

CHINA HAS BECOME AN INTERNATIONAL center for semiconductor assembly and testing mainly due to relatively cheap labor and easy access to a huge market (Figure 1). With increasing facility and capital equipment investment, the demand for soft skills in complex production management is also growing rapidly because the profitability is highly dependent on

process efficiency.

The complexity of managing semiconductor production has been well recognized. In the semiconductor industry, the ratio of mean cycle time to the sum of process time, called actual-to-theoretical ratio, may range between 2.5 to 10 — more than most other industrial processes. The larger the ratio is, the more complex

the underlying process tends to be.

Traditionally, wafer fabrication is much more complicated than assembly and testing. However, the complexity of assembly manufacturing has been increasing exponentially since the introduction of multiple-chip products, which consist of several chips or spacers in one package. The technology requires conducting two »

National Semiconductor's Singapore assembly and testing facility uses the latest equipment for automated manufacturing.

# Doing by virtual experimenting

BY MIKE TAO ZHANG AND YANFENG WANG

*A reactive scheduling system in semiconductor manufacturing*

## doing by virtual experimenting

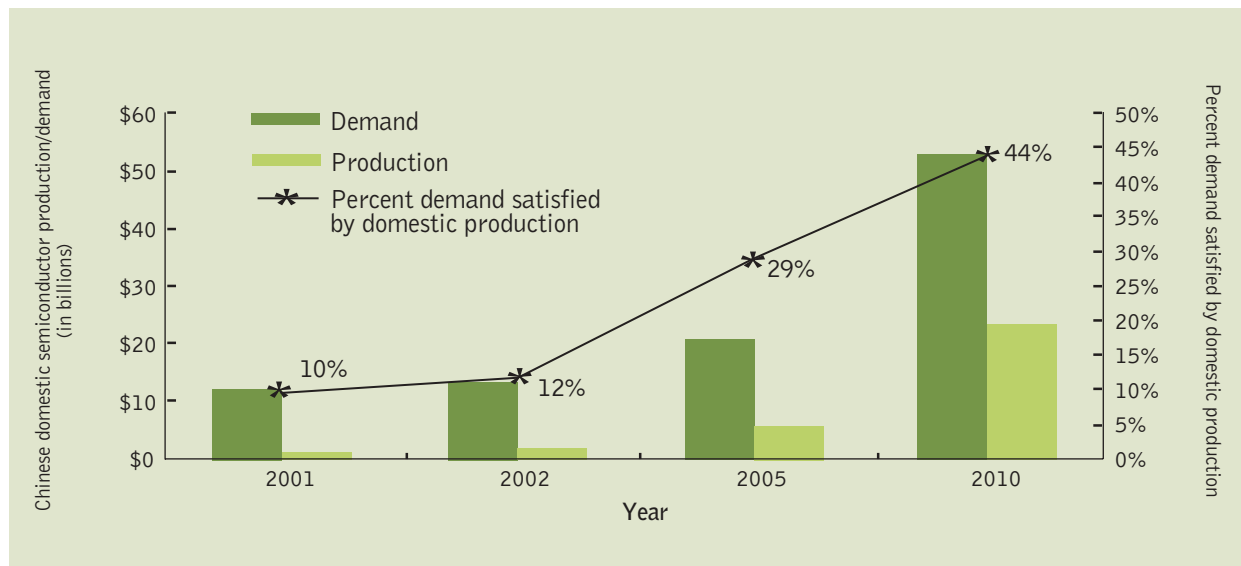


Figure 1. Semiconductor demand and domestic production in China is booming.

Source: Applied Materials Corporate Marketing Estimate, CCID

specific operations multiple times; therefore, the re-entrant workflow is a major characteristic of assembly and testing.

The first step to improving the effectiveness of complex assembly and testing systems is understanding production dynamics with the re-entrant workflow, which is most easily accomplished by simulating what-if scenarios. Better understanding gained from what-if analysis helps management make the right decisions consistently and consciously.

Although simulation could potentially generate enormous benefits, the people that can apply simulation technology to draw insightful conclusions are still a scarce resource, especially in an emerging market such as China.

Experimental analysis in simulation is normally trial-and-error and depends on the modeler's experience, intuition, and framework to lead the design of experiment. Therefore, it has become important in the semiconductor industry to know how to extend simulation technology to less experienced users who have limited or no knowledge of simulation while still allowing them to make better decisions through experiments. The answer is to structure the simulation experiments in a repeatable and controllable manner, which provides a user-friendly and easy operational interface.

A reactive scheduling system (RSS) has been proposed by a number of academic researchers and industrial practitioners. RSS addresses scheduling as a problem of maintaining a prescriptive solution over time and emphasizes objectives (for example, solution continuity and system responsiveness) that relate directly to effective development and use of schedules in dynamic environ-

ments.

The main advantage of an RSS is the effective and rapid response to production system changes that allows it to be used anytime a new schedule is needed. A sound simulation-based scheduling system lets operators conduct a series of scenario analyses quickly, decide the best job sequence, and apply the results to the shop floor operation. This is called doing by virtual experimenting, and it has proven to be an effective approach to disseminate the power of simulation technology in the semiconductor industry.

### The objective

Usually, two phases are required to develop an RSS. The objective of Phase 1 is to leverage simulation technologies to gain a better understanding of the production operations, to identify key system parameters, and to learn the effect of the interactions among these factors on machine utilization, cycle time, and line capacity.

The objective of Phase 2 is to design a scheduling tool that uses simulation techniques to allow users to adjust key parameters through a structured interface to generate a good job schedule in a short time.

In Phase 1, the modeler studies the factory floor practice and develops a simulation model to capture production dynamics. This model is then validated using historical data such as lot release, shipping date, cycle time, and machine utilization. In Phase 2, the validated analytical simulation model can be transformed into a more user-friendly and precisely operational scheduling tool.

Simulation-based scheduling is an alternative to seeking an optimized solution in the presence of large variation. By modeling as deterministic, the system avoids the NP-hard optimization problem — the complexity class of decision problems that take an unacceptably long time to solve and thus require alternative methods such as simulation to find near-optimal solutions. Instead, the system provides options to key decision parameters and allows users to choose the best one quickly according to predefined performance measures.

One advantage of a simulation-based scheduling system is that the scheduling results are easy to understand. The system mimics the behavior of the actual system intuitively, and most of the scheduling logic embedded in the system is what operators and planners routinely use. Therefore, there won't be a great surprise in the final results. In practical use, it is very important for a system to be easily understood by users. An optimal solution will be regarded as a better alternative only if it can be shown to be a realistic one that can reproduce common sense.

Another advantage is that it is very efficient to generate a number of what-if schedules. Normally, the real production is much more dynamic than the system could possibly capture in a timely manner. It requires the scheduling tool to be responsive to the system change. The responsiveness refers not only to the time to generate output but also the time to adjust input to reflect the system change. Therefore, the system is usually integrated with a work-in-progress tracking system to download current workflow status.

The time to schedule one week's worth of demand can be easily reduced up to 90 percent in semiconductor assembly manufacturing by migrating from manual processes to the RSS. Besides, the schedule time horizon can also be significantly extended up to 20 times with considerable accuracy.

The RSS also provides the following:

- A consistent scheduling strategy and logic that leads management to a more consistent and predictable scheduling style.
- More visibility to the scheduling horizon that leads to fewer device conversions, better tool allocation, improved machine utilization, and shorter cycle time.
- A sound foundation for further productivity improvement. The system allows people by virtual experiments to understand the relationship among key parameters and how they affect the factory production. Productivity improvement opportunities can be identified and ranked, and then the parameter changes (for example, optimal lot size) can be driven to achieve the most productivity improvement with the least investment.

## The RSS

The semiconductor assembly and test manufacturing system is called RESTLESS for the acronym of its characteristics:

- Re-entrant flow
- Early disruption forecast unobtainable
- Sizing lot
- Time constraints between steps
- Large disparity of tool performance
- Evolving line item mix
- Sequence-dependent conversions
- Shared operators or technicians

The assembly line mixes a number of devices, each of which follows different assembling routes. On the routing, two main operations — die attach (DA) and wire bonding (WB), are repetitively performed in a re-entrant manner. With a different device mix and release schedule, either of these operations could become the constraint of the assembly line.

Operations from DA to WB have been regarded as the bottleneck of the assembly line. If these operations are not scheduled effectively, the result will be longer cycle time, higher work-in-process accumulation, more frequent conversions, and most important, capacity degradation.

The RSS usually has the following decision variables (which operators can adjust to achieve performance goal): lot sizing, lot release, and device mix.

The performance goal of the RSS usually has the following objectives (which operators attempt to achieve by adjusting the aforementioned decision variables): on-time shipping, cycle time, and tool utilization.

Given the complexity and highly dynamic production environment, it is almost impossible to draw consistent and unambiguously defined rules to guide the schedule making. No clear underlying causes that drive the production have been identified and articulated. Therefore, exact duplication of the production operation into a computer simulation model becomes extremely difficult. Thoroughly understanding and summarizing production dynamics becomes critical for the simulation model to capture the system in essence, not just in appearance.

## The simulation logic

The simulation model should recognize that the key scheduling difficulties for semiconductor assembly and test manufacturing are tool allocation and production synchronization between DA and WB.

**Two-wheel drive.** Simply speaking, DA and WB are like two

## doing by virtual experimenting

wheels of a car: a driving wheel and a driven wheel. If the constraint is mainly in WB, it requires that WB drive the production rhythm to ensure the smooth flow between upstream and downstream. Meanwhile, WB is driven by the cutoff time at each major operation, such as WB<sub>1</sub>, WB<sub>2</sub>, WB<sub>3</sub>, etc., where WB<sub>1</sub>, WB<sub>2</sub>, and WB<sub>3</sub> are the first, second, and third operations, respectively, at WB due to the re-entrant working process. The same logic is applied to DA<sub>1</sub> through DA<sub>4</sub>.

The cutoff time at each step in DA and WB plays an important role in deciding the process sequence. The safety time buffer is predefined as a constant and stored in the model database, while cutoff time is equal to the due date minus the buffer.

The overall scheduling goal is to ensure on-time processing while minimizing the number and frequency of conversion. This simulation model is primarily based on heuristic rules for optimization. A production lot has higher priority if it has less slack time than required. Otherwise, when all devices have sufficient slack before the cutoff time, the priority is given to the devices that will require fewer conversions.

**Penalty.** Production dynamics is the result of balancing multiple driving factors that may have opposite impact. To make a better trade-off, decision makers have to assign implicitly or explicitly the weight of importance to each factor and consider the overall outcome in an aggregated fashion. Therefore, the simulation model introduces the concept of penalty, which is an aggregation of a number of weighted parameters. Then the lot is sequenced based on this penalty.

**Reducing the number of conversions.** Conversions consume tool capacity and increase cycle time. The simulation model assigns a (penalty) weight to the lot if the conversion is in need — that is, if the device in the lot has a different device name or step number than that of the previous lot when the lot enters into a certain machine queue. The weight is defined according to the conversion time and complexity.

**Downstream/upstream synchronization.** Synchronization

### SIMSOL BLOOMS IN MAY

Don't miss this year's Simulation Solutions Conference, which will be held in conjunction with the IIE Annual Conference in Atlanta May 18-19. Tracks focus on basic simulation skills; manufacturing; health care; transportation and military applications; supply chain, material handling, and distribution; and services and business processes. Go to [www.simsol.org](http://www.simsol.org) to register.

between steps would reduce the number of conversions on downstream machines. Because of vast differences in run rate between DA and WB, it becomes a very complicated decision to allocate the appropriate number of tools to the appropriate devices at the appropriate time to ensure the smooth continuity between upstream and downstream production. How can the system ensure no long idleness due to starvation at fast machines and WIP accumulation at slow machines in the re-entrant production? This is an open problem still under extensive research.

The good policy is subject to the trade-off among multiple objectives. It is also a dynamic function of ever-changing WIP distribution. Therefore, the policy requires constant update and, more importantly, may require looking forward at future WIP variation and making decisions based on constant forecasting.

**Tool device compatibility check.** The simulation model recognizes the existence of a variety of disparate tools as well as device restrictions on some tools. The simulation logic always allocates devices with minimal penalty.

**WB/DA device/tool selection.** A device with no conversion requirement will be directed to the appropriate tool when idle. Otherwise, the system constantly checks the WIP and the status of WB/DA to determine the appropriate device to schedule according to the following priority:

1. Device with the greatest lateness
2. First-in, first-out for unscheduled device
3. Move devices from slower to faster tools if required
4. No device selected

The right tool for the chosen device should be selected for a conversion. The tool chosen has to meet the following criteria:

- The tool has been idle longer than the predefined amount of time, which is roughly equivalent to the conversion time.
- The tool passes the tool-device compatibility check.
- The tool is the one that incurs the minimal penalty for the selected device.
- The search follows priority sequence by penalty ranking.

**WB adequacy.** An adequate number of WB tools should be allocated to a device. That means these tools could complete all current WIP of that device before the cutoff time with a certain amount of safety time buffer. If not, extra tools should be allocated to meet the cutoff time requirement.

As seen in Figure 2, if the following condition holds, the number of tools allocated to a specific device in WB is not sufficient:

Current time + Used time + Allocation time + Safety time > Cutoff time

Where

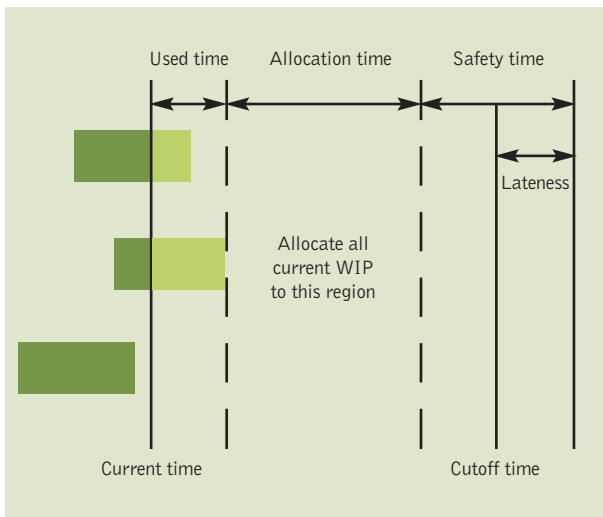


Figure 2. WB capacity allocation

- Current time refers to the moment that logic block makes the check.
- Used time is the average amount of time that WB tools will need to finish processing the already allocated lots.
- Allocation time refers to the amount of time required to spread evenly the current WIP over the already assigned tools. It is based on the theory that to complete WIP in the minimal make span, the workload has to be evenly distributed.
- Safety time is a buffer that refers to the amount of time the lot needs to be done ahead of the cutoff time.

**DA adequacy.** Because of the run rate discrepancy between DA and WB, the simulation deploys a simple yet useful rule of thumb to determine the adequacy of DA quantity. This rule is only evaluated when the downstream WB operation is in process.

If the subsequent WB process is started, the number of required DA tools is defined by the number of WB divided by the WB/DA ratio, where the WB/DA ratio is a global parameter that can be changed via the user interface. Its default value is five, mainly determined by the average run rate ratio between DA and WB. If its value is less than two, allocate one DA.

## SEMICONDUCTOR SALES ON THE RISE

According to the research firm Gartner Inc., worldwide semiconductor sales grew 23.4 percent to \$218.5 billion in 2004. The Associated Press reported that Intel Corp. topped all chip companies in sales for the 13th consecutive year with an estimated \$30.5 billion, while Samsung Electronics came in second with \$15.6 billion.

The rationale is to indicate the scheduling preference in which WB and DA should have a balanced capacity allocation to ensure on-time shipping of the finished goods.

## The conclusion

The simulation-based reactive scheduling system provides a doing-by-virtual-experimenting approach to conducting shop floor scheduling and what-if analysis in the fast-changing complex semiconductor assembly and test environment.

Simulation technology is introduced to mimic the behavior of the actual system intuitively, incorporating heuristic rules that improve the current practice while maintaining familiarity to users.

The RSS provides a solid foundation and extensibility for future work. It is an excellent tool for sensitivity analysis of the driving factors of factory physics, helping management identify critical potential areas for productivity improvement. Simulation is also valuable in forecasting the impact of parameter changes such as material availability and product mix variance. And RSS clarifies cost analysis in terms of overall equipment effectiveness, cost of ownership, and cost per good unit.

The main purpose of applying these Semiconductor Equipment and Materials Institute standards is to integrate all important information, including capacity, demand, yield, cost, and throughput, into a standard measurement to drive the company's ultimate goal of profitability. ~

## For further reading

Hasenbein, John, Silpa Sigireddy, and Robert Wright, "Taking a Queue from Simulation," *Industrial Engineer*, August 2004  
 Smith, Stephen F., "Reactive Scheduling Systems," *Intelligent Scheduling Systems*, Kluwer Publishing, 1994

*Mike Tao Zhang is a manager of the industrial engineering department at Intel Shanghai and a visiting associate professor of industrial engineering at Tsinghua University in China. He holds master's and doctoral degrees in industrial engineering and operations research from the University of California-Berkeley. His research interests are production planning and scheduling and optimal supply chain management.*

*Yanfeng Wang is the founder and CEO of Edgestone Information Technologies (Shanghai) Inc. of China. He holds a Ph.D. in manufacturing engineering from Boston University. His research and business interest is to improve operation efficiency and effectiveness through modeling and simulation.*