Chapter 7
Production Planning Approaches

In this chapter, we discuss production planning approaches for semiconductor manufacturing. Planning is on the highest level of the PPC hierarchy. Planning approaches provide important input for the order release schemes discussed in Chap. 6. We start by describing short-term planning approaches. Spreadsheet modeling and simulation are used in this situation.

Then, we continue by describing master planning approaches in semiconductor manufacturing. They are used to assign production quantities to different facilities in different periods of time for a horizon of several months. Weekly time periods are considered. Simulation-based performance assessment of master planning approaches is briefly discussed. Next, we discuss capacity planning approaches. In contrast to master planning, these approaches deal with a longer planning horizon and monthly time periods. We discuss only deterministic planning approaches for master and capacity planning. Then, we present enterprise-wide planning approaches. In this situation, we consider a planning horizon of several years and quarters as periods. We also deal with the question of whether or not it is beneficial to open new facilities. Deterministic and stochastic settings are described for enterprise-wide planning problems.

One typical assumption in planning approaches is a fixed CT; however, the CT is load-dependent. Therefore, we discuss different possibilities to model load-dependent CT within planning approaches. We consider CT-TP curves, iterative simulation, and finally clearing functions.

7.1 Short-Term Capacity Planning

In this section, we start by discussing the motivation of spreadsheet-based and simulation-based short-term capacity planning. We then make the first approach more concrete for wafer fabs. Spreadsheet-based short-term capacity planning approaches are discussed for back-end facilities. Finally, short-term capacity planning based on discrete-event simulation is described.

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_7,

7.1.1 Motivation

Spreadsheet-based capacity planning models are ubiquitous. From the early days of Lotus 123 and Quattro Pro to today's Microsoft Excel-based tools, many planners and other wafer fab personnel have developed their own spreadsheet capacity model to make important decisions with regard to near-term capacity needs in the wafer fab (see Occhino [216] and Ozturk et al. [224]). While they can and do vary in size, complexity, level of detail, focus area, and accuracy or validity, spreadsheet-based tools are widely accepted methods for short-term capacity analysis in both wafer fabs and assembly and test facilities.

The typical goal of any short-term capacity planning spreadsheet model is to calculate the expected utilization of one or more machines or machine groups under some amount of demand or loading. While this utilization calculation is often needed to assess the feasibility of a proposed machine loading scenario or to justify the need for additional wafer fab equipment, the underlying mathematics take a variety of forms. The ways in which utilization is defined or calculated and then reported by wafer fab personnel often differ due to one or more modeling assumptions and/or the contingency factors used by the analyst. While spreadsheet-based short-term capacity analyses are predominantly used throughout the front-end and back-end facilities worldwide, the models quite often contain static, deterministic data inputs that are updated on some sort of periodic or as-needed basis. While these updates can be automatically made using SQL queries into corporate data sources (cf. Witte [323] for such an approach), even the most up-to-date information being included in the model will still produce only a static, deterministic estimate of machine group capacity utilization.

Unfortunately, in the absence of some fairly sophisticated queueing network analysis (which is rare in the short-term capacity planning models), spreadsheet-based capacity analysis is unable to accurately model and predict dynamic performance measures associated with the planned capacity levels, such as ACT, WIP levels, and CT variability. While this may not always be of interest to managers conducting strategic capacity analyses such as yearly or five-year plans, short-term capacity analysis often is interested in expected out dates for products/jobs currently in the manufacturing line both in the front-end and the back-end. Discrete-event simulation can help to provide a dynamic perspective for short-term capacity planning.

7.1.2 Spreadsheet-Based Approaches for Wafer Fabs

The basic approach for short-term capacity planning in wafer fabs typically requires some of the following set of machine-specific input data for each machine group w that we desire to analyze:

- Number of machines contained within machine group w, denoted by $Q(w)$.
- Percentage of time that machines in machine group w are available on average for processing wafers, denoted by $\mathrm{Av}(w)$.

- Percentage of average available time each time period that machines in machine group w are processing wafers, denoted by $\text{Eff}(w)$.
- Total number of hours per time period that machines in machine group w are scheduled for production, denoted by $\text{SH}(w)$.

Multiple, different efficiency values can also be used in place of a single parameter estimate. For example, some wafer fabs track operator efficiency, machine loading efficiency, and other measures. In this case, the collection of efficiency measures, all of which are defined from 0 % to 100 %, would be multiplied together to compute an overall efficiency measure for the machine in question. With these machine-specific inputs, the total number of expected productive hours $\text{PH}(w)$ can be computed for each machine group w as follows:

$$\text{PH}(w) := Q(w)\text{Av}(w)\text{Eff}(w)\text{SH}(w). \tag{7.1}$$

For example, an etch machine group containing eight machines, each of which is scheduled 24 h per day, seven days per week, has a historical availability due to both scheduled and unscheduled downtime events of 92 %. In addition, the corporate policy is to assume an 85 % productivity efficiency, which relates to the company's desired minimum amount of idle time on the machine, and a 90 % load efficiency, which pertains to how fully the machine is typically loaded with regard to maximum load size. In this case, we obtain:

$$\text{Eff}(w) = (0.85)(0.9) = 0.765. \tag{7.2}$$

Applying Eq. (7.1) results in the expected number of productive hours per week given by

$$\text{PH}(w) = (8)(0.92)(0.765)(168) = 945.9. \tag{7.3}$$

Once the available machine group productive hours are known, the next step is to characterize how the machine group is impacted, i.e., visited, by demand for a specific product that is made according to some specified process flow or route i, i.e., route-specific information.

Given the reentrant nature of front-end wafer fabrication processes, it is important to capture a number of inputs to properly characterize a machine group's route-specific information. These inputs should include recipe-based parameters as different recipes are visited various numbers of times in a typical manufacturing route i. In addition, since the speed or processing rate of a machine can be recipe-dependent, this too should be taken into account.

With this in mind, route-specific inputs often contain some of the following inputs for recipe r:

- Number of times the current recipe r is visited for route i, denoted by NV_{ir}.
- Rate at which recipe r processes wafers, expressed in wafers per hour, on route i, denoted by UPH_{ir}.
- Percentage of recipe r wafers that must be reworked on route i, denoted by RWP_{ir}

In the case where multiple recipes are specified for a given route, typical capacity analyses aggregate the inputs to calculate a total number of visits, as well as average UPH and RWP values. Average UPH is calculated using a visits-weighted harmonic mean. A harmonic mean is required when any time rates, i.e., some quantity per unit time, are to be averaged, such as units per hour. The weighted harmonic mean of j positive real numbers $n_1,\ldots,n_j$ associated with weights $w_1,\ldots,w_j$ is defined as follows:

$$H = \sum_{k=1}^{j} w_k \Big/ \sum_{k=1}^{j} \frac{w_k}{n_k}. \tag{7.4}$$

Assume route i has j different recipes that machine group w encounters and that recipe $l, l = 1,\ldots,j$ is visited NV_{il} times by machine group w. A visits-weighted harmonic mean is used to calculate the average UPH, denoted by AUPH, for machine group w on route i with $\sum_{l=1}^{j} \mathrm{NV}_{il}$ total visits as follows:

$$\mathrm{AUPH}_i(w) := \sum_{l=1}^{j} \mathrm{NV}_{il} \Big/ \sum_{m=1}^{j} \frac{\mathrm{NV}_{im}}{\mathrm{UPH}_{im}(w)}. \tag{7.5}$$

The average RWP, denoted by ARWP, is a visits-weighted arithmetic mean that is given by

$$\mathrm{ARWP}_i(w) := \sum_{l=1}^{j} \left(\mathrm{NV}_{il}\ \mathrm{RWP}_{il}(w)\right) \Big/ \sum_{m=1}^{j} \mathrm{NV}_{im}. \tag{7.6}$$

Given the machine- and route-specific inputs, a short-term capacity analysis can be performed to determine the maximum number of wafers that machine group w can feasibly process in some desired period of time for route i with $\sum_{l=1}^{j} \mathrm{NV}_{il}$ total visits by machine group w on route i. This is called the maximum number of wafer starts per time period (MaxWSPT) and is determined as follows:

$$\mathrm{MaxWSPT}_i(w) = \frac{\mathrm{PH}(w)\mathrm{AUPH}_i(w)(1-\mathrm{ARWP}_i(w))}{\sum_{l=1}^{j} \mathrm{NV}_{il}}. \tag{7.7}$$

It follows that a similar analysis across the different routes to which machine group w is assigned will provide a range of $\mathrm{MaxWSPT}_i(w)$ values for the different routes. From this point, performing a short-term analysis across all machine groups used on a given route i will reveal the true maximum number of wafer starts per week that each route i can feasibly accommodate in terms of available capacity. This is equal to the minimum $\mathrm{MaxWSPT}_i(w)$ value for all machine groups visited on route i.

Finally, now that machine group- and route-specific parameters are known, the demand placed on the machine, i.e., demand-specific information, is the last piece of information required to compute the machine group's utilization. This information is typically specified in terms of the following inputs:

- The number of wafer starts planned for route i in the period, denoted by PS_i.
- The number of days in the period for which the analysis is being conducted, denoted by ND

Clearly, some unit conversion may be required to convert PS_i to a weekly quantity based on ND. Once this is reconciled, machine group w's capacity loading can be calculated as follows:

$$\mathrm{CLP}(w) = \sum_{i \in \mathrm{Routes}} \frac{\mathrm{PS}_i}{\mathrm{MaxWSPW}_i(w)}. \tag{7.8}$$

It is possible, even desirable, to maximize $\mathrm{CLP}(w)$ for each machine group w to 100 %, as this quantity is related to utilizing productive hours rather than total hours. Recall that we previously used various efficiency and availability factors to derate total hours down to the expected available number of $\mathrm{PH}(w)$. Therefore, a 100 % value for $\mathrm{CLP}(w)$ does not mean the machine group is always busy, i.e., 100 % utilized, but rather that the machine group is completely utilizing all planned available productive hours. It follows that because not all machine groups are required for processing wafers on all routes, wafer fab personnel are able to quickly analyze a variety of starts scenarios in such a spreadsheet-based capacity planning tool. After analyzing corporate planning's demand statement, the capacity planning tool reports the expected machine utilization levels if such a plan were implemented. Once the user appropriately adjusts the starts plan to make it capacity-feasible, subsequent discussions are typically had with sales and marketing to see if any additional products for which capacity is available to start can be sold. If so, additional starts are analyzed within the capacity planning tool with the goal of maximizing the number of machine groups for which $\mathrm{CLP}(w)$ attains its maximum 100 % value. In this way, the corporation's overall goal of maximizing profits is pursued via the appropriate, feasible utilization of available production capacity.

The above described methodology for front-end short-term capacity planning was used at Micrel Semiconductor to provide greater visibility into the hidden factory associated with Micrel's front-end wafer fabs. This hidden factory refers to the incremental manufacturing capacity that exists within a given wafer fab that is not being realized due to a combination of misleading machine performance assumptions and suboptimal wafer starts plans. By specifying accurate, up-to-date machine performance inputs to the capacity model, Micrel's wafer fabs provide both wafer fab and corporate planners a clear view of the amount of route-specific wafer starts that can be accommodated in their wafer fab. Similarly, by comprehending current market demands and forecasted orders, Micrel's planners provide Micrel's wafer fabs with capacity-feasible wafer starts plans that maximize the manufacturing capacity within each Micrel wafer fab.

7.1.3 Spreadsheet-Based Approaches for Back-End

In contrast to front-end processes, back-end processes are characterized by linear, rather than reentrant, process flows. CT in back-end facilities is generally measured in days rather than weeks. Although back-end machine groups are typically visited only a single time during a process flow, additional complexities exist in the back-end, such as the need for auxiliary handler equipment in final test processes (cf. the description in Sect. 2.2.3) and the fact that device outs, i.e., shipments to customers, are the typical demand-specific inputs, rather than wafer starts, that make short-term capacity planning for back-end facilities non-trivial.

Front-end wafers are sent to a sorting process that evaluates each individual die's functionality and marks defective dies in an electronic wafer map. This map electronically records the good and bad dies on the entire wafer. The map travels electronically to the assembly area with the wafer, and the assembly equipment reads the map such that it knows what good dies to assemble and then send on to final test. Back-end short-term capacity planning is complicated by the fact that hundreds of wafers from the front-end turn into hundreds of thousands of individual dies that are to be packaged as functional ICs, memory products, communications modules, or other products.

Spreadsheet-based tools are prevalent across back-end facilities. Similar to the front-end discussion in Sect. 7.1.2, back-end capacity planning requires machine-specific, route-specific, and demand-specific inputs. While the machine-specific input parameters are quite similar, back-end processes for electrical testing of individual dies, for example, require slightly different route-specific information. Furthermore, back-end demand-specific information may be specified in a variety of units of measure, such as wafers for the sort process or thousands of dies for the assembly and test processes. However, the same approach can be taken in order to estimate equipment utilization and/or capacity loading.

Consider the final electrical testing phase of the back-end process, a common bottleneck operation. The equipment associated with this step includes not only the tester but an accompanying handler at a minimum and potentially a load board (cf. Sect. 2.2.3). Route-specific inputs for a short-term capacity analysis for the electrical test of die d may include the following:

- The amount of time (in seconds) required for the tester to locate/navigate to the die being tested, denoted by $\mathrm{IT}(d)$
- The amount of time (in seconds) required to test an individual die, denoted by $\mathrm{TT}(d)$
- The number of test programs that an individual die must undergo, for example, room temperature test, elevated temperature test, etc., denoted by $I(d)$

- The number of locations on the die to be tested, denoted by $S(d)$
- The probability that a die being tested functions properly, i.e., yield, denoted by $Y(d)$

Consider device d that is to be electrically tested. If we assume $O(d)$ individual die outs are required by customers for device d, then the total number of test insertions required is calculated as

$$\mathrm{TI}(d) := O(d)I(d)/Y(d). \tag{7.9}$$

Now that the total number of test insertions is determined, the total amount of tester and handler time required to electrically test device d, denoted by total test hours (TTH), is calculated as

$$\mathrm{TTH}(d) := \mathrm{TI}(d)\,(\mathrm{IT}(d) + \mathrm{TT}(d))\,S(d)/3600. \tag{7.10}$$

In Eq. (7.10), the 3,600 value in the denominator is used to convert seconds into hours. This total number of hours required for final testing to produce $O(d)$ good customer units out of device d would then be summed up with all other devices that require similar back-end equipment in order to compute the capacity loading for each machine group and handler group using the previously defined machine- and handler-specific inputs.

An interesting reality in short-term capacity planning for back-end facilities is the comprehension of both tester and handler capacity requirements. Consider the following resource requirements resulting from a capacity analysis:

- Device LM001 requires 27.5 h of tester T_1 and handler H_1 time.
- Device JWF223 requires 42.5 h of tester T_1 and handler H_2 time.
- Device SJM11 requires 30.0 h of tester T_2 and handler H_1 time.

Clearly, this product mix results in a total of $27.5 + 42.5 = 70.0$ h of required tester T_1 time and 30.0 h of tester T_2 time. However, in terms of the handling resources, a total of $27.5 + 30.0 = 57.5$ h of handler H_1 time is required, in addition to 42.5 h of handler H_2 time. In this case, assuming a 24-h work day and an 85 % efficiency factor, the required number of resources needed to produce these desired device outs in a single day would be calculated as follows:

- Number of required testers T_1: $\lceil \frac{70}{24(0.85)} \rceil = 4$.
- Number of required testers T_2: $\lceil \frac{30}{24(0.85)} \rceil = 2$.
- Number of required handlers H_1: $\lceil \frac{57.5}{24(0.85)} \rceil = 3$.
- Number of required handlers H_2: $\lceil \frac{42.5}{24(0.85)} \rceil = 3$.

Therefore, the utilization of the tester and handler resources must be carefully computed as above so that accurate short-term capacity planning is performed that produces effective estimates of capacity loading and/or resource utilization. This is necessary when one considers the fact that an insufficient

amount of tester or handler resources can limit back-end test capacity. Although typically not the case, additional auxiliary resources, such as load boards and/or operators, can also limit capacity. If this is potentially the case in the back-end facility being analyzed, then a similar analysis should be performed on those resources as well, as they also can be modeled in terms of machine-, route-, and demand-specific inputs. We note that all the calculations described in this section can be performed by spreadsheets.

7.1.4 An Integrated Approach Using Simulation

Often, a result of short-term capacity analysis is the expected out dates for products/jobs currently in the BS both in the front-end and back-end. Discrete-event simulation can help to provide a dynamic perspective for short-term capacity planning. While calculating such a WIP Flush projection can be done in a spreadsheet using historical estimates for expected process step CT, many analysts have turned to discrete-event simulation methods as they are designed to accommodate many of the uncertain realities present in wafer fabs that spreadsheet models do not readily model, such as machine failures (cf. Sect. 3.2.8).

While spreadsheet-based models do include machine availability assumptions, stating that a machine is available 92 % of the time, for example, is simply a high-level (but necessary) assumption. This is different from the capability provided in simulation models to specify both machine TTF and TTR distributions. The same 92 % availability can be modeled as mean TTF of 100 h and mean TTR of 8 h, for example, in a simulation model when TTF and TTR are assumed to be exponentially distributed with rate parameter $\lambda = 0.010$ and $\lambda = 0.125$, respectively.

A validated simulation model of the wafer fab can be automatically populated with the current WIP at each process step and the status of each machine group such that a short-term WIP Flush analysis can be conducted to estimate the day and time one or more jobs of interest are expected to exit the BS. This can be especially useful when customers are calling to ask when their requested products will be available to them. In addition, wafer fabs sometimes perform WIP Flush runs to estimate if they will be able to make their quarterly shipment goals to their back-end facilities.

It is important to note that many of the inputs required to build a valid simulation model can also be found in spreadsheet-based capacity analysis models, and as such, one powerful technique for performing short-term capacity analysis studies is an integrated approach that utilizes the strengths of both methods. First, the spreadsheet model can be used to accurately determine resource levels in terms of number of machines, operators, and/or other capacitated resources that are required to make some desired quantity of goods. The resulting resource levels can then be fed into the simulation model, along with process flow, equipment, and demand information, and this proposed BS and BP configuration can be simulated to ascertain the

resulting dynamic performance of the wafer fab in terms of CT and WIP levels.

While this integrated approach has proven quite valuable, an additional level of model utility can be achieved when the results of the simulation runs are used to change some of the underlying assumptions and/or inputs contained in the spreadsheet capacity analysis model. The analyst can fine-tune the desired performance of the wafer fab under study by using both modeling approaches in this interactive fashion. This can be especially important considering the ability of a spreadsheet model to compute required investment levels regarding new equipment acquisition and personnel hiring decisions. By using both spreadsheet- and simulation-based short-term capacity analysis methods, a greater level of insight and understanding may be afforded to the analyst conducting the study.

7.2 Master Planning

Master planning (MP) is somewhere between short-term capacity planning and more strategic capacity planning. It deals with determining appropriate wafer quantities for several products, several production sites, and several periods of time.

A master plan typically has a horizon of six months divided into weekly time buckets. Since market demand is not entirely known when planning a couple of weeks or months ahead, we have to distinguish between firm customer orders and additional forecasts. Explicit customer requirements are confirmed, postponed, or reduced by the order management process based on available supply. On the other hand, the demand planning process performed every month by sales and marketing departments tries to foresee the rest of the market needs. Both are main inputs of MP (see Vieira [313]).

In the following, we describe a model for master planning as proposed by Ponsignon and Mönch [245]. The resultant model is called MPSC for abbreviation. We start by presenting the related index information:

$p = 1,\ldots,P$: product index
$t = 1,\ldots,T$: time index
$k = 0,\ldots,k_{\max}$: index for measuring capacity consumption
$m = 1,\ldots,m_{\max}$: facility index
$b = 1,\ldots,b_{m,\max}$: bottleneck index for facility m

We assume that P products can be processed in $m_{\max}$ facilities consisting of $\mathrm{ih}_{\max}$ in-house locations and $\mathrm{sc}_{\max}$ subcontractor sites. The total number of bottleneck work centers associated with all facilities is represented by $b_{\max}$. We assume that each bottleneck is assigned to exactly one facility and that each facility has at least one bottleneck. This assumption is reasonable because planned bottlenecks, caused by very expensive machines, exist in all wafer fabs. Clearly, $\sum_{m=1}^{m_{\max}} b_{m,\max} = b_{\max}$ holds. In case of subcontractors, we model only one bottleneck, i.e., we set $b_{m,\max} = 1$. The quantity T stands for the planning horizon measured in periods. We use one week as the length of

a single time bucket. We assume for simplicity reasons that all products have the same cycle time of $k_{\max}+1$ weeks.

The following parameters are part of MPSC:

B_{p0} : initial backlog of product p at the beginning of the first period
$C_{mbt}^{\min}$: minimum utilization of bottleneck b in facility m in period t (in hours or pieces)
$C_{mbt}^{\max}$: maximum available capacity of bottleneck b in facility m in period t (in hours or pieces)
cc_{pmbk} : capacity consumption of one wafer of product p when this product is processed in facility m at bottleneck b and the completion period is k periods ahead
$d_{pt}^{(fc)}$: additional forecast demands for product p at the end of period t
$d_{pt}^{(o)}$: confirmed orders for product p at the end of period t
hc_{pt} : inventory cost for holding one wafer of product p during period t
I_{p0} : initial inventory level of product p at the beginning of the first period
lc_{pmt} : location cost when product p is processed in facility m in period t, i.e., fixed costs
mc_{pmt} : cost to produce one wafer of product p in facility m in period t, i.e., variable costs
rev_{pt} : expected revenue per wafer for satisfying additional demands of product p in period t
udc_{pt} : cost due to unmet confirmed orders for one wafer of product p postponed from period t to period $t+1$
$x_{pmt}^{(i)}$: initial number of wafers of product p to be completed at the end of period t in facility m, i.e., WIP started before the first period of the model
α : large number

The following decision variables are used within the model:

x_{pmt} : number of wafers of product p to be completed at the end of period t in facility m
$s_{pt}^{(fc)}$: sales quantity of additional forecast demand of product p in period t
$s_{pt}^{(o)}$: sales quantity of confirmed orders of product p in period t
B_{pt} : backlog of confirmed orders of product p at the end of period t
I_{pt} : inventory level of product p at the end of period t
u_{pmt} : binary indicator variable for occurrence of fixed production costs of product p in facility m in period t

The model can be formulated as follows:

$$\max \sum_{p=1}^{P} \sum_{t=1}^{T} \left\{ rev_{pt} s_{pt}^{(fc)} - hc_{pt} I_{pt} - udc_{pt} B_{pt} - \sum_{m=1}^{m_{\max}} (mc_{pmt} x_{pmt} + lc_{pmt} u_{pmt}) \right\} \tag{7.11}$$

subject to:

$$I_{p,t-1} - s_{pt}^{(o)} - s_{pt}^{(fc)} + \sum_{m=1}^{m_{\max}} \left(x_{pmt} + x_{pmt}^{(i)} \right) = I_{pt},\ p = 1,\ldots,P,\ t = 1,\ldots,T, \quad (7.12)$$

$$s_{pt}^{(o)} + B_{pt} = d_{pt}^{(o)} + B_{p,t-1},\ p = 1,\ldots,P,\ t = 1,\ldots,T, \quad (7.13)$$

$$s_{pt}^{(fc)} \le d_{pt}^{(fc)},\ p = 1,\ldots,P,\ t = 1,\ldots,T, \quad (7.14)$$

$$C_{mbt}^{\min} \le \sum_{p=1}^{P} \sum_{k=0}^{\min(k_{\max},T-t)} cc_{pmbk} \left(x_{pm,t+k} + x_{pm,t+k}^{(i)} \right) \le C_{mbt}^{\max},$$

$$m = 1,\ldots,m_{\max},\ b = 1,\ldots,b_{m,\max},\ t = 1,\ldots,T, \quad (7.15)$$

$$x_{pmt} \le \alpha u_{pmt},\ p = 1,\ldots,P,\ m = 1,\ldots,m_{\max},\ t = 1,\ldots,T, \quad (7.16)$$

$$x_{pmt} \ge 0, s_{pt}^{(o)} \ge 0, s_{pt}^{(fc)} \ge 0, I_{pt} \ge 0, B_{pt} \ge 0,\ p = 1,\ldots,P,$$

$$m = 1,\ldots,m_{\max},\ t = 1,\ldots,T, \quad (7.17)$$

$$u_{pmt} \in \{0,1\},\ p = 1,\ldots,P,\ m = 1,\ldots,m_{\max},\ t = 1,\ldots,T. \quad (7.18)$$

The objective is to maximize the overall difference between the revenues and the sum of costs. The first term in the objective function (7.11) models the revenues for fulfilling additional forecast demands. The costs for holding inventory are modeled by the second term. The third term refers to penalty costs for backlogged customer orders. The fourth and fifth terms represent variable and fixed production costs, respectively.

Constraint (7.12) represents the flow balance in every period and for every product. The inflows are the initial inventory, the production quantities, and the WIP inventory; the outflows are the sales quantities related to confirmed orders and forecasts and the ending inventory. Constraints (7.13) and (7.14) relate sales quantities to market demand. Backlog is allowed only for customer orders. In case of additional forecasts, we only consider a maximum bound. The capacity restrictions for every bottleneck in each period are defined in constraints (7.15) with minimum and maximum utilization limits. The overall loading is calculated by taking production quantities and WIP inventory of all products into account. We assume $\sum_{p=1}^{P} cc_{pmb0} > 0$ to ensure that there is at least one product p such that $\sum_{k=1}^{\min(k_{\max},T-t)} cc_{pmbk} > 0$ for all $t = 1,\ldots,T$, b, and m. Inequalities (7.16) set the binary variable u_{pmt} to 1 whenever there is a positive production for the considered product, location, and time period. On the other hand, $u_{pmt} = 0$ leads to $x_{pmt} = 0$. It makes sure that an additional facility is used only when it is necessary. Nonnegativity and binary conditions are defined by constraints (7.17) and (7.18).

It is shown in [245] that this problem is NP-hard because a knapsack problem can be reduced to a special case of it. Therefore, efficient heuristics are proposed in [245]. A product-based decomposition heuristic and a GA are described. The product-based decomposition procedure is similar to fix-and-optimize approaches in lot sizing. It can be summarized as follows.

Product-based Decomposition (PD)

1. Initialize the objective function value by $f_{\text{curr}} := 0$.
2. Sort the products with respect to the index I_p in descending order, where we define
$$I_p := \sum_{t=1}^{T} udc_{pt} d_{pt}^{(o)}, \quad p = 1, \ldots, P. \tag{7.19}$$
3. Decompose the set of all products into n disjoint subsets $P_1, \ldots, P_n$ of equal size, only the last subset might have a different size, such that products with similar I_p values are part of the same subset or in consecutive subsets.
4. Solve MPSC given by objective function (7.11) and constraints (7.12)–(7.18) for the current product subset P_i by taking the actual maximum capacity limits into account and by setting the minimum capacity bounds to zero. Increment f_{curr} with the objective value of the current subproblem.
5. Decrease the maximum capacity limits as follows:
$$C_{mbt}^{\max} := C_{mbt}^{\max} - \sum_{p \in P_i} \sum_{k=0}^{\min(k_{\max}, T-t)} cc_{pmbk} \left(x_{pm,t+k} + x_{pm,t+k}^{(i)} \right) \tag{7.20}$$
for each $m = 1, \ldots, m_{\max}$, $b = 1, \ldots, b_{m,\max}$, and $t = 1, \ldots, T$.
6. As long as any product subset has not been considered, increment the index i of the current product subset P_i and go to step 4, else return f_{curr}.

In step 3, the quantity n is determined by some preliminary computational experiments in such a way that the subproblems in step 4 can be solved to optimality by a MIP solver.

We see that the minimum capacity limit is ignored in the PD algorithm. Consider that a minimum utilization threshold leads to an artificial increase of production quantities for products of the first subset. As a result, the remaining capacity may not be sufficient for other subsets. That is why an a posteriori repair scheme where the bottleneck usage in each time period is checked and increased in the case that the minimum bound is not met is proposed by Ponsignon and Mönch [245].

Some computational results for $P \in \{50, 100, 200\}$ and $m_{\max} \in \{8, 12\}$ are shown in Table 7.1. We provide the ratio of the objective function values obtained by PD and by the MIP. The total number of considered problem instances is 120. The MIP is solved using the commercial solver CPLEX. The number of products within each subproblem is four, i.e., we have $n := \lfloor P/4 \rfloor$. The maximum computing time for the MIP is 30 min per problem instance, while the average computing time of PD for $P = 50$, $P = 100$, and $P = 200$ is 10, 15, and finally 30 min, respectively.

We can see from Table 7.1 that the MIP gap increases quickly when the number of products gets larger. Up to 100 products, PD behaves similar to the MIP, but PD clearly outperforms the MIP for $P = 200$.

Table 7.1: Computational results for MPSC

P	ih_{max}	sc_{max}	PD/MIP ratio	Average MIP gap
50	6	2	0.9775	0.0318
50	8	4	0.9753	0.0185
100	6	2	0.9824	0.1441
100	8	4	0.9762	0.0786
200	6	2	1.1343	0.7361
200	8	4	1.1090	0.4008

More computational results, including results for the proposed GA, can be found in [245]. Note that the GA is faster than PD, especially for large-scale problem instances. However, PD usually performs better from a solution quality point of view. Some computational results using heuristics for master planning in a rolling horizon setting can be found in Ponsignon and Mönch [244]. An architecture similar to those described in Sect. 3.3.2 is used. Feedback from the BS and the BP is taken into account with respect to backlog, inventories, and capacity each time an MPSC instance is solved.

7.3 Capacity Planning

In contrast to master planning, the planning horizon for capacity is usually one to three years. Capacity planning is therefore mid-term or long-term. Instead of weeks, usually months or even quarters are used as periods. Continuous decision variables are generally appropriate for production quantities. However, integer-valued decision variables come into play, when capacity expansion decisions are considered by purchasing new machines.

In the following, we present a multi-period capacity planning formulation that is due to Barahona et al. [22]. We start by introducing the following indices and sets that are used within the model:

$j = 1,\ldots,J$: operation index
$i = 1\ldots,I$: machine group index
$I(j)$: set of all machine groups that can perform operation j
$J(i)$: set of all operations that can be performed on machine group i
PT : set of primary machine groups
ST : set of secondary machine groups
$t = 1,\ldots,T$: period index
p : product index
P : set of all products

The following parameters are used within the model:

γ_{pt} : expected number of wafers completed per wafer started for product p in period t
d_{pt} : demand in wafers per day for product p in period t

b_{jpt} : number of passes, adjusted for yield, of operation j on product p in period t
μ_{it} : initial capacity for machine group i in hours/day in period t
c_{it} : unit capacity for machine group i in hours/day in period t
h_{ijt} : number of hours to process one wafer through operation j on machine group i in period t
m_{it} : cost of purchasing a new machine group i in period t
β_t : total budget available for buying new machines in period t
α_{pt} : upper bound for the unmet demand in wafers per day for product p in period t
q_1 : penalty for buying a primary tool
q_2 : penalty for buying a secondary tool

The following decision variables are used in the model:

U_{pt} : unmet demand for product p in wafers per day in period t
W_{pt} : number of wafers per day for product p that enter the wafer fab in period t
O_{jit} : number of wafers per day that require operation j on machine group i in period t
N_{it} : number of new machines bought for machine group i in period t

The capacity planning model can be formulated as follows:

$$\min \sum_{t=1}^{T} \sum_{p \in P} U_{pt} + \sum_{t=1}^{T} \left(q_1 \sum_{i \in PT} N_{it} + q_2 \sum_{i \in ST} N_{it} \right) \tag{7.21}$$

subject to:

$$\gamma_{pt} W_{pt} + U_{pt} = d_{pt}, \quad t = 1, \ldots, T, \ p \in P, \tag{7.22}$$

$$\sum_{p \in P} b_{jpt} W_{pt} = \sum_{i \in I(j)} O_{jit}, \quad j = 1, \ldots, J, \ t = 1, \ldots, T, \tag{7.23}$$

$$\sum_{j \in J(i)} h_{ijt} O_{jit} \leq \mu_{it} + c_{it} \sum_{\tau=1}^{t} N_{i\tau}, \quad t = 1, \ldots, T, \ i = 1, \ldots, I, \tag{7.24}$$

$$\sum_{i=1}^{I} m_{it} N_{it} \leq \beta_t, \quad t = 1, \ldots, T, \tag{7.25}$$

$$U_{pt} \leq \alpha_{pt}, \quad p \in P, \ t = 1, \ldots, T, \tag{7.26}$$

$$U_{pt} \geq 0, W_{pt} \geq 0, O_{jit} \geq 0, \ t = 1, \ldots, T, \ p \in P, \ i = 1, \ldots, I, \tag{7.27}$$

$$N_{it} \in \mathbb{N}, i = 1, \ldots, I, \quad t = 1, \ldots, T. \tag{7.28}$$

The objective (7.21) of the model is minimizing the sum of the total unmet demand and two penalty terms that discourage purchasing primary and secondary machines, respectively. Constraints (7.22) relate the demand for each product to unmet demand and production quantities. It is assumed that the

demand of all periods is satisfied with production from the same period. This assumption is reasonable because inventory is generally kept very low in semiconductor manufacturing. Note that $0 \leq \gamma_{pt} < 1$ models the occurrence of yield in semiconductor manufacturing. Constraints (7.23) determine the total number of wafers that require a specific operation distributed over all possible machines as the sum of the corresponding production levels. Constraints (7.24) ensure that the total production load on machine group i is smaller than the available capacity for this machine group measured in hours per day of production in a specific period. Budget constraints for purchasing new machines are given by constraints (7.25). Upper bounds for the unmet demand are set by constraints (7.26). Finally, constraints (7.27) and (7.28) express the fact that all decision variables are non-negative and that N_{it} is an integer for each machine group and each period.

Note that a two-stage stochastic programming formulation for a situation similar to that covered in model (7.21)–(7.28) is provided by Hood et al. [117] and Barahona et al. [22]. The first stage deals with capacity expansion decisions. The second stage is related to production decisions that can be made when the demand profile is known with certainty. Several demand scenarios with associated probabilities are provided to tackle the two-stage model similar to that described in Sect. 3.2.4. However, additional difficulties have to be resolved that are imposed by the integrality requirements for variable N_{it}.

We continue by briefly discussing the capacity optimization planning system (CAPS). CAPS is a decision-support system used by IBM for strategic planning of its semiconductor capacity (see Bermon and Hood [24]). It is based on linear programming. CAPS determines the product mix that maximizes profit given the existing machine capacity. At the same time, it is also able to determine the necessary capacity taking a given product mix into account. The unrelated parallel machines that are typical for wafer fabs are modeled in detail in the LP to determine a preferential order in which these machine groups are used.

While capacity planning for a single wafer fab is addressed in model (7.21)–(7.28) and by the CAPS model, a multi-facility situation is covered by the model proposed by Habla and Mönch [113]. The model is somewhat similar to the master planning model described in Sect. 7.2; however, assignment decisions to single wafer fabs are not taken. Consequently, integer-valued decision variables are not necessary. The objective is to maximize revenue for forecasted orders and minimize at the same time production costs, inventory holding costs, and costs for unmet committed orders. A quite general product structure is assumed to allow for modeling make-to-stock, assemble-to-order, and make-to-order production.

A detailed survey of strategic capacity planning approaches in semiconductor manufacturing is presented by Geng and Jiang [97]. A stochastic programming model for capacity planning in wafer fabs with uncertain demand and capacity is described by Geng et al. [98].

7.4 Enterprise-Wide Planning

Enterprise-wide semiconductor planning considers the allocation of products to wafer fabs and then routing the wafers with the ICs for testing. The tested wafers are routed to where they can be cut into individual chips and put in a package. The packages are then sent to final test facilities for testing and classification. The products are classified, i.e., binned, according to performance, and shipped to final inventory warehouses, or demand centers, for selling. Planning when to increase or decrease capacity at the production facilities as well as planning when and whether to build new facilities are some of the possibilities for these operations, for example, purchasing a new machine for one of the bottleneck machine groups in a wafer fab, building a new test facility in a new region, or subcontracting to a foundry. In the following, we present a MIP that is due to Stray et al. [293]. The model can be used to answer the following questions:

- What facilities should be built?
- What machines should be purchased?
- What products should be manufactured in which facilities?
- What demand should be met by subcontracting, and what demand should be left unmet in order to maximize profit?

The model is focused at a strategic level, and a typical instance of the problem covers a few years in several segments of perhaps three months per segment, i.e., quarters. The level of detail is deep enough to support decisions such as quarterly production amounts of each product in a company's product portfolio, including the routing of the product between facilities. The model does not attempt to schedule individual jobs of products within facilities. The model is presented below.

We use the following sets and indices in the model formulation:

$\mathbf{FAM}$: set of product families
$\mathbf{PKG}_p$: set of packages, where one set $\mathbf{PKG}_p$ is for each p in $\mathbf{FAM}$
$\mathbf{BIN}_{pq}$: set of bins for each product package q and family p
$\mathbf{BET}_b$: set of bins that can be sold as product with bin b characteristics
L : set of all location sets
L_F : wafer fab set
L_S : sort location set
L_M : assembly set
L_T : test set
L_D : demand center set
$\mathbf{MG}_l$: set of all machine groups in location l
p : index for product families
q : index for packages
b : index for bins

f, l : index for locations
i : index for machine types
t : index for time periods

In the remainder of this section, we use F, S, M, T, D as abbreviations for fab, sort, assembly, test, and demand center, respectively.
The following parameters are part of the model:

PBC_{lt} : cost of building facility l in period t
POC_{lt} : cost of operating facility l in period t
PRC_{lt} : cost of removing facility l in period t
MPC_{ilt} : cost of purchasing a single machine i in facility l in period t
MOC_{ilt} : cost of operating machine i in facility l through period t
SC_{pt} : cost for subcontracting one job of wafers of family p in period t
m_{il} : number of machines initially installed in machine group i and facility l
MAX_{il}^{S} : maximum number of machines allowed in machine group i in facility l
MAX_{l}^{T} : total number of machines allowed in all machine groups in facility l
α_{il} : maximum machine utilization for machine group i in location l
S_{il} : average downtime of machine group i in hours in location l over a period of length TPL
TPL : length of one period in hours
T : number of periods in the model
C_{plt} : fraction of product p in location l started in period t that finishes in period $t+1$
C_{pqlt} : fraction of product p and package q in location l started in period t that finishes in period $t+1$
Q_{plt} : yield of product p in location l in period t
Q_{fpqlt} : yield of the product p and package q in location l in period t, where f is the wafer fab in which the original wafer was manufactured
Q_{fpqblt} : resulting bins b of a product, depending on origin wafer fab f, family p, package q, location l, and time period t
G_{pq} : number of chips per wafer for family p and package q
T_{ipl} : total time product p takes to complete on machine group i in location l
D_{pqblt} : demand of a product p in package q and bin b at location l and period t
PC_{plt} : cost of starting product p at location l in period t
TC_{ldt} : transportation cost from l to d in period t
IC_{plt} : inventory cost for product p in location l and period t
PV_{pqblt} : sales price for product p in package q and bin b at demand center l in period t

PEN_{pqbt} : penalty for not meeting demand for product p in package q and bin b in period t
WLS_l : number of wafers in a job at wafer fabs and wafer sorts, by location l
CLS_l : number of chips in a job at assembly, test, and demand centers by location l
LBT_l : building time for location l
N : large number

All periods are of the same length in the model. The complementary fractions of C_{plt} and C_{pqlt} finish in period t. In order to incorporate yield at the test operations, the quantity Q_{fpqblt} is summed over q for all p, and this number has to be less than or equal to one.

The following decision variables are used within the model:

$X_{plt}^{S_1}$: number of jobs of product p to start in facility l in period t, $S_1 \in \{F,S,M,T\}$
$X_{fplt}^{S_1}$: number of jobs of product p produced in fab f to start in facility l in period t, $S_1 \in \{S,M,T\}$
X_{fpqlt}^{M} : number of jobs of product p, package q, produced in fab f to start in assembly facility l in period t
$W_{plt}^{S_1,S_2}$: number of jobs of product p to put in inventory before (B) or after (A) location l in period t, $S_1 \in \{A,B\}$, $S_2 \in \{F,S,M,T,D\}$
$W_{fplt}^{S_1,S_2}$: number of jobs of product p produced in fab f to put in inventory before (B) or after (A) location l in period t, $S_1 \in \{A,B\}$, $S_2 \in \{F,S,M,T,D\}$
$W_{fpqlt}^{S_1,S_2}$: number of jobs of product p, package q, produced in fab f to put in inventory before (B) or after (A) location l in period t, $S_1 \in \{A,B\}$, $S_2 \in \{F,S,M,T,D\}$
$W_{fpqblt}^{S_1,S_2}$: number of jobs of product p, package q, bin b, produced in fab f to put in inventory before (B) or after (A) location l in period t, $S_1 \in \{A,B\}$, $S_2 \in \{F,S,M,T,D\}$
$Y_{pldt}^{S_1,S_2}$: number of jobs of product p shipped between two locations l and d in period t, $S_1 \in L$, $S_2 \in L$
$Y_{fpldt}^{S_1,S_2}$: number of jobs of product p produced in fab f shipped between two locations l and d in period t, $S_1 \in L$, $S_2 \in L$
$Y_{fpqldt}^{S_1,S_2}$: number of jobs of product p, package q, produced in fab f shipped between two locations l and d in period t, $S_1 \in L$, $S_2 \in L$
$Y_{fpqbldt}^{S_1,S_2}$: number of jobs of product p, package q, bin b, produced in fab f shipped between two locations l and d in period t, $S_1 \in L$, $S_2 \in L$
Z_{pqbdt} : number of jobs of product p, package q, bin b, demand center d, sold in period t
ς_{pqbdt} : number of jobs of product p, package q, bin b, available at demand center d in period t

M^A_{ilt} : number of machines added to machine group i, location l, and period t
M^R_{ilt} : number of machines removed from machine group i, location l, and period t
Ω^A_{lt} : binary indicator variable for adding plant l in period t
Ω^R_{lt} : binary indicator variable for removing plant l in period t
S_{plt} : number of wafers subcontracted of each family p to each assembly operation l in each period t
M_{ilt} : number of machines in machine group i, location l, and time period t
Ω_{lt} : binary indicator variable for plant existence for location l in time period t

The objective function and the constraints of the model can be formulated as follows:

$$\begin{aligned}
&\max \sum_{t,p,q,b,d\in L_D} \mathrm{PV}_{pqbt}\ Z_{pqbdt} - \sum_{t,p,q,b,d\in L_D} \mathrm{PEN}_{pqbt}\left(D_{pqblt} - Z_{pqbdt}\right) \\
&- \sum_{t,p,l\in L_F} \left(\mathrm{PC}_{plt}\ X^F_{plt} + \mathrm{IC}_{plt}\ W^{AF}_{plt}\right) - \sum_{t,p,l\in L_F,d\in L_S} \mathrm{TC}_{ldt}\ Y^{FS}_{pldt} \\
&- \sum_{t,p,f\in L_F,l\in L_S} \mathrm{IC}_{plt}\ W^{BS}_{fplt} - \sum_{t,p,f\in L_F,l\in L_S} \left(\mathrm{PC}_{plt}\ X^S_{fplt} + \mathrm{IC}_{plt}\ W^{AS}_{fplt}\right) \\
&- \sum_{t,p,f\in L_F,l\in L_S,d\in L_M} \mathrm{TC}_{ldt}\ Y^{SM}_{fpldt} - \sum_{t,p,f\in L_F,l\in L_M} \mathrm{IC}_{plt}\ W^{BM}_{fplt} \\
&- \sum_{t,p,q,f\in L_F,l\in L_M} \left(\mathrm{PC}_{plt}\ X^M_{fpqlt} + \mathrm{IC}_{plt}\ W^{AM}_{fpqlt}\right) \\
&- \sum_{t,p,q,f\in L_F,l\in L_M,d\in L_T} \mathrm{TC}_{ldt}\ Y^{MT}_{fpqldt} - \sum_{t,p,q,f\in L_F,l\in L_T} \mathrm{IC}_{plt}\ W^{BT}_{fpqlt} \\
&- \sum_{t,p,q,f\in L_F,l\in L_T} \mathrm{PC}_{plt}\ X^T_{fpqlt} - \sum_{t,p,q,b,f\in L_F,l\in L_T} \mathrm{IC}_{plt}\ W^{AT}_{fpqblt} \\
&- \sum_{t,p,q,b,f\in L_F,l\in L_T,d\in L_D} \mathrm{TC}_{ldt}\ Y^{TD}_{fpqbldt} - \sum_{t,p,q,b,f\in L_F,l\in L_D} \mathrm{IC}_{plt} W^{BD}_{fpqblt} \\
&- \sum_{t,l\in L} \mathrm{PBC}_{lt}\Omega^A_{lt} - \sum_{t,l\in L} \mathrm{POC}_{lt}\Omega_{lt} - \sum_{t,l\in L} \mathrm{PRC}_{lt}\Omega^R_{lt} \\
&- \sum_{t,l\in L,i\in \mathrm{MG}_l} \mathrm{MPC}_{ilt}M^A_{ilt} - \sum_{t,l\in L,i\in \mathrm{MG}_l} \mathrm{MOC}_{ilt}M_{ilt} - \sum_{p,l\in L_A,t} \mathrm{SC}_{plt}S_{plt} \qquad (7.29)
\end{aligned}$$

subject to:

$$\sum_{t=1}^{r}\left\{\left(1-C_{plt}\right)Q_{plt}X^F_{plt} + C_{pl,t-1}Q_{pl,t-1}X^F_{pl,t-1} - \sum_{d\in L_S} Y^{FS}_{pldt}\right\} = W^{AF}_{plr},$$
$$p\in \mathrm{FAM},\ l\in L_F,\ r=1,\ldots,T, \qquad (7.30)$$

$$\sum_{t=1}^{r}\left(Y^{FS}_{pldt} - X^S_{lpdt}\right) = W^{BS}_{lpdr},\ l\in L_F,\ p\in \mathrm{FAM},\ d\in L_S,\ r=1,\ldots,T, \qquad (7.31)$$

$$\sum_{t=1}^{r}\left\{(1-C_{pst})Q_{fpst}X^{S}_{fpst}+C_{ps,t-1}Q_{fps,t-1}X^{S}_{fps,t-1}-\sum_{d\in L_M}Y^{SM}_{fpsdt}\right\}=W^{AS}_{fpsr},$$
$$f\in L_F,\ p\in \mathrm{FAM},\ s\in L_S,\ r=1,\ldots,T, \tag{7.32}$$

$$\sum_{t=1}^{r}\left\{\sum_{l\in L_S}Y^{SM}_{fplat}-\sum_{q\in \mathrm{PKG}_p}\frac{\mathrm{CLS}_a}{G_{pq}\mathrm{WLS}_f}X^{M}_{fpqat}\right\}=W^{BM}_{fpar},$$
$$p\in \mathrm{FAM},\ a\in L_M,\ f\in L_F,\ r=1,\ldots,T, \tag{7.33}$$

$$\sum_{t=1}^{r}\left\{(1-C_{pqat})Q_{fpqat}X^{M}_{fpqat}+C_{pqa,t-1}Q_{fpqa,t-1}X^{M}_{fpqa,t-1}-\sum_{d\in L_T}Y^{MT}_{fpqadt}\right\}$$
$$=\ W^{AM}_{fpqar},\ f\in L_F,\ p\in \mathrm{FAM},\ q\in \mathrm{PKG}_p,\ a\in L_M,\ r=1,\ldots,T, \tag{7.34}$$

$$\sum_{t=1}^{r}\left\{\sum_{l\in L_M}Y^{MT}_{fpqldt}-X^{T}_{fpqt}\right\}=W^{BT}_{fpqdr},$$
$$f\in L_F,\ p\in \mathrm{FAM},\ q\in \mathrm{PKG}_p,\ d\in L_T,\ r=1,\ldots,T, \tag{7.35}$$

$$\sum_{t=1}^{r}\left\{(1-C_{qlt})Q_{fpqblt}X^{T}_{fpqt}+Q_{fpqbl,t-1}C_{ql,t-1}X^{T}_{fpq,t-1}-\sum_{d\in L_D}Y^{TD}_{fpqbldt}\right\}$$
$$=W^{AT}_{fpqblr},\ f\in L_F,\ p\in \mathrm{FAM},\ q\in \mathrm{PKG}_p,\ b\in \mathrm{BIN}_{pq},\ l\in L_T,\ r=1,\ldots,T, \tag{7.36}$$

$$\sum_{t=1}^{r}\left\{\sum_{l\in L_T}Y^{TD}_{fpqbldt}-\varsigma_{pqbdt}\right\}=W^{BD}_{fpqbdr},$$
$$f\in L_F,\ p\in \mathrm{FAM},\ q\in \mathrm{PKG}_p,\ b\in \mathrm{BIN}_{pq},\ d\in L_D,\ r=1,\ldots,T, \tag{7.37}$$

$$\sum_{\tilde{b}\in \mathrm{BET}_b}\varsigma_{pq\tilde{b}dt}-\sum_{\tilde{b}\in \mathrm{BET}_b}Z_{pq\tilde{b}dt}\geq 0,$$
$$p\in \mathrm{FAM},\ q\in \mathrm{PKG}_p,\ b\in \mathrm{BIN}_{pq},\ d\in L_D,\ t=1,\ldots,T, \tag{7.38}$$

$$Z_{pqbdt}\leq D_{pqbdt},\ p\in \mathrm{FAM},\ q\in \mathrm{PKG}_p,\ b\in \mathrm{BIN}_{pq},\ d\in L_D,\ t=1,\ldots,T, \tag{7.39}$$

$$\sum_{p\in \mathrm{FAM}}\left\{T_{ipl}\left((1-C_{plt})X^{F}_{plt}+C_{pl,t-1}X^{F}_{pl,t-1}\right)\right\}\leq \alpha_{il}M_{ilt}(\mathrm{TPL}-S_{il}),$$
$$l\in L_F,\ i\in \mathrm{MG}_l,\ t=1,\ldots,T, \tag{7.40}$$

$$N\Omega_{lt}\geq X^{F}_{plt},\ p\in \mathrm{FAM},\ l\in L_F,\ t=1,\ldots,T, \tag{7.41}$$

$$\Omega_{lt}=\sum_{r=1}^{\max(t-\mathrm{LBT}_l,0)}\Omega^{A}_{l,r+\mathrm{LBT}_l}-\sum_{r=1}^{t}\Omega^{R}_{lr},\ t=1,\ldots,T,\ l\in L, \tag{7.42}$$

$$M_{ilt}=\sum_{r=1}^{t}(M^{A}_{ilr}-M^{R}_{ilr})+m_{il},\ i\in \mathrm{MG}_l,\ l\in L,\ t=1,\ldots,T, \tag{7.43}$$

$$M_{ilt}\leq \mathrm{MAX}^{S}_{il},\ i\in \mathrm{MG}_l,\ l\in L,\ t=1,\ldots,T, \tag{7.44}$$

$$\sum_{i\in \mathrm{MG}_l}M_{ilt}\leq \mathrm{MAX}^{T}_{l},\ l\in L,\ t=1,\ldots,T, \tag{7.45}$$

$$X^{S_1}_{plt} \geq 0, X^{S_1}_{fplt} \geq 0, X^{M}_{fpqlt} \geq 0, W^{S_1,S_2}_{plt} \geq 0, W^{S_1,S_2}_{fplt} \geq 0, W^{S_1,S_2}_{fpqlt} \geq 0, W^{S_1,S_2}_{fpqblt} \geq 0,$$
$$Y^{S_1,S_2}_{pldt} \geq 0, Y^{S_1,S_2}_{fpldt} \geq 0, Y^{S_1,S_2}_{fpqldt} \geq 0, Y^{S_1,S_2}_{fpqbldt} \geq 0, Z_{pqbdt} \geq 0, \varsigma_{pqbdt} \geq 0, S_{plt} \geq 0,$$
$$f \in L_F,\ l \in L,\ t = 1,\ldots,T,\ p \in \text{FAM},\ q \in \text{PKG}_p,\ b \in \text{BIN}_{pq}, \tag{7.46}$$
$$M_{ilt} \in \mathbb{N},\ M^A_{ilt} \in \mathbb{N},\ M^R_{ilt} \in \mathbb{N},\ t = 1,\ldots,T,\ l \in L,\ i \in \text{MG}_l, \tag{7.47}$$
$$\Omega_{lt}, \Omega^A_{lt}, \Omega^R_{lt} \in \{0,1\},\ t = 1,\ldots,T,\ l \in L. \tag{7.48}$$

The objective function (7.29) includes revenue generated from selling products, the costs of not meeting the demand, the production costs, and the costs for building and operating or removing facilities and machines. The first line of objective function (7.29) represents the revenue generated by meeting demand and the penalty for not meeting the demand. The second line indicates the wafer fab production costs, inventory carrying costs for finished wafers, and the costs for transporting wafers between the wafer fab and sort sites. The third line is related to the inventory carrying costs before sort, the products costs for sort, and the inventory carrying cost before assembly. The fourth through eighth lines represent the costs associated with the assembly and test operations, the transportation between these facilities, the transportation costs to the demand centers, and the inventory carrying costs at each of these facilities. The ninth line indicates the costs for building facilities, operating the facilities, and closing facilities, and the first two terms of the tenth line represent the costs for purchasing machines and operating them. Finally, the last term in the last line of expression (7.29) represents the costs for subcontracting the fabrication of wafers, but does not include any costs for establishing subcontract relationships.

The model contains network flow constraints, capacity constraints, product substitution constraints, demand constraints, production suppressing constraints, facility counting constraints, machine counting constraints, and constraints on the number of machines that can be purchased. The network flow constraints (7.30)–(7.37) enforce the material flow conservation, i.e., total inflow is equal to total outflow. The even-numbered network flow constraints deal with the production of products in a facility and the shipment of products to the next facility, while the odd-numbered network flow constraints deal with the balance of flow between the inflow of materials into a facility and the amount of products started for production.

Constraints (7.38) determine which products will be downgraded in order to meet demand, and the demand by product constraints are given by inequalities (7.39). Wafer fabrication is limited by constraints (7.40), and while they are not shown here, there are similar constraints sets for sort, assembly, and test.

In order to prevent production in a non-existent facility, constraints (7.41) are needed for fab production with similar constraints sets (not shown) for sort, assembly, and test. Constraints (7.42) keep track of the facilities that

are built and shutdown. In a similar way, constraints (7.43) keep track of the number of machines purchased and sold. Constraints (7.44) and (7.45) limit the number of machines in each machine group and place a limit on the total number of machines in a facility, respectively. Finally, constraints (7.46)–(7.48) are nonnegativity and integer restrictions for the decision variables.

Next, we discuss some computational results from Stray et al. [293]. We present a break-even analysis for a situation where demand ranges from very low to very high. Very low demand is related to a situation where the production facilities are running with substantial excess capacity. In the very high demand case, the production facilities are running with too little capacity. In addition, the amount of subcontracting is limited. Practically, this means setting the parameter demand level D_{pqblt} to an even level across all periods and then solving the problem. The solution is then analyzed, and the number of machines bought, wafers subcontracted, and what facilities were built are noted. The model is rerun with a different demand level, and the same characteristics are noted. After running enough problem instances with different demand levels, curves are generated and examined to see what the demand level is that makes the model go from current capacity to subcontracting, from subcontracting to buying new machines, and from buying machines to buying entire wafer fabs, assembly facilities, and test facilities. The network for this analysis is shown in Fig. 7.1.

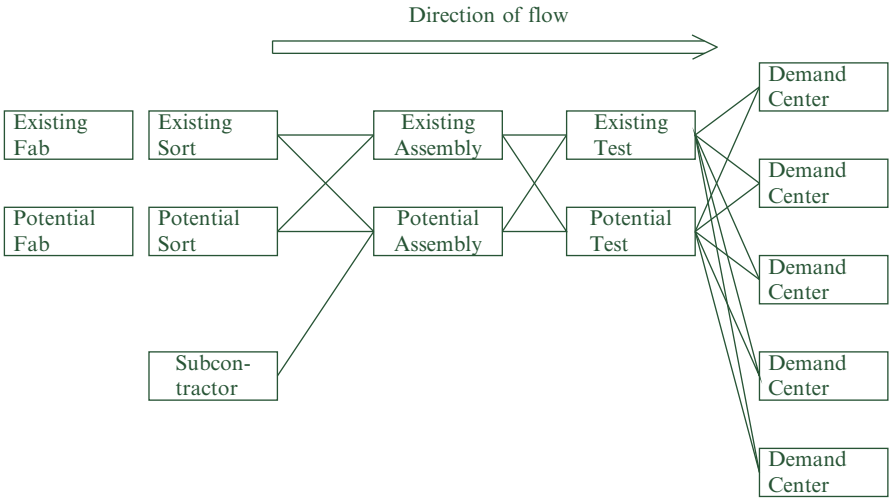

Figure 7.1: Example network

In this scenario, one wafer fab is already up and running, together with one sort, one assembly, and one test facility, and five demand centers. The capacity in this model is balanced so that the maximum capacity of each individual facility matches the maximum capacity of a single facility, preceding it in

the manufacturing supply chain. One sort facility can thus handle the output from one wafer fab, and one assembly facility can handle the output from a single sort facility.

There is also one foundry available with limited capacity. The product from the foundry is ready for the assembly operation. The price of wafers from the foundry is higher than the cost of producing them in-house as long as existing capacity is utilized. However, if a wafer fab has to be built, the cost per unit will increase. When the costs are high enough, subcontracting becomes an interesting option. All existing and potential wafer fabs have the same maximum capacity per period. The maximum capacity is defined as the capacity of a facility when the allowed maximum number of machines has been installed. Demand is varied evenly over the five demand centers.

There are two product families in the model, each divided into two packages. In the binning process, two different qualities result from each package. Wafer fabs are considered to have two bottleneck machine groups, while sort, assembly, and test have one each. The planning problem is NP-hard, because it contains knapsack- and facility location problem-type subproblems. Therefore, we will allow feasible solutions that are provably within 5 % of optimum. Figure 7.2 shows the obtained solutions for 20 different demand settings, varied along the x-axis, with subcontracting of wafers limited. For each demand scenario, the MIP was run using AMPL and CPLEX.

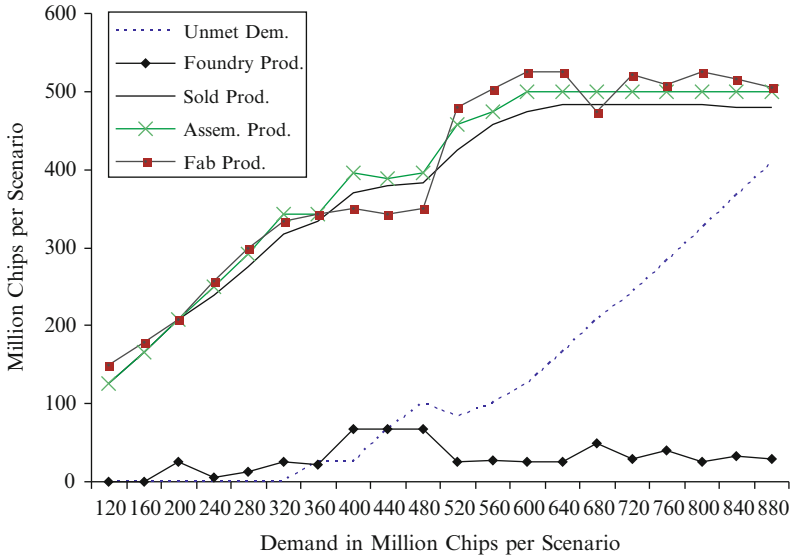

Figure 7.2: Optimal production solutions with limited subcontracting

In this analysis, a wafer fab is built at five hundred twenty million chips, two hundred million more than the maximum capacity of a wafer fab that is

three hundred twenty million chips. After the wafer fab is built, foundry production goes down and stays down comfortably under the allowed maximum. The reason why foundry production is not zero is that it covers for the production capacity that is lost during the building of the wafer fab. The dip in in-house production and increase in subcontracting at the six hundred eighty million chips demand scenario can only be explained by the 5 % allowed MIP gap, i.e., accepting the solution even though it is not optimal. The solution at six hundred forty million chips consisted of shutting down one of the wafer fabs in the last time period, replacing its capacity with subcontracting, saving the operation costs for the facility and its machines, and spending the money on the subcontracted wafers.

Rastogi et al. [258] present a stochastic version of the enterprise model of Stray et al. [293] where the total expected profit is maximized when product demand is uncertain. A two-stage, multiperiod stochastic MIP with recourse was developed to provide solutions that reduce the overall risk in planning (cf. Sect. 3.2.4). The first stage decisions include purchasing of machines at various production facilities, outsourcing production, or even construction of a new production facility depending upon the demand. The second stage (recourse) actions include increasing the internal capacity by purchasing machines at a premium as well as external capacity by subcontracting and cancellation of contracts for outsourcing made in the first stage.

The model provides information regarding the trade-offs between risk and expected short- and long-term returns. It is coded in AMPL and solved using CPLEX. When the uncertainty in demand increases, a more conservative approach is adopted, and the model displays an inherent tendency of no commitment, i.e., the capacity increment is negligible.

In addition to the uncertainty in demand, the effect of correlation between the demands of two products is studied. It is evident from the analysis, and also as stated by Simchi-Levi et al. [283], that positive correlation between the products, for example, increasing market size, involves higher risk compared to negative, for example, introduction of new products, or no correlation.

The usefulness of the model compared to the alternatives available was evaluated. The model was compared to the expected value model of Stray et al. [293] and to the perfect information case, which revealed that as the uncertainty in demand increases, the model improves its performance over the expected value model. However, the gap between the stochastic solution and perfect information solution also increases with the increment in variability of demand. By increasing the number of scenarios to map the uncertainty of demand, the results show that the efficiency of stochastic solutions increases. Adding uncertainty to the deterministic version of the model with multiple scenarios yielded more realistic and robust results, and analysis on correlation between multiple product demands resulted in unintuitive decisions for strategic make/buy problems.

7.5 Modeling of Load-Dependent Cycle Times

CT is load-dependent. It increases nonlinearly with resource utilization as known from queueing theory. This causes some problems in model formulations for production planning because CT information serves as a parameter of the models. At the same time, production planning approaches determine the load of the BS by determining release quantities. In this section, we discuss several methods to tackle this conflict. We study CT-TP curves, iterative simulation schemes, and finally clearing functions. For a detailed review of production planning models with load-dependent CT in manufacturing, we refer to Pahl et al. [226].

7.5.1 Cycle Time Throughput Curves

This section is an abridged version of Ankenman et al. [8]. CT-TP curves are often employed as decision-making tools in manufacturing settings (cf. Brown et al. [33]). A CT-TP curve displays the projected average CT plotted against TP rate, or start rate. These curves are useful for planning at both the strategic and tactical levels.

Decisions regarding the impact on CT of a 2 % increase in start rate can be widely different depending on the shape of the curve and the distance from the knee. For example, if a wafer fab has a curve as illustrated in Fig. 7.3 and is operating at the level of 22,000 wafer starts per month, it will experience only a minor change in average CT by ramping up an additional 500 wafer starts.

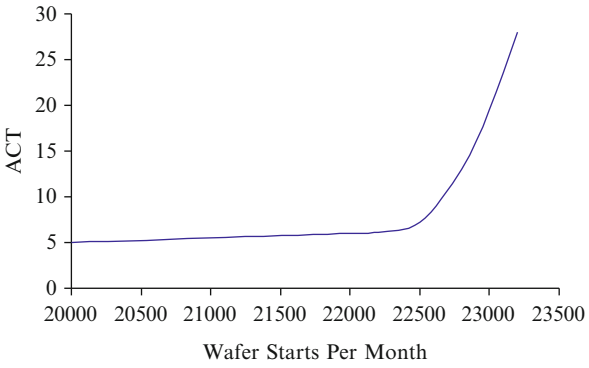

Figure 7.3: Sample CT-TP curve

Alternatively, if the wafer fab is operating on the same curve but is at 22,500 wafer starts per month, a 500-wafer start increase dramatically alters CT. In both cases, we called for a 500-wafer start increase, yet drastically different outcomes resulted from what seemed to be the same action. Man-

agement therefore needs to develop CT-TP curves if they are interested in predicting the impact of start rate changes on average CT.

Unfortunately, the simple collection and analysis of past TP history is insufficient for curve generation. It is unlikely that an operating wafer fab has experienced a sufficient number of changes along the same curve to allow creation of the curve. For example, a wafer fab seldom operates on the flat portion of the curve where equipment utilizations are in the less than 70 % range. It is also unlikely that the wafer fab has carefully ramped up production start rates over the most rapidly changing portion of the curve, so the estimation of the shape in this region becomes problematic. In fact, every time the wafer fab changes its dispatching policy or adds more equipment, it may not just be moving along the curve, it may in fact be shifting to an entirely new curve. The technique of empirical CT-TP curve generation requires the collection of large amounts of representative data. As a result, other than for the simplest of systems, simulation is the preferred method of data generation.

While simulation is the most common technique for generating CT-TP curves, the methods used to select the points to simulate and the effort to allocate to these points vary. Several different design points must be simulated to generate a CT-TP curve. A careful selection of the design points can lead to minimal simulation expense. Various authors have discussed methods for generating a CT-TP curve and how to select these design points (cf. Park et al. [230], Fowler et al. [86], and Yang et al. [325]).

Other authors have presented methods for determining an appropriate allocation of simulation effort to the design points of the CT-TP curve being simulated so as to obtain nearly equal absolute or relative precision (cf. Leach et al. [153]).

The method commonly used by practitioners to generate a CT-TP curve via simulation is to allocate an equal amount of simulation effort to each TP rate being simulated. This situation is shown in Fig. 7.4. As TP rate approaches capacity, the CT and the variance of CT (Var(CT)) increase. Figure 7.4 illustrates that by equally allocating simulation effort to all design points, yielding a CT-TP curve that is less precise as we approach capacity, a clearly undesirable characteristic. We consider single-product CT-TP curves. Note that when we say single product, we could be considering a CT-TP curve of a facility that produces only one product, or we could be focusing on one product out of many provided that the relative mix of the various products, as defined below, remains the same at all levels of the system's TP.

We define the following quantities:

$$\begin{aligned}
\lambda &:= (\lambda_1, \ldots, \lambda_K) && \text{: vector of start rates for } K \text{ products} \\
x &&& \text{: utilization of the bottleneck in the wafer fab, } 0 < x < 1 \\
\alpha &:= (\alpha_1, \ldots, \alpha_K) && \text{: product mix vector where } \alpha_k := \lambda_k / \textstyle\sum_{h=1}^{K} \lambda_h
\end{aligned}$$

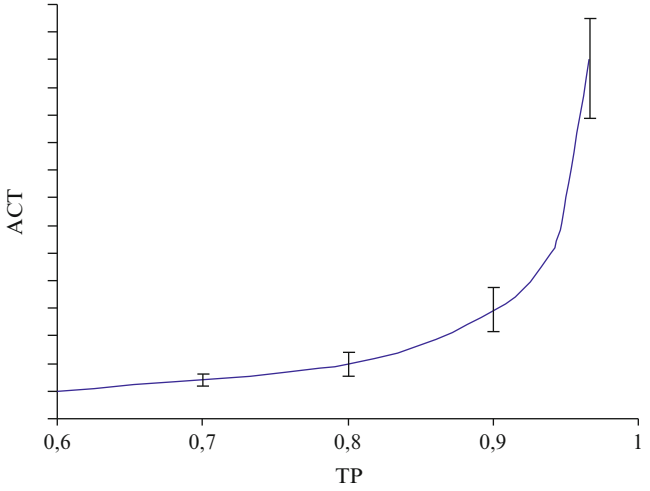

Figure 7.4: CT-TP curve using equal allocation of simulation effort

Without loss of generality, we will only consider the CT of product 1 and denote its steady-state CT as $C(\lambda) = C(x, \alpha)$, a random variable with unknown distribution that depends on the start rates. Notice that if we know the processing capacity of the bottleneck station, then specifying (x, α) is equivalent to specifying λ. We will drop the dependence on α in the single-product case. For the CT random variable to have a limiting distribution steady state, among other things, the system logic and driving inputs must not be changing over time. Generically, let

$$c_r(\lambda) = c_r(x, \alpha) := E(C^r(x, \alpha)),\ r \in \mathbb{N},\ r \geq 1 \tag{7.49}$$

be noncentral moments of the steady-state CT; we drop the subscript r when we refer to the mean, i.e., first moment.

To estimate moments of CT, we will make one or more replications of a typically large number of individual product CT values. Let $C_{ij}(x, \alpha)$ be the jth observed CT from the ith replication for $i = 1, \ldots, m(x, \alpha)$ and $j = 1, \ldots, l(x, \alpha)$. The quantity $m(x, \alpha)$ is the number of replications for given x and α, while $l(x, \alpha)$ denotes the length of a single simulation run for given x and α. Our steady state assumption corresponds to requiring that $C_{ij}(x, \alpha)$ converges in distribution to $C(x, \alpha)$ as $j \to \infty$ for any i.

The most straightforward way to generate a CT-TP curve via simulation is to select a fine grid of TP values, say $0 < x_1 < \cdots < x_d < 1$, and run simulation experiments at each one to estimate $c_r(x)$. We could, equivalently, select a grid of release rates λ that correspond to TP in a steady state. However, later when we fit CT-TP curves to the data, there are a number of advantages to standardizing TP so that system capacity always corresponds to a TP of 1.

Unfortunately, this approach has pitfalls. First, it requires a large number of simulation runs to develop a fine grid. A second repercussion of point-by-point CT-TP curve estimation is that some sort of interpolation is needed to estimate CT properties at TP values x that were not simulated. If the grid points are packed closely enough, then perhaps a simple linear interpolation is adequate. However, as mentioned earlier, exceptionally long runs may be required at the higher levels of TP, which argues against running simulations at a very fine grid.

In a series of papers, Kleijnen and van Beers (cf. Kleijnen and van Beers [143] and van Beers and Kleijnen [308, 309]) describe how the interpolation method of Kriging can be adapted to the output of discrete-event, stochastic simulations in general and queueing simulations in particular. In its simplest form, Kriging estimates $c(x)$ by a weighted average of the estimated ACT values at the grid points $x = (x_1, \ldots, x_d)$. Loosely speaking, the Kriging estimator gives more weight to CT estimates at grid points $x_h, h = 1, \ldots, d$ that are closer to the point x to be interpolated.

Because the Kriging approach is an interpolation method, it favors a finer grid, i.e., more design points x, than the queueing-motivated models we describe below. Furthermore, there is no guarantee that the Kriging estimator will exhibit known properties of the response function, for instance, that $c(x)$ is nondecreasing in x. However, the Kriging approach has the advantages that it is general purpose, it will not be subject to the lack of fit inherent in a poorly chosen meta-model, and it works largely without change for interpolating higher moments than the mean. Further, Kriging extends naturally to a multidimensional independent variable, like the product mix α.

Fowler et al. [86] investigated the use of variance reduction techniques based on common random numbers and antithetic variates (cf. the discussion in Sect. 3.3.1) in efficiently generating CT-TP curves that linearly interpolate a set of (TP, CT) points. In their paper, the term efficient reflects the capability to provide a simulation-based CT-TP curve with an acceptable precision and accuracy by using limited available resources. The goal was to generate CT-TP curves more economically, so the cost of analysis could be reduced, thus allowing companies to make better manufacturing capacity management decisions. The experimentation in their paper included simulating an M/M/1 queueing system (cf. Sect. 3.2.7) and a system with five stations in series, a special case of a Jackson queueing network. The results showed that common random numbers were effective when there was an adequate computing budget, but they introduce too much bias when the computing budget is not large enough. On the other hand, the results showed that antithetic variates were effective for small or large computing budgets.

Park et al. [230] use the D-optimality criterion (cf. Sect. 3.3.1) to choose design points for building the CT-TP curve. Since it concerns the confidence interval of the parameters of a model, D-optimality assumes that a model is specified for the response curve. Park et al. [230] suggest two nonlinear regression models, one of which is bowl shaped and is appropriate when using

batching policies such as the full batch policy MBSF or the minimum batch size policy MBS with a minimum greater than 1 (cf. Sect. 4.5 for a description of these two batching policies) that increase the CT at low levels of TP. The other model, which models the CT as a monotonically increasing function of the TP, is used when there is no batching or a greedy batch policy is employed. The two models are given below. Notice that both models have the CT exploding as the TP, x, nears the capacity β_2. If the TP is normalized to the capacity, then $\beta_2 = 1$. We obtain for the two models:

$$c(x) := \frac{\beta_1 x}{\beta_2 - x} - \beta_3, \tag{7.50}$$

$$c(x) := \frac{\beta_3}{x} + \frac{\beta_1 x}{\beta_2 - x} - \beta_4, \tag{7.51}$$

where $\beta_i \in \mathbb{R}$ are appropriate parameters of the models. Both of these models are generalizations of the CT-TP curve of a G/G/1 queue.

The experimental design for fitting these models is a selection of TP values at which exhaustive simulations are conducted and the steady state ACT value is recorded. Nonlinear regression is used to estimate the parameters, and thus the linear approximation to the variance/covariance matrix is used to approximate the D-criterion. The candidate design points are placed at regular intervals from zero TP to one, where one represents full capacity. The experimental design procedure recommended is a sequential procedure that starts with the minimum number of design points that are required to support the model, i.e., three in the case of model (7.50) and four in the case of model (7.51). The D-criterion can be expressed as a function of the location of the design points. The initial points are selected as the set of three or four in the case of model (7.51) candidate points that maximize the D-criterion. All the other candidate points are then ranked according to the D-criterion for entry into the design if needed. After simulations are conducted at the initial points, the model parameters are estimated. Each additional candidate point is added sequentially in the predefined order until the parameter estimates no longer change by an appreciable amount, i.e., 1 % was used as a stopping criterion. This method was validated through construction of a CT-TP curve for a wafer fab.

Another approach to building CT-TP curves was proposed by Cheng and Kleijnen [46], hereafter called the CK approach, where they generalized the CT-TP curve of the M/M/1 curve as shown below:

$$c(x) := f(x) \sum_{l=0}^{t} \beta_l x^l + \varepsilon(x) = \frac{\sum_{l=0}^{t} \beta_l x^l}{1 - x} + \varepsilon(x), \tag{7.52}$$

where $f(x) := 1/(1-x)$ and it is assumed that TP, represented by variable x, is scaled from zero to one. The quantity $\varepsilon(x)$ is an error term. Again $x = 1$ represents full capacity. The CK approach only deals with the case of no

batching or a greedy batch policy, and thus the model in (7.52) is a more general form of the model (7.50) proposed by Park et al. [230].

To fit model (7.52) to the simulation data, the CK approach develops a linear regression model since the only nonlinear part of the equation, $f(x)$, is known and can be dealt with through a transformation. The variance of the error term in model (7.53) depends on x as

$$\text{Var}[\varepsilon(x)] = [h(x)\sigma]^2, \tag{7.53}$$

where $h(x)$ is assumed known from asymptotic theory or other considerations.

The design of the experiment consists of the location of the design points $x = (x_1, \ldots, x_m)$ and the fraction of a total of N replications assigned to those points $\pi := (\pi_1, \ldots, \pi_m)$. The design is constructed to minimize a criterion called PM, which is a scaled version of the weighted-average variance of the estimated expected response over the TP range of interest.

The CK procedure for fitting the model (7.52) can be summarized as follows. Given $f(x)$, $h(x)$, a maximum value of t, and a fixed budget of N replications, find the optimal design (x, π) by minimizing PM. With the design points x fixed, carry out simulation experiments sequentially and adjust the allocation x. Once the total number of runs has been exhausted, use backward selection to decide the appropriate polynomial order of model (7.52) and obtain the fitted curve.

The CK method leaves open the question of how to specify $f(x)$ and $h(x)$, which affect the design of the experiment and, more importantly, the adequacy of model (7.52) to represent the true CT-TP curve. When these two functions are known, CK is highly effective and efficient, and works within a fixed budget. However, for complicated manufacturing systems, there is not likely to be sufficient information to infer such characteristics. In other words, obtaining good choices for $f(x)$ or $h(x)$, although not impossible, is difficult in practice. Further, we have strong empirical evidence from Allen [6] and Johnson et al. [132] that the $f(x)$ and $h(x)$ used by the CK method can be far from correct in realistic manufacturing simulations. Since model (7.50) used in Park et al. [230] is a specific instance of Eq. (7.52), the same weakness can be attributed to their method as well.

In summary, the procedures by Park et al. [230] and CK are both interesting and useful methods of experimental design for fitting a model such as is given in Eq. (7.52), but there may arise cases in practice where these models are not sufficiently accurate to produce useful CT-TP curves.

A precision-driven design of experiment strategy was proposed in Yang et al. [325] to sequentially build up simulation experiments for the efficient generation of CT-TP curves. It allows the user to specify a precision level and is able to provide a fitted curve with desired precision by running simulation. We summarize the method in the remainder of this section.

The estimation of the CT-TP curve is based on the two statistical regression models (7.54) and (7.55), the forms of which are both motivated by heavy

traffic queueing analysis and supported by extensive investigation of realistic manufacturing systems. One is called the expected CT (ECT) model:

$$c(x) = E[\overline{C}_i(x)] = \frac{\sum_{l=0}^{t} \beta_l x^l}{(1-x)^p}, \quad i = 1,,\ldots,m(x), \tag{7.54}$$

that characterizes the relationship between the expected CT and normalized TP x over a range of interest $[x_L, x_U]$. Unknown parameters are the polynomial coefficients $\boldsymbol{\beta}$, polynomial order t, and the exponent p. As explained earlier, the sample mean CT $\overline{C}_i(x)$ obtained from the ith simulation replication performed at x will be used as the data points to which the CT-TP models are fit. The variance of $\overline{C}_i(x)$ depends on x and is represented by the following variance model:

$$\text{Var}[\overline{C}_i(x)] = \frac{\sigma^2}{(1-x)^{2q}}. \tag{7.55}$$

Both σ^2 and q are unknown parameters. With the sample mean CT data $\{\overline{C}_i(x),\ i = 1,\ldots,m(x)\}$ at different values of x, the sample variance of $\overline{C}_i(x)$ can also be estimated over x, from which the variance model (7.55) can be fitted. With the estimated parameter $\hat{q}$, transforming the response $\overline{C}_i(x)$ by multiplying by $(1-x)^q$ will yield a constant variance and result in a standard nonlinear regression model:

$$c(x)(1-x)^q = E[\overline{C}_i(x)(1-x)^q] = (1-x)^{q-p} \sum_{l=0}^{t} \beta_l x^l = (1-x)^r \sum_{l=0}^{t} \beta_l x^l, \tag{7.56}$$

where $\boldsymbol{\beta}$, t, and the exponent r are unknown parameters. Thus, given a $\{\overline{C}_i(x),\ i = 1,\ldots,m(x)\}$ dataset, the model fitting is performed in two steps:

1. Fit the variance model (7.55) and obtain the q estimate.
2. Use the estimated parameter $\hat{q}$ to stabilize the variance for the original observations $\overline{C}_i(x)$ and then fit model (7.54).

The estimators of the ECT model (7.54) are obtained indirectly by noting that the coefficients $\boldsymbol{\beta}$ in model (7.54) coincide with those in (7.56), and p is estimated by the difference between the q and r estimates.

The goal is to obtain a precisely estimated CT-TP curve that helps manufacturers decide at what TP they should run the system. Thus, Yang et al. [325] evaluate the goodness of the fitting by the relative error achieved on the ECT response estimators. Since the curve fitting is based on the nonlinear regression performed on models (7.55) and (7.56), variance estimates can be obtained on the estimated parameters in Eqs. (7.55) and (7.56). Since model (7.54) is derived indirectly from Eqs. (7.55) and (7.56), a conservative variance estimate can be inferred for $\hat{c}(x)$, the ECT predicted at x under some empirical approximation (see Yang et al. [325]). Yang et al. [325] let the user specify a target precision, say $\gamma\%$, which is defined as the relative error on the ECT estimator:

$$\gamma\% := \frac{\sqrt{\text{Var}[\hat{c}(x)]}}{\hat{c}(x)}. \tag{7.57}$$

Once fitted curves have been obtained, the relative error on the ECT estimate $\hat{c}(x)$ can be approximated for any TP x over $[x_L, x_U]$. The user can choose to check the precision achieved at a TP level of particular interest or at a number of points in $[x_L, x_U]$ before they declare that a fitted curve with desired precision has been generated.

For the efficient estimation of the CT-TP models presented above, design of experiments methodologies is developed to collect simulation data sequentially. The experimental design consists of the location of design points, the TP levels at which simulations will be executed, the allocation of computational effort, and the number of simulation replications assigned to each design point. The best choice of experimental design depends on the true ECT and variance curves, which are unknown at the stage of designing experiments. In light of this, Yang et al. [325] approach the design of experiments problem in a sequential manner. The model curves are estimated ever more precisely as more simulation data are obtained, and further experimentation is guided by the current best estimate of the models. This design and modeling process is continued until the prespecified precision $\gamma\%$ is achieved on the ECT response estimator.

To demonstrate the effectiveness of the Yan procedure, Yang et al. [325] applied it on a number of systems to generate their corresponding CT-TP curves. The systems explored included analytically tractable queueing models and realistic semiconductor manufacturing systems. For simple queueing models such as M/M/1/FIFO, M/M/1/SPT, and M/M/1/LPT, the true CT-TP curves can be derived analytically, and hence the quality of the simulation-based model estimation can be evaluated easily. The real wafer fab considered is provided by the MASM Lab testbed (see Fowler and Robinson [83]). Since the true underlying curve is unknown in this case, nearly true ECT estimates were obtained by running simulations until the standard error of the expected CT estimates were essentially zero. These estimates provide a benchmark to which the ECT estimates obtained from the Yan method are compared. All the computational experiments show that the Yan method is able to generate high-quality CT-TP curves with desired precision. Comparisons were also performed that show that the Yan approach can be more efficient than the procedure proposed by Cheng and Kleijnen [46].

The focus of this section has been on CT as a function of TP, but product mix (PM) can also affect CTs even if the overall system TP is unchanged. However, fitting CT-TH-PM surfaces via simulation is a much more challenging problem and is beyond the scope of this section. More details of this problem are presented in Yang et al. [326].

Note that considerable simulation effort is necessary to determine meaningful CT-TP curves. These curves are valid for all possible TP situations

however, it is not evident to see how these curves can be used in some production planning approaches. In the next two sections, we will describe two more methods. The first method, iterative simulation, tries to determine CT values that are appropriate for certain released quantities. The second method, the clearing function approach, is similar to the CT-TP curve approach; however, it tries to find the relationship between load and CT and also covers the incorporation of the clearing function into production planning approaches.

7.5.2 Iterative Simulation

The first iterative procedure for production planning in semiconductor manufacturing was proposed by Hung and Leachman [120]. An LP model is formulated that requires estimated lead times, i.e., cycle times, F_{pl} for a job of product p to reach process step l after being released into the wafer fab. It is shown by Irdem et al. [124] that an unambiguous convergence of the approach of Hung and Leachman is hard to achieve. This is true even for situations where the demand is constant for all products over the planning horizon. Because of the limitations of the Hung and Leachman approach, we discuss a second formulation that when used within an iterative simulation setting shows a consistent convergence.

We describe an LP model for production planning that is due to Kim and Kim [139]. Recently, this formulation is extended to a production planning situation in semiconductor manufacturing by Irdem et al. [124]. The actual workload profiles on each machine over the planning periods are taken into account. Therefore, the effective loading ratio $e_{pk(g,t)}$ is introduced in [139]. This quantity is defined as the proportion of the start quantity of product p released in a period $g \leq t$ that contributes to the workload at machine group k in period t. Furthermore, the effective utilization u_{kt} of machine group k in period t is taken into account. The quantity u_{kt} is defined as the proportion of the total capacity of machine group k that is available to process the start quantities during period t. The adjusted capacity of machine group k in period t is obtained by simply multiplying the capacity C_{kt} by u_{kt}. It is clear that the effective loading ratios and the effective utilization can be used to model load-dependent cycle times.

Following Irdem et al. [124], the corresponding LP model is formulated. The following indices and index sets are used:

$t = 1, \ldots, T$: period index
p : product index
P : set of all products
$k = 1, \ldots, K$: machine group index
$l = 1, \ldots, l_p$: process step index for wafers of wafer type p

The parameters used within the model are:

a_{plkt} : average machine hours for process step l of a single wafer of wafer type p on machine group k processed in period t
C_{kt} : hours of machine group k available in period t
r_{pt} : unit revenue from product p in period t
c_{pt} : unit incremental production cost of product p in period t
h_{pt} : unit inventory holding cost for product p in period t
b_{pt} : unit backlog cost for product p in period t
d_{pt} : demand for wafer type p in period t
$e_{pk(g,t)}$: effective loading ratio of product p on machine group k in period t due to starts in period g
$e_{pM(g,t)}$: effective loading ratio of product p on the last processing machine in period t because of starts in period g
u_{kt} : effective utilization of machine group k in period t
B_{p0} : initial backlog for wafer type p
I_{p0} : initial inventory for wafer type p

The following decision variables are used in the model:

X_{pt} : release quantity for wafers of type p in period t
Y_{pt} : output quantity for wafers of type p in period t
I_{pt} : units of product p in inventory of finished goods at the end of period t
B_{pt} : units of product p backlogged at the end of period t

The production planning model can be formulated as follows:

$$\max \sum_{p\in P}\sum_{t=1}^{T}(r_{pt}Y_{pt}-c_{pt}X_{pt}-h_{pt}I_{pt}-b_{pt}B_{pt}) \tag{7.58}$$

subject to:

$$\sum_{p\in P}\sum_{g=1}^{t}\sum_{l=1}^{l_p} e_{pk(g,t)}a_{plkt}X_{pg} \le u_{kt}C_{kt}, \quad t=1,\ldots,T,\ k=1,\ldots,K, \tag{7.59}$$

$$Y_{pt}+I_{p,t-1}-B_{p,t-1}+B_{pt}=d_{pt}+I_{pt}, \quad p\in P,\ t=1,\ldots,T, \tag{7.60}$$

$$\sum_{g=1}^{t} e_{pM(g,t)}X_{pg}=Y_{pt}, \quad p\in P,\ t=1,\ldots,T, \tag{7.61}$$

$$X_{pt}\ge 0, I_{pt}\ge 0, B_{pt}\ge 0, \quad t=1,\ldots,T,\ p\in P. \tag{7.62}$$

The objective function (7.58) is related to profit. The objective function value is the difference of revenue and the sum of production, inventory holding, and backlog costs. Constraints (7.59) model the resource capacity, whereas constraints (7.60) are material conservation equations. The release-output relationship is expressed by constraints (7.61). Finally, nonnegativity conditions

for decision variables are taken into account by constraints (7.62). As typical in production planning models, we do not use integer variables.

The iterative procedure proposed by Kim and Kim [139] can be formulated as follows (cf. Sect. 3.2.8 for the general principle of iterative simulation). Note that we use the term KK procedure as an abbreviation.

KK Procedure

1. Initialize the counter for the current iteration $\text{curr} := 1$ and select the maximum number of iterations $\text{iter}_{\max}$. Calculate initial effective loading ratios $e_{pk(g,t),\text{curr}} := e^{(0)}_{pk(g,t)}$ and machine utilizations $u_{kt,\text{curr}} := u^{(0)}_{kt}$ using a steady state simulation. The period demands are used as release quantities within the simulation.
2. Solve the LP (7.58)–(7.62) using $e_{pk(g,t),\text{curr}}$ and $u_{kt,\text{curr}}$ to determine release quantities $X_{pt,\text{curr}}$ and wafer output quantities $Y_{pt,\text{curr}}$.
3. Use a prescribed number of independent replications of a simulation run to obtain updates for $e_{pk(g,t),\text{curr}}$ and $u_{kt,\text{curr}}$, taking the release quantities $X_{pt,\text{curr}}$ of step 2 into account. Take the average for $e_{pk(g,t)}$ and u_{kt} over all simulation runs to determine $e_{pk(g,t),\text{curr}+1}$ and $u_{kt,\text{curr}+1}$. Collect also output quantities $SY_{pt,\text{curr}}$, where S indicates that these quantities are determined from the simulation.
4. If $\text{curr} < \text{iter}_{\max}$, then set $\text{curr} := \text{curr} + 1$ and go to step 2. Otherwise, the iterative scheme terminates.

The mean absolute deviation between $Y_{pt,\text{curr}}$ and $SY_{pt,\text{curr}}$ is used to measure convergence of the KK iterative procedure. A relatively small number of iterations is generally enough. It is shown in [124] by extensive simulation experiments that the KK procedure shows a consistent convergence behavior. Hence, it seems that this procedure has some potential for being applied in practice.

7.5.3 Clearing Functions

Clearing functions are used to model the relationship between the expected output of a manufacturing system and the WIP inventory. These functions have the advantage that they are able to capture the nonlinear relationship between resource utilization, i.e., load, and CT.

Clearing functions were proposed for the first time by Graves [108]. The following linear function f is used in [108]:

$$Y_t = cW_t, \tag{7.63}$$

where $c > 0$ is a constant, called the proportional factor, and Y_t is the output at the end of period t. Finally, W_t is a measure for the WIP at the beginning of period t. An infinite capacity assumption is a consequence of this model,

because the manufacturing system is assumed to be able to complete the amount cW_t even when W_t is very large. The major drawback of this model is that the planned CT is fixed based on Little's law (cf. Eq. (3.21) in Sect. 3.2.7), even when W_t is changing. Therefore, CT values are not appropriate taken into account in model formulations that use this clearing function.

Later, nonlinear clearing functions were proposed that take the finite capacity of the BS into account. The general idea of a clearing function f is introducing a clearing factor $c(W)$ to obtain a clearing function f of the form:

$$f(W) := c(W)W, \ W \geq 0. \tag{7.64}$$

The clearing factor is a nonlinear function of the WIP W. Clearly, we have $f(0) = 0$ for each clearing function of this form.

Kamarkar [136] proposed the following clearing function:

$$f(W) := \frac{C_1 W}{C_2 + W}, \ W \geq 0, \tag{7.65}$$

where C_1 and C_2 are positive constants. This clearing function is a nondecreasing, concave function of WIP. We can see easily that C_1 is the maximal possible output that is obtained for $W \to \infty$. It represents the maximum capacity. The quantity C_2 is a user-specific parameter controlling the curvature of the clearing function. The following clearing function is due to Srinivasan et al. [292]:

$$f(W) := C_1(1 - \exp^{-C_2 W}), \ W \geq 0, \tag{7.66}$$

where again C_1 and C_2 are positive constants. By considering $W \to \infty$, we obtain that C_1 is the maximum possible output. By using the expression $\exp x = \sum_{k=0}^{\infty} x^k/k!$, it is evident that the clearing function (7.66) is of the form (7.64). Note that the two constants C_1 and C_2 can be determined, in principle, by fitting the function to empirical data.

In the remainder of this section, we assume that the clearing function f is concave and $f(0) = 0$ holds. Furthermore, it is a smooth function with the property $\frac{df(W)}{dW} > 0$, i.e., f is monotone increasing.

In the following, we want to derive an LP formulation similar to the model (7.58)–(7.62). Therefore, we have to incorporate the clearing function into the LP model. Following Asmundsson et al. [13], we replace the capacity constraints (7.59) in a single-stage multi-product situation with the product set P and T periods by

$$\sum_{p \in P} \xi_{pt} Y_{pt} \leq f_t \left(\sum_{p \in P} \xi_{pt} W_{pt} \right), \ t = 1, \ldots, T, \tag{7.67}$$

where the following notation is used:

Y_{pt} : total production quantity of product p in period t
ξ_{pt} : amount of resource (machine time) required to produce one unit of product p in period t
W_{pt} : WIP of product p in period t
f_t : clearing function for period t

As stated in [13], there is no link between the mix of WIP available in the period and the corresponding production in the capacity restriction (7.67). To avoid this problem, the overall clearing function is decomposed. We obtain:

$$\xi_{pt} Y_{pt} \leq Z_{pt} f_t \left(\sum_{p \in P} \xi_{pt} W_{pt} \right), \; p \in P, \; t = 1, \ldots, T, \tag{7.68}$$

$$\sum_{p \in P} Z_{pt} = 1, \; t = 1, \ldots, T, \tag{7.69}$$

where the new decision variable $Z_{pt} \geq 0$ represents the allocation of the expected TP represented by the clearing function among the different products.

The capacity constraints (7.68) still have the disadvantage that f_t has the total WIP as argument and not the WIP for a specific product. To solve this problem, it is assumed in [13] that the expected TP between products is proportional to the mix of products represented in the WIP in period t. We obtain

$$\sum_{p \in P} \xi_{pt} W_{pt} = \frac{\xi_{pt} W_{pt}}{Z_{pt}}, \; p \in P, \; t = 1, \ldots, T. \tag{7.70}$$

The quantity $\frac{\xi_{pt} W_{pt}}{Z_{pt}}$ can be interpreted as the extrapolated total WIP in period t. Using Little's law (cf. Eq. (3.21) in Sect. 3.2.7), it is shown in [13] that this extrapolation is exact when all products have the same average CT at the resource. The resultant capacity constraints substituting the right-hand side of Eq. (7.70) into capacity constraints (7.68) are

$$\xi_{pt} Y_{pt} \leq Z_{pt} f_t \left(\frac{\xi_{pt} W_{pt}}{Z_{pt}} \right), \; p \in P, \; t = 1, \ldots, T, \tag{7.71}$$

$$\sum_{p \in P} Z_{pt} = 1, \; t = 1, \ldots, T. \tag{7.72}$$

This is called the allocated clearing function (ACF) formulation. To obtain a tractable LP formulation, we replace the partitioned clearing function (7.71) by a set of linear constraints using outer approximations. Because f is concave, it can be approximated by the convex hull of a set of linear functions $\alpha^c \xi_{pt} W_{pt} + \beta^c$, where $c = 1, \ldots, C$ denotes the individual straight line, i.e., segment, in the approximation. The quantity α^c denotes the slope of the linearized clearing function for segment c, whereas β^c is the intercept

of the linearized clearing function for segment c. We use $\alpha^C = 0$ to model the fact that the maximum throughput is reached. In addition, we have $\alpha^C < \ldots < \alpha^2 < \alpha^1$ and $\beta^1 = 0$. An individual clearing function is assigned to each resource $k = 1, \ldots, K$. The capacity constraint is linear because we have

$$Z_{pt} f_t \left(\frac{\xi_{pt} W_{pt}}{Z_{pt}} \right) = Z_{pt} \min_c \left\{ \alpha^c \frac{\xi_{pt} W_{pt}}{Z_{pt}} + \beta^c \right\} = \min_c \left\{ \alpha^c \xi_{pt} W_{pt} + \beta^c Z_{gp} \right\}. \quad (7.73)$$

Next, we present a corresponding LP formulation following [12, 123] with products $p \in P$ and periods $t = 1, \ldots, T$. Setup times and consequently lot sizing effects are not modeled. For simplicity reasons, we model only a single stage multiproduct system; however, extensions to multistage situations are presented in [12, 123]. The following indices and index sets are used in the resultant LP:

$t = 1, \ldots, T$: period index
p : product index
P : set of all products

The parameters used in the model are:

c_{pt} : unit production cost of product p in period t
ξ_{pt} : amount of resource (machine time) required to produce one unit of product p in period t
h_{pt} : unit inventory holding cost of product p in period t
b_{pt} : unit backlog cost of product p in period t
w_{pt} : unit WIP holding cost of product p in period t
d_{pt} : demand for product p in period t
α^c : slope of the linearized clearing function at segment c
β^c : intercept of the linearized clearing function at segment c
W_{p0} : initial WIP for wafer type p
B_{p0} : initial backlog for wafer type p
I_{p0} : initial inventory for wafer type p

The following decision variables are used within the model:

X_{pt} : release quantity for wafers of type p in period t
Y_{pt} : output quantity for wafers of type p in period t
W_{pt} : WIP quantity for wafers of type p over period t
I_{pt} : units of product p in inventory of finished goods at the end of period t
B_{pt} : units of product p backlogged at the end of period t
Z_{pt} : fraction of capacity that is used by product p in period t

Now, the model can be formulated as follows:

$$\min \sum_{p \in P} \sum_{t=1}^{T} \left\{ c_{pt} Y_{pt} + h_{pt} I_{pt} + b_{pt} B_{pt} + w_{pt} W_{pt} \right\} \quad (7.74)$$

subject to:

$$W_{pt} = W_{p,t-1} - Y_{pt} + X_{pt},\ p \in P,\ t = 1,\ldots,T, \tag{7.75}$$

$$B_{p,t-1} + I_{pt} + d_{pt} = Y_{pt} + I_{p,t-1} + B_{pt},\ p \in P,\ t = 1,\ldots,T, \tag{7.76}$$

$$\xi_{pt} Y_{pt} \leq \alpha^c \xi_{pt} W_{pt} + \beta^c Z_{pt},\quad p \in P,\ t = 1,\ldots,T,\ c = 1,\ldots,C, \tag{7.77}$$

$$\sum_{p \in P} Z_{pt} = 1,\ t = 1,\ldots,T, \tag{7.78}$$

$$X_{pt} \geq 0, Y_{pt} \geq 0, W_{pt} \geq 0,\ I_{pt} \geq 0,\ B_{pt} \geq 0,\ Z_{pt} \geq 0 \quad p \in P,\ t = 1,\ldots,T. \tag{7.79}$$

The objective function (7.74) is based on the production, backorder, WIP, and finished good inventory costs. Constraints (7.75) model the WIP flow. The inventory balance equations are given by constraints (7.76). Constraints (7.77) are related to capacity represented by the linearized clearing function. Constraints (7.78) ensure that the fractions of capacity used by a single product sum up to one. Finally, nonnegativity conditions for decision variables are taken into account by constraints (7.79).

Computational experiments with the linearized ACF multi-stage approach for wafer fabs are described in [12, 123]. Job releases obtained by the ACF formulation are smoother and lead consequently to better overall CT performance compared to production planning approaches based on the fixed CT assumption.

There are several ways to derive clearing functions. The first approach consists in using steady-state or transient queueing models to determine clearing functions analytically [13]. This approach has some limitation in complex manufacturing systems such as wafer fabs. The second approach is estimating clearing functions from empirical data. The empirical data can be collected using discrete-event simulation. The overall procedure from [12, 123] can be summarized as follows.

Estimating a clearing function (ECF)

1. Generate randomly demand realizations that correspond to different bottleneck utilization levels.
2. Determine job releases using a production planning approach that takes the demand realizations from step 1 as input. Alternatively, job releases can be determined based on the demand and some simple backward calculation to obtain starting times for the jobs.
3. For each release plan from step 2, perform repeated simulation runs of the BS and BP using myopic dispatching as FIFO to determine pairs (W_t^k, Y_t^k) for each period t and each machine group k.
4. Determine C_1 for the functional forms (7.65) or (7.66) from the empirical data from step 3. Use a nonlinear least-square fitting technique to find the remaining parameter C_2.

5. Perform a piecewise linearization procedure, i.e., a nonlinear optimization, to find the segments with the corresponding slopes and intercepts (cf. Irdem [123] for details). Three segments are generally appropriate.

It is obvious that the ECF procedure is time-consuming because of the repeated simulation runs and because of running the different optimization procedures.

So far, a clearing function is constructed for each resource separately. The resulting output capacity is allocated to the different products. A different approach is proposed by Kacar and Uzsoy [134]. A clearing function is estimated for each product based on the release quantities and WIP levels of this product and other products in a certain number of periods using multiple regression.

In conclusion, it seems that estimating clearing functions from empirical data is far from being a trivial task. The computational methods and the resultant effort are similar to the case of CT-TP curves.

Overall, it seems that, from a real-world implementation point of view, the iterative simulation approach requires the least effort among the three methodologies discussed in this section.