Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

SIMULATION-BASED SCHEDULING OF PARALLEL WIRE-BONDERS WITH LIMITED CLAMP&PADDLES

Daniel Quadt

Infineon Technologies AG
Wernerwerkstr. 2
D-93049 Regensburg, GERMANY

ABSTRACT

We present a scheduling procedure for the wire-bonding operation of a semiconductor assembly facility. The wire-bonding operation typically consists of a large number of unrelated parallel machines and is typically one of the bottlenecks in an assembly facility. The scheduling procedure is able to handle setup times, limited fixtures (clamp&paddles) and non-zero machine ready-times (initial work in progress). It is based on a simulator that generates a schedule and a Simulated Annealing approach to optimize the schedule. Some preliminary results from an implementation in a large assembly facility are given.

1 INTRODUCTION AND PROBLEM DESCRIPTION

The semiconductor manufacturing process is usually divided into wafer fabrication, wafer probe, assembly and final test. We consider the wire-bonding operation, which is part of the assembly process. After slicing the wafer and attaching the individual dies on a leadframe, the wire-bonding operation connects the contacts on the die with the contacts on the leadframe to ensure the electrical path between the two. The wire-bonding operation is typically one of the bottlenecks in an assembly facility. A number of wires has to be soldered per die, and the process requires more time than other assembly operations. To reach the desired output, several wire-bonding machines ('wire-bonders') are used in parallel. A large assembly facility comprises more than 150 parallel wire-bonders. The machines may be divided into machine-groups, and each machine-group is able to process a certain sub-set of products at a specific processing time (unrelated parallel machines). A priority may be assigned to each machine-group and product combination, indicating that it is more favorable to produce a product on a certain machine-group than on others. A setup time is incurred when changing a machine from one product to another. Setup times are typically sequence-independent. Setups should also be avoided because of quality issues and lost product units for machine-calibration.

A number of products ('devices') have to be produced with weekly due dates. The product units are grouped to jobs of varying quantities, which may have different priorities, for example because they belong to different customer orders. 'Jobs' are frequently called 'lots' in the semiconductor industry. Lots are not divisible and lot pre-emption is not allowed.

A fixture ('clamp&paddle') is required to produce a product on a machine. The type of clamp&paddle depends on the product dimensions—i.e., its package—and the machine-group. In some cases, different machine-groups may use the same type of clamp&paddle for a certain product, e.g., when both machine-groups are manufactured by the same vendor. In the environment considered in this paper, production runs 24 hours a day, 7 days per week. Together with the relatively long processing times, which may be longer than 18 hours per lot, this implies that the machines may not be assumed to be empty at the beginning of the planning horizon. In contrast, the scheduling procedure has to take non-zero machine ready-times into account.

One of the objectives of the scheduling procedure is to minimize the number of setups. However, a setup-optimal schedule on the wire-bonding operation may lead to bad schedules on adjacent operations (Quadt and Kuhn 2006), e.g., incurring a high number of setups or long flow times. Therefore, the adjacent operations have to be considered as well and the overall objective is to minimize the total number of setups under a user-given flow-time constraint.

As a summary, the problem at hand is a large scheduling problem with unrelated parallel machines, setup times, limited sub-resources (clamp&paddles) and non-zero machine ready-times (initial work in progress).

The remainder of the paper is organized as follows: Section 2 gives a brief overview of existing literature. Section 3 describes a simulator of the assembly facility that is used to generate a schedule. The simulator considers one of

the weekly due dates at a time. It uses two parameter-sets that determine the schedule. These parameters are modified by a Simulated Annealing procedure in order to optimize the schedule. The algorithm is repeated for all due dates. The Simulated Annealing approach is presented in Section 4. Section 5 covers the extension of the algorithm to non-zero machine ready-times, limited clamp&paddles and intra-week due dates. Some preliminary results from an implementation in a large assembly facility are illustrated in Section 6. The paper closes with a summary and conclusions in Section 7.

2 LITERATURE REVIEW

Parallel machine scheduling problems have been studied by a number of authors, including Lee and Pinedo (1997), Meyr (2002) and Luu et al. (2002).

Some authors consider the special characteristics of a semiconductor 'back-end', which combines the assembly and the final test processes. Yin et al. (2004) present a back-end scheduling heuristic based on two priority rules, one for the prioritization of lots, the other to assign lots to machines. Quadt and Kuhn (2003) present a hierarchical production planning approach for semiconductor back-ends comprising lot-sizing and scheduling. Some authors have focused on the wire-bonding operation because of its importance and complexity: Potoradi et al. (2002) employ a simulation tool to create an initial wire-bonder schedule and improve the schedule with postprocessing steps. A mathematical model and a rule-based heuristic to schedule the wire-bonding operation is presented by Tovia et al. (2004).

Quadt (2005) discusses some experience with a commercially available scheduling tool and gives a brief overview of a back-end scheduling procedure. The core of this procedure is a scheduling algorithm for the wire-bonders. In the current paper, we illustrate this algorithm in more details and extend it to handle limited clamp&paddles.

3 SCHEDULING SIMULATOR

A deterministic simulator is used to generate a schedule on the wire-bonding operation and later for all preceding operations. It considers one of the weekly due dates at a time. Machine down-times are incorporated on an average basis by prolonging the process times. In this section, the simulator is presented without a clamp&paddle limitation. The limitation will be included in Section 5.

The simulator uses a sequence of devices and a target number of wire-bonders per device. These values are generated initially and optimized by a Simulated Annealing approach, which will be explained in Section 4. The devices are considered in the given sequence and all lots of the current device (and the current due date) are assigned before the next product is considered. The following steps are performed for each device:

1. Generation of Wire-Bonder-List
2. Selection of Wire-Bonders
3. Loading of Wire-Bonders

3.1 Generation of Wire-Bonder-List

The first step of the simulator is to generate a list of eligible machines for the current device. The list contains all wire-bonders that are technically able to produce the device, and will be sorted by (1) setup-state, (2) makespan (ascending), and (3) priority of the associated machine-group (higher priority first). 'Setup-state' refers to the last device produced on a machine. If the last device on a machine is the same as the current device, no setup has to be performed. These machines are considered first. All other machines are considered afterwards. It would be possible to add additional categories if setup times or costs were sequence-dependent, as in Yin et al. (2004).

3.2 Selection of Wire-Bonders

Let m' be the target number of wire-bonders for the current device. If m' is larger than the number of machines in the wire-bonder-list, we reduce m' to that number. We tentatively select the first m' machines of the wire-bonder-list.

It may not be possible to produce the demand volume of the current device on these m' machines because of capacity limitations. Therefore, we estimate the available capacity. This can be done by calculating the available production time on each of the m' wire-bonders (weekly due date minus makespan) and multiplying it with the processing speed (e.g., in product units per hour) on the machine. The sum of these values is an estimation of the available capacity in product units of the current device. It is only an estimation because it assumes that the lots can be split, which is not the case.

We have to add more machines if the available capacity does not allow to produce the demand volume of the current device. Therefore, we iteratively take the next wire-bonder from the list and add the additional capacity. We do this until the available capacity is equal or greater than the demand volume. Let m be the new number of selected machines.

3.3 Loading of Wire-Bonders

The simulator iteratively assigns one lot at a time. The lots are considered in order of (1) their priority (highest priority first) and (2) their quantity (largest quantity first). The 'largest quantity' rule is similar to the longest processing time (LPT) rule, because all lots belong to the same product

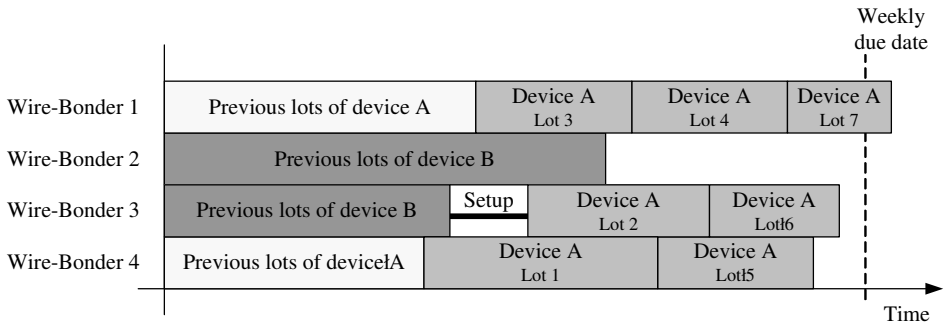

Figure 1: Example of the Wire-Bonder Loading

and thus will be produced with the same processing speed on any given machine. The LPT rule is commonly used for parallel machine scheduling problems. However, the largest quantity rule is easier to apply in our case because the processing times may differ for different machines and the machine-assignment is not yet known.

We iteratively repeat the following sub-steps until all lots of the current device are assigned:

1. Pick the machine with the lowest makespan among the m selected wire-bonders.
2. Load the next lot in the above sequence on the wire-bonder
3. Update the makespan of the wire-bonder

This procedure does not guarantee that all lots meet their due dates because the available capacity has only been estimated in the wire-bonder-selection step. However, this is usually not a problem in practical settings as the wire-bonding operation is not the last operation in an assembly facility. The wire-bonding due dates are typically set artificially, based on a cycle-time offset from the final operation. Hence, the following operations are able to make up for a short delay on the wire-bonders. In cases where even a small delay on the wire-bonders poses a problem, one may employ a safety buffer-time to offset the wire-bonding due dates.

Figure 1 gives an example of the simulator. In the example, the target number of wire-bonders for device A is $m' = 2$. Wire-Bonders 1 and 4 are selected initially because of their setup-state. A third wire-bonder (Wire-Bonder 3) has to be added because of the capacity limitations. The lots are iteratively loaded on the wire-bonder with the lowest makespan, i.e., the first lot is loaded on Wire-Bonder 4, the second on Wire-Bonder 3, the third and fourth on Wire-Bonder 1, and so on.

3.4 Evaluation of Schedule

The main objective on the wire-bonding operation is to minimize the number of setups or the setup time. However, a schedule cannot be evaluated on the wire-bonding operation alone. A schedule with few setups on the wire-bonders may incur a high number of setups on the other operations or a long flow time between the operations (Quadt and Kuhn 2006). Therefore, the simulator is extended to the adjacent operations. In our case, the simulator covers all operations from the material release in the die bank (the first operation of an assembly facility) until the wire-bonding operation. Only the die-attach operation is modelled accurately. The other operations, e.g., wafer-mounting, usually do not constitute a bottleneck and are only modeled as accurately as necessary. For all operations, the simulator uses the chronological sequence of lots as determined by the wire-bonders and loads the lots in this sequence. On the die-attachers, it additionally uses a setup-avoidance strategy to batch lots of the same device on a machine. This strategy allows some local re-sequencing of lots as long as a user-given flow-time between operations is not exceeded.

The overall schedule is evaluated by its total number of setups on the die-attach and the wire-bonding operation. Additional measures may easily be added if necessary. An evaluation-score is computed and handed over to the Simulated Annealing procedure described in the following section.

4 SIMULATED ANNEALING APPROACH

The scheduling simulator centers around a device-sequence, which determines the order in which the devices are scheduled, and a target number of wire-bonders per device. We first describe how these values are initialized and then how they are optimized using a Simulated Annealing approach (Reeves 1995, Michalewicz and Fogel 2004). Both steps require a lower and an upper bound for the target number of wire-bonders.

A trivial lower bound for the target number of wire-bonders is given by a single machine, because we have to use at least one wire-bonder for each device. More limiting lower bounds may be used on a case-by-case basis. An upper bound is given by the number of lots to be produced for the device. In our case, another upper bound can be

derived from the processing speed on the wire-bonders in relation to the processing speed on the die-attach operation: Because of external requirements, at most one 'die-attacher' per device shall be used. At the same time, the output on the wire-bonders shall not exceed the output on the die-attachers. Thus, the number of wire-bonders is upper-bounded by the processing speed ratio of these two operations. We use the processing speed of the highest priority machines for this calculation.

4.1 Initialization

The basic idea is to set the initial device-sequence and target number of wire-bonders in a way that the resulting schedule can as much as possible make use of the existing setup-states. Therefore, we take a snapshot of the machine configuration prior to the scheduling procedure and count the number of wire-bonders set up for each device. The higher this number, the earlier the device will appear in the initial device-sequence. The target number of wire-bonders is also initialized with this value. This shall ensure that the right devices are scheduled early and that the machines may continue production without a setup to another device.

No machine is currently set up for the remaining devices. Here, the idea is to alternate high- and low-volume devices in order to evenly spread out the low-volume devices over the planning horizon. We accomplish this by sorting the devices by their total production time for all lots on the highest priority machine. We then alternately pick the device with the longest and the shortest production time. The target number of wire-bonders for these devices is initialized with the mean of its upper and lower bound.

4.2 Optimization with Simulated Annealing

A Simulated Annealing procedure is used to optimize the device-sequence and the target number of machines. The procedure iteratively either changes the device-sequence or the target number of wire-bonders. The decision which of the two is modified is drawn at random.

If the number of wire-bonders is modified, we pick one of the devices at random and select a new target number between the lower and upper bound, again at random. If the device-sequence is altered, we choose a 'from'- and a 'to'-position at random and move the device from the 'from'-position to the 'to'-position.

The new parameter-set is evaluated with the simulator and is accepted if its associated schedule is better than the previous schedule. If the schedule is worse, the parameter-set is still accepted with a certain probability. The Simulated Annealing approach prescribes that the probability of acceptance decreases ('anneals') the more iterations are performed. In the beginning, the probability is relatively large to overcome local optima, while at the end, it is rather small to deeply explore the most promising area of the search space. When a new parameter-set is rejected, the previous parameter-set is recovered and used as a basis for another iteration of the Simulated Annealing procedure. If a new parameter-set is accepted, it becomes the new basis for the next iteration.

The algorithm terminates after it has reached a specified number of iterations and reports the best schedule found. It is repeated for all weekly due dates.

5 FEATURES AND EXTENSIONS OF THE ALGORITHM

In this section we describe how the algorithm handles non-zero machine ready-times and how it can be extended to consider limited clamp&paddles and intra-week due dates.

5.1 Non-Zero Machine Ready-Times

Non-zero machine ready-times imply that a machine cannot immediately be loaded at the beginning of the planning horizon. One reason may be that previously assigned lots have to be completed before newly loaded lots can be started, i.e., there is initial work in progress (WIP) on some or all machines.

In our case, the machine ready-times are estimated by calculating the remaining production time for lots that are already running on a machine or that have been issued to a machine. The scheduling procedure will not modify the assignments of such lots, and thus the first lot to be loaded by the scheduling procedure can start after these lots are completed.

We initialize the machine makespan with the expected machine ready-time to incorporate this information in the scheduling procedure. Once this is done, non-zero machine ready-times are handled inherently by the algorithm, because it uses the makespan to sort the wire-bonder-list and in the machine-loading step.

5.2 Limited Clamp&Paddle Availability

The algorithm can be adjusted to consider a limited clamp&paddle availability. There are typically a number of clamp&paddles of the same type. A clamp&paddle limitation implies that only a limited number of machines can be used in parallel. The required clamp&paddle for a device depends on the machine-group. Hence, the limitation may restrict the number of machines of a certain machine-group, while other machine-groups may still have remaining clamp&paddles for the device. Further, some devices may use the same clamp&paddle on a certain machine-group and some machine-groups may use the same clamp&paddle for a certain device. The former is for example the case if the devices are of the same product family ('package'), i.e., they

have the same shape. The latter is sometimes the case when the machine-groups are manufactured by the same vendor.

We adjust the machine-selection step to include the clamp&paddle limitation. The machine-selection step has to ensure that the selected machines may be used in parallel without exceeding the available clamp&paddles. This can be done by adding a preliminary step, which scans the wire-bonder-list in the given sequence and decides whether to keep or to delete a machine from the list. A machine will be kept if it currently uses the correct clamp&paddle needed for the device—i.e., if the correct clamp&paddle is already mounted on the machine. A machine will also be kept if the required clamp&paddle is still available, i.e., not all of the required clamp&paddles are in use. All other wire-bonders are deleted from the list. We keep track of the number of additionally required clamp&paddles of each type and delete all machines that would exceed the limitation. Afterwards, the machine-selection step can continue with the selection of the first m' machines of the list as described in Section 3.2.

In addition, the clamp&paddle configuration has to be initialized before running the scheduling procedure and the machine-loading step has to keep track of the clamp&paddle assignment to machines.

5.3 Intra-Week Due Dates

Another extension of the algorithm considers intra-week due dates. Intra-week due dates may be due to various reasons, e.g., because of urgent customer orders.

Intra-week due dates can be incorporated in the machine-selection step. A simple approach is to consider device-specific due dates for the calculation of the available capacity on the selected machines. As a result, an earlier intra-week due date may lead to a higher number of selected wire-bonders.

A limitation of this approach is that all lots of a device have to have the same (intra-week) due date. To overcome this, the device could be split into several products—one for each due date—which would be considered separately.

6 PRELIMINARY RESULTS

The procedure is currently used to schedule a large assembly facility on a daily basis. Some major performance indicators are shown in Table 1. The first column shows the performance of the line previous to the implementation of the scheduling system. At that time, the line was planned with a manual planning system that involved lengthy and tedious spreadsheet calculations. The second column shows the results of the scheduling algorithm as calculated by the simulator, and the third column the actual performance of the line when using the schedule.

Table 1: Results of the Scheduling Procedure

	Line with manual planning	Simulated schedule	Line with schedule
Setups	18%	17%	26%
WIP	100%	117%	144%

The setup-figures show the daily number of setups on the wire-bonding operation in relation to the total number of wire-bonders. Thus, a value of 18% implies that 18% of the wire-bonders are set up per day. The work in process (WIP) has been scaled to the line performance with manual planning. The figure includes all lots that have been released to the line, until a lot completes the wire-bonding operation. We use the WIP as a substitute for flow time (Little's Law). All figures are averages.

The results of the simulation are close to the line performance with manual planning. The higher WIP is mainly due to relatively long safety buffer-times, which have been added to facilitate the schedule adherence in the line. We plan to reduce these buffers in the future.

The line performance when using the schedule is worse than the results of the simulation. This is mainly due to the relatively poor performance for work in process (WIP) lots, i.e., all lots that have already been released to the line when the scheduling procedure is invoked. WIP-lots are not rescheduled with the above procedure, but are assigned in the same sequence and on the same machine as in the previous schedule. This is motivated by a number of reasons, e.g., to reduce schedule nervousness and to align the previous schedule with the new one. However, the line is usually not able to follow the schedule on some machines. This may be due to various reasons, e.g., because a preceding operation does not deliver the required device on time or because the wire-bonder itself encounters a technical problem. In cases of poor schedule adherence, the number of setups in the line is much higher than necessary. Thus, the next step is to extent the scheduling system so that it includes and reschedules WIP-lots.

7 SUMMARY AND CONCLUSIONS

We have presented a scheduling procedure for the wire-bonding operation of a semiconductor assembly facility. The procedure consists of a scheduling simulator for the wire-bonders and the preceding operations. A Simulated Annealing approach is used to optimize the schedule. The procedure is used on a daily basis to create a schedule in a large assembly facility. Initial results show that the procedure leads to acceptable results, but is currently not superior to the manual planning approach employed previously. A number of improvements potentials have been identified, foremost

the re-optimization of WIP-lots. These will be included in the next project phase.

ACKNOWLEDGMENTS

The author would like to thank Mr. Chua Koon Min and the whole team of Infineon MAL LP for their continuous support as well as Production Modeling Corporation, USA, for many fruitful discussions.

REFERENCES

Lee, Y. H., and M. Pinedo. 1997. Scheduling jobs on parallel machines with sequence-dependent setup times. *European Journal of Operational Research* 100:464–474.

Luu, D. T., E. L. Bohez, and A. Techanitisawad. 2002. A hybrid genetic algorithm for the batch sequencing problem on identical parallel machines. *Production Planning & Control* 13 (3): 243–252.

Meyr, H. 2002. Simultaneous lotsizing and scheduling on parallel machines. *European Journal of Operational Research* 139 (2): 277–292.

Michalewicz, Z., and D. B. Fogel. 2004. *How to solve it: Modern heuristics*. 2nd ed. Berlin: Springer.

Potoradi, J., O. S. Boon, S. J. Mason, J. W. Fowler, and E. M. Pfund. 2002. Using simulation-based scheduling to maximize demand fulfillment in a semiconductor assembly facility. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 1857–1861.

Quadt, D. 2005. Scheduling a semiconductor back-end: From theory to practice. In *Proceedings of the 3rd International Conference on Modeling and Analysis of Semiconductor Manufacturing*, ed. P. Lendermann, J. Fowler, and A. K. Gupta, 194–200: Singapore Institure of Manufacturing Technology (SIMTech), ISBN 981-05-4194-5, Singapore.

Quadt, D., and H. Kuhn. 2003. Production planning in semiconductor assembly. In *Proceedings of the Fourth Aegean International Conference on Analysis of Manufacturing Systems*, ed. C. T. Papadopoulos, 181–189: University of the Aegean, Samos Island, Greece.

Quadt, D., and H. Kuhn. 2006. Batch scheduling of jobs with identical process times on flexible flow lines. *International Journal of Production Economics*. Forthcoming.

Reeves, C. R. 1995. *Modern heuristic techniques for combinatorial problems*. London: McGraw-Hill.

Tovia, F., S. J. Mason, and B. Ramasami. 2004. A scheduling heuristic for maximizing wirebonder throughput. *IEEE Transactions on Electronics Packaging Manufacturing* 27 (2): 145–150.

Yin, X.-F., T.-J. Chua, F.-Y. Wang, M.-W. Liu, T.-X. Cai, W.-J. Yan, C.-S. Chong, J.-P. Zhu, and M.-Y. Lam. 2004. A rule-based heuristic finite capacity scheduling system for semiconductor backend assembly. *International Journal of Computer Integrated Manufacturing* 17 (8): 733–749.

AUTHOR BIOGRAPHY

DANIEL QUADT holds a Ph.D. in Production and Logistics Management. He received the Best Dissertation Award from the German Operations Research Society (GOR) and was a finalist for the INFORMS George B. Dantzig Best Dissertation Award. He joined Infineon Technologies AG in 2004 as a project leader in the area of Operations Research and Factory Logistics. One of his projects is to develop a scheduling system for a backend facility. His e-mail address is <daniel.quadt@infineon.com>.