Chapter 2
Workload and Cycle Time in the Production Unit

Our description of the PPC problem in Chap. 1 identified the effective management of cycle times as a critical link between the planning level and the realized performance of the production units it seeks to coordinate. Most of the PPC systems prevalent in industry today approach this issue through planned lead times and maximum capacity loading, assuming that as long as the capacity loading does not exceed the agreed-upon maximum level, the production units will be able to complete work within the planned lead time with high probability. This chapter argues that reliance on exogenous planned lead times represents a significant drawback of this approach because cycle times through a production unit are, in fact, an outcome of the work release decisions made by the PPC system. Since this dependence between cycle times and work release decisions lies at the center of the problems addressed in this volume, we now discuss the relationship between a production unit's workload and cycle time in more detail.

2.1 Preliminaries

Per Chap. 1, we follow Bertrand et al. (1990) in viewing a production system as a network of production units—groups of production resources such as machines and personnel that must perform specific tasks (e.g., particular operations on particular production orders by a specified due date) and can exhibit different material flow structures such as job shop, flow shop, manufacturing cells, etc. Detailed scheduling and resource allocation decisions within the production unit are not visible to, or subject to the control of, the planning level. Hence the construction of optimization models for planning releases into production units, the primary concern of this volume, must begin with a viable model of an individual production unit that permits anticipation of their behavior by the planning level. Since we seek optimization models that are applicable to a wide variety of manufacturing environments, we must seek general laws describing the behavior of production units. Laws of this

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_2

type have been studied extensively in the field of production economics (Fandel 1991; Hackman 2008), which views production as a transformation process that converts input factors (such as labor and machines) into goods and services for internal or external customers.

Our focus is on the relationship between the work release decisions made at the planning level and the performance measures, particularly work-in-process (WIP) inventory levels, cycle times and output, of the individual production units. Since the planning level seeks to ensure that supply from the production units matches demand for the final products in some "optimal" way, the cycle time, the delay between work being released into the production unit and the completion of its processing at that production unit, takes on a critical role. Hence we are primarily interested in the time dimension of the relationship between factor input, whose timing is determined by work release, and the time the work is completed and output of the finished product occurs.

Since, as argued in Chap. 1, the primary actionable decision of a PPC system is the quantity and timing of work releases into the production units, the evolution of resource workloads over time is determined by decisions at the planning level. Queueing models (Buzacott and Shanthikumar 1993; Curry and Feldman 2000; Hopp and Spearman 2008), which represent production systems as networks of queues, provide useful tools for examining the consequences of planning decisions on the WIP levels, cycle times, and output realized at the production units.

2.2 Insights from Queueing Models

A production unit consists of one or more workcenters, groups of possibly nonidentical machines that are managed on the shop floor as a unit. For simplicity of exposition, we shall frame our discussion in terms of a single production resource, such as a machine, whose behavior can be modeled as a queueing system. While production units may have multiple machines and complex structures within themselves, the problem of how to anticipate their behavior at the planning level is the same in its essence, although the resulting queueing models are more complex. Beyond a certain level of complexity, simulation models are required to describe the behavior of many production units as discussed in later chapters.

We consider a single-machine workcenter modeled as a queueing system, closely following the development in Chap. 8 of Hopp and Spearman (2008). Production orders, which we shall refer to as jobs to avoid confusion with lot-sizing models, are released to the production unit and—possibly after being processed at some workcenters that are not modeled explicitly—arrive at the workcenter under consideration according to some stochastic process. The interarrival times between jobs follow a known probability distribution $F_a(.)$ with mean t_a and squared coefficient of variation (SCV) c_a^2. The effective processing times of the jobs, which incorporate the effects of disruptions such as setup times, machine failures, and scrap, are independent of their arrival times and follow a known probability distribution $F_e(.)$ with

mean t_e and SCV c_e^2. Hence the average arrival rate is $\lambda = 1/t_a$ and the average service rate $\mu = 1/t_e$. The cycle time spent by a job in this queueing system consists of the time it spends in the queue and the time to complete its processing (including setup), and is a random variable determined jointly by the two probability distributions $F_a(.)$ and $F_e(.)$. A well-known result (Kingman 1961; Hopp and Spearman 2008) states that the steady-state expected cycle time T of this *G/G/1* queue (Kendall 1953) is approximated by

$$T = \frac{\left(c_a^2 + c_e^2\right)}{2} \frac{u}{1-u} t_e + t_e \tag{2.1}$$

where the average utilization of the resource is given by $u = t_e/t_a$. Equation (2.1) suggests that the expected cycle time is influenced by four quantities: the variabilities of the arrival and service processes, expressed by c_a^2 and c_e^2, respectively; the mean effective processing time t_e; and the average utilization u of the workcenter, which, in turn, is jointly determined by t_a and t_e. The effect of the average utilization u is of particular interest for production planning models. The release decisions made by the planning level that specify how much work to release to a given production unit in a planning period determine the mean arrival rate of work $\lambda = 1/t_a$ to the workcenter.

Figure 2.1 shows the behavior of the average cycle time T per Eq. 2.1 as the average utilization u and the variance term $C = \left(c_a^2 + c_e^2\right)/2$ vary. T increases nonlinearly with u, eventually tending to infinity as u approaches 1. This behavior shows that the planning level's work release decisions affect the average cycle time; T is endogenous to the planning decision, not an exogenous parameter unaffected by the planning process.

Another important observation from Eq. 2.1 is that T is also affected by the variability c_a^2 in the material flow into the workcenter and the variability c_e^2 of the production process itself. The influence of c_a^2 is particularly important since the arrival

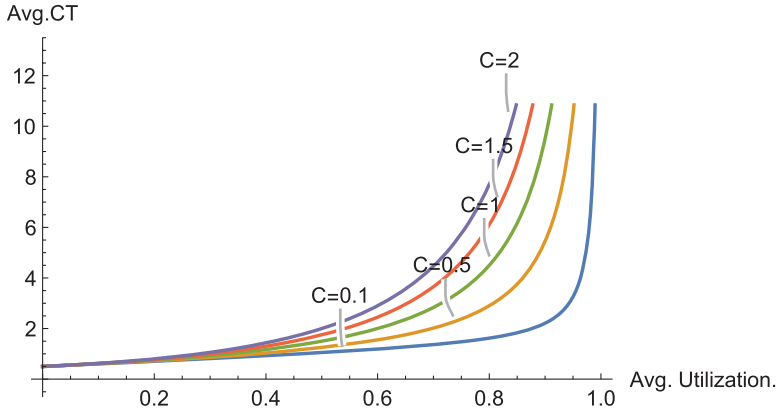

Fig. 2.1 Behavior of average cycle time T of a *G/G/1* queue

process to a workcenter is determined by the departure processes from the upstream workcenters that provide its inputs. Hence the average cycle time T at a workcenter is affected by how the workcenters upstream of it are managed; variability at upstream workcenters will affect the performance of those downstream, as discussed by Hopp and Spearman (2008) and Godinho Filho and Uzsoy (2014). This functional relationship between average WIP and output, average flow time, and possibly other performance measures of the production unit is referred to as the *characteristic* or *operating curve* in the literature (Aurand and Miller 1997; Schoemig 1999) and is often estimated by simulation (Yang et al. 2006).

Representing our basic workcenter as a *G/G/1* queue allows us to invoke another fundamental queueing result. In a production context, the number of customers in the queueing system (in the queue or at a server) at a given point in time corresponds to the amount of work in process inventory (WIP) at the workcenter, which is a random variable we shall denote by WIP, with $W = E[\text{WIP}]$ denoting the expected WIP level expressed as number of customers or, in our context, jobs. If WIP is measured in units of the product or amount of work (standard hours) the queueing relationships given below must be modified accordingly. Following standard queueing analysis, let us also assume that we observe the system over a long period of time, such that the average rate of arrivals to the production unit is equal to its average processing rate. Thus the system is stable with no unbounded increase in the WIP quantity, and the expected throughput rate X of the system, the average rate at which completed work leaves the workcenter, will be $X = \lambda = 1/t_a$. Under these conditions, Little's Law (Little 1961; Hopp and Spearman 2008) gives

$$W = XT = \frac{T}{t_a} \tag{2.2}$$

This expression has several important implications. For the purposes of managing a production system to achieve a given average throughput rate X, the average WIP level W and average cycle time T are directly proportional. A given throughput rate X can be achieved either by controlling the average cycle time T to achieve a desired average WIP level W or by controlling the average WIP level W to achieve an average cycle time of $T = W/X$. Loosely speaking, the former approach is associated with "push" systems such as MRP, where work is released into the production unit to meet due dates derived from customer orders or from forecasts of future demand. The latter is associated with "pull" systems such as the kanban system used in the Toyota Production System (Sugimori et al. 1977; Liker 2004). An excellent discussion of the distinctions between, and relative merits of, push and pull systems is given by Hopp and Spearman (2004).

Combining Eqs. (2.1) and (2.2), the expected WIP level of the steady-state queue is

$$W = \frac{T}{t_a} = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)\frac{t_e}{t_a} + \frac{t_e}{t_a} = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u^2}{1-u}\right) + u \tag{2.3}$$

The average utilization u can be interpreted as the long-run fraction of time the resource will be busy, and thus producing output. Using the average WIP level W as a measure of the resource's workload, i.e., the amount of work available for it to process, and solving for u in terms of W yields a quadratic in W whose nonnegative solution is given by

$$u = \frac{-(W+1) + \sqrt{(W+1)^2 + 4(C-1)W}}{2(C-1)} \quad \text{for } C \neq 1 \tag{2.4}$$

where $C = \left(c_a^2 + c_e^2\right)/2$. When $C = 1$, representing an *M/M/1* queue with exponentially distributed interarrival and service times, Eq. (2.4) takes the simpler form

$$W = \frac{u^2}{1-u} + u = \frac{u}{1-u} \tag{2.5}$$

yielding

$$u = \frac{W}{(W+1)}. \tag{2.6}$$

As shown in Fig. 2.2, for given values of t_e and C, u is a monotonically nondecreasing concave function of W; as the average WIP level W increases, u increases at a decreasing rate. Intuitively, the higher the average WIP level W in the system, the lower the probability $(1-u)$ that the resource will be idle due to lack of work; maintaining a given average throughput requires maintaining a certain average WIP level in the production unit.

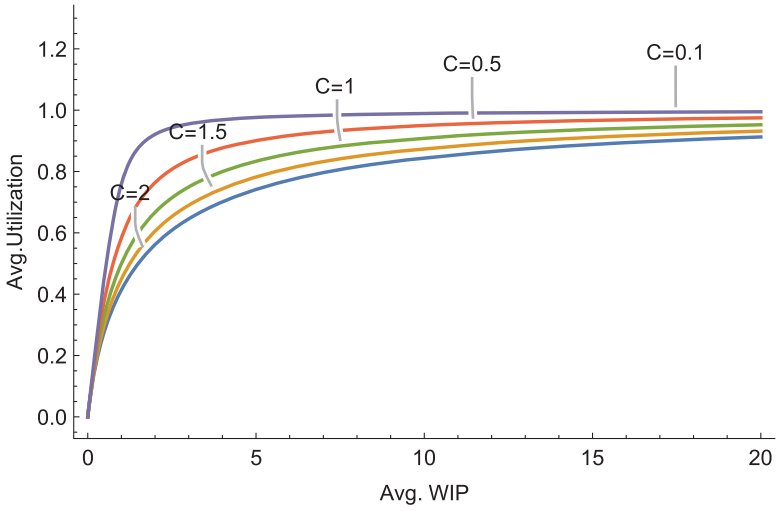

Fig. 2.2 Average utilization as a function of average WIP

Relationships similar to Eq. (2.4) between the expected throughput of a queueing system and its expected WIP level can be derived analytically for a variety of queueing models, under steady-state or transient behavior (Selçuk et al. 2007; Asmundsson et al. 2009; Missbauer 2009). When closed-form analytical expressions are not available, empirical relations can be postulated by fitting an appropriate functional form to data obtained from either industrial observations (Häussler and Missbauer 2014) or a simulation model (Kacar et al. 2012; Kacar and Uzsoy 2015). We shall refer to these functions as *clearing functions*, since they represent the ability of the workcenter to process, or clear, some fraction of its workload in a planning period. They are the central construct of interest to this volume, discussed in Chaps. 7 and 8.

Equations (2.1) and (2.2) together determine the relationship between average WIP and average cycle time. Substituting Eq. (2.4) or Eq. (2.6) for u into Eq. (2.1) yields

$$T = \begin{cases} \dfrac{t_e(C-1)\left(1-W+\sqrt{(W+1)^2+4(C-1)W}\right)}{2C-1+W-\sqrt{(W+1)^2+4(C-1)W}} & \text{for } C \neq 1 \\ t_e(W+1) & \text{for } C = 1 \end{cases} \tag{2.7}$$

The average cycle time increases linearly for the *M/M/1* case where $C = 1$. For $C < 1$ the slope is smaller for low WIP levels in both single- and multiple-server systems since there will be (almost) no queueing delay for low WIP levels.

Together, Eqs. (2.4) and (2.7) imply that given the queueing characteristics of the production unit, once the average WIP level is determined, the average utilization and cycle time are determined as well. This observation motivates the Workload Control framework presented in Chap. 4.

A number of caveats are, however, in order. The discussion above assumes that the given input rate λ unambiguously determines the utilization of the workcenter. This is not the case in the presence of sequence-dependent setup times, since in this case the distribution of the effective processing time depends on the sequence in which the jobs are processed. If jobs are released without considering this issue, some form of batching and sequencing must be performed within the production unit to manage setup times. More WIP in the production unit gives its management more options to optimize the job sequence with respect to setups, reducing the total setup time for the given production quantities as average WIP increases. Several papers have examined the relationship between average WIP level and total setup time per period (Kekre 1984; Kim and Bobrowski 1995; Missbauer 1997; Thürer et al. 2012). Since these savings in setup time reduce the utilization required to produce a given output, the relationship between WIP and output illustrated in Fig. 2.2 is also affected. Informal production control rules applied at the shop-floor level, such as those that adapt the processing rate to the WIP level (e.g., Agnew (1976)), might also affect the operating curve.

The queueing analysis above suggests that the cycle time of a job through a production unit is a random variable whose distribution is affected by the utilization u of the resources. The planned lead times used for order release planning by the PPC system are based on estimates of the cycle times through the production units making up the production system, so it is important to understand the structure of these cycle times. We now turn to this discussion.

2.3 Structure of Cycle Times in Production Units

The cycle time of a production order (job) through a production unit is the time elapsing between its release and its completion and is the sum of the cycle times of all operations performed on this order, accounting appropriately for any overlaps in time. The cycle time of the k'th operation of a job (a *throughput element* in Wiendahl 1995: 41 ff.) is usually defined as the time from the completion of the previous operation $k-1$ to the completion of operation k and consists of any necessary delay between the completion of operation $k-1$ and the start of operation k (such as curing time for a painting operation, or transportation time between locations), queueing time and the setup and processing time of operation k. In discrete manufacturing, the interoperation time, defined as the time from the completion of the previous operation $k-1$ to the start of operation k, consists mainly of waiting time due to queueing at capacitated resources and is often substantially higher than the operation time. Empirical studies report the ratio of operation time (raw process time in the terminology of Hopp and Spearman) to cycle time as about 0.1 in mechanical engineering (Wiendahl 1995: 37f.), and about 10% in the CD/DVD manufacturing system in Sect. 1.2.2. This is consistent with queuing-theoretical results where at high utilization the queuing time constitutes by far the greater part of the average cycle time in (Eq. 2.1).

Hence the variance of the cycle times is mainly determined by the variance of the waiting times, which is often fairly high in queueing systems. In the *M/M/1* queue, the conditional waiting time given that the server is busy on arrival is exponentially distributed. For the *G/G/1* queue, the waiting time distribution depends on the distributions of the interarrival and service times (Shortle et al. 2018: 320 ff.). In line with these analytical results, the empirical distribution of the cycle times at a workcenter often exhibits high variance, as illustrated in Fig. 2.3. The positive skewness due to very long cycle times experienced by a small fraction of the orders is typical of many production environments, and can be caused both by time-varying WIP levels at a workcenter and by expediting or delaying orders by dispatching; Ehteshami et al. (1992) illustrate the effect of expediting in the context of semiconductor wafer fabrication. Four priority classes can be distinguished in the figure; some orders are deliberately delayed for the reasons given in the legend. However, even the cycle times of the normal orders exhibit high variance, making it difficult to derive planned lead times from observed cycle times.

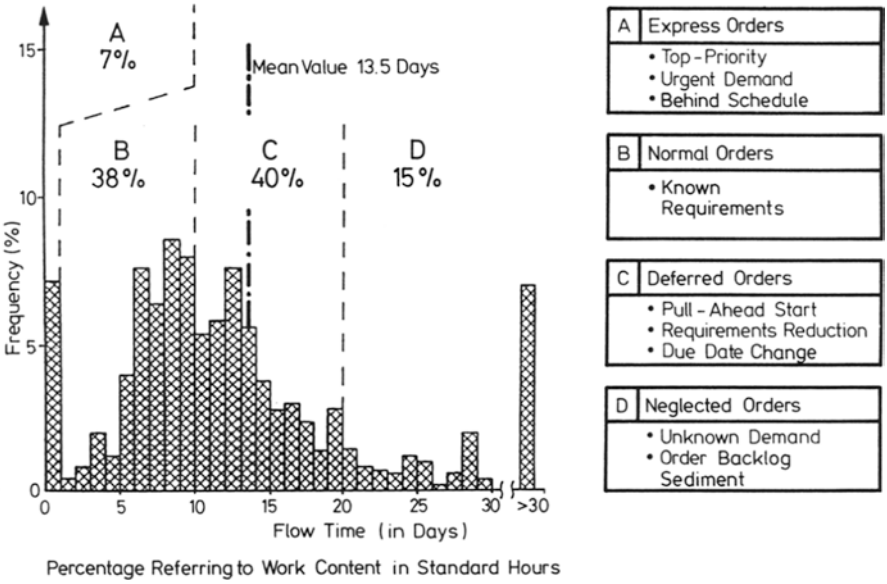

Fig. 2.3 Distribution of the weighted cycle time of the orders processed at a lathe workcenter over 16 weeks; Wiendahl (1995): 30

The mean cycle time at a workcenter, which is of major concern in this book, is usually defined as the mean of the distribution of the individual job cycle times, which is also the standard definition in the scheduling literature (Pinedo 2012). Wiendahl (1995: 55ff) recommends using the weighted mean cycle time

$$\bar{T}_m^W = \frac{\sum_{j \in \Im} T_{jm} a_{jm}}{\sum_{j \in \Im} a_{jm}} \tag{2.8}$$

since it is less sensitive to priorities at the dispatching level than the unweighted mean cycle time. In Eq. (2.8) T_{jm} denotes the observed cycle time of order j at workcenter m, a_{jm} the processing time (including setup time) of order j at that workcenter, and $\Im$ the set of all operations represented in the observed sample of orders. This quantity represents an estimate of the average cycle time of each hour's worth of work processed at the workcenter in a certain time interval, the definition used in Fig. 2.3.

The importance of realistic lead times for the planning level and the large contribution of waiting time to the observed cycle times, at least at bottleneck workcenters, makes management of the waiting times an essential task for shop-floor management, and the derivation of accurate planned lead times from them crucial to effective operation of the planning level.

2.4 From the Production Unit to the Goods Flow Problem

Having described the behavior of a generic production unit, we are now in a position to relate the conceptual model of PPC systems developed in Chap. 1 to the vital statistics of our production unit: WIP, throughput, and cycle time. Since the production units are managed autonomously to meet the output targets determined by the planning level, the planning level must be able to estimate the impact of its requests, i.e., planned releases and output, on the ability of the production unit to meet them in a timely and cost-effective manner. Per Sect. 2.1, such a model must recognize the nonlinear relationship between average throughput X and average cycle time T as approximated by Eqs. (2.3), (2.7), or some similar relation. The task of the planning level is to release production orders into the production units such that they can carry out the processing necessary to meet demand in time. This requires coordination of activities across multiple production units across time. This, in turn, requires both effective management of the cycle times within each production unit to coordinate the timing of production with demand, and planning and control of the production-inventory system according to the product structure, including the determination of desired stock levels at the various stock points over time. Much of the complexity of the PPC task results from the interference of these two modeling and control tasks, and it is not surprising that PPC systems in practice emphasize one or the other of these tasks in order to keep complexity manageable. We now turn to the PPC frameworks that provide the basis for the developments presented in this book.

References

Agnew C (1976) Dynamic modeling and control of some congestion prone systems. Oper Res 24(3):400–419

Asmundsson JM, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning models with resources subject to congestion. Nav Res Logist 56(2):142–157

Aurand S, Miller P (1997) The operating curve: a method to measure and benchmark manufacturing line productivity. In: 1997 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop ASMC 97 Proceedings, IEEE, Cambridge, pp 391–397

Bertrand JWM, Wortmann JC, Wijngaard J (1990) Production control: a structural and design oriented approach. Elsevier, Amsterdam

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs

Curry GL, Feldman RM (2000) Manufacturing systems modelling and analysis. Springer, Berlin

Ehteshami B, Petrakian R, Shabe P (1992) Trade-offs in cycle time management: hot lots. IEEE Trans Semicond Manuf 5(2):101–106

Fandel G (1991) Theory of production and cost. Springer, Berlin

Godinho Filho M, Uzsoy R (2014) Assessing the impact of alternative continuous improvement programmes in a flow shop using system dynamics. Int J Prod Res 52(10):3014–3031

Hackman S (2008) Production economics. Springer, Berlin

Häussler S, Missbauer H (2014) Empirical validation of meta-models of work centres in order release planning. Int J Prod Econ 149:102–116

Hopp WJ, Spearman ML (2004) To pull or not to pull: what is the question? Manuf Serv Oper Manag 6(2):133–148

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Kacar NB, Uzsoy R (2015) Estimating clearing functions for production resources using simulation optimization. IEEE Trans Autom Sci Eng 12(2):539–552

Kacar NB, Irdem DF, Uzsoy R (2012) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. IEEE Trans Semicond Manuf 25(1):104–117

Kekre S (1984) The effect of number of items processed at a facility on manufacturing lead time. Working Paper Series. University of Rochester, Rochester

Kendall DG (1953) Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann Math Stat 24(3):338–354

Kim S-C, Bobrowski PM (1995) Evaluating order release mechanisms in a job shop with sequence-dependent setup times. Prod Oper Manag 4(2):163–180

Kingman JFC (1961) The single server queue in heavy traffic. Math Proc Camb Philos Soc 57(4):902–904

Liker J (2004) The Toyota way: 14 management principles from the world's greatest manufacturer. McGraw-Hill, New York

Little JDC (1961) A proof of the queueing formula L =λw. Oper Res 9:383–387

Missbauer H (1997) Order release and sequence-dependent setup times. Int J Prod Econ 49:131–143

Missbauer H (2009) Models of the transient behaviour of production units to optimize the aggregate material flow. Int J Prod Econ 118(2):387–397

Pinedo M (2012) Scheduling. Theory, algorithms, and systems. Springer, New York

Schoemig AK (1999) On the corrupting influence of variability in semiconductor manufacturing. In: Proceedings of the Winter Simulation Conference, P. A. Farrington, H. B. Nembhard, D. T. Sturrock, G. W. Evans (eds), 1999, pp. 837–842, IEEE

Selçuk B, Fransoo JC, de Kok AG (2007) Work in process clearing in supply chain operations planning. IIE Trans 40(3):206–220

Shortle JF, Thompson JM, Gross D, Harris CM (2018) Fundamentals of queueing theory. Wiley, Hoboken

Sugimori Y, Kusunoki K, Cho F, Uchikawa S (1977) Toyota production system and Kanban system: materialization of just-in-time and respect for human system. Int J Prod Res 15(6):553–564

Thürer M, Silva C, Stevenson M, Land M (2012) Improving the applicability of workload control (Wlc): the influence of sequence dependent setup times on workload controlled job shops. Int J Prod Res 50(22):6419–6430

Wiendahl HP (1995) Load oriented manufacturing control. Springer, Heidelberg

Yang F, Ankenman B, Nelson BL (2006) Efficient generation of cycle time-throughput curves through simulation and metamodeling. Nav Res Logist 54(1):78–93