

SLIM: Short Cycle Time and Low Inventory in Manufacturing at Samsung Electronics

Robert C. Leachman • Jeenyoungh Kang • Vincent Lin

Department of Industrial Engineering and Operations Research, University of California at Berkeley, Berkeley, California 94720-1777

IBM Korea, Inc., MMAA Building, Dogok-dong, Gangnam-gu, Seoul, Korea 467-12

Leachman and Associates LLC, 5870 Carmel Way, Union City, California 94587

leachman@ieor.berkeley.edu • kangjy@kr.ibm.com • vincent@leachmanandassociates.com

SLIM is a set of methodologies and scheduling applications for managing cycle time in semiconductor manufacturing. SLIM includes methodology for calculating target cycle times and target WIP levels for individual manufacturing steps, heuristic algorithms for factory floor scheduling, and optimization-based capacity analysis. Between 1996 and 1999, Samsung Electronics Corp., Ltd., implemented SLIM in all its semiconductor manufacturing facilities. It reduced manufacturing cycle times to fabricate dynamic random access memory devices from more than 80 days to less than 30. Considering the decline of selling prices for dynamic random access memory devices, SLIM enabled Samsung to capture an additional \$1 billion in sales revenue compared to the revenue it would have realized had cycle times not been reduced. (*Manufacturing: performance, productivity. Industries: computer, electronics.*)

Samsung Electronics Corp., Ltd. (SEC) is a leading merchant of dynamic random access memory (DRAM) devices, static random access memory (SRAM) devices, and other advanced digital integrated circuits. SEC has sustained about a 20 percent market share of the vast DRAM market every year since 1995. In terms of unit volume, SEC is the largest manufacturer of digital integrated circuits in the world. SEC's advanced memory-chip products require more than 400 fabrication steps performed in multibillion-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, SEC fabricates more than 300,000 silicon wafers per month (about half six-inch wafers and half eight-inch) and employs over 10,000 people. SEC's chip assembly and test site in Onyang, South Korea, employs over 4,000 people. SEC also operates a large semiconductor fabrication plant in Austin, Texas, under the auspices of its US semiconductor subsidiary, Samsung Austin Semiconductors (SAS).

Cycle time is the industry's term for the manufacturing flow time, that is, the elapsed time from the release of a lot of blank silicon wafers into the fabrication process until completion of the devices that are fabricated on those wafers. The entire semiconductor manufacturing process may be divided into four stages: wafer fabrication, electrical die sort, device assembly, and device testing. Depending on device and wafer sizes, each wafer processed through the fabrication stage contains 400-800 identical memory devices that are cut out of the wafer and packaged and tested in the assembly and testing stages. Most of the cycle time is consumed in the wafer-fabrication stage.

Since 1992, the competitive semiconductor manufacturing (CSM) program at the University of California at Berkeley has been benchmarking the performance of semiconductor fabrication plants (fabs) around the world. The CSM program documents 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 first two authors were part of the CSM team visiting the site. The CSM program found that SEC achieved excellent yields and excellent productivity of equipment and labor, but its manufacturing cycle time was the worst of 29 fabs in the CSM survey. Line 3's cycle time was about 35 percent longer than the average of those surveyed and more than double the leading-edge performance of 2.0 days per layer of circuitry. The Berkeley team presented SEC manufacturing management with its 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 reducing cycle time had just become a priority.

The year 1995 was memorable for manufacturers of DRAM devices. Soaring worldwide demand, coupled with tight capacity, pushed prices of four megabit DRAMs upwards in a sellers' market unlike any seen

SEC had become a victim of its own success.

before or since in the memory-chip business. DRAM producers, including SEC, enjoyed record profits. As a consequence, existing DRAM manufacturers and new entrants invested heavily in new fabrication capacity. At the end of 1995, the market turned. Prices collapsed in early 1996, and the market quickly became a buyers' market.

Recognizing the impending shift, in December 1995, Y. W. Lee, president of SEC's semiconductor business, informed the SEC semiconductor manufacturing department of the urgent need to reduce cycle time. The huge work-in-process inventory was going to lose value rapidly. President Lee also recognized that customers would become much more selective about their DRAM suppliers. The CSM survey's finding that SEC's cycle times were noncompetitive implied that 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. Leachman and Associates submitted a proposal for a one-year project to SEC in February 1996, which it accepted. A project team consisting of staff from both Leachman and Associates and SEC was formed. SEC gave the project the acronym SLIM (short cycle time and low inventory in manufacturing) and initiated it in March 1996. The project was extended for four more one-year periods, ultimately ending in June 2001.

The Problem

Wafer fabrication is one of the most complex manufacturing processes. Lots of 25 silicon wafers pass through hundreds of manufacturing steps in which 18 to 25 layers of integrated circuitry are fabricated on the surface of the wafers. Each layer involves visits to specialized processing equipment that perform photolithography, diffusion, etching, ion implantation, chemical-vapor deposition, cleaning, ashing, measurement, and other processes in a specified sequence. Batch sizes at various equipment types range from one wafer to 150 wafers.

Process equipment is unavailable during preventive maintenance, engineering work, unplanned repairs, and requalification. The manufacturing process is very delicate and somewhat unstable, especially for new devices and new process technologies. Individual lots or process steps may be put on hold while engineers try to restore process stability. Of a group of seemingly identical machines, the engineers may qualify only a few successfully to perform a particular step on a particular device, and the list of qualified machines may change as the engineers struggle to achieve process controllability.

These factors make the flow of work in process (WIP) through the factory more turbulent than it is for other products. To attain reasonable productivity, the manufacturing process needs significant WIP. For example, in the 1995 CSM survey, the best-performing memory-device fab achieved an average cycle time of about 2.0 days per layer of circuitry (Leachman and Hodges 1996). The intrinsic cycle time, that is, the sum of times required for machine processing and material

handling at each step, was about 0.8 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 type of equipment may perform as many as 30 different manufacturing steps to fabricate a given device, each equipment type must perform a large variety of steps. For most machine types, some productivity is lost during changeovers between different process steps, and frequent changeovers make sustaining process control difficult. Moreover, imbalances in the distribution of WIP through the manufacturing process arising from the turbulence described above can be mitigated or exacerbated, depending on the choice of which process step to perform. Thus, improving production scheduling can reduce cycle time.

Generally, in SEC fabs, the machines performing photolithography, known as *steppers*, are the most heavily utilized and thus form bottlenecks. (The photolithography machines must “step” across the surface of the wafer making many exposures to cover the entire surface.) Steppers are heterogeneous. Different

More than 3,000 people attended training sessions.

makes and models have varying capabilities to perform various photolithography steps accurately. Moreover, to achieve controllability, the process engineers may qualify only a small subset of a particular make and model to perform a certain step on a certain device. Adding even more complexity, the engineers sometimes base the list of suitable machines for a given lot at certain photo steps on which stepper processed the lot at some preceding photo step days or weeks earlier.

In this situation, determining the capacity of a fab line is complex and dependent on the product mix. It is difficult to plan changes in the mix of the devices produced and keep the factory running at maximum output without causing a WIP buildup at any steppers or losing process stability. For the same reason, it is also difficult to plan how many steppers to qualify to perform each step. Especially difficult to manage are transitions between generations of DRAMs, when the

volume of a new-generation device must be ramped up as the volume of the older-generation device decreases. In the most unfortunate cases, no one realizes certain steppers are overloaded until WIP backs up. Thus, improving production planning can reduce cycle time as well.

In its planning and scheduling practices, SEC's starting point was not too different from that of many other semiconductor manufacturers at the time. It based production planning on an aggregate capacity expressed for each fab line and did not analyze the qualifications for performing particular steps on particular devices of each stepper. It established a monthly target quantity for each device in each fab line. The scheduled input of blank wafer lots roughly corresponded to the fab output schedule shifted backwards in time by a target cycle time. But it did some batching of fab input lots into groups corresponding to the largest batch size of all equipment types, and it input some lots early to sustain maximum productivity.

The company used a lot-dispatching system known as MSS (manufacturing scheduling system) in the fab lines. MSS prioritized production lots waiting for processing at each equipment type based on the age of the lot versus a target for that age. (This is equivalent to the familiar least-slack dispatching rule.) But SEC could not follow this priority list too closely because that would cause an excessive number of equipment changeovers. More influential was the so-called cutoff-schedule method of managing manufacturing. Managers prepared a monthly target output quantity for each device in each fab line. Manufacturing supervisors identified what WIP had to make it out of the factory in the current month to meet the target output, and then they identified the steps to operate to flush this WIP. (The process steps needed for each device were said to fall within the cutoff for the current month's fab outs; hence the name cutoff schedule.) Operators then set up the equipment to run the identified devices through those steps. They set up the equipment for devices and steps for the subsequent month's output only if no WIP remained for the current month's target.

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 when we started the project in early 1996.

Project Strategy

SEC management imposed certain constraints on the project. The company's leadership yields and wafer throughput could not drop as we reduced cycle time. We could have reduced cycle time easily by reducing all production levels or by relaxing the constraints that process control imposed on machine allocation. Instead, we were charged with devising methods by which the fabs would work smarter. Even after enforcing these constraints, SEC management still viewed major reductions in WIP levels as risky, so we were instructed to initiate the SLIM project working on only two of the six fab lines at Kiheung. If the project succeeded with these two lines, SEC could propagate the methods to the other fab lines.

SEC establishes policies for managing production by consensus. Manufacturing managers, supervisors, equipment leaders, and equipment operators all must understand and agree with the logic used to set schedules and make work decisions. Schedules proposed by systems were and are subject to override by the people actually controlling the factory. Factory staff had to understand the basic methodology we proposed for scheduling and had to be convinced of its value. To accompany any changes in the methodology for managing production, we needed to provide education and training.

The urgency for reducing cycle time also constrained our approach to the SLIM project. We would have to improve the sophistication with which cycle time was managed incrementally, enabling SEC to benefit from reducing cycle time 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.

In addition to forming the SLIM project, SEC asked its process engineers to assess process flows for opportunities to reduce cycle time, especially by qualifying more machines to perform bottleneck process steps. SLIM helped guide this engineering activity.

SLIM Principles

Most semiconductor companies manage production under the lot-dispatching paradigm (managing the cycle times of production lots). A newly released lot of wafers is assigned a due date equal to its release date plus a target cycle time. To schedule lots on the factory floor, computers prioritize the lots waiting at each equipment bay. Most commonly, they base this prioritization 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 a different paradigm. Instead of using due dates for lots, it relies on a target fab-out schedule for each device. This output schedule is a continuous-time schedule; for example, if the fab-out schedule is

Reducing cycle times brought many benefits.

expressed in terms of output quantities each day, then it is assumed that one quarter of the quantity of a particular device scheduled on a particular day 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 objects in SLIM are device-steps, not individual lots. Based on a review of actual output to date versus the target fab-out schedule and of actual downstream WIP versus target WIP, SLIM establishes production targets and priorities for each device at each step. We call the paradigm driving SLIM the WIP-management paradigm.

The WIP-management paradigm has several advantages compared to the lot-dispatching paradigm:

(1) It is easier to control the number of setups scheduled. Typically, the SLIM algorithms schedule each device to be set up at a step once per shift (or more times if the volume requires parallel machines) to meet production targets for that shift. Line staff and process engineers find such schedules much more acceptable than lot-dispatch schedules.

(2) In wafer fabrication, lots of the same device will get out of order relative to the order in which they were released. Many time-consuming inspections and

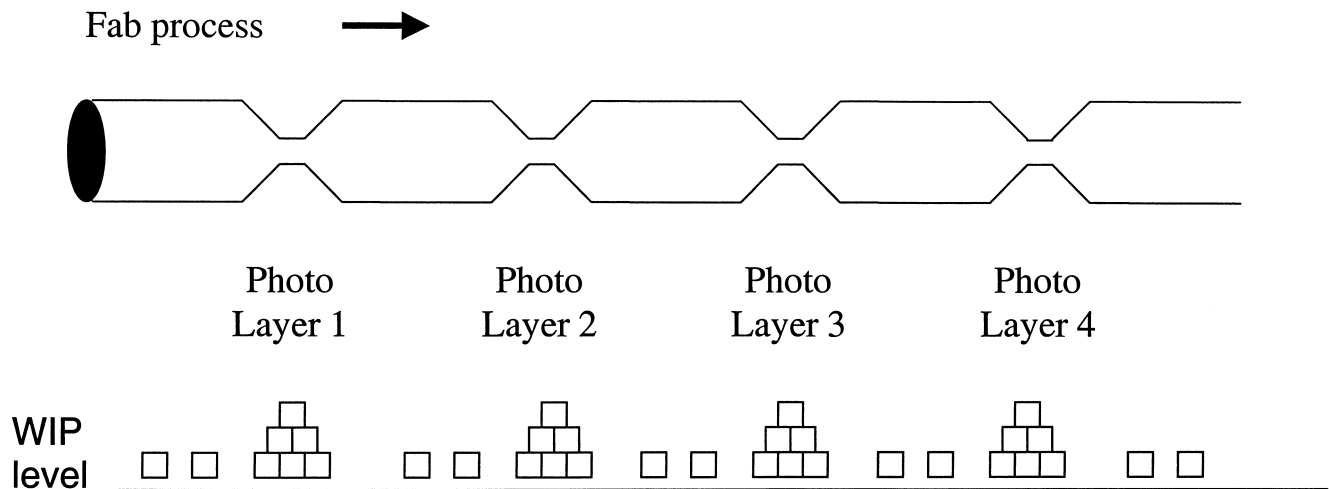


Figure 1: The pipe represents the production line, and its width represents the maximum flow rate or capacity at various process steps. SEC fabs were designed so that the photo machines are the bottlenecks, and other machines have surplus capacity. Thus the pipe in the figure narrows at each photo step. When all the machines are up and the process is in control, the photo machines are the bottlenecks, and the largest concentration of WIP is at those points.

tests are performed on samples of the stream of lots undergoing fabrication. Moreover, individual lots may be placed on hold when process-control issues arise. Under lot-based logic, the due-date performance of each lot is equally important. The lot-dispatching paradigm tries hard to put lots of the same device back in order, which is often unnecessary to meet the fab-out schedule for the device. This logic sometimes directs operators to process lots of a device whose WIP is ahead of schedule rather than lots of other devices whose WIP is behind schedule. Thus the lot-dispatching paradigm can compromise the fab-out schedule. SLIM never compromises the fab-out schedule.

(3) If lots are scrapped or the fab-out schedule changes, all the due dates for lots become incorrect in dispatching systems. Since SLIM incorporates online analysis of the downstream WIP and the fab-out schedule, its targets and priorities never become stale.

Determining Target Cycle Times for Individual Process Steps (SLIM-M)

Managing WIP effectively depends on establishing an efficient WIP profile. For a fab with a constant rate of

output, Little's formula indicates that determining target WIP levels is equivalent to determining target cycle times (Nahmias 2001). We took great care in developing SLIM to establish target cycle times for the steps in fabricating each device.

The fragile nature of advanced semiconductor process technology renders the actual distribution of WIP through the fabrication process dynamic. When all equipment is operating and all process steps are under control, WIP will normally be concentrated in the photo area; typically, 30 to 40 percent of total WIP (Figure 1).

When serious and lasting process or equipment trouble arises in some other area besides photo, the supply of WIP decreases downstream, particularly at the next photo step (Figure 2). Thus at some point, the disruption will lower stepper utilization and fab throughput will drop. While other machines have excess capacity and can run faster than the fab output rate, the steppers cannot. Losses of photo throughput cause losses of fab throughput that cannot be recovered. Clearly, some buffer WIP is needed at photo steps to insulate the steppers from upstream disruptions.

If we watched an accelerated animated film of the fab over time, we might first see the equipment and

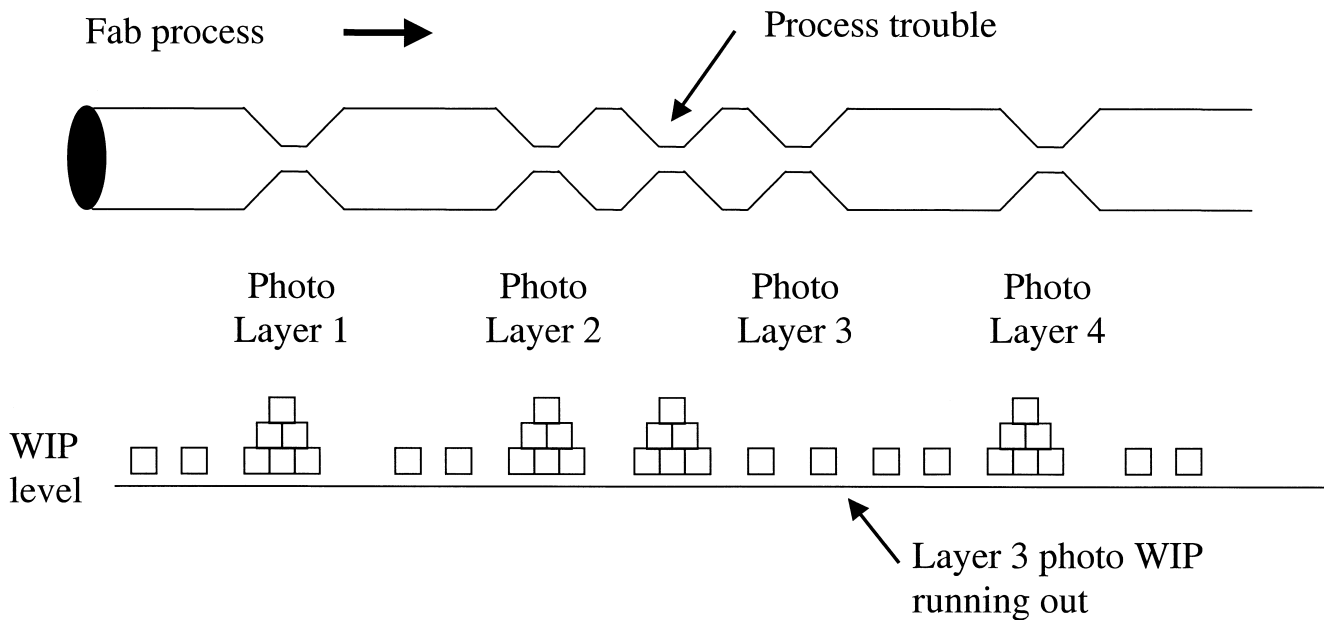


Figure 2: The case of process or equipment trouble at a nonphoto manufacturing area of the fab is depicted by a constriction of the pipe at a nonphoto step. If this condition persists, the WIP at a photo machine can become exhausted. A buffer is needed, proportional to the risk of trouble in that layer. For example, if Layer 2 of the process experiences more trouble than Layer 3, the bottleneck step immediately following Layer 2 should be awarded a larger buffer than the bottleneck step following Layer 3.

process running well, with WIP concentrated in the photo area. Then with a disruption, the WIP population would shift upstream from one or more photo steps. Next, with the problem resolved, the WIP population would migrate back downstream to the photo steps. Then another disruption would occur, and so on. The philosophy underlying SLIM is to distribute WIP to put 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 WIP as possible to the photo bottleneck, to be prepared for 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, that is, 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 the total buffer time for the device, that is,

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

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 .

To establish target cycle times, SLIM allocates all of the budgeted buffer time to bottleneck steps. That is, it sets the target cycle times for nonbottleneck steps equal to the intrinsic cycle times. It sets the target cycle times for bottleneck steps equal to the intrinsic cycle times plus an allocation of the total buffer time. This reflects the philosophy that we should provide the maximum insulation for the bottleneck steps that the process and equipment allow. SLIM allocates WIP to provide buffers proportionate to the amount of disruption each bottleneck step may suffer. The objective is to insure similar utilization of all steppers.

The amount of trouble that actually occurs in a stretch of fabrication process is not easy to measure. As a proxy in SLIM, we measure 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 occurring in this stretch. We allocate buffer time to the downstream bottleneck step in proportion 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 preceding bottleneck step (or at fab start if there is no preceding bottleneck step). Then for bottleneck steps $j = 1, 2, \dots, NB_i$ performed on device i , the target cycle time is

$$TCT_{ij} = ICT_{ij} + BT_{ij},$$

where

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

For all nonbottleneck steps j performed on device i , we set

$$TCT_{ij} = ICT_{ij}.$$

The SLIM methodology automatically adjusts the target cycle times as the process technology evolves. If SEC engineers mitigate the problems in a particular stretch of the process, 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 one downstream.

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 (Lu et al. 1994). Because these methods average over the transient WIP distributions and the distributions when all processes and equipment are working well, these targets call for less concentration of WIP at the bottleneck than SLIM. We would expect lower utilization of the bottleneck equipment 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 (Fordyce et al. 1992). Using this method, one 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 are the bottleneck, this does not seem wise.

Determining Target WIP for Process Steps (SLIM-M)

We developed an integral generalization of Little's formula to translate target cycle times into target WIP levels. If SEC had a constant fab-out rate and anticipated no loss in yield, 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. When the fab-out schedule varies over time, and if the fab achieves exactly the target cycle times and the fab-out schedule, the WIP population at step j should equal the target fab outs in the interval $[TCTFO_{ij-1}, TCTFO_{ij}]$, where $TCTFO_{ij}$ denotes the sum of target cycle times for all steps subsequent to the completion of step j on device i (Appendix).

Short-Term Production Targets and Priorities (SLIM-M)

SLIM analyzes the actual downstream WIP versus the fab-out schedule to determine production targets for short horizons, such as an eight-hour production shift. To understand this, suppose actual fab-outs to date conform exactly to the target fab-out schedule. Suppose also that SEC plans no loss in yield from completion of step j to fab out of device i . Now consider the progress of step j relative to a time horizon of eight hours (that is, 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 a downstream deficit of device i exists. An amount of WIP equal to this deficit should be processed through step j during the current shift 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 because it may be infeasible to process this amount for a variety of reasons: Not enough WIP may be supplied to step j during the shift to complete the IPQ; not enough qualified machines may be available to complete the IPQ

in one shift; or the IPQ could be negative. (In the Appendix, we give a mathematical formula for the IPQ for the more general case in which SEC plans for yield losses at various steps.)

SLIM prepares dispatching priorities for the various device-steps processed by a given equipment type based on the IPQ targets. By dividing the IPQ by the average fab-out rate of the device over the interval $[0, TCTFO_{ij} + 0.33]$, it converts the wafer quantity into units of time. We change the sign of the result so that a negative sign indicates lateness. We call the result the schedule score (SS), reflecting how many shifts early or late step j will be if the fab completes no lots for that step during this shift. For example, a schedule score of -1.0 means the output of the step is behind schedule by an amount equal to one shift's normal production of that device. The IPQ quantity for the step indicates how many wafers need to be processed during this shift to put the step on schedule by the end of the shift.

The schedule score and the ideal production quantity form the basic foundation of SLIM's machine-scheduling algorithms. SLIM prioritizes the device-steps according to their SS. Once SLIM assigns a device-step to a machine, it does not interrupt processing of that device-step until either the IPQ is completed or the available WIP is exhausted. Since the IPQ is a target quantity for a production shift, SLIM tends to schedule device-steps for setup once per machine per shift.

In the extreme case that every lot is a different product, prioritizing device-steps by schedule score becomes equivalent to prioritizing 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. The SS-IPQ schedules are demonstrably much better, both in terms of maintaining the target WIP profiles and in terms of the number of device-step setups.

Scheduling Nonbottleneck Machines (SLIM-L)

In online scheduling of nonbottleneck machines, we take three concerns into account. The first concern is

maintaining the target WIP profile for each device, as reflected in the SS and IPQ scores. The second is maintaining an adequate supply of WIP to fully utilize the machines at the next downstream bottleneck step. For this second concern, consider the case of two device-steps with comparable schedule scores. Suppose that for device-step A the supply of WIP at the next downstream bottleneck step is fairly high, and the amount in the pipeline leading to that step is also high. On the other hand, for device-step B the supply of WIP at the next downstream bottleneck step is low, and the amount in the pipeline leading to that step is also low. Since the machines qualified to perform the bottleneck steps are somewhat inflexible, the machines that perform the next bottleneck step for device-step B may be underutilized. Thus, we prefer to dispatch device-step B .

SLIM includes a metric for assessing the supply of WIP to the next downstream bottleneck step, the *balance index* (BI) (Appendix). The balance index for a given device-step expresses the difference between actual downstream 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 that 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 in terms of keeping the bottleneck machines working at full capacity.

To integrate both concerns, we define priority levels in SLIM-L corresponding to ranges of schedule score and balance index (Table 1). We give priority to device-steps in level 5 over device-steps in level 4 and so on. Within each priority level, we further prioritize device-steps according to least schedule score.

The third concern SLIM-L addresses is to limit the

	BI < -0.5	-0.5 < BI < 0.5	BI > 0.5
SS < -16	5	4	3
-16 < SS < 16	4	3	2
SS > 16	3	2	1

Table 1: SLIM prioritizes nonbottleneck device-steps according to priority levels that integrate priority scores for maintaining the target WIP profile (SS) and for maintaining bottleneck utilization (BI). Device-steps in level 5 are prioritized ahead of those in level 4, and so on.

number of device-step setups scheduled to a level acceptable to line staff. This is done in two ways: First, it will not change a machine that is working on a device-step with a positive IPQ to run a different device-step until either the WIP of that device-step is exhausted or the IPQ becomes negative. Second, it limits the number of parallel machines set up to process a particular device-step to MNM, the minimum number needed to complete the IPQ by the end of the shift (Appendix). For example, suppose 10 lots of the highest priority device-step are in available WIP and there are eight machines. Because processing times are short, machines become ready for dispatching frequently. Suppose the MNM for this device-step is three. Lot dispatching systems might assign all eight machines to be set up to run this device-step. SLIM will set up only three (unless the other machines would be idled). From the point of view of the line staff, the SLIM schedule is much more rational.

Scheduling Bottleneck Machines (SLIM-S)

For the photo area, structuring the scheduling problem as one of selecting device-steps to assign to machines as they become idle, as is done in SLIM-L, is unwise because of tight capacity and 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 having idle machines and unassigned WIP that are not compatible.

In SLIM, a scheduling algorithm (SLIM-S; "S" for stepper) is applied periodically to develop Gantt-chart schedules for every stepper. This algorithm assigns as much as possible of the currently available WIP, addressing the following concerns: first, to achieve maximum utilization and minimum setups of the steppers, since they are the bottleneck machines, and second, to maintain the target WIP profile for each device, as reflected in the SS and IPQ scores.

SLIM-S is a three-pass algorithm. In the first pass, it examines current setups to see if the assigned device-step has fulfilled its IPQ; if not, it assigns more WIP as available. In the second pass, it plans new setups so as

to complete the IPQs of as many device-steps as possible. In the third pass, it assigns the remaining WIP to achieve full utilization. In each pass, it selects steppers for WIP assignment so as to minimize setups and to conform to machine qualifications. It uses a least-candidate-WIP rule when steppers are qualified for a limited number of device-steps. By making the assignment to the machine that has the least total WIP to choose from, we reduce the likelihood that it will become idle for lack of suitable WIP. When qualifications are quite flexible, the algorithm uses the most-available-time rule to accelerate schedule recovery.

During the course of the shift, more WIP will arrive at the photo area. Another algorithm, sub-SLIM-S, considers this WIP for insertion into the Gantt chart schedules of the machines. In principle, we could run this algorithm on a transaction basis every time new WIP arrives, but in practice at SEC, we run it every 10 minutes. Sub-SLIM-S inserts into the Gantt-charts of the current SLIM-S schedule WIP arrivals that are urgent (for 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.

Another issue in scheduling the steppers concerns the masks used to print the patterns on the wafer. Sometimes for a given device-step, masks may be fewer than qualified steppers, and sometimes, different steppers require different masks for the same device-step. SLIM-S and sub-SLIM-S schedule the use of masks as well as steppers.

Scheduling Diffusion Batches (SLIM-D)

Diffusion furnaces accommodate one to six lots for a fixed-duration machine cycle. Typically, several device-steps have identical furnace-process steps. These factors must be considered in deciding what device-steps to include in a run of a particular furnace process-step and whether to run the furnace step partially full or wait for more WIP to arrive.

Like SLIM-L, SLIM-D is an online, transaction-based algorithm. It uses the same priority ordering of device-steps as SLIM-L. For each device-step, SLIM-D formulates tentative furnace batches considering the WIP at hand. It first selects lots of the device-step under

consideration; if insufficient to fill the furnace, it considers lots of the next highest priority device-step that are compatible in the same furnace run, and so on, until it forms a full furnace batch from the available WIP. For each furnace step, it must form a batch that exceeds a minimum batch size (MBS). Otherwise, it postpones dispatching the device-step until additional compatible WIP arrives.

Scheduling the Release of New Lots (SLIM-I)

The SLIM-I algorithm schedules the release of new lots into the fab periodically, such as once per shift or once per day. Given a choice of time horizon (shift or day), SLIM-I calculates an IPQ for the starting step for each device. That is, for each device, we first calculate what quantity we need to reach the target WIP level. We round up this value to an integer multiple of the standard batch size for the release of the device. Next, we estimate the workload on the qualified steppers over the estimated lead time from release to arrival at the first photo step plus the length of the horizon. If the machines have no remaining capacity, we block the release. Otherwise, we schedule the release quantity calculated. SLIM-I thus incorporates the ideas of constant WIP control (Hopp and Spearman 1996) and bottleneck workload regulation (Wein 1988).

Line Simulation (SLIM-F)

We incorporated all of the SLIM scheduling modules in a discrete-event simulation model, SLIM-F. We feed the SLIM-F simulation with initial conditions corresponding to the actual factory state (WIP status, equipment status, machine-arrangement tables, target out schedule, and so forth). We simulate factory operation up to a user-specified horizon.

Systems technology engineers use SLIM-F to test changes in the SLIM logic and to monitor factory use of SLIM by comparing actual WIP movement to simulated movement. A discrepancy between the two indicates that use of SLIM has faltered because of a failure in data maintenance, a software error, or the factory staff refusing to accept the SLIM schedules. SLIM managers then seek the cause and take corrective action.

SEC also uses SLIM-F to support maintenance and engineering activities. For example, 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 when particular machines will be idle or underutilized, when preventive maintenance would be the least disruptive. 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 simulated future WIP photo levels, process engineers can determine when they must requalify a machine to avoid a serious impact on cycle times.

SLIM Planning Logic

The effectiveness of the SLIM floor-scheduling logic depends on an assumption that the target fab-out schedule is consistent with the bottleneck machine capacities. We need a detailed analysis comparing fab-out demands with machine capacities.

Many factors make analyzing capacity at the SEC fabs challenging. Three important factors are particularly important. First, the product mix is highly dynamic. For even a single type of device, such as a 64M synchronous DRAM, 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 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. The product mix thus changes continually. A static capacity-analysis tool cannot accurately assess the workloads on individual machines over time.

Second, SEC often changes the fab-out schedule inside the target cycle time, sometimes dramatically, as it revises forecasts of customer demand. Such changes mean that the planning model must explicitly schedule WIP movement; to be accurate, it cannot assume that initial WIP should simply continue to move through its process flow according to prespecified cycle times.

The third factor is that the machines of a general type (such as steppers) are not homogeneous. A variety of makes and models exist, and limited numbers of machines are qualified to perform each device-step. In the most challenging case, what machines are qualified to perform a certain device-step is a function of what machines were used on that device during a previous step. An accurate planning model therefore cannot treat the machines of a particular type as a homogeneous group with an aggregate capacity. Instead, the model must calculate feasible allocations of WIP to individual machines.

We formulated a linear programming (LP) model to cope with these challenges. It includes variables for the

If SEC had not reduced cycle times, it would have missed over \$1 billion in sales revenue.

release of new lots into the fab and for 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. We defined additional variables to route new releases and initial WIP through alternative machines. We use constraints to express the capacity of individual machines in each period and the consistency of WIP movement through different steps in different time frames. We also developed constraints to express back-orders or finished-goods inventory for different classes of demands. Objective functions minimize back-orders and finished-goods inventory. We developed an application to generate this formulation from the SLIM database and, after optimization using a commercial software package, to prepare user reports: the SLIM-O (O for output) capacity-analysis tool.

SLIM-O incorporates a much more precise capacity analysis than is typical in LP models. Lead-time parameters are time-varying and noninteger, and constraints relating production variables to workloads and demands are formulated at noninteger 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. Lin (1999) and Leachman (2002) give details of the SLIM-O LP formulation.

SEC relies on SLIM-O for transition planning, that is, planning for the period when it is ramping up a new-generation DRAM device while ramping down an older-generation DRAM device. The fab goes through a learning curve of cycle time in the first months of a new device's existence. Initially, few machines are qualified for the various manufacturing steps, as the engineers struggle to achieve a robust and controllable process. As time goes on, they must qualify more pieces of equipment to increase production volumes and shorten cycle times. SLIM-O is designed to identify and plan the qualifications of steppers and other equipment needed to relieve bottlenecks and to speed up production and the learning curve for the new device.

We implemented another version of SLIM-O for analyzing capacity in the device assembly and testing areas. The primary technical issue here is the considerable tooling needed for certain steps and the limited compatibility of different types of tools. In addition to its other features, this version of SLIM-O includes variables assigning workloads to alternative sets of compatible tools.

SLIM Implementation

SEC formed interdisciplinary project teams, each focusing on SLIM systems development, training, and implementation in two or three large production lines. Ultimately, it formed seven such teams. Members of each team included staff from SEC Manufacturing, Photo Engineering, Total Productivity (an IE department), Production Control and Systems Technology (MIS), and staff from Leachman and Associates (L and A). On every team, some team members from both SEC and L and A had graduate-level OR/MS training.

Each team set up an online database supporting SLIM and identified any gaps in electronically stored data (intrinsic cycle times, process times, machine qualification tables, and so forth) so they could begin industrial-engineering efforts and data-interface efforts immediately. Fortunately for the SLIM project, SEC

was data rich compared to other semiconductor companies. When the project began, the SEC total productivity team already maintained a fairly complete database of intrinsic cycle times and process times for most device-steps and efficiency and availability metrics for the major equipment types in all fab lines. Nonetheless, the SLIM teams needed to improve and maintain data.

The complexity of the various SLIM algorithms and the required links to other systems determined who developed the software. SEC Systems Technology coded all of the user interfaces, database queries, and interfaces to factory-floor execution systems, and it coded and maintained the SLIM-M, SLIM-L, SLIM-S, and SLIM-I formulas and algorithms that we had developed. Leachman and Associates coded and maintained SLIM-O and SLIM-F. We used various commercial software packages in SLIM. In particular, we used the CPLEX optimization software from ILOG to solve linear programs in SLIM-O, and we used the APF software tools from the AutoSimulations division of Brooks Automation to implement the SLIM-F simulation, to convey SLIM-L and SLIM-S scheduling decisions to the factory floor, and to update the SLIM database with factory-floor status.

Managers, manufacturing staff, and engineering staff needed training in the SLIM principles and logic if they were to accept and use SLIM. The SLIM teams ran classes for practically every manufacturing department employee (including all the supervisors and operators staffing every production shift), all of the managers, and many of the process and equipment engineers. Ultimately, more than 3,000 people attended training sessions.

Each team made a phased implementation of the SLIM scheduling modules in an effort to reduce cycle times as early as possible. They implemented SLIM-M first, so that production managers and supervisors could begin maintaining the target WIP profile and reducing WIP to levels consistent with the target cycle times. Next, they implemented SLIM-I, SLIM-L, SLIM-S, and SLIM-D. Typically, they first displayed schedules in an advisory mode, including the online SLIM-S Gantt chart schedule for all the steppers in the photo area and the SLIM-L online device-step priorities (SS, IPQ, how many machines to set up) in other areas.

Later, they fully integrated SLIM-L and SLIM-S with the factory-execution system in Kiheung Lines 6, 7, and 8 and in the SAS fab in Texas, reducing the need for human involvement in scheduling to an on-exception basis.

They next implemented SLIM-O, which helped SEC process engineers to reduce cycle times; in some cases, SEC engineers reallocated machines to device-steps and qualified additional machines for bottleneck device-steps identified by SLIM-O. These changes reduced cycle times.

Finally, they implemented SLIM-F. By simulating fab operation in previous periods, SEC staff could monitor use of SLIM. By simulating fab operations in future periods, they could schedule maintenance and engineering in light of their impacts on cycle time.

As the teams implemented each module, they reduced cycle times. As they reduced actual cycle times, they further reduced target cycle times, generally aiming for about 0.2 days less than the actual cycle times per circuitry layer. Generally, each SLIM team reduced cycle time significantly within three months. They generally took about two years to achieve full benefits (Table 2).

From the initial team formed in early 1996 to reduce cycle times in two fab lines, the SLIM project gradually expanded to embrace almost all SEC semiconductor manufacturing facilities. Throughout the project, SEC executives gave the SLIM project teams strong support. They impressed upon every manufacturing and engineering employee the importance of the project, and the SLIM teams generally received excellent cooperation. The president of the semiconductor business (Y. W. Lee) personally reviewed and approved the SLIM plan and budget each year of the project. 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 status of the project and provide direction.

SLIM Results and Economic Benefits

Cycle times fell steadily as we implemented and refined SLIM. From 4.0 days or more per layer of circuitry at the start of 1996, cycle times in SEC's flagship

Period	Project Activity	Highlights of Results
March 1996–February 1997	SLIM team formed for Kiheung fab lines 4 and 5. L&A staff on site one week per month, plus one L&A staff member full-time on site 6/96–8/96.	Cycle time per layer (CTPL) for lines 4 and 5 reduced from 4.5 to 2.0 days.
March 1997–February 1998	SLIM team formed for Kiheung lines 6 and 7. L&A staff on site one week per month, one L&A staff member on site four additional days per month.	CTPL for lines 4 and 5 reduced to 1.5–1.6. CTPL for Lines 6 and 7 reduced from 3.3–3.5 to 2.5.
March 1998–February 1999	SLIM teams formed for Kiheung lines 2, 3, and 8, for three fab lines at Bucheon, Korea, and for the SAS fab line in Austin, Texas. L&A posts full-time staff in Korea.	CTPL in lines 6, 7 and 8 reduced to 1.3–1.6. Line 8 sustained a CTPL of 1.3 for three months in 1999 while operating at full capacity.
March 1999–March 2000	SLIM teams formed for Kiheung Electrical Die Sort (EDS) and for Assembly and Test operations at Onyang, Korea.	EDS on-time delivery (OTD) improved from 70 to 96 percent. Test OTD improved from 74 to 98 percent.
April 2000–June 2001	Focus on simulation, scheduling of maintenance and engineering, and planning the ramp-up of new devices.	

Table 2: This chronology of the SLIM project includes highlights of the project results. Ultimately, seven SLIM teams were formed, each reducing cycle times in two or three manufacturing lines.

DRAM fabs dropped to 1.3 to 1.6 days per layer by early 1999. Considering that SEC's intrinsic cycle times were 0.9 to 1.1 days per layer, this represented a decrease in cycle times from about 4X to about 1.5X.

Immediately after implementation of SLIM-M, L, and S, factory managers typically observed a redistribution of fab WIP. The percent of total fab WIP in the photo area typically rose by 10 to 15 percent. This enabled higher utilization of the fairly 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 21-shift simulation of Line 4 performed shortly before its implementation, the SLIM-S algorithm achieved about 12 percent higher wafer throughput in the photo area than the actual performance in the line. This improvement alternatively could be translated into a further reduction of the WIP level needed to sustain current throughput. After making engineering improvements to mitigate bottlenecks identified by SLIM-O, SEC attained a further reduction in WIP or a further increase in throughput or both.

From 80 days or more to fabricate DRAMs in late 1995, the SLIM teams brought these cycle times down to about 30 days by the end of 1998 (Figure 3). We passed a major milestone in November 1998, when

SEC reduced the cycle time for fabricating third-generation 64M DRAMs in Line 7 to 27 days (about 1.35 days per layer of circuitry) and reduced the cycle time for electrical die sort following wafer fabrication to three days. In doing this, the semiconductor manufacturing division met a challenge laid down by SEC CEO J. Y. Yun earlier in 1998: Make 64M DRAMs in 30 days. SEC manufacturing managers had been doubtful that they could achieve such performance at high volume, but Executive Yun was sure they could. He was right.

Reducing cycle times brought many benefits. The accuracy of the sales forecasts used in production planning improved, thereby reducing inventory levels. The lead times to customers were reduced, transforming a competitive disadvantage into a competitive advantage. Reduced cycle times facilitated more rapid feedback of the results of engineering experiments as well as more rapid feedback of customer suggestions for product modifications, allowing SEC to improve products and process designs. Such benefits are real but difficult to quantify in dollar terms. But reducing cycle time results in another large but often overlooked economic benefit.

Prices for DRAMs and other memory devices generally drop dramatically over time (Figure 4). Generally, prices fall steeply during the first half of devices'

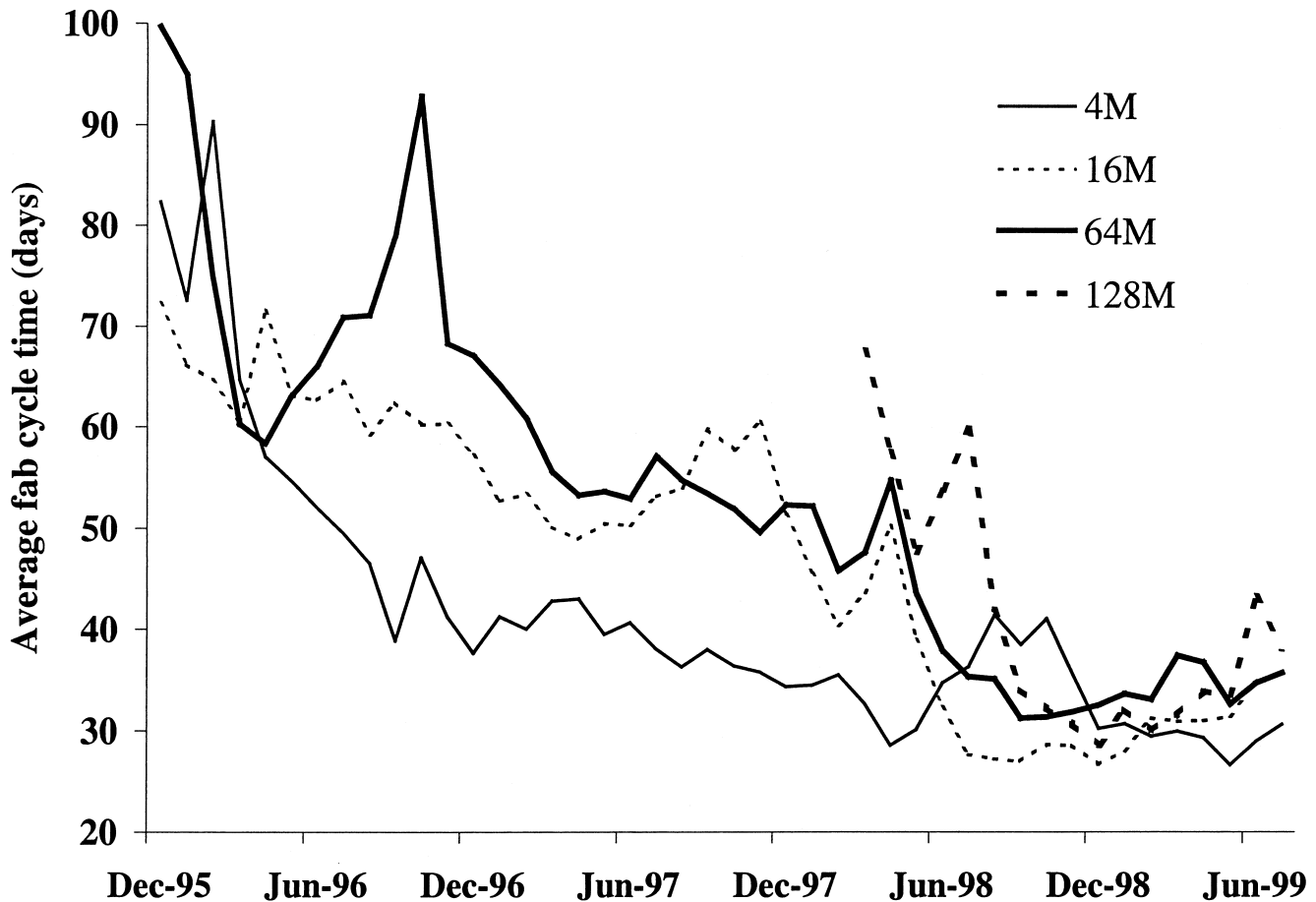


Figure 3: This graph of SEC fabrication cycle times shows trends in average cycle times for the fabrication of 4M, 16M, 64M, and 128M DRAM product families at the Kiheung fabs. The SLIM project began on lines 4 and 5 (4M and some 16M production) in March 1996. The project was extended to lines 6 and 7 (16M and some 64M production) in March 1997 and to all SEC fab lines in March 1998. SLIM dropped DRAM cycle times from more than 80 days to less than 30.

life, and then they stabilize at a low level until end of life. Competitive supply and ever-growing obsolescence drive these price declines. For any DRAM device, new and improved versions follow it in the design-and-manufacturing pipelines of SEC and its competitors, and the price of the original device falls. SEC tries to offer new and improved versions of DRAMs as early as possible, to always operate fabrication lines at full capacity, and to sell off all output at market-clearing prices. Because of the rapid obsolescence, SEC does not hold inventory of finished DRAMs.

When prices are falling fast, SEC cannot simply

make wafer starts earlier to increase revenues, since it deploys new products as soon as possible and the fabrication lines operate at capacity. But shortening cycle times can enable SEC to make more sales earlier before prices decline further, thus greatly increasing revenues.

Taking SEC's monthly average selling prices and fabrication volumes from March 1996 to December 2000 as given, we calculated the sales revenue SEC would have gained with hypothetical manufacturing cycle times compared to its revenues with the cycle times it actually achieved using SLIM. We assumed that the trends in SEC's DRAM selling prices would

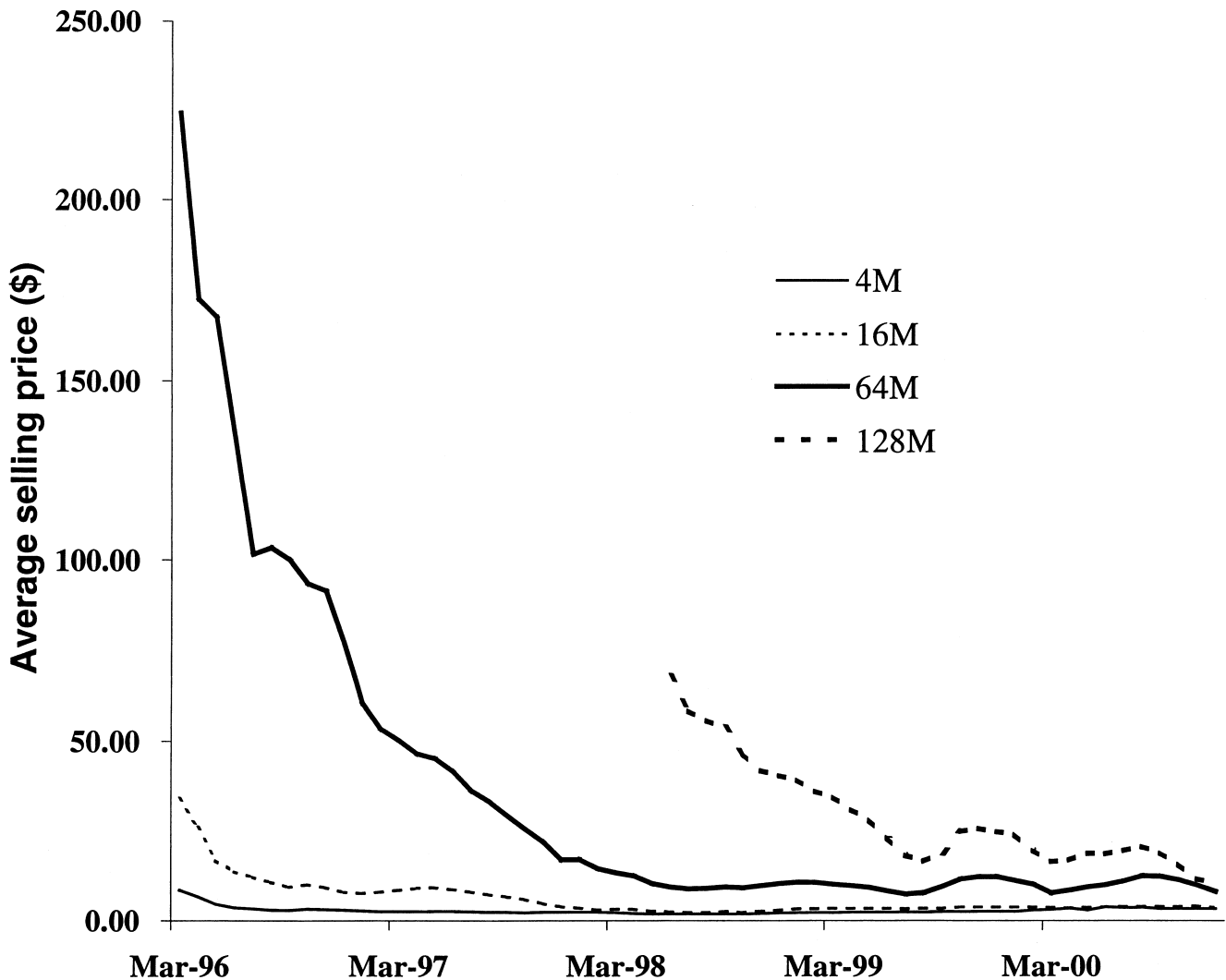


Figure 4: SEC's selling prices for the period 1996 to 2000 for the 4M, 16M, 64M, and 128M DRAM product families declined steeply during the first half of device life and then stabilized at a low level until end of life.

not be affected by its reductions in cycle times. While SLIM may have helped SEC to supply its 20 percent of the global market for DRAMs a month or so earlier than it would have otherwise, we believe that this is not one of the primary factors explaining market prices. Moreover, to offset this assumption, we calculated revenue differences only for the SEC fab lines in Kiheung that produced DRAMs during the SLIM project, even though the benefits from reducing cycle times in the SAS fab, the Bucheon fab lines, and the other Kiheung fab lines also were significant.

To understand our calculation, suppose manufacturing cycle times in month X were extended 15 days. Then sales of half of the manufacturing output in that month would be shifted into month $X+1$. If selling prices were lower in month $X+1$, revenue would be lost (or if prices were higher, revenue would be gained). To compute the revenue gains from SLIM, we assumed that the cycle times achievable without SLIM were bounded below by 4.0 days per layer of circuitry. That is, in periods when actual cycle times were below 4.0 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 four-days-per-layer cycle time.

We calculated the total sales revenue over this period for DRAMs produced by the Kiheung fabs to be \$21.9 billion. We calculated the gain in DRAM revenue afforded by reduced cycle times to be \$954 million, that is, about a 4.4 percent increase in sales revenues. Since SEC produces some non-DRAM memory products, we also calculated the revenue gain prorated for the entire output of the Kiheung fabs producing DRAMs, assuming DRAM output displaced the non-DRAM output in each fab on a wafer-per-wafer basis. The result in this case is \$1.133 billion. (Generally, prices of non-DRAM memory products are higher than those of DRAMs and also fall quickly, so we believe the revenue gains we calculated for this second case are conservative.) Even for this conservative estimate, the numbers are huge. If SEC had not reduced cycle times, it would have missed over \$1 billion in sales revenue that it gained using SLIM.

Other DRAM manufacturers reported heavy losses between 1996 and 1998, and many left the market entirely or sharply curtailed their participation. SEC alone reported profits every year and generated enough internal cash flow to expand its capacity aggressively. During the SLIM project, SEC's DRAM market share rose from 18 to 22 percent. According to President Y. W. Lee, "SLIM is essential to the success of SEC in the semiconductor business."

Conclusion

In only three years, SEC transformed itself from the worst-cycle-time semiconductor manufacturer to the best. It now regards factory-floor scheduling and WIP management as a core competence. This transformation required concerted effort and enlightened thinking on the part of the entire manufacturing and engineering organization. SEC management and staff changed the way they think about manufacturing.

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 brings new challenges to those trying to simultaneously achieve high yield, high throughput, and low cycle time. In effect, semiconductor manufacturing seems to be an intellectual black hole—despite any amount of brain power thrown at it, it always offers challenges for doing it better.

Acknowledgments

In addition to the authors, L and A employees participating in the SLIM project included Sejung Kim, Dr. Sooyoung Kim, Dr. Younghoon Lee, Jerome Levadoux, and Jeonghoon Mo. University of California-Berkeley students Jingliang Chen and Payman Jula developed the SLIM-F simulations. J. H. Lim of SEC served as project director in the first year of the SLIM project; in subsequent years, Y. I. Kim was the SLIM project director. Chang-Ho Choi, retired executive director of finance and administration for the semiconductor manufacturing division of SEC, initially championed a SLIM project involving Leachman and Associates, and he was instrumental in convincing other senior managers at SEC to start up and subsequently to sustain SLIM.

With the exception of SLIM-O, we formulated all of the SLIM principles and algorithms in response to the discussions of the SLIM teams, which then approved them. Thus, all of the members of the SLIM teams contributed to the design and development of SLIM. In the final analysis, SLIM embodies the best efforts of SEC managers and engineers to manage cycle time.