Chapter 6
Order Release Approaches

In this chapter, we provide an overview of order release approaches for semiconductor manufacturing. Order release is between production planning and scheduling in the PPC hierarchy. The fundamental concepts of push-based order release and pull-based order release are presented, along with a comparison of these two key methods and their implementation in a variety of production environments. Next, we present two seminal wafer fab-specific order release approaches, namely starvation avoidance by Glassey and Resende [100] and workload regulation by Wein [318].

After discussing subsequent order release methods that followed these first two key approaches, we demonstrate the interaction between order release and scheduling. A DSBH-type heuristic (cf. Sect. 5.4.6) is used to compare three different order release methods from the literature under a variety of wafer fab operating conditions.

Next, we present the findings of a large-scale order release study that was conducted at a large, global semiconductor manufacturer in order to answer important questions relating to both the timing and quantity of order release into their wafer fab.

Finally, we present a MIP model for optimizing order release into wafer fabs that seeks to improve the utilization of machines in the constraint machine group. This optimization model determines both the timing and quantity of order releases into a wafer fab on a weekly basis.

6.1 Push Versus Pull Approaches

In this section, we start by discussing push and pull approaches for order release. Moreover, we compare these two important classes of order release schemes.

L. Mönch et al., *Production Planning and Control for Semiconductor Wafer Fabrication Facilities*, Operations Research/Computer Science Interfaces Series 52, DOI 10.1007/978-1-4614-4472-5_6,

6.1.1 Push Approaches for Order Release

From the early years of material requirements planning (MRP) in the 1960s through the early to mid-1980s, many followers of MRP/MRP II methods simply released all orders for which material and planning were available according to the calculated order release date (see Wight [322]). This push approach amounted to material planners starting into their respective systems what they hoped to get out without any recognition or understanding of the BS capacity and/or potential BS congestion. Such push approaches quite often led to large WIP levels and high CT values in the presence of capacity constraints. It is not surprising therefore that dispatching rules were the primary means of production control and the subject of much research in the 1970s and 1980s as there was much WIP to control and dispatch.

Faced with excessive amounts of WIP, many companies responded by limiting the daily release of material to some fixed levels that were based on production goals. However, this release-limiting approach did not properly comprehend BS capacity and congestion, i.e., it was still a push philosophy. Typically, the release-limiting policies only marginally improved WIP levels as compared to the earlier order release methods, as companies initially overestimated their own production capacity. However, as these same companies worked to better understand their own capacity and potential congestion issues, they achieved greater WIP management success because they gradually began to adjust order release rates to appropriate levels.

6.1.2 Pull Approaches for Order Release

A new collection of order release strategies based on the concept of pulling work into a BS that was ready for it versus pushing work into the same BS, regardless of its ability to accept the work, began to emerge in the early 1980s due to the following:

1. Internal recognition that current release practices lacked intelligence
2. The appearance of Japanese management concepts like just-in-time (JIT)
3. The development of bottleneck-based methods like the optimized production technique (OPT) (cf. Jacobs [125])

Initially, some industries attempted to adopt JIT philosophies to reduce WIP by implementing Kanban cards—a signaling mechanism between different points in the manufacturing process that visually indicates when a new order can be released into the BS or when an existing WIP job can be moved to a subsequent/downstream process step—or some other limited or fixed WIP approach.

Concurrent with the growth of interest in JIT and Kanban strategies, bottleneck resource scheduling philosophies gained prominence. These methodologies revolved around first identifying bottleneck machines and/or processes and then using the identified bottlenecks as central points of focus for production control strategies. Both OPT and Goldratt's theory

of constraints [104] are representative. The drum-buffer-rope approach to production control by Goldratt and Fox [105] suggests that:

1. The slowest paced (bottleneck) process provides the pace of a system or production line (drum).
2. The bottleneck should be tied to the entry points of the system (ropes).
3. The bottleneck should be always provided with a time-phased inventory of work (buffer) that guards it from being idle.

In the 1990s and early 2000s, considerable interest grew in the CONWIP methodology (see Spearman et al. [290]) for limiting the inventory levels of a manufacturing system. Although conceptually similar to early input/output control ideas from the 1970s, CONWIP focuses on WIP control rather than TP control and can also be viewed as a generalization of Kanban. CONWIP is a simple, robust control philosophy based on a fundamental understanding of the relationship between the WIP in a BS and its TP. This relationship is visually captured in a production system characteristic curve that can be either developed through a simulation study or approximated analytically. An example for a characteristic curve is shown in Fig. 6.1.

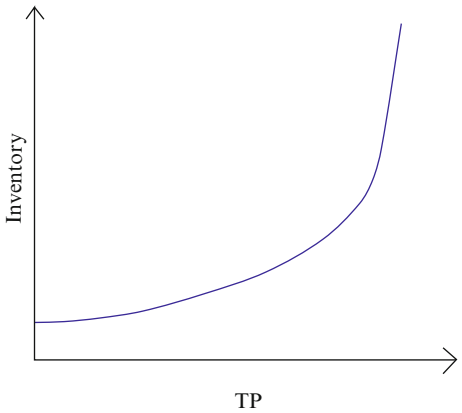

Figure 6.1: Example of manufacturing system characteristic curve

Once the characteristic curve is developed, one selects a target WIP level for a desired TP rate of the BS. Then, efforts are made to keep WIP at or below this target level in the BS and measure the resulting TP to validate the previously developed characteristic curve. WIP targets must be adjusted if the TP of the BS does not meet the desired goals. Clearly, the specification of the order release rate into the wafer fab directly impacts performance according to Little's law (cf. Eq. (3.21)). The combination of setting the value of the TP rate λ and specifying a desired target inventory level, represented by the WIP, leads to an effective estimate of expected CT via this

important governing law of factory physics. Spearman et al. [289] describe the development of a CONWIP-based hierarchical production planning and control system for a circuit board manufacturer's facility.

6.1.3 Comparing Push Versus Pull Approaches

As described in Sects. 6.1.1 and 6.1.2, push order release approaches are motivated by a company wanting to produce some desired quantity of goods, while pull-based methods are based on the knowledge of what actually can possibly be produced in light of existing capacity constraints and/or congestion issues. The superiority of pull systems is well documented in the literature for a variety of production environments employing dispatching policies under varying levels of due date tightness [256, 262].

Unfortunately, research findings do not always make their way into practice, as the presumed need for and protection of large amounts of WIP have not been easily overcome in some industries and companies, even when these same companies purport to follow JIT philosophies.

6.2 Tailored Approaches for Wafer Fabs

Order release and dispatching have received a fair amount of attention from both semiconductor manufacturing researchers and practitioners alike. One of the earliest papers that focused on semiconductor manufacturing workload control is by Dayhoff and Atherton [61]. Although they focused solely on dispatching, their concept of signature analysis is embedded in much of the semiconductor manufacturing-focused research that followed, which concerned the impact of workload control in wafer fabs. Two seminal works of note, the starvation avoidance method of Glassey and Resende [100] and Wein's workload regulation technique [318], were the primary catalysts that launched a flurry of order release methods for wafer fabs. We discuss these two important papers in the next two subsections. Then we focus on more recent order release methods.

6.2.1 Starvation Avoidance

The starvation avoidance (SA) method of Glassey and Resende [100] focuses on a single bottleneck work center and calculates a virtual inventory measured over a lead time in order to regulate order release. This virtual inventory comprises all work in the BS that potentially can reach the bottleneck within the prespecified lead time. The lead time is the time required for jobs to arrive to the bottleneck the first time after order release in a single-product environment. In a multiproduct system, the worst case time among the products to arrive to the bottleneck is used.

An important challenge in the SA method is tracking jobs that are recirculating in the system as is the case in semiconductor manufacturing.

One needs to detect when each recirculating job will move again/next within the lead time window. The virtual inventory calculation also includes a mechanism to account for failed machines at the bottleneck machine group. A target value for the virtual inventory level is determined using inventory concepts and safety stock considerations. The primary idea of SA is to determine an acceptable level of risk of the bottleneck machine group running out of work, i.e., starving.

Glassey and Resende [100] compare SA to four order release methods: random starts, uniform starts, a simple input/output approach based on the idea of constant WIP, and a simplified version of the workload regulation method of Wein [318]. The workload regulation method is described in more detail in Sect. 6.2.2. Glassey and Resende [100] also present a hybrid dispatching method to boost the efficacy of SA that was subsequently enhanced and expanded by Leachman et al. [156]. In addition to the hybrid dispatching method, they use simple FIFO and SRPT dispatching (cf. Sect. 4.2.1 for the corresponding priority indices) in their study.

Unlike Wein [318], the SA study concentrated more on the order release process rather than investigating a large number of dispatching rules. In fact, Glassey and Resende [100] suggest that dispatching decisions seem to have little impact when uniform job releases are used. However, the comparisons of SA both to the simple input/output approach based on constant WIP and to Wein's method for a simple wafer fab with 12-step process flows contain no mention of statistical significance.

One of the issues practitioners can have with the SA method is that it is both conceptually and computationally more complex than other available approaches and it requires global information about the wafer fab inventories. A companion paper by Lozinski and Glassey [168] provides details on performing the necessary calculations and implementing the approach. The superiority of SA over other simpler methods has never been adequately verified. In fact, at least two subsequent attempts [54, 99] suggest the opposite conclusion. However, the concept has a strong intuitive appeal and is inherent in much of the subsequent research and development of workload control and production control software for semiconductor manufacturing.

6.2.2 Workload Regulation

The second seminal paper in semiconductor manufacturing workload control was the Wein [318] work that introduced the concept of workload regulation (WR). The workload-regulating input method for order release is conceptually similar to other bottleneck methods that compute the load destined for the bottleneck machine group and then strive to maintain this target level of loading. The WR method computes machine group load in terms of the number of hours of work in front of the bottleneck machine group

rather than expressing the workload as a number of jobs or wafers. Wein [318] introduces several modified versions of dispatching rules and introduces a new dispatching method called workload balancing.

The bottleneck-oriented order release methodology that Wein created is not that different from the OPT concept. Similarly, expressing bottleneck workload in terms of hours, rather than jobs or items, had been used previously as an input/output control variable. The primary theoretical contributions of the seminal Wein [318] paper pertain to dispatching based on workload balancing. Of potentially even more importance, however, are the following two points:

1. The WR approach taken by Wein focuses on the semiconductor manufacturing environment.
2. The examination of four order release methods, i.e., random starts, uniform starts, a version of input/output control, and a bottleneck approach, combined with a number of dispatching rules is based on a rigorous design of experiments using a significant testbed model.

The simulation study of Wein [318] compares a number of order release strategies using three variations of a realistic semiconductor wafer fab model that was developed using actual wafer fab data. Similar to Glassey and Resende [100], Wein concludes that order release with a 30–40 % change in desired performance is more important than dispatching that leads to less than 10 % change. However, Wein's statistical results reveal an important interaction between order release and dispatching decisions. Finally, both the SA and the WR studies support the conclusion that pull-based order release strategies are preferable to push-based methods.

Most major semiconductor manufacturers are aware of Wein's WR work, have embraced the WR concept philosophically, and have developed control systems around the method. This is most likely due to the fact that the information associated with and the computational requirements of implementing this approach are modest as compared to SA. While most of the information is determined from the jobs being released, one also needs to estimate the relationship between the desired workload target and BS TP. Such an estimate is often determined using a wafer fab simulation model and/or a queueing network approximation of the wafer fab.

It is interesting to note that there is no mention in either Glassey and Resende [100] or Wein [318] of how dispatching decisions at batch processes, such as diffusion ovens and wet sinks in etch, are treated. The dispatching methods cited in these studies do not seem to apply, and given the stated interaction between order release and dispatching, it follows that a similarly strong interaction may exist with batching disciplines as well.

6.2.3 Subsequent Order Release Methods

While both SA and WR were developed for single bottleneck systems, both have been extended to multiple bottleneck environments by a number of authors [17, 54, 99, 156, 166, 167]. When we consider the dispatching aspect of flow control, it follows that as order release approaches become more effective, dispatching decisions will have a diminishing effect on BS performance. However, dispatching can still be a means to assist the flow control process by smoothing input flows to bottlenecks to prevent starvation and clogging. For example, dispatching strategies may prioritize jobs for processing on a given machine that are required at a key downstream step while deprioritizing jobs for those whose downstream steps already have sufficient amounts of WIP in front of key tools. The work of Wein [318] and Leachman et al. [156] are examples of this trend.

Miller [184] describes an IBM wafer fab simulation model and its application to the study of flow control policies for reducing CT. Through effective flow control, WIP was reduced 30 % in concert with a reduction in CT of 25 %. This is even more impressive when one considers that at the same time, TP modestly increased. These reductions were achieved using a very simple closed-loop order release method similar to CONWIP. Miller [184] also concludes that when queues are reduced by better order release practices, dispatching becomes less important. A simulation study of a packaging line at IBM Bromont by Chandra and Gupta [44] uses an order release strategy similar to WR in that the release quantities of different products into the packaging line are determined so that total manufacturing lead time is minimized, subject to satisfying product demands. The release quantities are considered for the bottleneck, which happened to be the last batch station of the line.

A case study by Martin-Vega et al. [170] provides an interesting example of applying the general JIT philosophy to a photolithography area in a wafer fab. Although a mention is given to Kanban, the authors achieve WIP reduction by physically limiting and redesigning buffer spaces and by prioritizing specific operations. Leachman [154] is a suggested reference for readers desiring a discussion of production planning and scheduling practices in the semiconductor industry as well as for additional discussion of workload control implementations.

It should be noted, however, that exceptions to the viewpoint that order release, when done well, is more important than dispatching do exist. Lu et al. [169] introduce fluctuation smoothing policy-type dispatching rules (cf. Sect. 4.2.1). Their dispatching policies compare favorably with WR in Wein's same wafer fab setting in that they produce more than 10 % reductions in both the ACT and Var(CT). Given the variety of wafer fab environments, order release strategies, and dispatching approaches, the only thing that is clear is that no one specific approach or method exists that is best for all semiconductor manufacturing environments or conditions.

In order to overcome some of the performance problems associated with CONWIP and workload regulation-based rules during product mix changes, Rose [264] introduces the constant load rule, CONLOAD. It is claimed that while pull-based approaches are capable of maintaining appropriate inventory levels in a wafer fab based on the current BS status, they unfortunately can suffer from not comprehending the wafer fab's current and/or desired product mix. By taking into account the associated additional load that is placed on a single machine or a group of machines due to a pending order release decision, more informed order release decisions can be made based on a desired bottleneck machine group loading threshold. This threshold is calculated as the product of the desired bottleneck machine group's utilization and the number of machines in the bottleneck machine group. A simulation-based study concludes that CONLOAD outperforms CONWIP, a workload regulation-based approach called CONWORK, and a simple push methodology in terms of producing and maintaining a desired level of bottleneck machine group utilization while providing a smooth evolution of fab WIP over time. An additional study by Rose [266] reveals that CONWIP-based order release methods can help to reduce the variability in both WIP and CT. However, it is confirmed that this reduced variability may come at the price of increased mean values of both WIP and CT.

Later, Bahaji and Kuhl [16] present multiobjective composite dispatching rules for both an application-specific integrated circuit (ASIC) wafer fab and a low-mix, high-volume wafer fab. The proposed composite dispatching rules utilize a combination of values based on current BS and individual job status, such as the processing time of a job, the job's arrival time at the current process step, the amount of work present in queue at the job's next process step, and a job's accumulated CT as compared to its theoretical processing time, i.e., the job's current flow factor. The authors conduct a rigorous statistical analysis of both wafer fab environments using an AutoSched AP simulation model for five different performance measures of interest based on MASM lab testbed dataset 5 (cf. Fowler and Robinson [83] for a description of these models). After analyzing four proposed approaches and ten competing methods from the literature, Bahaji and Kuhl [16] find that their composite dispatching approaches outperform both fixed-interval push order release and a CONWIP policy in terms of producing superior ACT, the lowest amount of variability in CT, and meeting required job due dates.

Finally, Qi et al. [250] examine the impact of production control methodologies and other BS factors on both the ACT and VAR(CT), as well as average lateness, WIP, and wafer fab output, at a Chartered Semiconductor wafer fab. A full factorial design of experiments that examines three order release methodologies in concert with three dispatching rules and three greedy batching policies reveals that the proposed WIPLOAD control (WIPLCtrl) job release methodology nicely balances fab performance across all of the performance measures of interest. In addition, a Markov process-based analysis of the behavior of WIPLCtrl using a model of a transfer line

system is presented in [251]. After defining WIPLOAD as the sum of the remaining processing times of all jobs in the BS, Qi et al. [250] introduce a control policy that only releases new jobs into the wafer fab when some desired reference WIPLOAD level is not being met. This reference level may be prescribed by the wafer fab's manufacturing manager according to some desired level of TP. The AutoSched AP simulation model of the Chartered Semiconductor wafer fab contained many realistic BS factors such as machine breakdowns. Thorough experimentation conducted suggests the efficacy of their WIPLCtrl approach for a variety of wafer fab output levels.

6.3 Interaction of Order Release and Scheduling

In this section, we start by discussing the scheduling heuristic and the order release approaches used. Then we describe the experimental setting and present computational results. Finally, we discuss some conclusions from the interaction study.

6.3.1 Scheduling Approach and Order Release Schemes

Given this background on the evolution and importance of order release in wafer fabs, we now investigate the influence of three order release strategies on the performance of a popular job shop scheduling heuristic. Order release schemes and scheduling are usually treated independently. There is only little known on the interaction of order release schemes and sophisticated scheduling approaches. The interaction of a scheduling approach and an order release scheme is discussed in a sequence-dependent setup situation by Ashby and Uzsoy [11].

As described in Sect. 5.4, the SBH is a decomposition-based heuristic that solves the job shop scheduling problem iteratively by solving a sequence of machine scheduling subproblems and then determines the overall shop schedule via a disjunctive graph. Mason et al. [172] modify the SBH for complex job shops as exemplified by semiconductor wafer fabs. Batch-processing machines and reentrant process flows are modeled by adding additional arcs to the disjunctive graph. In turn, unfortunately, the size of the graph increases significantly with a large scheduling horizon $h := \tau_\Delta + \tau_{ah}$, and as a result, runtime performance can be poor and software application memory requirements can be large.

To effectively investigate the interaction of order release and scheduling, we consider the two-layer, distributed approach for wafer fab scheduling DSBH that is described in Sect. 5.4.6. We use an order pool to collect jobs released for production prior to their release to the BS as a new ingredient. The overall situation is shown in Fig. 6.2.

Within DSBH, the SBH is applied separately for each work area due to the decoupling effect of the top ICA layer. Clearly, the performance of the DSBH

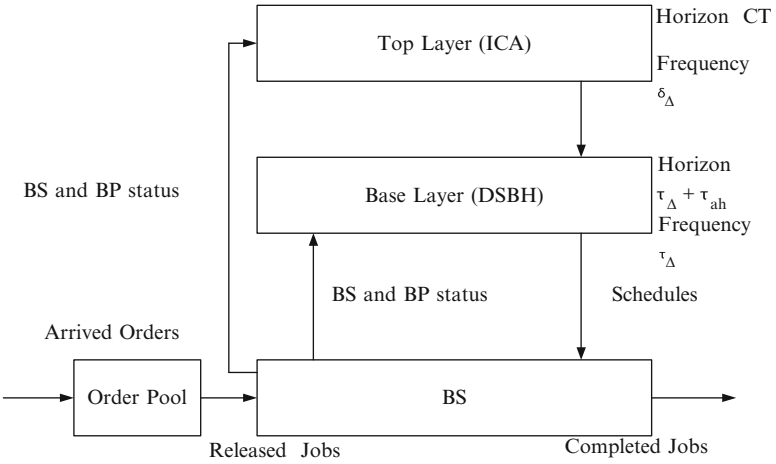

Figure 6.2: Interaction of DSBH and order release

can be improved if a schedule for one or more work areas already exists. This leads to the IDSBH scheme as described in Sect. 5.4.6. Given this background, we use the DSBH approach to investigate the interaction between the push, CONWIP, and CONLOAD order release strategies and scheduling.

The push strategy releases jobs into the BS as required by customer due dates. Only simple capacity considerations are taken into account during job release, and the release time r_j for job j is calculated by a simple backward calculation based on some desired flow factor $\mathrm{FF} \geq 1$:

$$r_j := d_j - \mathrm{FF} \sum_{i=1}^{n_j} p_{ji}. \tag{6.1}$$

The CONWIP order release strategy requires a characteristic curve of the wafer fab that provides the relationship between WIP and the production rate of the wafer fab, i.e., number of jobs produced/output per day. Once the WIP level corresponding to the desired production rate is determined, this amount of WIP is set as the CONWIP quantity. Then, a new job is released into the fab each time a job completes its processing such that the target WIP level is achieved. Finally, the CONLOAD strategy also requires the use of a characteristic curve. The workload of the wafer fab is measured as the sum of the processing times at each remaining process step for all released jobs. We obtain

$$\mathrm{WL} := \frac{1}{n\ \mathrm{CT}} \sum_{j=1}^{n} \sum_{i=k_j+1}^{n_j} p_{ji}, \tag{6.2}$$

where we assume that job j has completed all of its processing through process step k_j and that target cycle time is given by CT. In addition, the total number of jobs released into the fab that have not yet completed their processing is denoted as n. It is easy to see that Eq. (6.2) reveals that $\mathrm{WL} \in [0,1]$ when $\mathrm{CT} := \mathrm{FF}\sum_{i=1}^{n_j} p_{ji}$ is used.

6.3.2 Experimental Setting and Computational Results

We use the simulation framework described in Sect. 3.3.2 to analyze the interaction of order release and scheduling for a simulation model that is derived from the MiniFab model (cf. the description in Fig. 3.4). The new model contains three work areas. Each of them contains the machinery of the MiniFab model. The process flows are organized into two mask layers.

We focus on different performance measures of interest. The ACT and AWT measures are considered. In addition, we use TP for the wafer fab within the simulation horizon T that is defined in this situation as follows:

$$TP := |\{j|0 \le r_j, C_j < T\}|. \tag{6.3}$$

We also consider the average WIP in jobs during the simulation horizon as a performance measure. In addition to the three order release strategies described, we also vary the loading of the BS, the distribution of job weights, and the desired wafer fab FF used in setting job due dates. We compare the performance of the DSBH to FIFO dispatching using two different weight distributions for jobs in terms of the probability that a given job will have a specific weight value. We have

$$D_1 := \begin{cases} w_j = 1,\ p_1 = 0.5 \\ w_j = 5,\ p_2 = 0.35 \\ w_j = 10,\ p_3 = 0.15 \end{cases} \tag{6.4}$$

and

$$D_2 := \begin{cases} w_j = 1,\ p_1 = 0.5 \\ w_j = 2,\ p_2 = 0.45. \\ w_j = 10,\ p_3 = 0.05 \end{cases} \tag{6.5}$$

The two weight distributions differ in that D_1 has a small number of jobs that have a high weight and a large number of jobs that have a medium weight as compared to D_2. Distribution D_2 represents a wafer fab in which a very small portion of the jobs have a high weight and the remaining jobs have a small weight.

Figure 6.3 shows the relationship between WIP and wafer fab TP for our simulation model of interest. From our initial simulation runs, we see that the DSBH leads to a higher WIP level for a fixed TP value than pure FIFO dispatching does. Based on the relationships displayed in Fig. 6.3, we define specific TP levels of interest. For example, we use $\lambda_1 = 14$ jobs per day in our

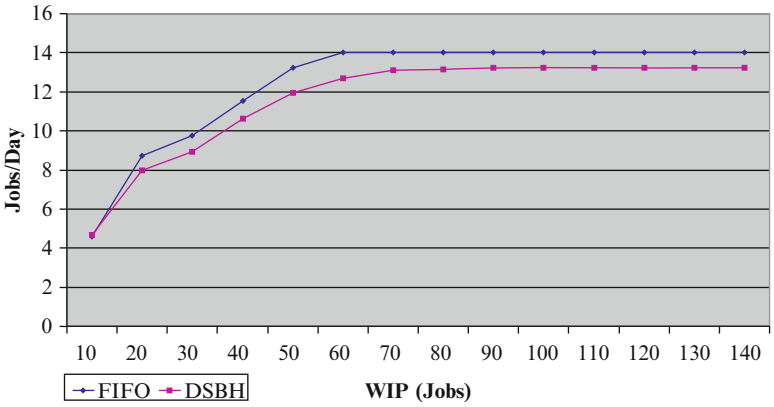

Figure 6.3: Simulated characteristic curve

model as the TP rate obtained by a WIP of 80 jobs. We refer to this situation as high wafer fab loading. Furthermore, a WIP of 60 jobs leads to a TP rate of 13.3 jobs per day in a moderately loaded wafer fab, while the low load case, i.e., 40 jobs in WIP, leads to a TP rate of 11.5 jobs per day. Additionally, we obtain a very highly loaded BS by increasing the job release rate that leads to a highly loaded BS. We use $\mathrm{WL} = 0.76$ for the highly loaded case and $\mathrm{WL} = 0.78$ for the very highly loaded case for the CONLOAD strategy. In this situation, we simply set the target CT as the raw processing time, i.e., the sum of the processing time of all process steps of a job. Finally, we use these desired wafer fab TP rates as the job release rates for the push order release strategy.

For each performance measure of interest, we compute the performance ratio of the DSBH-obtained result to the result derived by pure FIFO dispatching. In this way, any performance ratio greater (less) than one denotes superior DSBH performance for objectives that we wish to maximize (minimize). Of the four performance measures of interest, the only one that we wish to maximize is TP. Otherwise, we seek to minimize AWT, ACT, and WIP.

In all experiments, we simulate 180 days of wafer fab operations once an appropriate amount of warm-up time has elapsed to initialize the wafer fab. We do not consider any machine failures in our experimentation and employ a scheduling time horizon of $h = 2$ and $\tau_{ah} = 0\,\mathrm{h}$. Tables 6.1 and 6.2 present the results for the model comparison of DSBH scheduling and FIFO dispatching under all three order release strategies for the high and very high load cases.

We use P, CW, and CL for abbreviation for the push, CONWIP, and CONLOAD order release schemes, respectively. In the case of a highly loaded wafer fab, the FIFO dispatched system is stable, while the DSBH results suggest increasing WIP levels, which consequently produce large CT values.

Table 6.1: Computational results for AWT and ACT

			AWT			ACT		
Load	Weight	FF	P	CW	CL	P	CW	CL
High	D_1	1.3	0.98	0.65	0.70	1.56	1.12	1.18
		1.5	1.11	0.59	0.57	1.61	1.14	1.10
		1.7	1.82	0.62	0.43	1.82	1.15	1.04
	D_2	1.3	1.00	0.83	1.02	1.51	1.11	1.21
		1.5	1.61	1.41	0.87	1.47	1.11	1.11
		1.7	2.35	0.97	0.71	1.52	1.12	1.04
Very high	D_1	1.3	0.39	0.44	0.62	1.06	1.02	1.19
		1.5	0.33	0.37	0.48	1.06	1.02	1.12
		1.7	0.35	0.32	0.46	1.14	1.02	1.12
	D_2	1.3	0.58	0.60	0.85	1.04	1.01	1.18
		1.5	0.58	0.58	0.70	1.06	1.02	1.08
		1.7	0.64	0.54	0.62	1.12	1.01	1.04

Table 6.2: Computational results for TP and WIP

			TP			WIP		
Load	Weight	FF	P	CW	CL	P	CW	CL
High	D_1	1.3	0.98	0.97	0.96	1.60	1.06	1.07
		1.5	0.97	0.96	0.95	1.81	1.06	0.93
		1.7	0.96	0.95	0.94	2.08	1.08	0.87
	D_2	1.3	0.98	0.98	0.98	1.50	1.04	1.14
		1.5	0.98	0.98	0.97	1.44	1.01	1.05
		1.7	0.98	0.97	0.96	1.63	1.03	1.07
Very high	D_1	1.3	0.98	0.98	0.97	1.14	0.99	1.12
		1.5	0.98	0.96	0.96	0.99	0.98	1.14
		1.7	0.96	0.97	0.95	1.07	0.97	0.85
	D_2	1.3	0.98	0.99	0.97	0.93	0.99	1.12
		1.5	0.98	0.98	0.98	0.93	0.98	1.02
		1.7	0.98	0.98	0.97	0.93	0.98	1.00

The best improvement occurs in the cases of tight, i.e., FF = 1.3, and moderate due dates, i.e., FF = 1.5, for the push scheme. Furthermore, we find no significant difference between the two job-weighting distribution schemes. Therefore, it appears that only in a very congested wafer fab would the use of DSBH be warranted under a push order release strategy; otherwise, FIFO dispatching is advisable.

However, in a highly loaded system, i.e., 80 jobs in WIP, the use of the DSBH in combination with CONWIP order release can lead to an AWT reduction of 30 % or more as compared to the FIFO dispatching scheme. Finally, Tables 6.1 and 6.2 suggest that the DSBH method with CONLOAD outperforms FIFO dispatching with respect to AWT in almost all situations.

It follows that it is useful to combine DSBH scheduling with a CONLOAD order release strategy in highly loaded wafer fabs. For additional experimentation and results, we refer to Mönch [191].

6.3.3 Conclusions from the Interaction Study

Our experimentation does not reveal any significant difference between the three order release strategies in the low, and moderate-loaded cases. In every case, pure FIFO dispatching outperforms the DSBH scheduling method for the experimental wafer fab model under study (see Mönch [191]). For all experimental levels of FF and job weight distribution, we find that ACT increases and TP decreases when DSBH is used. Further examination of the results confirms that this behavior is caused by low-quality scheduling decisions being produced in the DSBH subproblems.

For a highly loaded BS, the use of either CONWIP- or CONLOAD-type order release strategies in combination with DSBH scheduling appears to be quite useful for reducing AWT. Finally, the push order release strategy can be applied in conjunction with the DSBH in a very highly loaded BS to produce reductions in AWT. However, CONWIP performance is superior to that of both CONLOAD and push in the most congested wafer fab case, while CONLOAD outperforms push. Note that we only consider the case of continuous job arrivals in these experiments, i.e., newly arrived jobs are released from the order pool on a regular, fixed time interval basis such as every two, three, or four hours. However, one can investigate additional job release schemes characterized by daily or weekly release frequencies. In this situation, we expect reduced due date-based performance for both CONWIP and CONLOAD order release strategies, as the time that a job spends waiting in the BS will be shifted to waiting time in the order pool prior to being released into the fab.

In the future, it is important to investigate the connection between order release decisions and the anticipated scheduling decisions of the DSBH to allow for release of new jobs into the wafer fab based on the anticipated load at bottleneck machines caused by both newly released and current WIP jobs. The next section describes an order release case study at an actual wafer fab that further investigates the frequency and size of order releases into the wafer fab.

6.4 A Large-Scale Order Release Study

In this section, we start by describing the overall situation. We discuss the results of the release timing study. Finally, the findings of the release quantity study are presented.

6.4.1 Overall Situation

A global semiconductor manufacturer commonly releases new wafer jobs into one of its wafer fabs both during morning and evening shifts, but never during their night shift. The jobs are released one at a time, with the job releases being carefully spread out across the time period spanning the morning and evening shifts. Upper-level management at this company commissioned a simulation-based case study to examine how different job release policies could potentially impact wafer fab performance. Simulation is a popular method for conducting such case studies as a high-fidelity model can mimic wafer fab operations quite effectively without ever having any impact on current wafer fab operations and output. The simulation-based order release study investigated two specific questions. First, the study examined the impact of releasing wafer jobs into the wafer fab around the clock, i.e., during all three production shifts, as compared to the current two-shift release policy. We refer to this question as the release timing case study. Next, management was interested in understanding the impact of releasing similar products as groups, called trains, of jobs into the wafer fab rather than spreading out individual job releases over time. We refer to this question as the release quantity case study.

6.4.2 Release Timing Case Study

Consider five different job release plans, denoted as Case 1 through Case 5, that each release an equal fraction of a given week's job starting from Sunday through Saturday. The cases differ in terms of at what time(s) during each day jobs are released.

Figure 6.4 portrays the job release distribution for Case 1 along with the proportion of each day's job releases that enter the factory during 2-h time intervals.

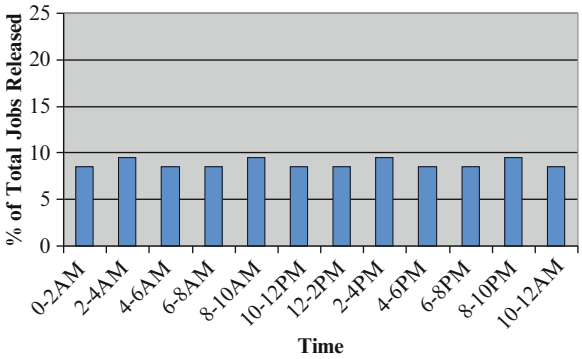

Figure 6.4: Order release distribution for Case 1

It is important to note that jobs are released individually at evenly spaced time intervals within each 2-h time block in Fig. 6.4 according to the total number of job releases planned for the time block.

In Case 2, job release only occurs during two time blocks per day. The situation is depicted in Fig. 6.5. All jobs originally released between midnight and noon in Case 1 are now scheduled for release after the morning shift change, i.e., between 6:00 and 7:00 am. Furthermore, all jobs originally scheduled for release between noon and midnight in Case 1 are rescheduled for release after the evening shift change, i.e., between 2:00 and 3:00 pm. Within each of the two job release time blocks in Fig. 6.5, individual jobs are released uniformly over time.

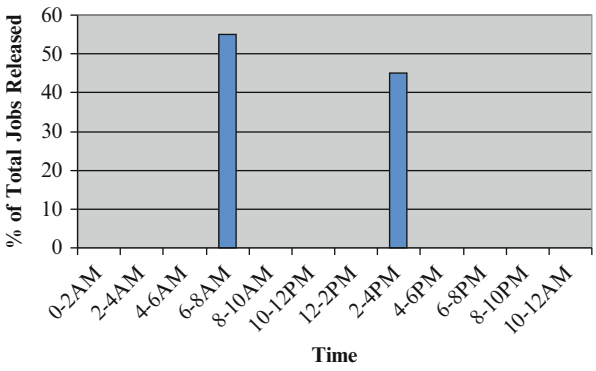

Figure 6.5: Order release distribution for Case 2

Case 3 redistributes daily job releases into equal numbers of jobs every 2 h for each day. This is shown in Fig. 6.6. Jobs are released at each even-numbered hour throughout the 24-h day, i.e., 12 times per day. As the number of jobs scheduled each day is not necessarily evenly divisible by 12, the number of jobs released at 8:00 pm and 10:00 pm will be potentially less than the other job releases to account for beginning- and end-of-day effects.

Case 4 job releases follow the semiconductor manufacturer's current order release policy in that job release occurs only during the morning and evening shifts. The policy is shown in Fig. 6.7.

Individual job releases are distributed evenly for each of these two shifts, with all jobs released during the first half of the day, i.e., between midnight and noon in Case 1, being scheduled for release at evenly spaced time intervals during the morning shift, i.e., between 6:00 am and 2:00 pm. All jobs originally scheduled for release during the second half of the day, i.e., between

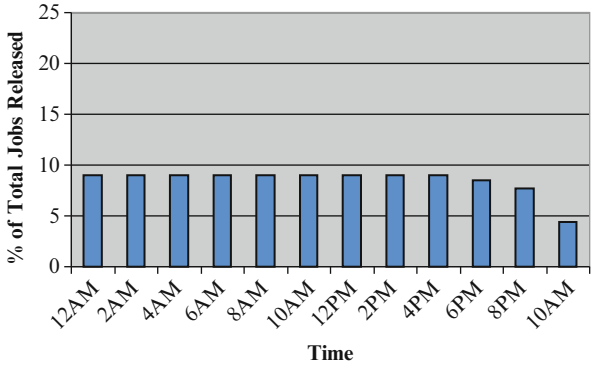

Figure 6.6: Order release distribution for Case 3

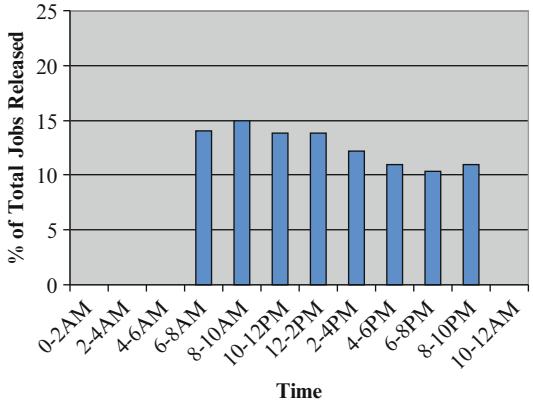

Figure 6.7: Order release distribution for Case 4

noon and midnight in Case 1, are rescheduled for release at evenly spaced time intervals during the evening shift, i.e., between 2:00 and 10:00 pm in Case 4.

Finally, Case 5 job releases occur only during the morning and evening shifts. The jobs released into the wafer fab in Case 5 are released in groups only at specific even-numbered hours during these two production shifts. This is depicted in Fig. 6.8.

The semiconductor manufacturer's validated AutoSched AP simulation model was used with representative job starts data to examine the five order release cases previously discussed. Each simulation replication was run for a period of three years, with the first year of results being discarded to mitigate any potential for initialization bias. Given the complexity inherent in the company's simulation model, each simulation run required approximately 12 h

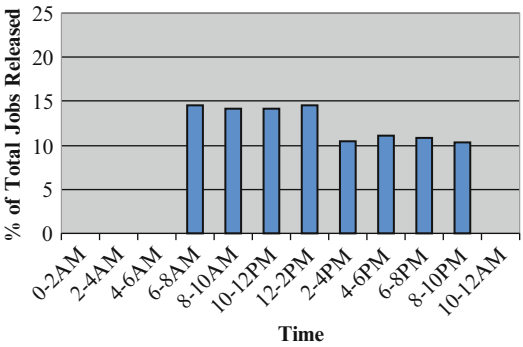

Figure 6.8: Order release distribution for Case 5

of wall clock time. In order to measure the potential for on-time delivery compliance, a due date offset equivalent to three times of each job's raw processing time was used.

A summary of all simulation replications made for the five job release cases is given in Table 6.3 in terms of TP, expressed as number of jobs completed per day; ACT, expressed as a multiple of the raw processing time, i.e., as flow factor FF; and the percentage of jobs completed on or before their due date, denoted by OTD(%). As stated previously, Case 4 is the most accurate characterization of the company's current job release policy. Case 3 was identified by the semiconductor manufacturer as the most appropriate alternative approach for comparison purposes with Case 4. Finally, Case 5 is quite similar to Case 4, except that instead of releasing jobs every m minutes during the morning and evening shifts, it only releases jobs into the wafer fab every 2 h. For these reasons, it was decided to focus the detailed results analysis on these three cases rather than on all five options.

Table 6.3: Simulation results for release timing case study

Compare	TP (Jobs)	ACT (FF theoretical)	OTD(%)
Case 1	34.808	3.180	82.400
Case 2	34.848	3.150	86.900
Case 3	34.853	3.080	92.000
Case 4	34.850	3.110	90.000
Case 5	34.854	3.130	89.100

A paired t-test analysis of Cases 3 and 4 revealed the following with 95 % confidence:

1. The ACT value of jobs in Case 3 is shorter than the ACT value of Case 4 jobs.

2. The on-time delivery performance of Case 3 is superior to that of Case 4.

Furthermore, no significant statistical difference was found between the TP of Cases 3 and 4. As Case 3 was determined to be statistically superior to Case 5 in all three performance measures of interest, the semiconductor manufacturer's study found compelling evidence to try an alternative job release strategy that was determined to have the potential to improve operational performance of its wafer fab.

6.4.3 Release Quantity Case Study

In the previous release timing case study, the jobs to be individually released into the wafer fab were furnished by the semiconductor manufacturer in a particular desired order without any regard being given to the product type of each job. In the release quantity case study, both Cases 3 and 4 are examined further by arranging the list of jobs to be released into the BS into groups, i.e., trains of multiple jobs less frequently, according to product type. A potential benefit of the train approach is that early batch process steps will be able to make fuller batches as sufficient quantities of production jobs will be available within a shorter time horizon.

Figures 6.9 and 6.10 present the release time distribution for Case 3 trains of jobs and Case 4 in the release quantity case study, respectively. The release time distributions for the trains of jobs (TofJ) of the two cases are denoted for abbreviation as TofJ3 and TofJ4, respectively.

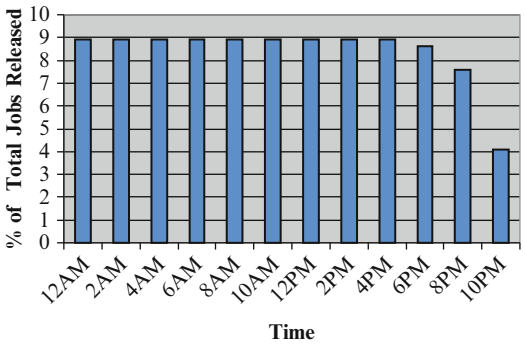

Figure 6.9: Order release distribution for Case 3 trains jobs

An initial analysis of the job trains contained in both the TofJ3 and TofJ4 simulation model inputs led the semiconductor manufacturer to suggest the establishment of an upper bound on the number of jobs that can be present

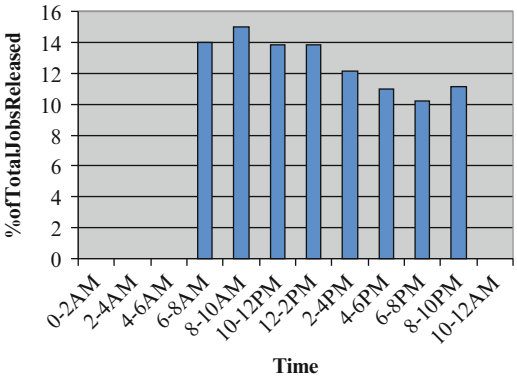

Figure 6.10: Order release distribution for Case 4 trains of jobs

in a single train. The limit of no greater than ten jobs per train was established for both cases and with the idea that a train of ten jobs or less would not unnecessarily overload any fab machine group. We use the notation TofJ3_NGT10 and TofJ4_NGT10, respectively. Finally, the job trains established for Case 4 were examined for two additional upper-bound limits on train size. The TofJ4_NGT_OvenBatch case was created to restrict Case 4's job train size to the maximum load size, measured in jobs, of the first diffusion oven process step contained in each product's process flow. Similarly, the TofJ4_NGT6 case restricts the length of the Case 4's jobs to six jobs.

As was the case in the previous case study, the semiconductor manufacturer's AutoSched AP simulation model was used, and each simulation replication was run for a period of 3 years, with the first year of results being discarded to mitigate any potential for initialization bias. Table 6.4 shows the results of all simulated cases. We note that each TofJ case has a higher ACT value and a lower OTD(%) value than the original Cases 3 and 4 except for TofJ4_NGT_OvenBatch. Although releasing jobs in trains can significantly increase CT values and reduce OTD(%) compared to current practice, the resulting performance appears to depend on the size of the trains formed. In some cases, the semiconductor manufacturer was presented with the opportunity to actually improve performance, should they choose to use oven batch-sized trains.

6.5 Optimization-Based Order Release

In addition to simulation-based methods for order release analysis/planning, mathematical optimization-based approaches can be used. Such an approach is taken, for example, by Missbauer [185]. In this section, we discuss an optimization-based approach for order release in semiconductor manufacturing.

Table 6.4: Simulation results for release quantity case study

	TP (jobs)	ACT (FF theoretical)	OTD(%)
Case 3	34.853	3.080	92.000
Case 4	34.850	3.110	90.000
TofJ3	34.858	3.240	78.200
TofJ4	34.857	3.250	76.800
TofJ3_NGT10	34.846	3.170	85.800
TofJ4_NGT10	34.854	3.180	84.700
TofJ4_NGT_OvenBatch	34.855	3.100	93.100
TofJ4_NGT6	34.855	3.250	76.500

Consider a wafer fab that is interested in planning starts for some period of time, for example, the next week. At the current time, WIP exists in the wafer fab at a variety of locations, i.e., at different process steps, with each job in WIP containing some number of wafers of a predefined technology, process, and device type. The process step at which each job is located corresponds to some process flow that the job is required to follow. During this process flow, we focus on the photolithography process steps and the loading of the potential bottleneck machines of the wafer fab, the photolithography steppers (cf. the description in Sect. 2.2.3). New job starts into the wafer fab are released to meet customer demands for a specific product. They are characterized in terms of their technology, process, and device. However, the wafer fab has potentially more than one option, i.e., job type designation, that it can make on new job starts that directly corresponds to the specific steppers that will be visited during critical photolithography layers.

The option designation for all new job starts directly impacts the wafer fab loading, as the choice of any device d's option 1, for example, can require the job to visit the first stepper two times and the second stepper six times at the eight critical layers of the process flow. However, designating device d's job release as option 2 alternately can result in five visits to the first stepper, two visits to the second stepper, and one visit to the third stepper for the eight critical layers. Simulation-based optimization is used by Mönch et al. [201] to solve a similar load-balancing problem for steppers in an ASIC wafer fab.

In this way, effective release job option designations are an important way in which the photolithography capacity can be utilized most effectively. The problem is further complicated when one considers that jobs can be released today, tomorrow, or on any other day within the desired job release horizon. However, management may dictate that specific jobs and/or specific job option designations must be started on a specific day.

We now provide a MIP formulation for the wafer release optimization problem. The only wafer fab capacity constraints being represented in this model are photolithography steppers. Without loss of generality, we use MES data to collect information on the expected CT values of all other process

steps for each process flow and model job travel through the BS in terms of daily, i.e., 24-h, movements from a given process step to the expected location of the job after 24 h of wafer fab operations.

The following indices and index sets will be used within the model:

t : technology index
p : index of process flows
d : index of devices
o : index of job type designation options for new job starts
s : index of process steps in process flow p of technology t
l : index of existing jobs in WIP that are following a specific process flow
e : index of photolithography machines
h : index of production days in the planning horizon
T : technology set
$P(t)$: set of process flows of technology t
$D(t)$: set of devices of technology t
$O(t)$: set of job type designation options for new job starts of technology t
$S(t,p)$: set of process steps in process flow p of technology t
$L(t,p)$: set of all existing jobs in WIP that are following process flow p of technology t
E : set of all steppers
H : set of all production days in the planning horizon

In addition, for ease of reading, we define $K(t,p) := \{t\} \times P(t) \times S(t,p) \times L(t,p)$, $R(t,p) := K(t,p) \times H$, and $J(t) := \{t\} \times P(t) \times D(t) \times O(t) \times H$ for abbreviation. The following parameters will be used within the model:

α : first day job starts can be made
β : last day job starts can be made
σ_{tps} : number of step s in process flow p of technology t
π_{tps} : $\begin{cases} 1, & \text{if process step } s \text{ in process flow } p \text{ of technology } t \text{ is a photolithography step} \\ 0, & \text{otherwise} \end{cases}$
ω_{tpsl} : processing rate (in wafers per hour) for fab job l at photolithography step s in process flow p of technology t
ξ_{tpsd} : processing rate (in wafers per hour) for any job of device type d at photolithography step s in process flow p of technology t
ϕ_e : number of available processing hours per day for stepper e
τ_{tpl} : number of wafers of job l in initial WIP that follow the process flow p of technology t
η_{tp} : number of the last step in process flow p of technology t
δ_{tpl} : number of the process step in process flow p of technology t at which job l is initially located

$\mu_{tp,\delta_{tpl},1}$: number of the process step in process flow p of technology t at which job l will move to 24 h after being at its initial location δ_{tpl}

γ_{tps} : number of the process step in process flow p of technology t at which any job will move to 24 h after being at process step s

$$\kappa_{tps} : \begin{cases} 1, & \text{if process step } s \text{ in process flow } p \text{ of technology } t \text{ can have WIP present during the daily job location assessment} \\ 0, & \text{otherwise} \end{cases}$$

$$\psi_{tpss'} : \begin{cases} 1, & \text{if step number } s' < \gamma_{tps} \\ 0, & \text{otherwise} \end{cases}$$

$$\varepsilon_{tplse} : \begin{cases} 1, & \text{if stepper } e \text{ is qualified to process existing job } l \text{ at process step } s \text{ in process flow } p \text{ of technology } t \\ 0, & \text{otherwise} \end{cases}$$

$$\upsilon_{tpdose} : \begin{cases} 1, & \text{if stepper } e \text{ is qualified to process job releases of device } d\text{'s option } o \text{ at process step } s \text{ in process flow } p \text{ of technology } t \\ 0, & \text{otherwise} \end{cases}$$

ρ_{tph} : number of wafers for process flow p in technology t that must be released on day h

θ_{tp} : number of wafers for process flow p of technology t to be released during some days $\alpha \leq h \leq \beta$

$$\lambda_{tpdo} : \begin{cases} 1, & \text{if jobs of device } d\text{'s option } o \text{ are qualified for release in process flow } p \text{ of technology } t \\ 0, & \text{otherwise} \end{cases}$$

The following decision variables are required in the model:

I_{tpsl} : initial WIP (in wafers) for job l at its initial step s in process flow p of technology t

M_{tpdoh} : number of wafers in WIP from a new release of device d's option o in process flow p of technology t on day h

N_{tpdsoh} : number of wafers in WIP at process step s on day h in process flow p of technology t from a new release of device d's option o

V_{tpslh} : WIP (in wafers) for job l at process step s in process flow p of technology t at the end of day h

Q_{tpselh} : total hours of workload associated with existing job l, which follows process flow p of technology t that is assigned to the stepper e at process step s on day h

$R_{tpdseoh}$: total hours of workload associated with new job releases of option o of device d, which follows process flow p of technology t that is assigned to stepper e at process step s on day h

U_{eh} : loading (utilization) of stepper e on day h
W_{tpdoh} : integer number of 25 wafer jobs of device d's option o released in process flow p of technology t on day h
X_{tpdoh} : integer number of wafers of device d's option o released in process flow p of technology t on day h
Y_h : maximum projected daily loading of any stepper on day h
Z : maximum number of job starts on any day during the starts horizon

The order release optimization model can be formulated as follows:

$$\min w_1 Z + w_2 \sum_{h \in H} Y_h \tag{6.6}$$

subject to

$$I_{tpsl} = \tau_{tpl}, \quad \left\{(t,p,s,l) \in K(t,p) | \sigma_{tps} = \delta_{tpl}\right\}, \tag{6.7}$$

$$V_{tpsl1} = I_{tp,\delta_{tpl},l}, \ \left\{(t,p,s,l) \in K(t,p) | \sigma_{tps} = \mu_{tp,\delta_{tpl},1}\right\}, \tag{6.8}$$

$$V_{tpslh} = V_{tp,\mu_{tp,\delta_{tpl},h-1},l,h-1}, \left\{(t,p,s,l) \in K(t,p), h \in H | h>1, \sigma_{tps}=\mu_{tp,\delta_{tpl},h}\right\}, \tag{6.9}$$

$$\sum_{\{e \in E, r \in S(t,p) | \pi_{tpr}=1, \ \varepsilon_{tplre}=1, \ \delta_{tpl} \le \sigma_{tpr} < \mu_{tp,\delta_{tpl},1}\}} \omega_{tprl} Q_{tprel1} \ge \sum_{\{a \in S(t,p) | \pi_{tpa}=1, \ \delta_{tpl} \le \sigma_{tpa} < \mu_{tp,\delta_{tpl},1}\}} \sum_{b=\sigma_{tpa}} \psi_{tpab} I_{tp,\delta_{tpl},l}, \tag{6.10}$$

$$\left\{(t,p,s,l) \in K(t,p), \sigma' \in \{1,\ldots,\eta_{tp}\} | \pi_{tps}=1, \sigma_{tps}=\sigma', \delta_{tpl} \le \sigma_{tps} < \mu_{tp,\delta_{tpl},1}\right\},$$
$$\sum_{\{e \in E, r \in S(t,p) | \pi_{tpr}=1, \ \mu_{tp,\delta_{tpl},h-1} \le \sigma_{tpr} < \mu_{tp,\delta_{tpl},h}, \ \varepsilon_{tplre}=1\}} \omega_{tprl} Q_{tprelh} \ge \sum_{\{a \in S(t,p) | \pi_{tpa}=1, \ \mu_{tp,\delta_{tpl},h-1} \le \sigma_{tpa} < \mu_{tp(\delta_{tpl})h}\}} \sum_{b=\sigma_{tpsa}} \psi_{tpab} V_{tp,\mu_{tp,\delta_{tpl},h-1},l,h-1}, \tag{6.11}$$

$$\left\{(t,p,s,l,h) \in R(t,p), \sigma' | h>1, \pi_{tps}=1, \sigma_{tps}=\sigma' \le \eta_{tp}, \mu_{tp,\delta_{tpl},h-1} \le \sigma_{tps} < \mu_{tp,\delta_{tpl},h}\right\},$$

$$M_{tpdoh} = 25 W_{tpdoh}, \ \left\{(t,p,d,o,h) \in J(t) | t \in T, h \le \beta, \lambda_{tpdo}=1\right\}, \tag{6.12}$$

$$N_{tpdsoh} = M_{tpdoh}, \left\{(t,p,d,o,h) \in J(t), s \in S(t,p) | \sigma_{tps}=1, h \le \beta, \lambda_{tpdo}=1\right\}, \tag{6.13}$$

$$N_{tpdsoh} = \sum_{\{r \in S(t,p) | \kappa_{tp,\sigma_{tpr}}=1, \sigma_{tpr} \le \gamma_{pr}\}} N_{tpdro,h-1},$$

$$\{(t,p,d,o,t) \in J(t), s \in S(t,p) | t \in T, \kappa_{tp,\sigma_{tps}} = 1, \sigma_{tps} > 1, \lambda_{tpdo} = 1\}, \quad (6.14)$$

$$\sum_{\{e \in E | v_{tpdose}=1\}} \xi_{tpsd}\ R_{tpdseoh} \ge$$

$$\sum_{\{a \in S(t,p) | \pi_{tpa}=1,\ \sigma_{tpa} < \gamma_{tps1},\ \kappa_{tp,\sigma_{tpa}}=1\}} \{b | b=\sigma_{tpa},\ \sigma_{tpa} < \gamma_{tps},\ \sigma_{tps} < \gamma_{tpa}\} \psi_{tpab}\ N_{tpdao,h-1},$$

$$\{(t,p,d,o,h) \in J(t), s \in S(t,p) | t \in T, \pi_{tps} = 1, \lambda_{tpdo} = 1\}, \quad (6.15)$$

$$\phi_e U_{eh} \ge \sum_{t \in T} \sum_{p \in P(t)} \sum_{\{s \in S(t,p) | \pi_{tps}=1\}} \{l \in L(t,p) | \varepsilon_{tplse}=1,\ \mu_{tp,\delta_{tpl},h-1} \le \sigma_{tps} \le \mu_{tp,\delta_{tpl},h}\} Q_{tpselh}$$

$$+ \sum_{t \in T} \sum_{p \in P(t)} \sum_{d \in D(t)} \sum_{\{s \in S(t,p) | \pi_{tps}=1\}} \sum_{\{o \in O(t) | \lambda_{tpdo}=1,\ v_{tpdose}=1\}} R_{tpdseoh},$$

$$\{e \in E, h \in H\}, \quad (6.16)$$

$$\sum_{p \in P(t)} \sum_{\{o \in O(t) | \lambda_{tpdo}=1\}} X_{tpdoh} \ge \rho_{tdh},$$

$$\{t \in T,\ d \in D(t),\ h \in H | \alpha \le h \le \beta\}, \quad (6.17)$$

$$\sum_{p \in P(t)} \sum_{\{o \in O(t) | \lambda_{tpdo}=1\}} \sum_{\{h \in H | \alpha \le h \le \beta\}} X_{tpdoh} = \sum_{\{h' \in H | \alpha \le h \le \beta\}} \rho_{tdh'} + \theta_{td},$$

$$\{t \in T, d \in D(t)\}, \quad (6.18)$$

$$25 W_{tpdoh} \ge X_{tpdoh},\ \{(t,p,d,o,h) \in J(t) | t \in T, h \le \beta, \lambda_{tpdo} = 1\}, \quad (6.19)$$

$$Z \ge \sum_{t \in T} \sum_{p \in P(t)} \sum_{d \in D(t)} \sum_{\{o \in O(t) | \lambda_{tpdo}=1\}} W_{tpdoh}, \quad \{h \in H | \alpha \le h \le \beta\}, \quad (6.20)$$

$$Y_h \ge U_{eh}, \qquad \{e \in E, h \in H\}, \quad (6.21)$$

$$I_{tpsl} \ge 0, M_{tpdoh} \ge 0, N_{tpdsoh} \ge 0, V_{tpslh} \ge 0, Q_{tpselh} \ge 0, U_{eh} \ge 0, Y_h \ge 0, Z \ge 0,$$
$$\{t \in T, p \in P(t), d \in D(t), s \in S(t,p), o \in O(t), e \in E, l \in L(t,p), h \in H\}, \quad (6.22)$$

$$W_{tpdoh},\ X_{tpdoh} \in \mathbb{Z}_+, \quad \{(t,p,d,o,h) \in J(t) | t \in T\}. \quad (6.23)$$

We seek to minimize the weighted sum of the maximum daily starts on any single day and the sum of the maximum daily loading on each stepper over the entire planning horizon. Clearly, weights are required to properly balance the two objective function components in terms of their dimensionality, i.e., units of measure, along with the desired importance the user prefers to specify for each individual objective function component. The objective function that combines these two key performance measures is given by expression (6.6). In this objective function, both $w_1 \in \mathbb{R}_+$ and $w_2 \in \mathbb{R}_+$ are weights that can be specified according to the user's preference regarding the importance of each objective function component in relation to each other.

We assume that a certain number of jobs exist currently in the wafer fab. Without loss of generality, we assume that the MES or other database can be queried to ascertain the current location, i.e., process step, of each existing job in WIP as well as other job-specific attributes such as its associated technology, process, and the number of wafers in the job. Constraints (6.7) establish the value of I_{tpsl} based on the initial MES information.

Constraints (6.8) and (6.9) model the movement of the jobs existing in the initial WIP through each job's own respective process flow. Using MES data, individual process step CT values are obtained and then aggregated to establish the expected location of each job 24 h from the current time. While constraints (6.8) make this calculation based on each initial step number δ_{tpl} of each job, constraints (6.9) recursively project the rest of each job's 24 h daily movements using the same MES data based on the idea of a daily job location assessment.

Constraints (6.9) establish the daily wafer fab location in terms of the process step for each existing job during the planning horizon, while constraints (6.10) and (6.11) determine the assignment of the processing item associated with each photolithography process step to each qualified stepper by considering the wafer processing rate of each stepper at the process step. While constraints (6.10) perform this computation for the first day of the planning horizon, constraints (6.11) recursively compute this quantity for all subsequent days of the planning horizon. In this way, the total hours required to complete each existing job at each photolithography process step are completely allocated to one or more steppers. Therefore, rather than assigning a specific stepper to process a specific existing job at a given process step, we ensure that the total workload associated with the process step is allocated to one or more steppers. This approach allows for reducing the complexity of the model as typical binary assignment decision variables are not required.

Constraints (6.12) use the primary decision variable W_{tpdoh} for the number of jobs released into the wafer fab of a given type on a given day to establish the number of new wafer starts by a qualified technology-process-device option combination each day over the starts horizon for all new jobs released into the fab. The number of wafers released found in constraints (6.12) is subsequently used to establish the initial WIP at the first process step of each valid process flow in constraints (6.13).

Next, the initial WIP at the first process step of each valid process flow as determined in constraints (6.13) for the new wafer starts is recursively projected to where it is expected to move based on CT data from the MES that is mapped into a parameter for every day in the planning horizon, i.e., daily job location assessment, in constraints (6.14).

Constraints (6.14) establish the daily fab location in terms of the process step for each new job release during the planning horizon, and constraints (6.15) determine the assignment of the processing time associated with each photolithography step to each qualified stepper by considering the device-specific processing rate of each stepper at the process step. This calculation is analogous to the one in constraints (6.11), which focuses on existing jobs in the wafer fab.

With constraints (6.10), (6.11), and (6.15), all qualified steppers have some amount of assigned workload for each photolithography process step on a day of the planning horizon. Constraints (6.16) sum up all of these workload requirements and compute individual daily stepper loading percentages based on the available hours per day for each stepper.

If it is specified that some desired number of wafer starts must be started on a specific day during the starts horizon, constraints (6.17) ensure that at least this desired number of wafers is started on that day.

Finally, constraints (6.18) ensure that all starts demand is satisfied. Constraints (6.19) compute the number of 25 wafer jobs to be started over the starts horizon for each valid type of starts designation from the individual wafer starts decision variable.

Constraints (6.20) are used to compute the maximum number of job starts on any single day during the horizon in which new job starts can be made. These constraints are necessary to establish the value of one of the two objective function variables. Next, constraints (6.21) set the value of the second objective function variable, the maximum daily loading of a stepper during any day of the planning horizon.

Finally, constraints (6.22) describe the nonnegativity requirements for each decision variable, while constraints (6.23) require positive integer values for the two decision variables relating to wafer and job starts.

We now consider the case of a wafer fab's starts planner who is interested in making job release plans for the next work week. The current BS status in terms of stepper quantities, the current location of all jobs, and the anticipated wafer releases for the next week in terms of quantities and device types based on customer demand forecasts for the quarter are input quantities of the optimization model. While some of this demand is for a specific quantity of wafer releases on a specific day for one or more devices, much of the demand is general demand in the form of 100 wafers of device type D_1 that is of process P_1 of technology T_1 and should be started the next week.

Furthermore, the starts planner knows that due to an update from the photolithography process engineers, available option designations for these 100 wafers are options O_2 and O_3. All other such information is also available

and properly input into the model to develop a starts plan for the week that seeks to both:

1. Minimize the maximum daily loading of a photolithography stepper during any day of the planning horizon
2. Minimize the maximum total number of job starts that are made on any day during the next week

Table 6.5 presents example optimization model results detailing the job release plan for the upcoming week at the wafer fab.

Table 6.5: Example of weekly job starts recommended by Model (6.6)–(6.23)

Technology	Process	Device	Option	Day number	Jobs to start
AA	AAQ	7C65	O73	6	1
AA	AAQ	7C69	O73	1	1
AA	L8C	7AA6	O83	5	1
ALP11	S8DI	8C24	O72	1	1
ALP11	S8DI	8C25	O81	5	1
ALP11	S8DI	8C25	O83	3	4
ALP11	S8DIN	8F26	O82	7	1
ALP11	S8Q	8C38	O72	4	1
ALP11	S8Q	8C39	O83	3	1
ALP11	S8TMC	8C27	O83	3	2

Based on these recommended starts prescribed by the model, Table 6.6 indicates the expected stepper loading for the upcoming two weeks as well. It is important to note that while job starts are being made over some planning horizon such as a week, the overall planning horizon of the model must be at least the length of the longest expected CT of any device's process flow, as it is important to properly model the transitions of both existing jobs in the wafer fab as they work their way through the wafer fab and new job releases.

Finally, as this model will be run potentially on a weekly basis to plan weekly job releases, a new BS status is imported into the model each week to ensure that any unplanned events/changes in the wafer fab over the previous week are properly accounted for in the latest model, whether it be new technologies/process flows or new stepper processing rates and/or availabilities.

The order release optimization model (6.6)–(6.23) can be expanded and customized for the needs of a specific wafer fab. For example, Cypress Semiconductor, a global semiconductor manufacturer that designs, develops, manufactures, and markets high-performance, mixed-signal, programmable solutions for a wide variety of customers, operates an 80,000-square-foot wafer fab in Bloomington, Minnesota, called Fab 4. Fab 4 uses an

Table 6.6: Example photolithography stepper loadings by day

Day	iLine1	iLine2	iLine4	jLine1	jLine2	jLine3	jLine4	Alpha1	Alpha4
1	91	74	87	84	82	87	85	78	64
2	51	72	78	75	78	60	86	59	88
3	78	89	67	71	52	51	53	60	53
4	87	56	86	69	63	78	55	54	70
5	80	67	53	54	58	70	82	68	89
6	83	88	92	55	78	63	56	89	79
7	69	81	69	77	62	90	55	79	73
8	75	88	79	67	71	85	73	85	56
9	77	86	62	82	89	53	77	68	50
10	78	52	89	58	88	60	72	77	77
11	86	78	60	78	83	80	56	72	78
12	73	85	67	79	88	76	71	88	57
13	85	86	83	86	85	82	84	52	85
14	54	51	71	70	65	60	51	58	91

optimization-based approach for planning weekly order release. Fab 4's order release model, in addition to similar functionality to this base model, contains Cypress's own proprietary, company-specific constraints and additional objective functions that allow Fab 4 to effectively load its machines under a wide array of product mix scenarios. Currently, a practical-sized instance of the MIP model (6.6)–(6.23) can be solved to within 1 % of the optimal solution in less than two minutes on a desktop computer using commercially available optimization software.