



**Method and Apparatus for Synchronizing  
a Substrate Processing System**



\*166166\*

**BACKGROUND OF THE INVENTION**

5           1.     Field of the Invention

The invention relates generally to the field of microelectronic fabrication. More particularly, the invention relates to a technique of optimizing the throughput, reliability, and quality of microelectronic fabrication by synchronizing the individual processing and transportation mechanisms within a fabrication system.

10           2.     Discussion of the Related Art

In the process of manufacturing a semiconductor device such as an integrated circuit, numerous steps of micro-fabrication are performed to form a device. These steps are performed serially on the individual items of manufacture in individual modules; the items of manufacture are transferred between modules by transport mechanisms such as robots. In order to achieve desirable throughput, reliability, and fabrication quality, several conditions must be met:

1)           The delivery and removal of the substrate to and from the process modules, as well as the transportation of the wafer between modules, must be accomplished in a timely manner. This timely delivery and removal of substrate is achieved when the flow of substrate is maintained in a periodic and synchronized manner. If periodicity and synchronization are not maintained, the process results will be inconsistent from substrate to substrate, and the expected throughput may be reduced.

2)           It is desirable to transport the substrate in similar process flow paths to avoid inconsistency in process results due to variations in the process history of the substrates.

3)           It is imperative to ensure that the articles of manufacture do not spend any pre-process or post-process time idling in modules

where critical processes are performed. The addition of pre-process or post-process time in these modules degrades not only the throughput but also the process results. For example, in an IC fabrication system, if a substrate is not immediately transferred from the spin coat module to a bake module to thermally cure a photo-resist film layer, the resulting film thickness will be unpredictable. If it is impossible to totally eliminate pre-process and/or post-process times, they should be rendered as brief as possible, and any variations in these times cannot be allowed.

The inability to meet any or all of the above conditions come from the failure to resolve transport conflicts. Conflicts are situations wherein separate modules demand a robot within a time span insufficient for the robot to service these modules

One conventional solution to the concerns listed above is the addition of extra process modules and transportation resources. However, the size limitations and geometrical constraints of a track system limit the possibility of resolving the above difficulties by adding additional process modules or transportation resources.

The addition of dedicated transfer arms to transfer substrates between adjacent modules (hereinafter called Inter Bay Transfer Arms, or IBTAs) is another method used to improve throughput and eliminate some of the pre-process and/or post-process times. However, the addition of IBTAs also has serious drawbacks. Dedicated transfer arms complicate the tool and increase its cost, constrain the position of the modules, and cannot be used everywhere in the tool. As a result, the tasks of managing the substrate flow in the track system while maintaining both high throughput and quality and resolving all transport conflicts become unmanageable.

Another conventional solution is to assign a set of substrate transport priority rules. Prior to any robot move, the control system, also referred to as the software scheduler, verifies the status of substrates in different modules and makes transfer priority decisions based on these rules. However, to achieve high

throughputs, the scheduler may generate undesirable, unpredictable and variable pre-process and post-process times in critical modules, and the substrates may also be forced to follow different flow paths to complete their process cycle.

Heretofore, the requirements of conflict resolution, synchronization, quality, and path consistency referred to above have not been fully met. What is needed is a solution that simultaneously addresses all of these requirements.

## SUMMARY OF THE INVENTION

A primary goal of the invention is to provide a synchronized substrate-transporting algorithm that resolves any possible flow conflicts and thereby improves the performance and the throughput of a track system.

Another primary goal of the invention is to provide an algorithm that determines the optimal allocation of transportation resources for balancing the load amongst transport mechanisms.

Another primary goal of the invention is to enable flexible synchronization of the fabrication system, ensuring that substrate processing and transportation occur at regular, predictable intervals without diminishing the throughput.

In accordance with these goals, there is a particular need for an algorithm that synchronizes the fabrication system, so that each process or substrate transportation always occurs at precise point in the interval. Thus, it is rendered possible to simultaneously satisfy the above-discussed requirements of conflict resolution, synchronization, optimized throughput, and wafer quality, which, in the case of the prior art, are mutually contradicting and are not simultaneously satisfied.

A first aspect of the invention is implemented in an embodiment that is based on an algebraic transformation. A second aspect of the invention is implemented in an embodiment that is based on a genetic algorithm.

These, and other, goals and aspects of the invention will be better appreciated and understood when considered in conjunction with the following description and the accompanying drawings. It should be understood, however, that the following description, while indicating preferred embodiments of the invention and numerous specific details thereof, is given by way of illustration and not of limitation. Many changes and modifications may be made within the scope of the invention without departing from the spirit thereof, and the invention includes all such modifications.

## BRIEF DESCRIPTION OF THE DRAWINGS

A clear conception of the advantages and features constituting the invention, and of the components and operation of model systems provided with the invention, will become more readily apparent by referring to the exemplary, and therefore non-limiting, embodiments illustrated in the drawings accompanying and forming a part of this specification, wherein like reference characters (if they occur in more than one view) designate the same parts. It should be noted that the features illustrated in the drawings are not necessarily drawn to scale.

Figure 1 is a time line illustrating the points at which wafers are loaded into a cluster tool, wherein the points are separated by intervals of one sending period.

Figure 2 illustrates the relative and absolute pick-up times of three wafers at a process  $i$ , as well as the mantissas of the pick-up times, wherein the pick-up times are normalized in terms of the sending period.

Figure 3 illustrates the various modules and transportation modules in the cluster tool.

Figure 4 is a graph and corresponding table illustrating the module paths in the cluster tool.

Figure 5 is a table illustrating the recipe of a cluster tool.

Figure 6 is a graph and corresponding table of module pick-up times which arise in a sending period, wherein the sending period is broken into 6

sub-intervals, as there are six possible robot moves within a single sending period.

Figure 7 is a graph and corresponding table illustrating conflicts that arise between processes for access to robots, wherein conflicts are indicated when two modules assigned to a robot have pick-up demands within a single period.

Figure 8 is a graph and accompanying tables and matrices, which illustrate the insertion of queues to eliminate conflicts between modules for access to robots.

## DESCRIPTION OF PREFERRED EMBODIMENTS

The invention and the various features and advantageous details thereof are explained more fully with reference to the non-limiting embodiments that are illustrated in the accompanying drawings and detailed in the following description of preferred embodiments. Descriptions of well-known components and processing techniques are omitted so as not to unnecessarily obscure the invention in detail.

### **Synchronized, Conflict Resolving Schedulers**

An aspect of the present invention comprises a method for maximizing throughput and quality in a manufacturing system by scheduling events in the system in a periodic, predictable fashion which eliminates conflicts for system resources. An example of such a manufacturing system comprises a series of process steps 1, . . . , N, which are performed consecutively on individual units of manufacture. The individual process steps of the system are conducted in "modules", or "process chambers", and the series of steps is listed in a "recipe". The manufacturing system also includes resources for transporting the units of manufacture between modules in the series; these resources may include robots.

Conflicts may result between processes in the system when separate modules demand a robot within a time span which is insufficient for the robot to

service these modules. Additionally, it is desirable to schedule the system in a manner which exhibits periodicity, so that events in the system are synchronized to occur at periodic, predictable intervals. An embodiment of the present invention includes a technique of selectively scheduling delays in various steps of the manufacturing process in order to eliminate all such conflicts, as well enforce periodicity, without degrading throughput or quality of the system.

#### **An Example of Conflict-Resolving Synchronization: Wafer Cluster Tools**

An example of the type of manufacturing system described above is a wafer cluster tool. In a wafer cluster tool, the modules comprise process chambers, which are organized around a group of wafer transporting resources, or robots, to perform a sequence of process steps on the wafer. A wafer enters and exits the tool through a buffer called a load port. Once a robot retrieves a wafer from a load port, the wafer is transported sequentially through the series of modules specified in a recipe. The time period defined by a wafer's entrance to a module and the wafer's exit from the module is referred to as a module process time. This process time includes the time actually spent processing the wafer in the module as well as the overhead time required to prepare the wafer for processing and pick up.

( note that though the wafer cluster tool is described above as passing an individual wafer between modules, it will be apparent to one skilled in the art that the present invention is equally applicable to a wafer cluster tool in which a discrete set of wafers is passed between modules.)

In certain modules of the cluster tool, a delay in picking up the processed wafer may adversely affect on-wafer results; such modules are identified as "critical process modules," as they cannot tolerate delays. The module whose process time is longest amongst all modules in the cluster tool is identified as the "gating module"; the process time at this module determines the throughput of the cluster tool. Because the gating module determines the throughput of the cluster tool, it too cannot tolerate delays. The recipe for a wafer cluster tool lists the modules in sequential order, alongside their

respective process times. The time required by a robot to transport a wafer between two modules is referred to as its transport time.

### **Wafer Flow Management in the Cluster Tool**

5           Wafer flow management, i.e., the orchestration of wafer processing and wafer transporting in a cluster tool, determines both the throughput and the on-wafer results delivered by the system. Effective wafer flow management requires the simultaneous satisfaction of the following two conditions: a wafer which was just processed in the sending module and is now ready to move  
10           should do so when (1) the receiving module in which the wafer will subsequently be processed is empty; and (2) the robot assigned to transport wafers between those modules is available. In the prior art, condition (1) was satisfied by providing additional redundant modules. Such a solution, however, compromises condition (2) in two ways: (a) it results in an inadequate number  
15           of robots serving too many modules or (b) two or more modules may compete simultaneously for the service of a robot.

          When the two conditions listed above are compromised, delays in wafer pick-up result. If such delays occur at critical process modules, they adversely affect on-wafer results. And if such delays occur at the gating module, they  
20           slow down throughput. As such, it is imperative that the transport conditions listed above are guaranteed with respect to critical process modules and the gating module. In case (a) wherein more handling is required than the robots could provide, adding more robots can mitigate the situation. However, in case (b), the problem resides in the timing of the robot service request. While adding  
25           more robots can also alleviate case (b), this is an inadequate solution.

          Since the recipe prescribed for the cluster tool determines the timings of the robot service request, a fundamental solution to resolving the two conditions can arise from altering the wafer recipe to synchronize with wafer transport. A scheduling algorithm described herein performs such synchronization.

30           This scheduler described herein can be encoded in software executed by a computer, wherein the computer comprises a memory for storing the software,

and a CPU for executing the software. In an embodiment of the present invention, the scheduler may be used off-line from the manufacturing system to generate a pre-determined schedule for the system. Alternatively, the computer may be coupled to the manufacturing system so that the scheduler can update the operation of the system in real-time.

### Synchronizing Wafer Flow in the Cluster Tool

In an embodiment of the present invention, wafer flow is synchronized by sending individual wafers through the cluster tool at a constant rate. This rate, referred to as the tool's "sending rate", is expressed in number of wafers per hour, and paces the wafer flow with a periodicity equal to  $(3600/\text{sending rate})$  seconds. This period, referred to as the sending period of the cluster tool, is the heartbeat of the system. Individual wafer units are introduced to the system at intervals of one sending period. And in order to synchronize the cluster tool, all process and transport times are measured in units of sending period. Furthermore, to ensure that the same tasks can be repeated in succeeding periods, the robots in the cluster tool are scheduled to accomplish all service requests, hereafter referred to as "tasks", within a single sending period. As such, the synchronization of the cluster tool requires a determination of 1) the total number of tasks that are performed within a sending period and 2) the exact moment within a sending period that these tasks arise. These moments shall hereafter be referred to as the "timings" of the respective tasks.

The concepts of sending periods and synchronization are illustrated in Figure 1. The timeline 100 has an origin 102, which denotes the moment when the first wafer is loaded into the cluster tool. The timeline 100 is demarcated in units of one sending period 110. Each demarcation 104 106 108 indicates, respectively, the times at which the second, third, and fourth wafers are loaded into the cluster tool.

A principal characteristic of synchronization is periodicity: the present invention ensures that for each task  $i$ ,  $i=1, \dots, n$ , the pick-up times for any wafer undergoing that task are identical. Thus each task  $i$  in the cluster tool can

be associated with a relative pick-up time denoted  $T_i$ , where  $T_i$  is normalized in units of the sending period. Figure 2 depicts this feature of periodicity. Three wafers, wafer 1 208, wafer 2 210, and wafer 3 212 are depicted on the vertical axis 202. The horizontal line depicts the TIME axis 200. The origin of this axis 201 indicates the time at which wafer 1 is loaded into the cluster tool. The relative pick-up times  $T_i$  at task  $i$  200 are identical for each wafer. Because the wafers themselves are introduced at intervals of one sending period, the actual pick-up times are separated by units of one sending period.

Figure 2 also illustrates a distinction between relative and "actual" or "absolute" pick-up times. The relative pick-up time of a process  $i$  is denoted by  $T_i$  204. Since the relative pick-up time is measured from the time a wafer is introduced into the wafer cluster tool, the relative pick-up time is identical for each wafer, wafer 1 208, wafer 2 210, and wafer 3 212. The absolute pick-up time 214 is measured from the moment the first wafer was loaded into the cluster tool 201. Since the wafers are introduced at intervals of one sending period, it follows that for any wafer no.  $w$ , the absolute pick-up time of wafer  $w$  at module  $i$  is

$$(w-1) + T_i$$

This period  $(w-1)$  is illustrated in the figure 216.

Another parameter which is critical in synchronization is designated by the symbol  $\tau_i$ . The fraction  $\tau_i = T_i - \text{INT}(T_i)$ , where  $\text{INT}(T_i)$  is a function that rounds  $T_i$  down to the nearest integer, is the fraction of  $T_i$  that has elapsed since the beginning of the current sending period. These parameters 206 are also illustrated in Figure 2. Since the  $T_i$  values are identical for each wafer, and since the wafers are inserted at intervals of one sending period, the values of  $\tau_i$  206 are identical for each wafer. These fractions,  $\tau_i$ ,  $i = 1, 2, 3 \dots N$  comprise the timings of the tasks the robots must accomplish within a sending period.

The number of tasks  $N$  and the timings of these tasks constitute the load of the transport. Since  $T_i$  is the accumulation up to the  $i$ th module of the process times  $p_j$ ,  $j = 1, 2, 3 \dots i$ ; and the robot transport times

$t_j, j = 1, 2, 3 \dots i-1$ , it follows that for any wafer, the relative pick-up time at module  $i$  is:

$$T_i = \sum_{j=1}^i p_j + \sum_{j=1}^{i-1} t_j; \quad i = 1, 2, 3 \dots N$$

It also follows that the timing of the tasks,  $\tau = 1, 2, 3 \dots N$ ; is

$$\begin{aligned} \tau &= T_i - INT(T_i) \\ &= \sum_{j=1}^i p_j + \sum_{j=1}^{i-1} t_j - INT\left(\sum_{j=1}^i p_j + \sum_{j=1}^{i-1} t_j\right); \quad i = 1, 2, 3 \dots N \end{aligned} \quad (1)$$

Since transport times  $t_i$  are fixed for a given cluster tool, it is apparent from Equation (1) that the timing of robot task  $\tau$  is dependent solely on the process times  $p_i$  as prescribed by the recipe.

### Periodicity and Wafer Identification

The property of periodicity also enables the identification of wafers in the cluster tool. As elaborated infra, the synchronized scheduler ensures that 1) the wafers are loaded into the cluster tool in sequential order at intervals of one sending period, and 2) each wafer loaded into the cluster tool undergoes identical events at the same times, as measured relative to the moment they are loaded. A consequence of these two conditions is that wafers enter and depart each module in the cluster tool in the order they were originally loaded, at intervals of one sending period. As such, each wafer in a module can be identified simply by tracking the order in which they entered or exited that module. This feature of the synchronized scheduler is referred to as wafer identification, or wafer "tagging".

### Tagging and Module Paths

In an embodiment of the present invention, each wafer loaded into the cluster tool follows a particular "module path", i.e., a particular set of modules which correspond to the processes in the cluster tool. This feature of the present invention is illustrated in Figure 4. In this embodiment, each process in the cluster tool has one or more modules associated with it, wherein the wafers are

processed. The modules for each process are ordered in a sequence such that when wafers arrive at that process, they are placed in the corresponding modules in the sequential order (for e.g., if a process has two corresponding modules, the first wafer in the system goes to the first module, the second wafer goes to the second module, the third wafer enters the first module, the fourth wafer enters the second module, etc.) As a consequence, the total number of module paths that a wafer may follow is constrained to the least common multiple of the number of modules corresponding to each process

The embodiment described above is illustrated by example in Figure 4. Figure 4 shows a sequence of process steps, VP 400, VPC 402, CT 404, SB 406, SBC 408, PEB 410, PEBC 412, DEV 414, HB 416, HBC 418. A symbol of a process step appears for each module corresponding to the process step. For instance, the process CT 404 has three modules, and corresponding, the symbol of CT appears three times 404. Above each process step is the number of modules for that process step 420.

In this example, the least common multiple of the number of modules is:

$$\text{LCM}(2,2,3,3,3,3,3,4,2,2) = 12$$

Hence, the recipe for the cluster tool prescribes twelve module paths, which are listed 422. Each column in the table 422 lists the module number for that process step in the respective module path. As there are twelve possible path, every twelfth wafer follows the same module path. As such, by identifying a wafer and the order in which it was loaded into the tool, the present invention enables the determination of the module path followed by the wafer.

## **Adding Queues to Eliminate Conflicts for Transportation Resources**

If a recipe gives rise to simultaneous, competing service requests for particular robots, it would be desirable to resolve the conflicts not by adding more robots, but rather by modifying the recipe itself. One convenient scheme to modify the recipe is to introduce deliberate delays, hereafter called queues  $q_i$ , to the non-critical process steps in order to achieve timing which resolves conflicts without compromising the throughput or on-wafer results delivered by

the tool. Such a scheme, used in conjunction with Equation (1), is the basis for the "synchronous algorithm." To recap, a recipe as originally prescribed may introduce competing service requests which result in delays at critical process and gating steps, thereby degrading the wafer quality and throughput of the single-wafer cluster tool. The aim of "synchronous algorithm" is to insert intentional delays at non-critical process steps in order to ensure that no delays occur at critical process steps or gating steps, and thereby ensure guarantees of throughput and wafer quality.

### Solving For The Queues

We shall now demonstrate how to solve for the delays, or  $q_j$ . Let  $\tau$  be the timing of robot tasks as dictated by a prescribed recipe per Equation (1). By adding queues  $q_i$  to the process time  $p_i$  to modify the recipe, the new timing  $\tau^*$  is given by:

$$\begin{aligned} \tau_i^* &= \sum_{j=1}^i (p_j + q_j) + \sum_{j=1}^{i-1} t_j - INT \left( \sum_{j=1}^i (p_j + q_j) + \sum_{j=1}^{i-1} t_j \right) \\ &= \tau \cdot \sum_{j=1}^i q_j - INT \left( \sum_{j=1}^i (p_j + q_j) + \sum_{j=1}^{i-1} t_j \right) + INT \left( \sum_{j=1}^i p_j + \sum_{j=1}^{i-1} t_j \right) \quad (2) \end{aligned}$$

The objective is to find a set of queues  $q_i$  to be inserted at the non-critical process steps such that the time interval between any two modules  $k$  and  $m$ ,  $k=1, 2, 3, \dots, N$ ; and

$m = k, k+1, \dots, N$ , where module  $k$  and module  $m$  are assigned to have their wafers picked up by the same robot, is greater than the transport time of the robot. This would allow for time intervals sufficient for the robot to service all modules and thus avoid having to serve more than one module at a given time. However, the queues so derived must also be small enough to avoid excessive idling of modules. And there should be no queues at critical process modules or the gating module.

The set of queues will be solved for using Equation (2). This yields a system of linear equations  $(\tau^* - \tau) = \sum_{j=1}^N a_{ij} q_j$ ; where  $a_{ij}$  is a lower triangular matrix with  $a_{ij}=0$ , for  $i < j$ ; and  $a_{ij}=1$ , for  $i \geq j$ :

$$\begin{bmatrix} \tau_1^* - \tau_1 \\ \tau_2^* - \tau_2 \\ \tau_3^* - \tau_3 \\ \tau_4^* - \tau_4 \\ \vdots \\ \tau_{N-1}^* - \tau_{N-1} \\ \tau_N^* - \tau_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ \vdots \\ q_{N-1} \\ q_N \end{bmatrix} \quad (3)$$

5 The constraint that no delays should occur at critical modules is now applied to Equation (3). For example if module #3 and #4 are critical, Equation (3) should be modified to the linear equations as shown below.

$$\begin{bmatrix} \tau_1^* - \tau_1 \\ \tau_2^* - \tau_2 \\ 0 \\ 0 \\ \vdots \\ \tau_{N-1}^* - \tau_{N-1} \\ \tau_N^* - \tau_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & 0 & 0 & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 0 & 0 & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ \vdots \\ q_{N-1} \\ q_N \end{bmatrix} \quad (4)$$

10 In Equation (4) above, the timings  $\tau$  as prescribed by the original recipe are known. The target timings  $\tau^*$  are set to values which eliminate conflicts between all modules using the same robot, as described earlier. Thus LHS of Equation (4) are known values. The vector  $q_i$  is then solved for by pre-multiplying  $(\tau^* - \tau)$  with the inverse of the modified constraint matrix as shown in Equation (5) below. Adding this set of  $q_i$  to the corresponding module  
15 process time  $p_i$  will synchronize wafer transport with wafer process.

$$\begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ \vdots \\ q_{N-1} \\ q_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}^{-1} \times \begin{bmatrix} \tau_1' - \tau_1 \\ \tau_2' - \tau_2 \\ 0 \\ 0 \\ \vdots \\ \tau_{N-1}' - \tau_{N-1} \\ \tau_N' - \tau_N \end{bmatrix} \quad (5)$$

### An Application of the Synchronous Algorithm

Specific embodiments of the invention will now be further described by the following, nonlimiting example which will serve to illustrate in some detail various features of significance. The example is intended merely to facilitate an understanding of ways in which the invention may be practiced and to further enable those of skill in the art to practice the invention. Accordingly, the example should not be construed as limiting the scope of the invention.

The synchronization of wafer transport with wafer process in a cluster tool will now be illustrated using a specific example of a cluster tool. Figure 3 is a schematic for a conceptual plan view of a wafer processing apparatus. The resist coating and developing modules are identified as CT 300 and DEV 302 respectively. Also shown in Figure 3 are the different bake modules vapor prime (VP) 304, soft bake (SB) 306, post exposure bake (PEB) 308, and hard bake (HB) 310, as well as their corresponding chill modules. The arrows connecting adjacent bake and chill modules represent the inter bay transfer arms, IBTAs 312, that transfer the substrate between these modules. As a result, the locations of these bake modules constrain the location of their corresponding chill plate. The cassette end station (CES) robot 314 shown in the figure transfers substrates from and to the cassette end station. The stepper interface (SI) robot 316 transfers substrate from and to the stepper interface. The I/O module 318 is a buffer zone for the substrate transported to the stepper interface if and when becomes necessary. The main robot 320 is the means to transport the substrate between all other modules such as vapor prime chill (VPC) to resist coat (CT.)

Figure 4 is a schematic for the wafer process flow. As can be seen from the schematic, when the transport means is the IBTA the substrate will have only one option in the flow diagram. This is the case when the substrate is transported from a vapor prime bake 400 to vapor prime chill 402. However, when the transport means is the main robot the substrate could have several options. For example, when a substrate is removed from a resist coat module 404, it can be transported to any of the soft bake modules 406 shown in Figure 4.

## The Synchronization Algorithm

We will now demonstrate the application of the synchronization algorithm to this cluster tool as a series of four steps:

- **Step 1 Input The Recipe And Throughput Requirement.** This discussion refers to Figure 5. This step commences by inserting the recipe in a table 500. The first two columns 502 list the process steps in sequential order. The sending period, also referred to as the "system takt time" is also noted 504. Cycle times, also known as module takt times, are then calculated for each module in the "Module type" column 502 to ensure each module takt time is less than the system takt time. The module takt time for each process step is listed in a column 506. If not, redundant modules are added to reduce module takt time. As will be clear to one skilled in the art, for each process step:

$$\text{No. Modules Required} = \text{INT} (\text{Module takt time} / \text{System takt time})$$

In this example, most modules require one additional redundant module. The number of modules required for each process step is listed in a column 508.

- **Step 2 Determine The Load of The Wafer Transport.** As defined earlier, the "load" of a robot refers to the number of moves it is scheduled make as well as the times allotted for the robot to perform

these moves, as measured from the beginning of the current sending period. The determination of the robot load is depicted in the table 600 in Figure 6. This cluster apparatus of this example has twelve process steps. The timings of the twelve corresponding robot moves are determined as follows. Counting the time from the moment the wafer leaves the load port (cassette), cumulate the process times of each module and transport times up to the module of interest. For example, for the module code-named VP 602 (hereafter all modules are coded), it is  $(62+6) = 68$  sec; for module VPC 604, it is  $(68+65+5) = 138$  sec. The actual timings in seconds for all robot moves are listed in a column marked "Actual" 606. To determine the relative pick-up times  $T_i$ , the actual times are divided by the actual sending period. For example, dividing the actual pick-up times of VP and VPC by the actual sending period of 45 sec results, respectively, in normalized pick-up times 1.511 and 3.067. The normalized pick-up times for each of the twelve process steps are listed under the column " $T_i$  Normalized" 608.

Subtracting the integer portion of the  $T_i$  values results in the  $\tau_i$  values, i.e., the time the available time the respective modules have to be served by robots, as measured from the beginning of sending period, and expressed in units of one sending period. To illustrate, the  $\tau_i$  values of VP and VPC are 0.511 and 0.067, which indicates that VP must be serviced within .511 intervals of one sending period, and VPC must be serviced within .067 intervals of one sending period. The list of normalized  $\tau_i$  values are listed in a column 610. The diagram 612 is a pictorial representation of the load: a total of twelve moves and the time they have to be served within a sending period. Figure 7 shows the same information from another perspective. Since the robot transport times are around 5 & 6 seconds, the number of move one robot can make within a sending period of 45 sec is  $(45/6) \sim 7$ ; say 6 moves to be on the conservative side. Thus 6 vertical lines are drawn in the graph 700.

When the times the modules have to be served by a robot fall within one

interval, such as SBC 702, DEV 704, HB 706, and Cassette 708, all of which fall into one interval 710; it means that they are competing for the same move of the robot at a given time. To elaborate, if there are two or more tasks which use one robot, and if the  $\tau$  values of these tasks fall into one of the 6 intervals, it means that there is insufficient time for the robot to service each of the two or more robots.

These "conflicts" between the tasks for use of the robots are resolved as described in the subsequent steps.

• **Step 3 *Allocate The Transport Load.*** The first step to resolve conflicts in the manufacturing system is to allocate the load equally among robots in order to achieve a balanced transport load. Although conflicts are recipe dependent, assigning fewer loads per robot still reduces the chance for conflict. However, the possibility of balancing the load is dependent on the layout of the modules relative to robots. Poor layout limits accessibility to modules by robots and makes balanced loads difficult to attain. In this example, the layout is such that two robots, CES and SI, can each only serve two modules, leaving the bulk of the load to the main robot C-1 and three dedicated robots known as IBTA (Inter Bay Transfer Arm). The best allocation of transport load, under the constraints of the layout, is as shown in Figure 7. Six modules, VPC 712, CT 714, SBC 716, PEBC 718, DEV 720, and HB 722 are allocated to the main robot C-1 with three of them, SBC, DEV and HB competing for one move of the robot, as discussed earlier. These conflicts will be resolved by queuing as demonstrated in the next step.

• **Step 4 *Queuing For Synchronization.*** The information in the table 800 of Figure 8 is a summary of load allocation from Step 3. Only the six modules served by main robot C-1, i.e., VPC 802, CT 804, SBC 806, PEBC 808, DEV 810, and HB 812, need to be considered for queuing. The remaining modules should not experience conflicts since each has

dedicated robot, i.e., an IBTA, serving them. In the Target Column 814, the target timing profile is set for the 6 modules. For each of the modules listed in the table, a corresponding value is set for  $\tau^*$ , where  $\tau^*$  is an updated value for  $\tau$  which eliminates conflicts between tasks for robots. Since only 3 of these 6 modules, SBC 806, DEV 810, and HB 812 are in conflict, only two, DEV and HB, need to have timing targets different from the original values for  $\tau$  prescribed by the recipe. The timing targets are listed in a column 814, and are set such that the timing interval between any pair of the 6 modules is larger than robot transport time ( $=6/45 \sim 0.1333$ ). The differences between the target and the originally prescribed timing profiles are referred to as the gaps and are computed the column 816. These are shown pictorially in the graph 818 adjacent to the table. Another objective of the "synchronous algorithm" is to ensure that no delays are introduced at critical process steps. In this example, critical process steps are step 3 804, step 4 806, and step 7 809. No queues should be added to modules corresponding to these steps, i.e., the target timing for these modules should be the same as the prescribed values. The gaps computed in the Gap column 816 may now be substituted in to Equation (3) to solve for the queues that will close these gaps. However, to ensure zero delays at the critical process modules, the matrix 818 relating gaps and queues must be modified per Equation (4) to generate a modified matrix 820. Pre-multiplying the gaps from the Gap column 816 with the inverse of the modified matrix 822 produces the queues needed to close the gap 824. The solution for the queues is transferred to a Que Column 826. The solution, which is in units of the sending period, is converted to actual time in an Actual Que Column 828.

**Step 5 Check The Solution.** The queues determined in Step 4 are now added to the module processing time of the original recipe. This is to

verify if conflicts have been resolved. This is in fact the case as shown in the figure.

### Robot Assignment

Another aspect of the scheduling problem, which merits automation, is the assignment of robots to modules. For example, in step three listed above, a recipe was chosen which assigned a single robot between each pair of consecutive models; this allocation is shown in the recipe listed in figure 7. The allocation was chosen amongst many possible allocations.

In general, there is a need for an algorithm that determines an optimal robot assignment prior to the determination of queues. The need for such an algorithm will be demonstrated in the following example. Suppose we have a simplified track system, consisting of three modules, labeled Mod1, Mod2 and Mod3. Suppose we have two robots, Robot1 and Robot2, both of which can service all three modules. Let the sending period be designated by the variable SP. Suppose  $\tau_1=0.0$ ,  $\tau_2=0.6$ ,  $\tau_3=0.7$ , in units of the sending period, and suppose the robots can move in 0.3, in units of SP. There are four possible robot allocations:

| Assignment | Mod1 to Mod 2 | Mod2 to Mod3 |
|------------|---------------|--------------|
| 1.         | Robot1        | Robot1       |
| 2.         | Robot2        | Robot2       |
| 3.         | Robot1        | Robot2       |
| 4.         | Robot2        | Robot1       |

Upon inspection, only assignments 3 and 4 are viable. In assignments 1 and 2, the time interval between  $\tau_2=0.6$  and  $\tau_3=0.7$  is 0.1 sending periods, which is less than the 0.3 sending periods required for a robot to move. Hence the optimal allocations are, in this case, assignment #3 and #4; as the time intervals between the  $\tau$  values have sufficient distance, this robot assignment obviates the need to insert delays. Other criteria may also enter into the determination of an optimal robot assignment, for instance, balancing loads, increasing throughput. An

algorithm is necessary which would determine, in cases more complicated than the simplified example above, an optimal robot assignment.

One method of performing such an assignment is simply an exhaustive technique: generate all possible robot assignments, and determine the validity of each assignment, i.e., ensure that all modules assigned to a robot differ sufficiently in their  $\tau$  values to permit the robot to service them. The assignments thus generated may also be selected on additional criteria, such as load balancing.

#### 10 Solving for Updated Timings ( $\tau^*$ )

Another feature of the synchronization that merits automation is the derivation of the updated timings, given by  $\tau^*$ . To elaborate, in step 4 of the algorithm outlined earlier, the algorithm was fed updated values of  $\tau^*$ , where for any two modules which shared a robot, the respective  $\tau^*$  values differed by enough time to allow the robot to move between them. There is a need for an automated method of deriving these  $\tau^*$  values. One such technique is as follows:

For each robot with a conflict, take the  $\tau$  value for each of its modules.

For each combination of these  $\tau$  values, sort the  $\tau$  values from lowest to

20 highest. For each sorted list of  $\tau$  values:

- Proceed sequentially through the  $\tau$  values, from lowest to highest.
- Determine the difference between the given  $\tau$  value and the one that precedes it
- 25 • If the difference is less than the time allotted for the robot to move, increment the  $\tau$  value sufficiently.
- Go to next  $\tau$  value.

When this algorithm finds an updated set of  $\tau$  values for a given robot that eliminates conflicts, these become  $\tau^*$  values. It can be proven that if a conflict-free group of  $\tau^*$  values exist, the algorithm outlined above will find it.

30

### Genetic Algorithms

The synchronization, robot assignment, and derivation problems can also be solved by use of a genetic algorithm (GA). A GA is an iterative process that starts out with an initial population of genes, which encode possible states of the problem. This population is systematically improved with each iteration, through a process of selective breeding. To perform selective breeding, the GA needs to a) define the characteristics of a species, and b) judge a species' fitness.

**Characteristics of a Species** A specie is characterized by  $n$  genes. For our problem, we use two types of genes, one to represent the robot assignment, and the other, a queue segment. Consider the example used in the description of the earlier algorithm. The robot assignment will range from 1 to 4, indicating which robot will work on a particular module. Queue segment is also an integer indicating how many "time zones", i.e., robot move periods, are to be added to a module's arrival time in order to avoid conflict in robot assignment. In our earlier example, the modules arrive in six different time zones, as shown in figure 6 612. If a time zone sees the arrival of, for instance, five modules, a conflict results. The addition of a queue segment to one of the modules will push the arrival time to the next time zone and hence resolve the conflict.

**Fitness of a Species** We can measure fitness by the reciprocal of a species 'badness'. In turn, badness can be measured by a weighted sum of the degree of conflict and the number of added queue segments. An ideal species is one that has no added segments, and results in no conflict in robot assignment.

To derive the fitness function, we scan each time zone and count the number of redundant assignments for each robot. The results are summed for all robots and all time zones. Call this sum  $s$ . We proceed to count the number of added queue segments, and call it  $t$ . The fitness function is then

$$f(s,t) = \frac{1}{(1 + w_1 s + w_2 t)};$$

where the weights  $w_1$  and  $w_2$  are assigned according to the relative importance of  $s$  over  $t$ .

5       The foregoing description of a preferred embodiment of the invention has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise forms disclosed. Obviously, many modifications and variations will be apparent to practitioners skilled in this art. It is intended that the scope of the invention be defined by the following claims and their equivalents.

What is claimed

## CLAIMS

- 5 1. A method of optimizing throughput and performance in a wafer processing system, wherein the wafer processing system includes a plurality of modules for processing individual wafers in a sequence of wafers, and the wafer processing system includes a plurality of transportation resources for transporting a wafer unit from a first module in the plurality of modules to a subsequent module in the plurality of modules, the method
- 10 comprising:
- fixing a constant sending period for sequentially dispatching wafers into the wafer processing system, the sending period remaining fixed throughout the sequence of wafers;
- expressing times for pickup of the wafer from an  $i$ th module,  $(\tau_i)$ , in units of the sending period;
- 15 expressing a plurality of times  $(\tau_i)$  allotted for a wafer transporter to arrive at a module from initiation of a current sending period for transportation of a wafer from a module  $i$  to a module  $i+1$  as a portion of the sending period, and;
- selectively incrementing at least one in the plurality of times  $(\tau_i)$  to arrive at a solution of times  $(\tau_i)$  that remains fixed at constant values throughout the sequence of wafers, the
- 20 solution of times  $(\tau_i)$  assuring availability of the plurality of transportation resources.
2. The method of claim 1, wherein the plurality of modified times  $(\tau_i)$  are updated to accommodate changes in wafer processing system parameters.
- 25 3. The method of claim 1, wherein the plurality of modified times  $(\tau_i)$  are updated in real-time.
4. The method of claim 1, wherein the plurality of modified times  $(\tau_i)$  are updated to facilitate recovery of the wafer processing system, after deviation from nominal operating
- 30 conditions.

5. The method of claim 1, wherein the modified times ( $\tau_i$ ) are determined off-line from the operating of the wafer processing system.
6. The method of claim 1, wherein optimal increments for the plurality of pick-up times are  
5 determined by an iterative process.
7. The method of claim 1, wherein optimal increments for the plurality of pick-up times are determined by an algorithm in linear time.
- 10 8. The method of claim 1, wherein the difference between any two updated times ( $\tau_i^*$ ) is greater than or equal to the sending period divided by a time for one robot transfer.
9. The method of claim 8, wherein the difference is normalized in units of the sending period.  
15
10. The method of claim 1, wherein at least one of a plurality of individual scalars in a vector of delays is zero.
11. The method of claim 10, wherein a linear transformation is used to transform the vector  
20 of delays.
12. The method of claim 11, wherein the linear transformation is used to derive optimal pickup times.
- 25 13. An apparatus for scheduling allocation of a plurality of transportation resources between modules of a sequential wafer manufacturing process, comprising:
- a memory resource for storing an algorithm for scheduling the plurality of transportation resources;
  - a central processing unit coupled to the memory resource;
  - 30 an input/output interface coupled to at least one of the central processing unit and the memory resource, the input/output interface adapted, respectively, to input parameters describing the sequential wafer manufacturing process and to output schedules for allocation of transportation resources, the output schedules assuring availability of

transportation resources and determined such that a sending period and a plurality of times ( $\tau_i$ ) allotted for a wafer transporter to arrive at a module from initiation of a current sending period for transportation of a wafer from a module  $i$  to a module  $i+1$  for the sequential wafer manufacturing process remain fixed at constant values throughout a sequence of wafers.

5

14. The apparatus of claim 13, wherein the algorithm ensures availability of wafer transportation to minimize potential delays.

10

15. The apparatus of claim 14, wherein the memory resource and the central processing unit are external to the transportation resources.

16. The apparatus of claim 14, wherein the algorithm identifies at least one critical module that should not incur delays and exempts the at least one module from the addition of delays.

15

17. The apparatus of claim 16, wherein the algorithm employs a linear transformation to determine the appropriate delays to eliminate conflicts for the plurality of transportation resources.

20

18. A method for synchronizing a wafer processing system, wherein the wafer processing system comprises one or more ordered sequence of modules, the method comprising:

setting a constant sending period of the sequential processing system that remains fixed throughout a sequence of wafers, the constant sending period equaling an inverse of a sending rate;

25

expressing a plurality of times allotted for wafer transporters to arrive at corresponding modules of the one or more ordered sequence of modules as a plurality of portions of the sending period, each and every time allotted for wafer transporters remaining fixed at a constant value throughout a sequence of wafers.

30

19. The method of claim 18, where the wafer transporters are assigned to individual modules of the ordered sequence of modules in order to minimize conflicts between the modules for access to the wafer transporters.

20. The method of claim 18, where an ordered series of wafer sets is processed by the wafer processing system, and a rank of a wafer set within the ordered series of wafer sets identifies a module path followed by the wafer set.

- 5 21. The method of claim 20, wherein the module path followed by the wafer set is an ordered sequence of modules from the one or more ordered sequence of modules.

22. The method of claim 21, wherein the wafer processing system comprises one or more processes, and a number of modules corresponding to each of the one or more processes.

10

23. The method of claim 22, wherein the wafer processing system has a number of module paths equal to a least common multiple of the number of modules corresponding to each of the one or more processes.