to review the project status and to provide project direction.

**SLIM Results and Economic Benefits**

As SLIM was implemented and refined, cycle times were steadily reduced. Figure 12 displays the trends in average cycle time per mask layer at Kiheung Lines 4 through 8. As can be seen, substantial progress was made. During the first year of the SLIM project, cycle times in Lines 4 and 5 were brought down from about 4.5 days per layer to about 2 days per layer. Beginning in March, 1997, Lines 6 and 7 were added to the project scope. At the end of the second project year, cycle times in Lines 4 and 5 had been brought down from 2.0 to 1.5-1.6 days per layer, while cycle times in Lines 6 and 7 had been brought down from 3.3-3.5 to about 2.5 days per layer. During the third year of project, the project scope was expanded to all SEC fab lines. By March, 1999, cycle times in Lines 6, 7 and 8 had been brought down to the range of 1.3-1.6 days per layer, while Lines 4 and 5 maintained good cycle time performance, averaging about 1.7-1.8 days per layer. For several months in 1999, Line 8 achieved average cycle times of 1.31-1.34 days per mask layer across its entire product portfolio while sustaining maximum wafer throughput. This incredible performance surpasses that of any eight-inch advanced digital fab in the CSM survey, including the leading foundries, as reported in Leachman (2002).

Immediately after implementation of SLIM-M/L/S, factory managers typically would observe a redistribution of fab WIP. The percent of total fab WIP in the photo area typically would rise by about ten to fifteen percent. This enabled a higher utilization of the relatively inflexible steppers for a given amount of total fab WIP, or alternatively, a reduction in total fab WIP to sustain the same level of utilization. Moreover, the machine allocation logic in SLIM enabled an increase in utilization for a given level of WIP. For example, in a twenty-one-shift simulation comparison against actual performance in Line 4 shortly before its implementation, SLIM-S achieved about twelve percent higher wafer throughput in the photo area. This improvement alternatively could be translated into a further reduction of the WIP level needed to sustain a current throughput.

Figure 13 summarizes SEC cycle time performance a different way, providing the trends in average fab cycle time for production of 4M, 16M, 64M and 128M DRAMs at the Kiheung fabs. From 80 days or more to fabricate DRAMs in late 1995, cycle times had been brought down to about 30 days to fabricate DRAMs by the end of 1998. A major milestone was achieved in November, 1998, when the cycle time for fabrication of third generation 64M DRAMs in Line 7 was reduced to an average of 27 days (about 1.35 days per mask layer). Coupled with a 3-day cycle time for electrical die sort following wafer fabrication, the SEC Semiconductor Manufacturing Division had fulfilled a challenge laid down by SEC CEO J. Y. Yun earlier that year: Make 64M DRAMs in 30 days. SEC manufacturing managers were doubtful that they could achieve such performance at high volume, but Executive Yun felt sure it could be done. He was right.

Other DRAM manufacturers in the CSM survey achieved cycle times in the range 3-4 days per mask layer. A couple of foundries making advanced memory devices achieved cycle times below
two days per layer, but at much lower die yields. Investigation into their practices revealed that
many more photo machines were qualified to perform a typical device/step. This practice makes
it much easier to manage cycle time, but at the expense of much less process controllability and
consequent lower yields. See Leachman (2002).

Many benefits can be ascribed to cycle time reduction. Reduced flow times result in improved
accuracy of the sales forecasts used in production planning, thereby reducing inventory levels.
Lead times to customers may be reduced, perhaps affording a competitive advantage or staving
off a competitive disadvantage. Reduced cycle times may facilitate more rapid feedback of the
results of engineering experiments as well as more rapid feedback of customer suggestions for
product modifications, thereby resulting in more rapid deployment of improvements to product
and process designs. These kinds of benefits are very real but are difficult to quantify. But there
is another very large but often overlooked economic benefit of cycle time reduction.

Prices for DRAMs and other memory devices generally drop dramatically over time. Figure 13
displays the trends in average selling prices of DRAMs at SEC. Generally, prices fall steeply
during the first half of device life, then stabilize at a relatively low level until end of life.
Competitive supply and ever-growing obsolescence drive these price declines. During periods
when prices are falling fast, shorter cycle times can have a large economic impact, enabling sales
to be made earlier before prices decline further.

Taking SEC’s monthly average selling prices and fabrication volumes during the period March,
1996 – December, 2000 as given, calculations were made of the change in SEC sales revenue
resulting from hypothetical manufacturing cycle times compared to cycle times actually achieved
using SLIM. For example, suppose manufacturing cycle times in month \( X \) were extended fifteen
days. Then it is assumed that sales of half of the manufacturing output in that month would be
shifted into month \( X+1 \). If selling prices are lower in month \( X+1 \), there results a revenue loss (or
a revenue gain if prices are rising).
Figure 12. Cycle Time at SEC Fabs
Figure 13. DRAM Cycle Time

The graph shows the average fab cycle time in days for different memory capacities from 4M to 128M, with data points from December 1995 to June 1999. The cycle times vary over time, with some peaks and troughs, indicating fluctuations in manufacturing efficiency or demand.
Figure 14. DRAM Average Selling Prices
Three cases were calculated:

(1) Cycle times achievable without SLIM are bounded below by 4.0 days per mask layer. That is, in periods when actual cycle times were below 4 days per layer, we calculated changes in sales revenues assuming production output was delayed by an amount equal to the difference between actual cycle time and a 4-days-per-layer cycle time.

(2) Cycle times achievable without SLIM are bounded below by 3.0 days per mask layer.

(3) Cycle times achievable without SLIM are bounded below by 2.5 days per mask layer.

Table 2 displays the results. Revenue differences have been calculated only for the SEC fab lines in Kiheung producing DRAMs during the term of the SLIM project. Two cases have been considered: revenue gains only from actual DRAM output, and revenue gains from the entire fab output, assuming DRAM output displaced the non-DRAM output in each fab on a wafer-per-wafer basis. Generally, prices of the non-DRAM products are higher and falling just as fast, so these calculated revenue gains are conservative. Not included in the calculation are the SAS fab, the Bucheon fabs, or the Kiheung fabs not producing DRAMs, even though benefits from the cycle time improvements realized in those lines are also significant.

<table>
<thead>
<tr>
<th>Assumed cycle time achievement level without SLIM</th>
<th>Benefits from only DRAM production in Kiheung Lines 3-8</th>
<th>Benefits from DRAM production scaled to replace non-DRAM production in Lines 3-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTPL = 4.0</td>
<td>$954</td>
<td>$1,133</td>
</tr>
<tr>
<td>CTPL = 3.0</td>
<td>$236</td>
<td>$283</td>
</tr>
<tr>
<td>CTPL = 2.5</td>
<td>$54</td>
<td>$73</td>
</tr>
</tbody>
</table>

Notes: Calculated benefits reflect only the effect of declining average selling prices and do not include cost savings stemming from improved yields and other factors. Benefits for Line 3 are tabulated only for the period 3/96 - 8/96. Benefits for Lines 4 and 5 are tabulated only for the period 3/96 - 5/98. Economic benefits from cycle time reduction in SAS, in all Bucheon Lines and in Kiheung Lines 9 and 10 are not included.

Even for this conservative estimate, the numbers are huge. If SEC had not reduced cycle times from its level of performance just before SLIM, over $1 billion in sales revenue would have been missed, compared to the revenues SEC achieved using SLIM. If application of methodologies and software other than SLIM had enabled SEC to reduce cycle times to three days per layer – a better-than-average performance among DRAM manufacturers – SEC nonetheless would have missed more than $280 million in sales revenue enabled by SLIM. And if SEC somehow had been able without SLIM to reduce cycle times to two and one half days per layer – a
performance superior to other DRAM producers achieving comparable yields and wafer throughputs – SEC still would have been more than $70 million behind the sales revenue it obtained using SLIM.

In our judgment, the CTPL=3.0 case is a conservative valuation of the SLIM project. With intense focus, the SEC manufacturing managers of Lines 4/5 and of Lines 6/7 achieved 3.0 days per layer in May, 1996, and May, 1997, respectively, just before implementation of SLIM systems in those lines. The comment of one manager at that point was that he could not take the WIP down any lower without the support of better scheduling. He got it from SLIM.

Almost all DRAM manufacturers reported heavy losses during the 1996-1998 period, and many major producers either exited the market entirely or sharply curtailed their participation. In contrast, SEC alone reported profits every year and generated sufficient internal cash flow to aggressively expand capacity. SLIM is rightly considered to be a key factor enabling SEC to sustain its leadership in the DRAM market.

As impressive as these economic benefits are, the potential is even greater. The SLIM project was initiated too late to have a large impact on the sales revenues gathered from the 4M and 16M DRAM generations. Prices had already bottomed out by the second half of 1996. (In fact, for the 2.5 CTPL case above, the SLIM cycle times result in a calculated loss of more than $9 million on 16M DRAMs, because prices for this device were rising over most of the project term.) While the SLIM team was focusing on the lines fabricating 4M/16M DRAMs in the first year of the project, prices of 64M DRAMs fell from more than $200 to less than $50. By the time the attention of the SLIM team was directed to the fab lines producing 64M devices, a major opportunity for revenue gains had passed. Only the lower-volume 128M DRAM half-generation family of devices experienced the benefits of SLIM over its early life.

Looking forward, if SEC can sustain its efficient management of cycle times through the 256M DRAM generation that is now starting to ramp up, the economic returns promise to be staggering.

Summary

In the short span of three years, SEC transformed itself from the worst-cycle-time semiconductor manufacturer to the best. Factory floor scheduling and WIP management became a core competence at SEC.

This transformation required the concerted efforts and enlightened thinking on the part of the entire manufacturing and engineering organization. The way SEC management and staff thought about manufacturing fundamentally changed. This achievement also required the formulation and implementation of advanced planning and scheduling applications, including the following:

-- Methodology for calculating target cycle times and target WIP levels for individual manufacturing steps based on an integral generalization of Little’s Law and buffering of bottleneck steps according to measured variability in cycle times between bottleneck steps.
-- Heuristic algorithms for scheduling and dispatching based on WIP management rather than lot priorities. Living Gantt Charts are maintained for heterogeneous machines.

-- Optimization-based planning of semiconductor fabrication. Capacity analysis of individual process steps with static and dynamic machine assignment constraints, scheduling of WIP movement, and non-integer and time-varying target cycle times are rigorously treated.

The financial returns to Samsung from this project up through year 2000 exceeded one billion US dollars. The competitive advantage afforded to Samsung transformed the industry. As great as the returns were to Samsung, the financial pain to other memory chip manufacturers was worse. With Samsung flooding the market with new-technology memory chips before its competitors could, prices were forced down, and Samsung’s competitors could only sell their products for prices much lower than the prices Samsung enjoyed. The second largest Korean semiconductor manufacturers, Hyundai, was forced into bankruptcy. The third largest, LG, also ran into financial trouble and was merged with the remnants of Hyundai into a new firm named Hynix. The creation of Hynix, and restructuring after a second bankruptcy, required more than $3 billion in Korean government funding to keep it alive. IBM and Texas Instruments ceased memory chip production. Micron Technology and Siemens each lost more than $500 million. The great Japanese electronics companies who just a decade earlier dominated the memory chip business – NEC, Toshiba, Hitachi, Fujitsu, Mitsubishi – were basically eliminated as DRAM suppliers. DRAM market shares are measured in dollars. With Samsung enjoying higher average selling prices, and with competitors suffering, Samsung’s DRAM market share rose from 21% to more than 40% a few years after the SLIM project was completed. Subsequently, Samsung ported the SLIM scheduling methodology to its liquid crystal display business, assisting its growth of market share in that business.

Those of us from Leachman and Associates came away from the SLIM project thoroughly humbled by the challenges presented by semiconductor manufacturing. It is incredibly difficult to do well. Each new generation of process technology and equipment presents new challenges for simultaneously achieving high yield, high throughput and low cycle time. In effect, semiconductor manufacturing is an intellectual Black Hole – despite any amount of brain power thrown at it, it always offers challenges for doing it better.

Acknowledgments

Special recognition must be given to Chang-Ho Choi, formerly Executive Director of Finance and Administration for the Semiconductor Manufacturing Division of SEC, and currently the ranking executive for SEC’s manufacturing operations in Tijuana, Mexico. Director Choi initially championed a SLIM project involving the engagement of L&A by SEC, and was instrumental in convincing other senior managers at SEC to start up and to subsequently sustain SLIM. Without him, SLIM never would have happened.

In addition to the authors, L&A employees participating in the SLIM project teams included Jeonghoon Mo, Sejung Kim, Dr. Younghoon Lee, Dr. Sooyoung Kim and Jerome Levdoux. While in the employ of SEC, Dr. Lee served as SLIM Team manager during the first year of the
project. U C Berkeley graduate students Jingliang Chen and Payman Jula assisted the project by developing and testing the SLIM-F simulations of the SLIM logic.

All of the SLIM principles and algorithms were formulated by the authors in response to challenges posed and framed in the working group discussions with the SLIM Teams. The SLIM technical approach reflects inputs and ideas from many and is the result of a process that led to a consensus approval. Thus all of the members of the SLIM Teams made contributions to the design and development of SLIM, and in a sense they all should be considered as co-authors. In the final analysis, SLIM embodies the best efforts of SEC management and engineering staff to manage cycle time.

9. Evaluation of the Case Studies

Company A focused on the high-utilization photolithography area. It reduced variability by reducing setup times and run sizes and forcing more linearity. Company B effectively deployed the Negative Scheduling paradigm and excellent information systems to reduce variability. It reduced setups, it made production activity and engineering activity respond to trouble so as to minimize the impacts on production line linearity, and it reduced standard cycle time by collapsing steps using automation and rearranging the equipment layout to reduce transport time and facilitate more integration. At the planning level, it provided great stability by not rescheduling WIP, by phasing in mix changes gradually, and by ensuring the new releases were capacity-feasible and reticle-feasible. Company C coped with its custom-production environment by implementing and enforcing least-slack dispatching, thereby reducing variability in lot cycle times and permitting a substantial reduction in the quoted lead time. Moreover, company C provided good stability in its production planning, ensuring new lot starts were consistent with a capacity allocation to process technologies and never rescheduling WIP. Harris also provided strong stability from its planning system, not rescheduling WIP (but forcing it to keep moving according to target cycle times) and ensuring planned factory starts did not exceed equipment capacity remaining after capacity consumption by the WIP flush. Samsung developed shrewd logic for establishing target cycle times so as to protect the photolithography bottleneck according to statistics on the magnitudes of trouble occurring in the various portions of the fab process. Moreover, production targets followed at each stage reflected continuous re-balancing of the WIP so as to achieve target cycle times, target fab outs, and facilitate minimum variability in WIP supply to the photo bottleneck and hence maximum utilization. Intelligent logic for allocating lots among alternative machines also was implemented, thereby enabling higher throughput and engendering further opportunities for cycle time reduction.

While a wide variety of analytical techniques were successfully applied to reduce cycle time in these case studies, we can nonetheless identify some important commonalities. Successful cycle time reduction projects seem to require participation and leadership of the Production Team in the factories. Improvements inevitably required improved information, so strong support from the Information Systems Team is required to develop innovative new systems. It is important to note that none of the case-study systems could be fully specified before the cycle time reduction projects began. In every case, the scope of the systems expanded and the techniques and data
were refined as the project matured. In all cases, required information improvements included the following:

- Accurate data on cycle times, down to the step level, and data on variability in cycle times
- Fab starts planning must keep up with changes in demand yet always regulate workloads on high-utilization equipment
- There must be control or guidance of production activity and engineering activity so as to minimize variability and maintain performance to target cycle times and target output

Every project made significant and productive efforts to drive variability out of the fabrication process, or at least mitigate the amount of variability generated by unfavorable events.

10. References


