FLEXIBLE SEMICONDUCTOR MANUFACTURING

Modern high-tech industry is characterized by a number of trends that affect the way in which manufacturers produce finished products. In today's environment, manufacturers achieve competitive advantage by offering a variety of product types, a high level of product quality, and short lead times for customers (i.e., the time from which an order is placed until the product is received). In addition, there is constant pressure to innovate and customize product designs, thus resulting in compressed product life cycles and the need for new facilities to create the "next generation" of product. These trends contrast significantly with the traditional paradigm of mass production, which placed its emphasis on efficient production of high volumes of a standardized and relatively stable product type.

The semiconductor industry is a prime example of such trends. Manufacturers produce a variety of types of semiconductor-based products, or integrated circuits (e.g., memory chips or processors). At the same time, product quality is a critical consideration. In a manufacturing context, product quality translates to the ability to produce a given product design without defects. Semiconductor material is sensitive to the slightest contamination; hence, manufacturers employ extensive automation to ensure contamination-free, or "clean room," environments. To meet customer lead time demands, manufacturers focus on reducing cycle time, i.e., the time that it takes to produce a finished product once the raw material has been released for production, and its variability. From a planning perspective, reduced customer lead times and increased system responsiveness requires the continuous (re-)alignment of the planned production activity to externally imposed demand. Finally, manufacturers are faced with the constant challenge of having to adapt and produce new types of integrated circuits, given the rapid pace of technological advancement in product design.

The key question is how to handle the complexity associated with producing different product types, maintain an ability to adapt existing facilities and processes to manufacture new products, and still meet reasonable customer lead times. In producing multiple product types, a manufacturer must contend with a limited set of production resources and must allocate these resources to the production of each product type. The resource allocation problem can be a difficult one to solve, and the manufacturer often must make trade-offs in deciding which product has highest priority for a given set of resources. At the same time, a manufacturer must plan ahead to ensure that equipment to be purchased will be able to produce not only today's semiconductor products, but also tomorrow's.

In the modern manufacturing environment, these challenges typically are addressed through the concept of *flexibility*, or more specifically, *flexible automation*. In general, flexibility means the ability to adapt to new or different situations. A piece of equipment is said to be flexible if it can perform a number of different operations. A factory layout is said to be flexible if it can be reconfigured easily to accommodate changing production requirements. Likewise, a factory is said to be flexible if it can accommodate production of a variety of product types, or if it can switch to produce the next generation of product.

FLEXIBLE AUTOMATION AND SEMICONDUCTOR MANUFACTURING

In the early 1980's, manufacturers introduced the flexible manufacturing system as a way to enable efficient production of multiple product types, each having low-to-medium volumes of production. A flexible manufacturing system (FMS) is characterized by a number of automated process centers, or workstations, each of which performs transformation processes on a unit of material. These process centers are linked via an automated material handling system that is responsible for moving material between process centers. In general, the automated material handling system is flexible in that it does not require a fixed routing of material through the set of workstations. For example, it might be a robot that can move material between any two given workstations, or an automated guided vehicle network. In addition, an FMS may have a set of temporary storage buffers, where units of material may be stored between process operations, and containers that are used to transport material. Containers provide a standardized unit size and shape for handling by the automated material handling system. The whole system operates under a significant level of computerized control. The control system coordinates the various activities occurring in the system, with human operators needed only for a sub-set of activities (e.g., loading a new part into the system for processing). In general, this type of control is not trivial and requires extensive effort for successful implementation. Further information about flexible manufacturing systems is contained in Refs 1 and 2.

The flexible manufacturing concept was applied first to metal-cutting operations performed by stand-alone numerically controlled machines. A numerically controlled machine is flexible for two reasons – it can be programmed to perform different operations with the same cutting tool, and it can load a new tool to perform a different operation. The FMS concept extends machine flexibility by enabling a number of routings of material through a given set of workstations. An FMS can be a fairly large operation, with automated cells devoted to raw materials storage, fabrication, assembly and inspection. Each cell has an automated material handling system for transport within the cell, while cells are linked via another automated material handling system.

Flexible automation is well-suited to semiconductor manufacturing. Automation already is needed for intricate material processing requirements and for clean room production. Flexible automation is desirable due to the complex nature of the manufacturing operation. In semiconductor manufacturing, there are two major sub-systems: (1) wafer fabrication and probe and (2) device assembly and test. In wafer fabrication and probe, semiconductor wafers are made from raw material, and a variety of processes are performed on these disc-shaped wafers. These processes build up layers of integrated circuitry on the wafers. All told, a wafer may undergo several hundred processing op-

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 2007 John Wiley & Sons, Inc.

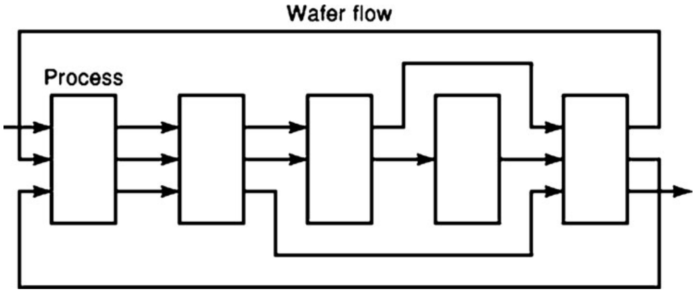

Figure 1. Representation of a small-scale version of a re-entrant flow line commonly found in semiconductor manufacturing. The figure illustrates material flow within a semiconductor fab, where wafers undergo a sequence of processes and then are sent back through the same processes. It should be noted that the wafers might not return though exactly the same sequence of processes, e.g., some processes might be skipped.

erations. These include such processes as surface cleaning, epitaxy, thermal oxidation, diffusion, ion implantation, photolithography, etching and deposition (3, 4). Moreover, the processing steps repeat, in that a wafer is sent through a sequence of process steps that build up a single layer of circuitry; then, it is sent back through the same set of operations to build the next layer. This type of material flow is called a *re-entrant flow*, and it introduces a great deal of complexity into the control of the material flow through the factory. Figure 1 shows a re-entrant flow through a set of operations. Between operations, wafers are transported in containers called cassettes. A cassette may hold up to 100 wafers, usually all of the same type. Wafer cassettes may be batched together during transport. Some processes are performed on batches of wafers, while others are performed on individual wafers.

In device assembly and test, the semiconductor wafers are sliced into smaller integrated circuit chips. Then they are assembled and packaged into electronic devices, and these devices are tested.

The primary emphasis in this article is on the fabrication system, often termed a *semiconductor fab*. In this environment, flexible automation is important for a number of reasons. First, a variety of wafer types may be in production, and wafers may revisit the same workstation, but need a slightly different process operation performed there. Hence, there is a need for equipment to exhibit flexibility in processing capability. Second, the industry is shifting from production of 200mm diameter wafers to 300mm wafers to gain improved chip yield from each wafer. Primarily for ergonomic reasons, the material handling system must be automated (i.e., 300mm wafer cassettes are too heavy for manual material handling). Moreover, these automated material handling systems must be flexibly automated to support re-entrant flows and to support routing a cassette to alternate workstations. Redundancy from multiple workstations performing the same function is typical, due to the high demand for semiconductor products and the corresponding need for high throughput. Third, equipment in a semiconductor plant is capital intensive. It is estimated that equipment costs comprise 75% of the investment cost for a new factory (5). A typical factory today costs several billion dollars. Therefore, there is a need to keep equipment utilized to the greatest extent possible. This can be achieved through flexibly automated material handling systems, which generally are more reliable than manual material handling. Fourth, technological innovation in product design means that factories must be prepared to manufacture new product designs. Moore's Law (6) provides ample evidence for the rapid pace of new product design. It is desirable that processing equipment, material handling systems and factory layouts be flexible enough to adapt to new production requirements. Finally, the computerized control applied by flexible automation creates the opportunity for computerized tracking and management of wip inventory. This is a crucial element needed to ensure that production is aligned to externally imposed demand and that product cycle times are minimized, to avoid lengthy customer lead times.

To a great extent, flexible automation has been enabled by advances in processing equipment and material handling hardware capabilities, open architecture controllers, and communications technology. At the same time, flexibly automated systems are complex to manage, due to (1) the large number of events occurring in the system, (2) the large state space associated with a factory, (3) the randomness associated with the factory (e.g., random machine failures), and (4) the difference in time scales associated with the various decisions needing to be made in the planning and scheduling of the factory. A key element needing to be addressed is the design of generic and easily reconfigurable frameworks and policies for flexible automation that ensure logically correct and near-optimal system performance.

THEORY AND CURRENT INDUSTRY PRACTICE

Production Objectives and System Performance

Semiconductor manufacturers are concerned with a number of specific measures that characterize the performance of the factory. Relevant measures have tended to focus on the quantity of finished product produced, but more recently on-time deliveries have become important. This is due to rapid declines in value for commodity semiconductor-based products when improved products (e.g., faster, more powerful) enter the market.

Historically, most of the improvements in system performance have been due to technological innovations. For example, decreased size of chip features enables an increased number of integrated circuits to be produced per wafer. There is some concern in the industry that increases in technological innovations associated with circuit design cannot by themselves sustain the current rate of performance improvement. Hence, manufacturers increasingly are looking to concepts such as flexible automation to improve overall factory performance.

Equipment utilization is a traditional measure of factory performance. Equipment utilization simply is the percentage of time spent by a piece of equipment in production. It excludes the amount of time spent (1) in a non-operational state (e.g., machine failure), (2) idle due to lack of material to process or lack of an operator to load material (i.e., the equipment is starved), (3) idle due to inability to unload material that has finished processing (i.e., equipment is blocked), and (4) setting up (i.e., changing configurations for a new process to be performed). A manufacturer typically desires high utilization to justify the large capital investment in automated equipment. Often, high utilization is achieved by having large amounts of wip inventory, which helps avoid starving.

Whereas utilization is a process-oriented measure, cycle time and throughput are product-oriented measures. As the industry has become more competitive, cycle time and throughput have eclipsed utilization as primary performance measures. Cycle time for a given batch of material is defined as the elapsed time between its release to the system and its completion. The minimum theoretical cycle time - typically used as a first-order approximation - is simply the sum of all processing times over the complete set of process steps. The actual cycle time may be an order of magnitude greater, due to waiting and transport times. Throughput is defined as the amount of finished product produced per unit time. In a wafer fab, this corresponds to the number of wafers produced per shift, for example. Closely related to these concepts are customer lead times, which also may be expressed as due dates. A customer order, represented by a batch of material in the system, may be assigned a due date. Failure to complete the batch of material by the due date results in a penalty, although this may be difficult to quantify. Predictability in lead times is important. Therefore, there is a large focus on reducing the variance of cycle times, in addition to reducing the mean. Finally, the amount of wip inventory itself is a performance measure. It was recognized in the 1980's that high levels of wip inventory tied up money that could be used elsewhere by a firm. To minimize this occurrence, manufacturers seek to avoid unnecessary WIP. Furthermore, in today's time-competitive manufacturing, it is recognized that reduced WIP levels can also enhance the system responsiveness, since they imply smaller batch cycle times, and faster switches of the production activity to different product types.

Even though the last remark implies a synergy between the objectives of reduced WIP and increased system responsiveness, it is still possible that the system performance measures described above can conflict with one another. For example, high levels of wip can help ensure good equipment utilization and good throughput. However, they are undesirable, and they also can cause congestion in the factory. Such congestion can lead to increased cycle times and cycle time variances, and the inability to meet due dates. Flexible automation, through appropriate management of wip inventory and flexible production capabilities to ensure balancing of production among equipment and to handle unexpected contingencies, can in theory create simultaneous improvement in several performance measures. For a simple example, consider a case in which two workstations can perform the same operation needed next by wafers in a cassette. Under a flexible routing scheme, the cassette could be taken to the one that is least utilized, hence avoiding a potential bottleneck at the other, reducing cycle time, and increasing utilization of the non-bottleneck workstation. There is great interest in application of flexible automation to semiconductor manufacturing.

Design of Flexibly Automated Production Systems

Often, for purposes of discussing a complex manufacturing system, the design and configuration aspects of the system are decomposed from the operation and control aspects. Design and configuration aspects encompass such things as factory layout, equipment selection and capacity determination, buffer space allocation and material handling system design. Operation and control encompass such things as production planning, order release, scheduling, real-time control, WIP inventory tracking, and operator task performance.

Layout and Material Handling System Design. Factory layout is tied closely to the material flow expected through a facility. The idea is to minimize the distance traveled by material between production processes. In doing so, material transport times should be minimized – leading to a possible reduction of the production cycle times - and material handling system (MHS) costs should be reduced. Traditionally, in the problem of aggregate layout, the factory is to be divided into a set of known departments, each of which performs a particular manufacturing function or process. The concept of departments is consistent with traditional practice in semiconductor manufacturing. Many processes are sensitive to contamination from other processes and hence must be isolated from them. Thus, semiconductor manufacturers have used a bay arrangement. In this type of layout, equipment devoted to the same type of process is situated in a large bay (i.e., department), which has a controlled environment and input and output chambers for entry and exit of wafer cassettes. Wafer cassettes travel between bays via an automated MHS, usually an overhead monorail system.

The layout problem initially is attacked through development of a block layout of departments within the factory. A block layout is judged by the material flow distances between departments.

Material flow distances are calculated using a from-to matrix that captures an estimate of the amount of material to be transported between each pair of departments. For example, f_{ij} would be the flow from department i to department j. This estimate is calculated from the expected routing of material between departments and the amount of material to be produced (i.e., that would travel between two departments). The distance d_{ij} between departments i and j often is assumed to be the distance between department centroids. Thus, the total material flow distances can be calculated as

$$\sum_i \sum_j f_{ij} d_{ij}$$

This composite material flow distance provides a metric by which different layouts can be judged. However, there are other considerations to judging a layout. Department shape is an important consideration, for example. This aspect strongly affects the detailed configuration of the material-handling activity, and the validity of the departmental "centroid" distances as a valid measure of it.

Indeed, in semiconductor fabs, where material-flow paths are known in significant detail, knowing the detailed shape and lay-out of the different departments can lead to much more accurate estimates of the expected traveling times. Flexibility for future expansion or reconfiguration is another important issue to be considered during the design of these environments.

Traditionally, there are two primary styles of bay layout and interbay material handling system design (7). The first is a perimeter configuration in which the monorail system traverses the perimeter of the facility. The bays are organized so that each faces onto the perimeter, and each has one or more pickup and deposit locations there where the monorail can load and unload cassettes. Typically, there are two monorail loops, one traversing clockwise and the other counter-clockwise. Additionally, the monorails have crossover turntables at certain points where a cassette can be transferred from one loop to another. The second one, more common now, is a spine configuration in which the monorail loop traverses in a narrow aisle through the center of the facility. Bays are located on either side of the loop, and each bay must face onto the loop. The monorail in a spine configuration usually has only one travel direction, but also has crossover turntables so that cassettes can change direction. Both systems promote the concept of routing flexibility in that, under computerized control, the monorail can deliver a cassette between any two bays as required.

Within a bay, the material handling systems are increasingly automated with the shift to 300mm production. The first component of an automated system typically is a stocker crane with a set of buffer locations for cassettes. These buffer locations accommodate WIP inventory. The stocker crane serves as the interface between the interbay MHS and the intrabay MHS. The crane receives cassettes delivered to the bay and delivers them to the intrabay MHS for wafer processing. Likewise, it receives cassettes whose wafers have been processed and delivers them to the interbay MHS for delivery to the next processing step. The intrabay MHS usually consists of person-guided track vehicles or an automated guided vehicle system. Both these types of systems support flexible routings of wafer cassettes through a series of process steps.

Flexible automation is starting to have a major impact on layout through the introduction of the integrated mini-environment, or cluster tool. A cluster tool is a flexible cell that combines two or more different processes in a controlled environment. Typically a robot inside the mainframe (i.e., an enclosing environment) transfers wafers between processes, which are housed in chambers. The cluster tool has one or more load-lock chambers through which wafer cassettes are loaded and unloaded. After cassette loading, the cluster tool performs a pump-down procedure. Then, wafers are sent through a set of single-wafer processes. Finally, after the wafers are placed back into the cassette, the cluster tool performs a pump-up procedure so that the cassette can be unloaded.

There is only one transfer of the wafer cassette through an input/output chamber. This has the effect of reducing cycle times due to (1) no need for multiple passes through input/output chambers, and (2) no need for long material transport times between processes. The second point is very important, due to the dramatic increase in the number of processing steps needed for wafers in recent years. Cluster tools have a significantly smaller footprint that the equipment arrangement used in the typical bay layout, and hence will have a major impact in reducing facility requirements. Also, cluster tools are a significant advance in flexible automation because they support process flexibility. The equipment manufacturer can attach specific process chambers to the mainframe to meet a semiconductor manufacturer's needs. Finally, the robot is programmable to allow for routing flexibility within the cluster tool. Figure 2 shows a schematic of a cluster tool.

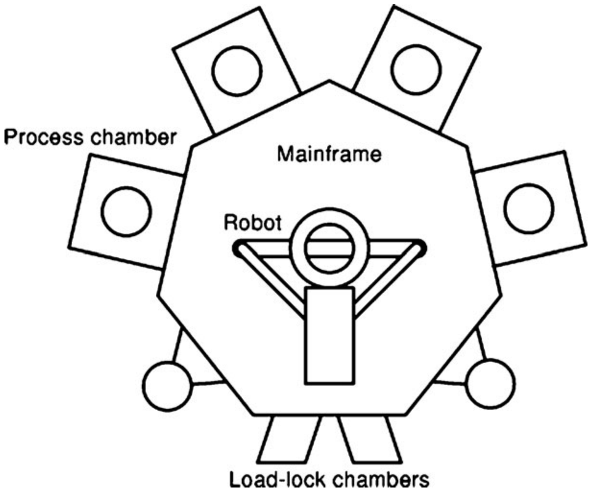

Figure 2. Schematic of a cluster tool. A cluster tool combines several processes into a single piece of equipment. The figure depicts a set of process chambers attached to an environmentally controlled mainframe. This mainframe houses a robot that moves individual wafers between process chambers. The particular types of processes assigned to chambers are determined by the needs of the semiconductor manufacturer.

Equipment Selection and Capacity Planning. Equipment selection is a critical element of system design. In terms of flexible automation, it is important to select equipment that can be adapted to perform processing operations for new designs of material. Much of the time, this is accomplished by use of masks for equipment such as photolithography. A mask imparts a particular pattern to a wafer. Different masks can be used on the same piece of equipment to perform different operations, though in the same process category.

Another important consideration in selecting equipment is how well it integrates with other types of equipment. In terms of hardware, this means the degree to which two or more pieces of equipment can be interfaced with one another for purposes of material transfer. A recent trend in this area is standardization, led by trade associations such as Semiconductor Equipment and Materials International (SEMI). To ensure that equipment provided by different vendors is compatible, SEMI has developed a number of standards for such things as wafer cassette transfer chambers, etc. These types of standards truly enable flexible automation from a hardware perspective.

Equipment capacity refers to the amount of equipment available for processing operations. Too little capacity for a given process can result in that process becoming a bottleneck; too little capacity for the entire system can mean that desired throughput levels are not achieved. On the other hand, too much capacity can result in low utilization and unneeded expense for equipment. Flexibly automated equipment can reduce the amount of specialized processing equipment needed, hence reducing the overall amount of equipment.

Another type of capacity to be determined for the system is the amount and location of buffer space. Buffers are where cassettes are stored in between operations, when they are not being transported by the material handling system. The storage locations of the stocker crane, described in the previous section, is a more concrete example of a semiconductor manufacturing system buffer. However, some limited buffering capacity might also exist at each separate process of a bay area or at a cluster tool chamber. In fact, the presence of a certain buffering capacity between any two processing stages is deemed desirable, since it decouples the operation of these processes, and prevents the effects of variability and operational contingencies to propagate throughout the entire system. In other words, the provision of limited buffering capacity between the system processes has a stabilizing effect. The detailed amount and configuration of the employed buffer space should be computed on the basis of cost and timing attributes of the system, and its estimation in the semiconductor manufacturing context constitutes a generalization of the more traditional "line balancing" problem (8). In an automated system, buffer space configuration and management is further complicated for two reasons: First, it must be specified precisely due to the hardware requirements needed for unattended operation; this enables an automated material handling system to effectively store and retrieve cassettes. Second, since there is a finite amount of buffer space that can be provided in each area, it falls to the control system to manage this finite set of buffers to ensure that they are not congested. In this context, and if used properly, buffer space can also support the flexibility required by the system operation. For example, high priority wafer cassette might be expedited to its next piece of equipment, while a lower priority could be stored in a buffer.

Process Planning. Each type of wafer has a sequential set of processing operations that must be performed on it to transform it from raw materials to a fully processed wafer ready for assembly and test. This set of operations is called the process plan, or the process log, or the recipe of the wafer. The process plan governs the routing of wafers through the system, as each operation can be performed only on equipment specified for that operation. To provide for flexible routing, the operations in the process plan must be supported by more than one workstation.

Development of a process plan for a given wafer type involves a number of steps. Given a set of existing equipment, an engineer must allocate processes to be performed to equipment. In doing so, the goal typically is to ensure some sort of balancing, so that one type of equipment is not over-utilized.

Metrology. Product quality is a key attribute needed by semiconductor manufacturers. Thus, wafers are subject to a number of inspection and measurement processes. In general, these procedures are called metrology. Metrology is performed by specialized equipment that measures, for example, the thickness of films that are deposited onto wafers during processing to ensure that they are within specified tolerances. Wafers not meeting tolerances may be sent back into the system for rework, or they may be scrapped.

In rework, wafers are sent back to previous sequence of processes, where a layer of circuitry is stripped off and then reapplied. An entire cassette of wafers may be judged defective and sent back for rework, or only some of the wafers may need rework. Flexible automation can enable this process to occur in an automated fashion, since flexibility is required to take the defective wafers back for reprocessing. If only some are to be reworked, it is left as an operational decision as to whether to split the contents of the cassette into two separate cassettes, or to hold the non-defective wafers at there current location until the defective ones have been reworked.

Operation and Control of Flexible Automation

The goals behind operation and control are to achieve good system performance in terms of cycle time, throughput, equipment utilization, etc. The type of control here is discrete-event control of logistics and systems, rather than continuous control of individual processes. In flexibly automated systems, the control problem is quite complex due to randomness in the system, the different time scales at which events occur, and the sheer size of the system state space, expressed by the operational condition of machines and amount of wip inventory. The typical approach is to decompose the overall problem of operation and control into a more manageable set of problems through hierarchical decomposition based on time scales.

Production Planning. Production planning is the problem of determining aggregate requirements for material to be released into a system and for labor, given a certain demand for finished goods. The standard approach to production planning is known as Manufacturing Resource Planning (MRP II). Traditionally, this methodology uses a series of calculations – known as MRP explosion - to determine order times and amounts of raw materials and release times and amounts. These calculations cascade from the desired factory output back through the various subsystems comprising the overall production system. They are based on the expected demand of finished goods, current WIP and finished goods inventory levels, and the bill of materials (i.e., the set of raw material required to make a finished good). In semiconductor manufacturing, an MRP system starts with the demand for integrated circuits and computes production requirements for assembly and test, and then for wafer fabrication and probe.

It should be noticed that the MRP explosion is an iterative process, performed on a "rolling-horizon" basis, i.e., production plans are computed and revised over a given time-window. In the past, this time window has typically been in the order of weeks or months. Currently, the increased emphasis on system responsiveness tends to compress this time frame to the order of days. From a more theoretical perspective, the minimum theoretical estimate of the production cycle time(s), introduced in our discussion on system performance measures, establishes a lower bound for the allowable MRP planning horizons.

A critical component in the "MRP explosion" calculations is the system capacity, i.e., its ability to produce the needed products. Traditional MRP systems typically did not include a model of this finite capacity; however, more current MRP systems are starting to include it. Leachman (9) discusses production planning for the semiconductor industry in extensive detail. Flexible automation can improve effective capacity, for example by increased throughput due to flexible routing on WIP through the system. In addition, flexible automation motivates the need for a flexible workforce – operators who are cross-trained on several types of equipment. This reduces the overall need for labor. Flexible automation can have a major impact on production planning by reducing inefficiencies, if estimates for increased effective capacity and reduced labor requirements can be captured in the production planning model.

Real-Time System Control. Once production requirements are set, factory management must determine a more detailed factory schedule to ensure that they are met. It is well known that the problem of computing an optimal production schedule for a job-shop or re-entrant flow environment is *NP-hard* (10). In practical terms, this means that computing such a schedule is intractable because, for all known algorithms, the computational time grows exponentially with the size of problem (number of wafer cassettes and number of machines). Scheduling is further complicated due to the highly stochastic nature of semiconductor manufacturing. Given a schedule (e.g., in Gantt chart format), it is impossible to know exactly the required makespan, since in this environment, processing and transportation times are characterized by significant variability. In fact, there are a number of events that could render such a pre-computed schedule infeasible. For example, a piece of equipment could fail or could require recalibration. Likewise, a "hot lot" could be introduced into the system. A hot lot is a batch of wafers that require expedited processing because it is part of a very important and time-critical customer order whose expedited processing might delay other wafers. Finally, two other types of disturbances have an impact: rework and engineering test jobs. If a wafer or set of wafers is found to be defective, it may be sent back for rework rather than being scrapped. An engineering test job is a wafer or set of wafers sent through a limited set of processes to test process capability (e.g., to determine if the processes need recalibration). At the same time, though, they use capacity that could be used by other wafers.

The generation of an effective and efficient production schedule for contemporary fabs is further complicated by the flexibility and the extensive levels of automation that are inherent in their operations. More specifically, the product and routing flexibility aspired for the contemporary fab operations can give rise to material flows with conflicting requirements and intricate behavioral patterns, like the so called *manufacturing system deadlock*, where a set of parts in order to proceed to require the allocation of resource currently held by some other part(s) in the set. On the other hand, the automated mode of the fab operations necessitates the a priori resolution of all these potential conflicts. This class of problems can be systematically addressed through the formal modeling, analysis and eventually control of the fab behavior in the context of Discrete Event Systems (DES) theory (11). Using formal modeling frameworks like Finite State Automata (FSA) and Petri Nets (PN), DES theory seeks to formally characterize the entire set of plausible behaviors generated by the plant, and eventually synthesize the necessary control logic that will restrain the system to an admissible behavioral space, which is free from any problematic behavioral patterns. Furthermore, this behavioral analysis and the ensuing control logic must be integrated into the applied scheduling policies, in order to ensure effective and efficient operation. This line of research has seen a number of advances in recent years (c.f., for instance, the work presented in (12, 13)); however, considerable work is needed to migrate the currently available theoretical results to the factory shop floor.

Hence, given the difficulties in computing a global production schedule and the current limitations of the relevant theory, manufacturers typically use a more pragmatic approach. First of all, they adopt operational patterns that tend to limit the underlying operational flexibility but are much simpler to analyze and control from a logical/qualitative standpoint. In other words, they tend to trade some of the plant efficiencies and productivity for operational simplicity and convenience. Moreover, they use simple scheduling policies, known as *dispatching rules,* that resolve conflict locally at each workstation, in a heuristic manner. Typically, these rules rely on an attribute of particular wafer batches (e.g., due date) or on the type of wafer to sequence jobs. The dynamic nature of the dispatching rule-based scheduling can make it more responsive to the various contingencies arising in the plant operation. On the other hand, it is recognized that, due to their more localized nature, distributed approaches might be suboptimal, i.e., they might fail to materialize the maximum throughput(s) that can be possibly supported by the system. Some important and practical considerations in semiconductor scheduling and dispatching are discussed in (14) and (15).

It is clear from the above that the support of effective and efficient semiconductor scheduling in the context of flexible automation is an open and challenging research issue. From a more technological standpoint, a major step in the direction of implementable efficient computerized control at the factory level is the Manufacturing Execution System (MES). Commercially available through a number of software vendors, MES provides a number of features such as tracking of wip inventory, scheduling of equipment maintenance, reporting of equipment failures, etc. In many implementations, the MES is implemented as an intermediate control level between production planning (MRP) and equipment-level control. Real-time tracking of wip inventory and equipment status, in particular, is a key part of flexible automation because it provides the information necessary for effective decision-making in flexible environments. For example, a decision in a flexible routing environment requires knowledge of the state of the workstations to which a wafer cassette might be routed (e.g., current utilization, calibration status, etc.).

Supervisory Control. Supervisory control at the equipment level is critical, not only for process monitoring, but also to ensure that the system is fully integrated. One problem is that manufacturing process equipment is developed by a number of different equipment vendors, and each vendor uses different communications protocols and control standards. Thus, it has been a major challenge to semiconductor manufacturers to integrate equipment as expected for an MES or other factory control system. This has led to the Generic Equipment Model (GEM), SEMI Equipment Communications Standard (SECS/SECS II) and the Computer Integrated Manufacturing (CIM) Framework, all available from SEMI. These create standardized interfaces to link equipment with the factory control system.

Although control is computerized to a great extent in 200mm fabs, there remain many tasks left to the responsibility of human operators. For example, human operators must load and unload material from much of the existing processing equipment. Additionally, human operators are responsible to a large extent for process monitoring. The automated material handling of 300mm fabs is likely to reduce or eliminate physical activities performed by operators. However, operators may be involved in some higher level supervisory control (e.g., job expediting to ensure due dates are met). Useful operator interfaces for these activities will need to be developed.

DESIGN METHODOLOGIES AND PERFORMANCE EVALUATION

As discussed previously, there are a host of problems encountered in the design and operation of flexibly automated semiconductor manufacturing systems. This section discusses several engineering methodologies that can be applied to solve these problems. Typically, these methodologies are supported by software packages that allow the engineer to develop models and perform analysis using a computer.

Optimization

Optimization methodologies have proven useful in facility layout and production planning. An optimization model seeks to maximize or minimize an algebraic objective function, subject to a set of algebraic constraints. For example, the objective function might be a cost or profit function. The constraints might be finite resource constraints. Both objective functions and constraints are functions of decision variables, which characterize the solution. Once the model is formulated, an algorithm is applied to perform the optimization. Algorithms are iterative in nature and may or may not provide optimal results. A good introduction to Optimization theory is provided in (16).

Facility layout, where the goal is to minimize material flow, is one application area. Constraints are in the form of department shapes and sizes and also could include that certain departments not be located near one another. These constraints are difficult to formulate as a set of linear, algebraic functions without resorting to requiring that some decision variables be binary or integer. Except in special cases, model formulations with this requirement (termed integer programming problems) are difficult to solve, i.e., there is no optimal algorithm that has a tractable computational time. Hence, engineers use heuristic algorithms to provide (hopefully) near-optimal solutions for facility layout. The decision variables from the formulation characterize the resulting layout.

Another area of application for optimization is production planning. In aggregate production planning, an optimization model can be formulated as a maximization of profit or revenue, or minimization of operational costs, subject to finite capacity constraints (e.g., material, labor, equipment) and demand constraints (e.g., produce at demand level). The decision variables characterize the quantity of products to be produced (in most cases in each period of a multi-period term), the amount of labor and material assigned to each, etc. Here, the integer requirement for decision variables often can be avoided. Most typically, the resulting formulations fall into the broad category of linear programming (9), for which there exist a number of algorithms that can produce an optimal solution within a reasonable computational time. Furthermore, a variety of software is available to solve linear programming and integer programming problems, and packages more customized to the specific application of production planning have also been developed.

Queueing Network Analysis

Queueing network analysis is based on the fundamental abstractions of servers, customers and queues, and it is the study of properties of a network of queues. Customers arrive at random intervals to a queue, where they wait for service by the server. Queueing analysis studies such properties as queue length, number of customers in the system, customer time in the system and time spent waiting for service. The characteristics of a queue are customer interarrival times, service times, queueing discipline (e.g., first-come-first-served) and number of parallel servers. A queueing network is a network of servers with queues, where there is a routing pattern between servers. This routing pattern is expressed as a set of routing probabilities r_{ij} dictating the probability that a customer leaving service at server i goes to server j, or possibly leaves the system. Customer interarrival times to the system and service times typically are random variables from a specified probability distribution. The classic distribution that supports closed-form solutions for system properties is the exponential distribution. For a comprehensive discussion of queueing network theory and its application to manufacturing, the reader is referred to (17).

In the semiconductor manufacturing context, customers of the generic queueing network structure can be used to represent wafer lots (in carriers or pods), and servers can be used to represent workstations. The routing probabilities represent the routings of wafer lots through the system. Routing flexibility can be modeled at an aggregate level via the routing probabilities. For example, if a wafer lot can go to one of two workstations for its next process (depending on whether rework needs to be done), the routing probabilities can be set accordingly. Flexibility also can be modeled by multi-class queueing networks, in which each customer belongs to a class (which represents a wafer product at a particular step in its manufacturing process).

Queueing network analysis provides rough cut estimates of various system properties and performance measures, mostly in the form of averages. This type of analysis can be used to determine buffer capacity (based on an average queue length). Also, it can be used to determine processing capacity at a workstation. If a queue for a given workstation is over-run, then the designer should increase its capacity. Estimates are rough cut because the typical queueing analysis assumes exponential service times, infinite buffer capacity for queues, and simple dispatching rules (e.g., first-come-first-served). In automated semiconductor manufacturing, process times rarely are exponential, and there is a limited amount of buffer space. Additionally, queueing network analysis does not account for automated material handling systems. There are software packages available that provide numerical analysis for systems that do not meet these assumptions. Presently, their modeling capabilities are rather limited.

More recently, manufacturing systems have been represented by stochastic processing networks, which are generalizations of queueing networks (18). Stochastic processing networks allow for a variety of shared resources such as equipment, operators, reticles (i.e., a mask used to etch patterns onto a wafer that can be shared among similar workstations) and other fixtures. A hierarchy of approximate models is used to analyze such systems. In particular, fluid approximations and heavy traffic (also known as Brownian motion) approximations have produced important insights in understanding how the performance of stochastic processing networks depends on different design and control parameters.

Discrete-Event Simulation

System complexity often requires detailed analysis that cannot be achieved through analytic approaches such as queueing network analysis. The typical approach used in this case is simulation modeling, which uses a more detailed model of system behavior. In simulation-based applications, specific events, such as routings between machines, are not modeled at an aggregate level, but rather at the level of the individual job. At this level, randomness is not modeled by the statistics of a probability distribution (e.g., mean and variance), but by the operational dynamics of a (computerized) random number generator. Execution of a simulation model occurs as a computer program that traces through a specific series of events (job movements, machine starts and completions, etc.) to determine estimates for overall system performance. Since a simulation model essentially is a computer program, the modeler can calculate any desired performance measure for a particular model execution.

Because of this, simulation modeling does not support closed-form or numerical approaches to determining estimates for system properties or performance. Rather, the modeler builds a simulation model and then performs a series of experiments to get performance estimates. The set of experiments usually requires multiple model executions (or replications) to ensure that the particular random numbers generated for one do not result in atypical results, and the experiments also are used to compare performance estimates for different system configurations or control policies. Due to the detailed level of modeling, the modeler is obligated to validate a simulation model, or in other words, to demonstrate that it is an accurate representation of the real system's behavior. This is a critical, but sometimes overlooked, activity in simulation modeling. Improperly validated models might lead to erroneous results, and expensive mistakes in system design. A good introduction on (discrete-event) simulation and its proper practice is provided in (19).

There are a number of commercially available languages for discrete-event simulation. Most of them use a process-oriented view of system behavior. This formalism uses a network of queues with customers as its underlying basis, but it adds additional constructs for the modeler to use. These additional constructs are helpful in modeling flexibility. For example, rather than routing probabilities, most languages provide a construct that allows the modeler to specify a specific rule that governs how flexible routing and dispatching occur in the real system. Customers (or jobs) can be assigned attributes that specify wafer type, so this data can be used in the routing and dispatching rules. Most simulation languages support explicit modeling of different material handling systems via specialized modeling abstractions. This is an important element needed to support modeling of flexible automation; however, explicit modeling of material handling systems is computationally expensive.

Simulation models can be time-consuming to develop, and they can also be time-consuming in execution (especially considering that multiple replications are needed). To address these limitations, one trend implemented in simulation packages is to separate the modeling of the factory production resources from the material handling resources. For rough-cut analysis, a less detailed simulation comprising just the production resources can be developed and executed, requiring less time than one integrated with the material handling system. Then, the material handling system can be added for more detailed analysis later. Another way to address these limitations is to model only the bottleneck resources (i.e., those whose design and performance matter most), and to represent the rest of the factory as a "black box."

One area in which simulation languages tend to be weak is their representation of integrated factory control (i.e., control beyond the level of a dispatching or routing rule for a single job).

Table 1. Semiconductor Trade Associations and Consortia Involved with Flexible Automation

International Sematech
Web address: www.sematech.org
International Technology Roadmap for Semiconductors (ITRS)
Web address: www.itrs.net
Semiconductor Equipment and Materials International (SEMI)
Web address: www.semi.org
Semiconductor Industry Association (SIA)
Web address: www.sia-online.org
Semiconductor Research Corporation (SRC)
Web address: www.src.org

Table 2. Trade Journals and Other Resources for Flexible Semiconductor Manufacturing

Modeling and Analysis of Semiconductor Manufacturing Laboratory
Arizona State University
Web address: www.fulton.asu.edu/~ie/research/labs/masm/
MIT Semiconductor Subway
Massachusetts Institute of Technology
Web address: www-mtl.mit.edu/semisubway/semisubway.html
Semiconductor Fabtech Online
Web address: www.fabtech.org
Semiconductor International
Web address: www.reed-electronics.com/semiconductor/
Solid State Technology
Web address: sst.pennnet.com/home.cfm

EVALUATION AND FUTURE RESEARCH

When the concept of flexibly automated production systems was first introduced, it was realized that the microprocessor and the emerging information technologies offered tremendous power for massive real-time analytical computation and data-processing. Indeed, considerable progress has been made regarding the processing capabilities of shop-floor equipment, as well as the supporting communication networks. Currently, it is possible to (re-)configure system workstations remotely through appropriate tooling and software so that they meet a variety of production requirements with small switching/set-up times, while computerized monitoring platforms known as Manufacturing Execution Systems (MES) provide (almost) real-time tracking of the shop-floor activity.

However, the manufacturing community still lacks the control paradigm that will master the complexities underlying the effective deployment and management of the operational flexibility provided by the aforementioned technological infrastructure. Hence, while the advantages and benefits of manufacturing flexibility have been understood, described and advertised at a conceptual level (20, 21), analytical characterizations that will allow the operationalization and evaluation of flexibility on the shop floor are missing. As a result, a number of past attempts to extensive deployment of flexible automation have failed (e.g., IBM Quick Turn Around Time (QTAT), (22)), and (most of) the current installations are operated in a very stiff and inflexible way (23, 24).

These problems are also imminent to the semiconductor manufacturing community. Among the efforts to address them, the most outstanding and long-lasting one is the work of the Modular Equipment Standardization Committee (MESC), a SEMI-sponsored group. MESC seeks to develop standardized, open-system architectures for integrated processing equipment and cluster tools. However, its work is focused mainly at the equipment control level, seeking to successfully interface components coming from many different vendors, through hardware and communication software standardization. Hence, while "multiprocessor control systems for cluster tools are an important step towards the 'island of automation' concept of computer integrated manufacturing (CIM)" (25), there are still a number of standing issues that must be addressed, in order to materialize the full potential of these environments in terms of operational flexibility and productivity enhancement. The rest of this section outlines these issues and it highlights the state of art and future directions of the relevant research.

Domain Analysis, Object-Oriented Simulation and Distributed Simulation

Like all major attempts to extensive automation, the starting point for effective modeling and analysis for flexibly automated production systems is the effective and rigorous characterization of the system components/entities and their behavior(s). These models must be detailed enough to capture all the relevant aspects of the system behavior and the entailed complexity, yet generic to allow for systematic analytical treatment of emerging control problems.

The emerging paradigm of object-oriented simulation (26), together with supporting software engineering techniques (e.g., domain analysis) provides a promising framework for the formal definition of the flexibly automated semiconductor fab. Enhanced with the capabilities of virtual reality technologies, object-oriented simulation platforms can provide a powerful tool for the systematic study of system behavior, as well as the evaluation and testing of existing or emerging operational policies and system designs. Such software product is often referred to as the "Virtual Factory." Object-oriented simulation platforms are useful because they provide extensive detail of system components and their real-time behavior, and they are remarkably close to the modeling abstractions typically used for mathematical analysis of Discrete Event System behaviors.

These models are being extended from modeling the factory itself, to modeling the entire supply chain. Such models represent factories, distribution centers, customers and transportation systems, and they often rely on distributed simulation technology (e.g., High-Level Architecture) to link sub-models together that physically execute on different computers (27). Such models may be built using commercially available simulation languages or using open-source simulation libraries developed in a high-level programming language such as Java^{TM}. To address problems with model execution speed, research into event-scheduling approaches to simulation, as opposed to process-oriented approaches, is finding methods that execute more quickly for models of highly congested systems such as semiconductor manufacturing (28). Other major challenges include using simulation for real-time problem-solving, developing plug-and-play interoperability for simulation models and supporting software, and convincing management to use simulation more extensively (29).

Resource Allocation and Structural Control of the Semiconductor Fab

The integrated processing (mini-)environments of semiconductor manufacturing can support, in principle, the automated concurrent handling of a number of wafer types through a set of reconfigurable processing tools, while maintaining consistently high throughputs and reduced cycle-times, and successfully coping with a number of operational contingencies. To address these requirements successfully requires logically correct and robust behavior of the system. The emerging control paradigm dealing with this class of problems is known as structural control (12, 30).

Within the scope of structural control for integrated processing environments, a primary issue is the resolution of the manufacturing system deadlock (31, 32). Specifically, due to the arbitrary routing of jobs through the system, and the finite buffering capacity of the system chambers, it is possible that a set of jobs becomes entangled in a circular waiting situation, in which each job is waiting for some buffering space on a workstation currently held by some other job(s) in this set. The formal modeling of the problem perceives the manufacturing system as a Resource Allocation System (RAS), where the system resources are the buffering capacity of the clustered chambers and material handlers. The applied analytical techniques are borrowed from the Qualitative Modeling and Analysis of Discrete Event Systems, with predominant approaches being based on Finite State Automata and Petri Net theory.

In fact, deadlock resolution and avoidance in flexibly automated production systems has been extensively studied in the past decade, with a richness of formal results. More specifically, the problem of designing maximally permissive deadlock avoidance policies for sequential resource allocation systems has been shown to be *NP-hard* in its general formulation (33, 34), but it has also been shown that, for a considerably large subclass of these systems with very practical implications for flexibly automated production and semiconductor manufacturing, maximally permissive deadlock avoidance can be obtained polynomially through one-step lookahead (35–37). Furthermore, for the remaining cases computationally efficient and provably correct policies have been developed (31,34,38). Additional work has sought to accommodate on-line routing flexibility in the policy design, and to exploit this capability for the effective response to operational contingencies (39), like machine outages and the appearance of "hot lots." The reader is referred to (12, 13) for a comprehensive discussion of the relevant theory, its current state of art, and directions of future research.

Regarding the implementation of the aforementioned set of results in the semiconductor manufacturing context, currently the main bottleneck is their dissemination in the relevant community and their integration in the emerging control software and practice. This is a non-trivial proposition since it implies that this community must accept the potential benefits to be materialized by a more flexible operation of the underlying production (mini)environments, and be willing to abandon its current conservative attitude on this issue (c.f. the relevant discussion in the section on real-time control of flexible automation). Beyond the complications arising from the human psychology and its inherent resistance to change, such a change of attitude is also a financially risky proposition, given the extremely high cost of modern fabs. Hence, the specification and successful implementation of some carefully chosen pilot projects seems to be the most natural next step regarding the aforementioned developments.

Performance Analysis and Control of Semiconductor Fabs

Given the high cost of a semiconductor fab, and the complexity of the material flow, it follows that the establishment of efficient resource allocation, in terms of throughput, resource utilizations and production cycle times, is of paramount importance.

The currently used distributed scheduling policies can be further divided in two broad classes: (1) *dispatching rules* that myopically sequence jobs waiting for some resource on the basis of some job attribute (e.g., remaining workload, due date, externally defined priority, etc.) (40, 41) and (2) *policies based on tracking of "optimal" target rates*, with the latter being computed through some optimizing "fluid" relaxation models (42, 43). The acceptance of all these policies is based on: (1) their relatively easy implementation, (2) their rather consistently good performance in current manufacturing settings and/or simulation studies, and (3) the emergence of a series of theoretical results establishing some robustness/stability properties (44–46). For an overview of the methodology pertaining to the design and evaluation of the aforementioned policies the reader is referred to (47).

An interesting open issue is the integration of the aforementioned policies with the structural control paradigm discussed in the previous section. More specifically, from an operational standpoint, popular dispatching rules and even target rate tracking policies can be easily adjusted to accommodate the logic of the applied deadlock avoidance policies. However, from a more theoretical standpoint, all the past results regarding the efficiency and the relevant performance of the aforementioned scheduling policies have been developed without taking into consideration the complications and tenements of the underlying structural control problem. Yet, this is an aspect that can have a strong impact on the resulting performance of the various policies. This issue was pertinently demonstrated recently in (46), where it was shown that bounding of the system WIP through a KANBAN mechanism can destabilize policies which appear to be stable under the assumption of infinite capacity buffers. A similar result regarding the (in-)stability of the Last Buffer First Serve policy in structurally controlled environments is reported in (48). This policy has been shown to be stable in the context of re-entrant lines with infinite buffering capacity (44, 45).

It follows, then, that the effectiveness of different distributed policies must be reconsidered in the context of structurally controlled flexibly automated discrete-part manufacturing environments. Popular dispatching rules and/or fluid models can be employed for the scheduling and dispatching modules, but the overall performance of the resulting scheme and the underlying system dynamics is an open research question. Simulation-based analysis making use of the Virtual-Factory platform(s) might be a good starting point for this analysis. From a more theoretical standpoint, the scheduling of the structurally controlled fab can be formally addressed in the analytical framework of Markov Decision Processes (MDP) (49). However, the super-polynomial size of the involved state spaces implies that this line of analysis can offer valuable qualitative insights but it is inherently limited in terms of providing practically computable and implementable policies. These practical complications can be potentially addressed in the context of the emerging paradigm of *approximate dynamic programming*. Generally speaking, approximate dynamic programming seeks to overcome the aforementioned complexities of the MDP theory by adopting a compact approximation of the value function that characterizes the optimal policies, which is built through simulation or other more computationally efficient approaches, like (approximate) linear programming. The reader is referred to (50) for a study that initiates the application of these ideas in the context of fab scheduling.

(Approximate) MDP theory can also be useful for characterizing the performance of any given scheduling policy. Furthermore, starting with the work of (51) on the performance evaluation of multi-class queueing networks, a theory for the generation of computationally efficient performance bounds has been developed. The reader is referred to (47) for its basic characterization and a more extensive listing of these results. It remains, however, to further validate and assess the quality of the obtained bounds, and their ability to effectively resolve the relative performance of the different policies. Also, the effective integration of this capability in the overall decision-making process is another practical issue that needs to be addressed.

A final issue concerns the effective modeling of the processing times involved in all the aforementioned analyses. In their basic characterization, most of the aforementioned theories assume exponentially distributed event times. Yet, it is well known that in most practical cases, the processing times experienced in the manufacturing shop-floor will not adhere to this assumption. Especially, in the highly automated environments of contemporary fabs, processing times tend to be more deterministically distributed. A typical approach to circumvent this complication is the approximation of the actual processing time distributions by Erlangs with an appropriate number of stages. The main question for this approach is whether a reasonably low number of stages would provide significant improvement on the model accuracy, compared to that obtained through the exponentiality assumption. The issue can be studied empirically, by comparing the analytically obtained results to those extracted through simulation.

Higher-Level Planning in Structurally Controlled Semiconductor Manufacturing

We envision the future semiconductor fab as a set of "universal" processors (processing tools), at each time point configured for a certain production run by the specific sets of tools/masks loaded in their magazines. Given that each station can hold a finite number of tools at each time, the problem that naturally arises is how to compute a time-phased reconfiguration plan that will allow the system to trace externally imposed demands for different products in the most efficient way (e.g., minimum inventory costs while attaining specific service levels). Notice that any solution addressing this problem automatically answers all the "classical" tactical planning problems formulated in (52). Also, any efficient algorithm addressing this problem can be effectively used for replanning system operations in the face of contingencies. Finally, resolving the problem of tactical planning from such a perspective would allow for the explicit consideration of all different modes of flexibility in the system operation (e.g., machine, routing, operations, volume, etc.) since the capability to reconfigure qualitatively and quantitatively the processing capacity of the different workstations is the main attribute on which these flexibilities are established.

The benefits of effectively exploiting system flexibilities and the open problems resulting from this requirement are extensively discussed in (53). Also, an initial effort to address the tactical (re-)planning problem in the simpler flow line setting is presented in (54). These problems must be revisited and re-modeled once the lower-level/real-time aspects of the system operation have been resolved. Currently, we are not aware of any research results along these lines. Traditional hierarchical planning and commercial MRP-like frameworks fail to address many of the real-time operational aspects of the flexibly automated shop-floor, and therefore, their results are characterized by infeasibility and/or considerable inefficiencies. Bridging the gap between real-time control and tactical and strategic planning units in tomorrow's flexibly automated semiconductor manufacturing remains a major research challenge.

ADDITIONAL INFORMATION

Table 1 lists some trade associations and consortia that have involvement in flexible automation in the semiconductor industry. Due to the large number of semiconductor manufacturers and equipment vendors, these are not listed. The items in Table 1 will provide information about manufacturers and equipment vendors. Table 2 lists trade journals and other resources of interest.

BIBLIOGRAPHY

1. J. Hartley, *FMS at Work*, New York: North Holland, 1984.
2. H. Tempelmeier and H. Kuhn, *Flexible Manufacturing Systems: Decision Support for Design and Operation*, New York: John Wiley & Sons, Inc., 1993.
3. T. Dillinger, *VLSI Engineering*, Englewood Cliffs, NJ: Prentice-Hall, 1987.
4. W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*, Reading, MA: Addison-Wesley, 1990.
5. S. Tandon, Challenges for 300 mm plasma etch system development, *Semiconductor Int.*, **21** (3): 75–91, 1998.
6. P. K. Bondyopadhyay, Moore's Law governs the silicon revolution, *Proc. IEEE*, **86** (1): 78–81, 1998.
7. B. Peters and T. Yang, Integrated facility layout and material handling systems design in semiconductor fabrication facilities, *IEEE Trans. Semicond. Manuf.*, **10** (3): 360–369, 1997.
8. S. B. Gershwin, *Manufacturing Systems Engineering*, Englewood Cliffs, NJ: Prentice Hall, 1994.
9. R. C. Leachman, Modeling techniques for automated production planning in the semiconductor industry, in T. A. Ciriani and R. C. Leachman (eds.), *Optimization in Industry*. Sussex, UK: Wiley, 1993, pp. 1–30.
10. M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, San Francisco: W. H. Freeman, 1979.
11. C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*, Boston: Kluwer Academic Publishers, 1999.
12. S. A. Reveliotis, *Real-Time Management of Resource Allocation Systems: A Discrete Event System Approach*, Boston: Springer, 2005.
13. M. Zhou and M. P. Fanti (eds.), *Deadlock Resolution in Computer-Integrated Systems,* Singapore: Marcel Dekker, Inc., 2004.
14. P. K. Johri, Practical issues in scheduling and dispatching in semiconductor wafer fabrication, *J. Manuf. Sys.*, **12** (6): 474–485, 1993.
15. P. R. Kumar, Scheduling semiconductor manufacturing plants, *IEEE Control Syst. Mag.*, **14** (6): 33–40, 1994.
16. W. L. Winston, *Introduction to Mathematical Programming: Applications and Algorithms*, 2nd Ed., Duxbury Press, 1995.
17. J. A. Buzacott and J. G. Shanthikumar, *Stochastic Models of Manufacturing Systems*, Englewood Cliffs, NJ: Prentice Hall, 1993.
18. J. G. Dai, Stability of fluid and stochastic processing networks, Miscellanea Publications No.9, January 1999, Centre for Mathematical Physics and Stochastics, Department of Mathematical Sciences, University of Aarhus, Denmark, 1999.
19. A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, 3rd ed., New York: McGraw-Hill, 2000.
20. J. Browne *et al.*, Classification of flexible manufacturing systems, *FMS Mag.*, **2** (2): 114–117, 1984.
21. A. K. Sethi and S. P. Sethi, Flexibility in manufacturing: a survey, *Int. J. Flexible Manuf. Syst.*, **2**: 289–328, 1989.
22. P. Singer, The driving forces in cluster tool development, *Semiconductor Int.*, **18** (1): 113–118, 1995.
23. S. B. Joshi *et al.*, Formal models for control of flexible manufacturing cells: physical and system models, *IEEE Trans. Robot. Autom.*, **RA-11**: 558–570, 1995.
24. F. F. Suarez, M. A. Cusumano and C. H. Fine, An empirical study of manufacturing flexibility in printed circuit board assembly, *Operations Res.*, **44**: 223–240, 1997.
25. M. E. Bader, R. P. Hall and G. Strasser, Integrated processing equipment, *Solid State Tech.*, **33** (5): 149–154, 1990.
26. S. Narayanan *et al.*, Research in object-oriented manufacturing simulations: an assessment of the state of the art, *IIE Trans.*, **20**: 795–810, 1998.
27. C. S. Chong, P. Lendermann, B. P. Gan, B. M. Duarte, J. W. Fowler and T. E. Callarman, Development and analysis of a customer demand driven semiconductor supply chain model using High Level Architecture (HLA), *Intl. J. Simulation and Process Modeling*, to appear.
28. L. W. Schruben and T. M. Roeder, Fast simulations of large-scale highly congested systems, *Simulation*, **79** (3): 115–125, 2003.
29. J. W. Fowler and O. Rose, Grand challenges in modeling and simulation of complex manufacturing systems, *Simulation*, **80** (9): 469–476, 2004.
30. S. A. Reveliotis, M. A. Lawley, P. M. Ferreira, Structural control of large-scale flexibly automated manufacturing systems, in C. T. Leonides (ed.), *Computer Aided and Integrated Manufacturing Systems: Techniques and Applications*, New York: Gordon & Breach, 1998.
31. Z. A. Banaszak and B. H. Krogh, Deadlock avoidance in flexible manufacturing systems with concurrently competing process flows, *IEEE Trans. on Robot. Autom.*, **RA-6**: 724–734, 1990.
32. R. A. Wysk, N. S. Yang and S. Joshi, Detection of deadlocks in flexible manufacturing cells, *IEEE Trans. Robot. Autom.*, **RA-7**: 853–859, 1991.
33. T. Araki, Y. Sugiyama and T. Kasami, Complexity of the deadlock avoidance problem, *Proc. 2nd IBM Symp. Math. Foundations Comput. Sci.*, Tokyo, Japan, 1977, pp. 229–257.
34. M. A. Lawley and S. A. Reveliotis, Deadlock avoidance for sequential resource allocation systems: hard and easy cases, *Intl. J. Flexible Manuf. Syst.*, **13**: 385–404, 2001.
35. S. A. Reveliotis and P. M. Ferreira, Deadlock avoidance policies for automated manufacturing cells, *IEEE Trans. Robot. Autom.*, **RA-12**: 845–857, 1996.
36. S. A. Reveliotis, M. A. Lawley and P. M. Ferreira, Polynomial complexity deadlock avoidance policies for sequential resource allocation systems, *IEEE Trans. on Automatic Control*, **42**: 1344–1357, 1997.
37. M. P. Fanti *et al.*, Event-based feedback control for deadlock avoidance in flexible production systems, *IEEE Trans. Robot. Autom.*, **RA-13**: 347–363, 1997.
38. M. Lawley, S. Reveliotis and P. Ferreira, The application and evaluation of Banker's Algorithm for deadlock-free buffer

space allocation in flexible manufacturing systems, *Intl. J. Flexible Manuf. Syst.*, **10**: 73–100, 1998.

39. S. A. Reveliotis, Accommodating FMS operational contingencies through routing flexibility, *IEEE Trans. Robot. Autom.*, **15**: 3–19, 1999.
40. S. S. Panwalkar and W. Iskander, A survey of scheduling rules, *Operations Res.*, **25**: 45–61, 1977.
41. J. H. Blackstone, D. T. Philips and G. L. Hogg, A state-of-the-art survey of dispatching rules for manufacturing job shop operations, *Int. J. Prod. Res*., **20**: 27–45, 1982.
42. A. Sharifnia, Stability and performance of a simple distributed tracking policy for production control of manufacturing systems, *IEEE Trans. Autom. Control*, **40**: 1109–1113, 1995.
43. D. Connors, G. Feigin and D. Yao, Scheduling semiconductor lines using a fluid network model, *IEEE Trans. on Robotics & Automation*, **RA-10**: 88–98, 1994.
44. S. H. Lu and P. R. Kumar, Distributed scheduling based on due dates and buffer priorities, *IEEE Trans. Autom. Control*, **36**: 1406–1416, 1991.
45. P. R. Kumar, Scheduling manufacturing systems of re-entrant lines, in D. D. Yao (ed.), *Stochastic Modeling and Analysis of Manufacturing Systems*, New York: Springer-Verlag, 1994, pp. 325–360.
46. T. I. Seidman and C. Humes, Jr., Some kanban-controlled manufacturing systems: a first stability analysis, *IEEE Trans. Autom. Control*, **41**: 1013–1018, 1996.
47. S. Kumar and P. R. Kumar, Queueing network models in the design and analysis of semiconductor wafer fabs, *IEEE Trans. Robotics & Automation*, **RA-17:** 548–561, 2001.
48. S. A. Reveliotis, The instability of the last-buffer-first-serve scheduling policy for capacitated re-entrant lines, *Proc. ACC '98*, Philadelphia, 1998, pp. 2780–2784.
49. M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc., 1994.
50. J. Y. Choi and S. Reveliotis, Relative value function approximation for the capacitated reentrant line scheduling problem, *IEEE Trans. Automation Science & Engineering*, **2**: 285–299, 2005.
51. S. Kumar and P. R. Kumar, Performance bounds for queueing networks and scheduling policies, *IEEE Trans. Autom. Control*, **39**: 1600–1611, 1994.
52. K. E. Stecke, Design, planning, scheduling and control problems of flexible manufacturing systems, in *Ann. Oper. Res.*, **3**, 1985.
53. K. E. Stecke and N. Raman, FMS planning decisions, operating flexibilities, and system performance, *IEEE Trans. Eng. Manag.*, **EM-42**: 82–90, 1995.
54. K. E. Stecke and I. Kim, A flexible approach to part type selection in flexible flow systems using part mix ratios, *Int. J. Prod. Res.*, **29**: 53–75, 1991.

Reading List

J. C. Ammons, T. Govindaraj and C. M. Mitchell, Decision models for aiding FMS scheduling and control, *IEEE Trans. Syst. Man. Cybern.*, **18**: 744–756, 1988.

D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 2nd Ed., New York: John Wiley & Sons, 1985.

W. J. Hopp and M. L. Spearman, *Factory Physics: The Foundations of Manufacturing Management*, 2nd Ed., Chicago: Irwin Press, 2001.

A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, 3rd ed., New York: McGraw-Hill, 2000.

M. McClellan, *Applying Manufacturing Execution Systems*, Boca Raton, FL: CRC Press LLC, 1997

W. E. Wilhelm and J. Fowler, Research directions in electronics manufacturing, *IIE Trans.*, **24** (4): 6–17, 1992.

DOUGLAS A. BODNER
SPYROS A. REVELIOTIS
RONALD L. BILLINGS
H. Milton Stewart School of Industrial and Systems Engineering Georgia Institute of Technology, 765 Ferst Dr. NW, Atlanta, GA, 30332-0205