Chapter 17
Aggregate Modeling of Manufacturing Systems

Erjen Lefeber and Dieter Armbruster

17.1 Introduction

Manufacturing systems can be modeled in several ways. In particular, during the design of a manufacturing system, discrete event modeling is an often used approach, cf. Banks (1998) and Cassandras and Lafortune (1999). Discrete event models often include a high level of detail. This high level of detail can be used to investigate the effect of all kinds of variables on the possible performance of the manufacturing system. However, when a manufacturing system is in operation, this model usually contains too much detail to keep all parameters up-to-date with the evolving current system. In addition, certain parameters cannot even be measured. Furthermore, running one scenario using a discrete event model takes several hours. Usually, discrete event models are only tailer made for answering specific problems. These models only contain part of the manufacturing system.

Another option might be to derive a less detailed model, in particular, for manufacturing planning and control or supply chain control. In this chapter, we discuss three classes of models, each at a different level of aggregation. We start with less detailed discrete event models based of effective process times (EPTs), where each workstation is modeled as a node in a queuing network. Next, in particular for the purpose of planning and control, we abstract from events and replace all discrete event queues with discrete time fluid queues. In addition, the throughput of each workstation is limited by a nonlinear function of the queue length, the clearing function, see also Chap. 16 of this book. Finally, we abstract from workstations and model manufacturing flow as a real fluid using continuum models. These models are scalable and suitable for supply chain control.

E. Lefeber (✉)
Department of Mechanical Engineering, Eindhoven University of Technology,
PO Box 513, Eindhoven, The Netherlands
e-mail: A.A.J.Lefeber@tue.nl

K.G. Kempf et al. (eds.), *Planning Production and Inventories in the Extended Enterprise*,
International Series in Operations Research & Management Science 151,
DOI 10.1007/978-1-4419-6485-4_17,

17.2 Effective Process Times

Building a discrete event model of an existing manufacturing system can be cumbersome, as manufacturing systems are prone to disturbances. Even though many disturbances can be modeled explicitly in highly detailed discrete event models, it is impossible to measure all sources of variability that might occur in a manufacturing system. In addition, highly detailed discrete event models are unsuitable for decision making due to their time-consuming simulation runs.

Instead of measuring detailed information, like raw process times, setup times, times to failures, times between repairs, operator behavior, etc., one can also try to measure the clean process time *including* other sources of additional waiting. This is the so-called *effective process time* (EPT), which has been introduced in Hopp and Spearman (2000) as the time seen by lots from a logistical point of view. In order to determine the EPT, they assume that the contribution of the individual sources of variability is known.

A similar description is given in Sattler (1996) where the EPT has been defined as all flow time except waiting for another lot. It includes waiting due to machine down time and operator availability and a variety of other activities. In Sattler (1996), it was also noticed that this definition of EPT is difficult to measure.

Instead of taking the bottom-up view of Hopp and Spearman (2000), a top-down approach can also be taken, as shown in Jacobs et al. (2001) and Jacobs et al. (2003), where algorithms have been introduced that enable determination of EPT realizations from a list of events. That is, instead of measuring each source of disturbances individually and derive an aggregate EPT distribution, one can also derive this EPT distribution from manufacturing data directly. In the remainder of this section, we illustrate for several situations how these EPTs can be measured from manufacturing data.

17.2.1 A Single Lot Machine, No Buffer Constraints

Consider a workstation consisting of one machine, which processes single lots (i.e., no batching) and assume that the Gantt chart of Fig. 17.1 describes a given time period.

- At $t = 0$, the first lot arrives at the workstation. After a setup, the processing of the lot starts at $t = 2$ and is completed at $t = 6$.
- At $t = 4$, the second lot arrives at the workstation. At $t = 6$ this lot could have been started, but apparently there was no operator available, so only at $t = 7$ the setup for this lot starts. Eventually, at $t = 8$, the processing of the lot starts and is completed at $t = 12$.
- The fifth lot arrives at the workstation at $t = 22$, processing starts at $t = 24$, but at $t = 26$ the machine breaks down. It takes until $t = 28$ before the machine has been repaired and the processing of the fifth lot continues. The processing of the fifth lot is completed at $t = 30$.

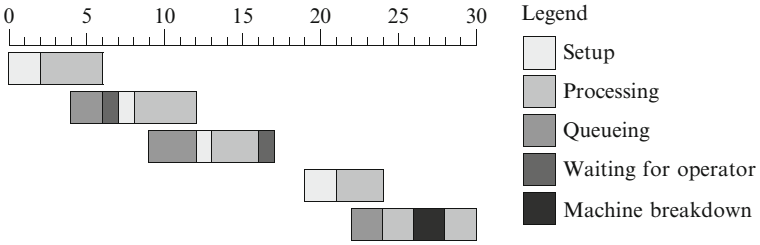

Fig. 17.1 Gantt chart of five lots at a single machine workstation

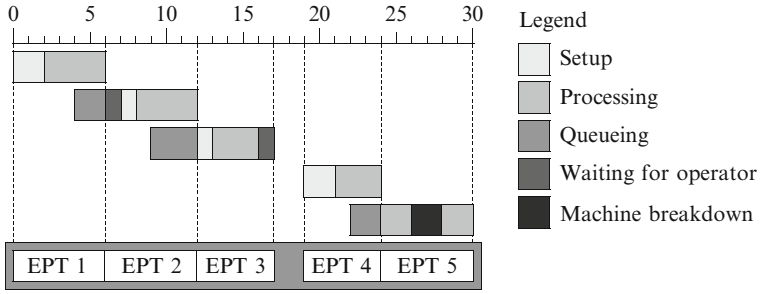

Fig. 17.2 EPT realizations of five lots at a single machine workstation

From a lot's point of view we observe:

- The first lot arrives at an empty system at $t = 0$ and departs from this system at $t = 6$. Its processing took 6 units of time.
- The second lot arrives at a nonempty system at $t = 4$ and needs to wait. At $t = 6$, the system becomes available and hence from $t = 6$ on there is no need for the second lot to wait. At $t = 12$, the second lot leaves the system, so from the point of view of this lot, its processing took from $t = 6$ till $t = 12$; the lot does not know whether waiting for an operator and a setup is part of its processing.
- The third lot sees no need for waiting after $t = 12$ and leaves the system at $t = 17$, so it assumes to have been processed from $t = 12$ till $t = 17$.

Following this reasoning, the resulting FPTs for lots are depicted in Fig. 17.2. Notice that only arrival and departure events of lots to a workstation are needed for determining the EPTs. Furthermore, none of the contributing disturbances needs to be measured.

In highly automated manufacturing systems, arrival and departure events of lots are being registered, so for these manufacturing systems, EPT realizations can be determined rather easily. These EPT realizations can be used in a relatively simple discrete event model of the manufacturing system, in this case a simple infinite FIFO queue. Such a discrete event model only contains the architecture of the manufacturing system, buffers, and machines. The process times of these machines

are samples from their EPT-distribution as measured from real manufacturing data, or most often from the distribution fitted to that data. There is no need for incorporating machine failures, operators, etc., as this is all included in the EPT-distributions.

Furthermore, the EPTs are utilization independent. That is, EPTs collected at a certain throughput rate are also valid for different throughput rates. Also, machines with the same EPT-distribution can be added to a workstation. This makes it possible to study how the manufacturing system responds in case a new machine is added, or all kinds of other what-if-scenario's.

Finally, since EPT realizations characterize operational time variability, they can be used for performance measuring as explained in Ron and Rooda (2005). Note that overall equipment effectiveness (OEE), which is widely used to quantify capacity losses in manufacturing equipment, directly relates to utilization, i.e., the fraction of time a workstation is busy. However, the performance of manufacturing systems is not only determined by utilization, but also by the variability in production processes. By only focusing on utilization, one may overlook opportunities for performance improvement by a reduction of variability. These opportunities are provided by measuring EPTs.

17.2.2 A Single Batch Machine, No Buffer Constraints

EPTs for equipment that serves batches of jobs have first been studied in Jacobs (2004) and Jacobs et al. (2006). Consider a workstation consisting of one machine, which processes batches of jobs and assume that the Gantt chart of Fig. 17.3 describes a given time period. As we know from the previous section, only arrivals

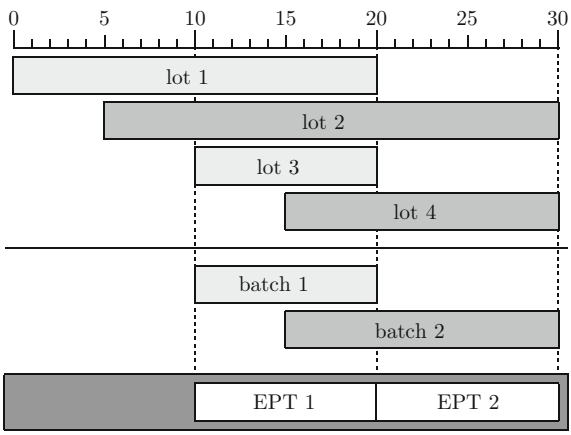

Fig. 17.3 Gantt chart of four lots (two batches) at a batch machine

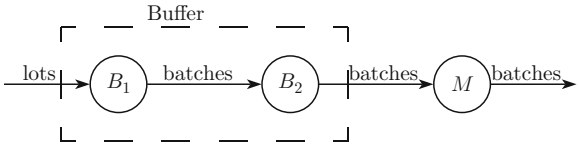

Fig. 17.4 Model of batch formation and queuing in front of a batch machine

and departures from jobs matter for determining EPTs, so in Fig. 17.3 we already abstracted from most disturbances. The only remaining issue is how to deal with the batching. For that purpose, we make a distinction between the policy for batch formation and the EPT of a batch. An other way of putting this is to assume that the buffer consists of two parts. A first part, B_1, in which lots are waiting to become batches, and a second part, B_2, where batches are queuing in front of the workstation, as depicted in Fig. 17.4. Taking this point of view, we can interpret Fig. 17.3 in the following way. At $t = 0$, the first lot arrives at the workstation in buffer B_1, waiting to become a batch with the third lot. At $t = 5$, the second lot arrives at the workstation in buffer B_1, waiting to become a batch with the fourth lot. At $t = 10$, the third lot arrives at the workstation in buffer B_1, resulting in the first batch to be formed. So at $t = 10$, the first batch moves from buffer B_1 to buffer B_2. At $t = 15$, the fourth lot arrives at the workstation in buffer B_1, resulting in the second batch to be formed. So at $t = 20$, the second batch moves from buffer B_1 to buffer B_2.

When we now look at the system consisting of buffer B_2 and the batch machine M, we have a system as we studied in the previous example. A system to which batches arrive, and which processes batches. The first batch arrives to this system at $t = 10$ and leaves the system at $t = 20$, the second batch arrives to this system at $t = 15$ and leaves the system at $t = 30$. Therefore, the first EPT runs from $t = 10$ till $t = 20$, the second EPT runs from $t = 20$ till $t = 30$.

Notice that using this approach, EPTs for batches only start as soon as a batch has been formed, or to be more precise: the batch that is processed finally. The period from $t = 0$ till $t = 10$, lot 1 was in the system and could have been processed as a batch of size 1. Therefore, one could argue that from the point of view of this lot, its EPT starts at $t = 0$. Also, one might say that as soon as lot 2 has arrived, a batch consisting of lots 1 and 2 could have been started, so the first EPT should have started at $t = 5$. This is not what we do, since we view batch formation as part of the way the system is controlled, not as a disturbance. As a result, we not only need to determine EPTs for batches, we also need to characterize the policy for batch formation. One way to deal with this is to include the policy for batch formation in the discrete event model which is actually being used in the manufacturing system under consideration. Another way to deal with this is to try to characterize the policy for batch formation in one way or the other, i.e., derive some "effective batch formation policy." The latter is still subject of current research.

As mentioned above, EPTs can also be used as performance measure. Notice that in case of batching, EPTs do not characterize capacity loss completely. Only the

capacity loss given by the batches is characterized, including variability. Capacity loss due to a bad policy for batch formation is not captured in the EPT. This should be derived by analyzing the (effective) batch formation policy. Notice that again only arrival and departure event of lots are needed for determining the EPTs of batches.

17.2.3 A Multimachine Workstation, No Buffer Constraints

So far, we only considered workstations consisting of a single machine. However, workstations consisting of several machines in parallel can also be dealt with, see, e.g., Jacobs et al. (2003), Jacobs (2004) and Jacobs et al. (2006). We do this in a similar way as we handled batching. That is, we view the decision of which lot is served by which machine again as part of the control system of the manufacturing system.

Consider a workstation consisting of two machines in parallel which both process single lots (i.e., no batching) and assume that the Gantt chart of Fig. 17.5 describes a given time period. Note that we abstracted from most disturbances like we did when we considered batching.

- At $t = 0$, the first lot arrives at the workstation. This lot is processed by Machine 1 and leaves this workstation at $t = 15$.
- At $t = 5$, the second lot arrives at the workstation. Even though Machine 2 is available, or at least not serving any job, this job is also processed by Machine 1 and leaves the workstation at $t = 25$.
- At $t = 10$, the third lot arrives at the workstation. This lot is processed by Machine 2 and leaves the workstation of $t = 30$.

The way we view this system, and is depicted in Fig. 17.6. We assume that the buffer consists of a dispatcher D which decides to which machine each lot will go. We assume that lots do not wait in this dispatcher, but immediately move on to a buffer in front of the machine at which they will finally be processed.

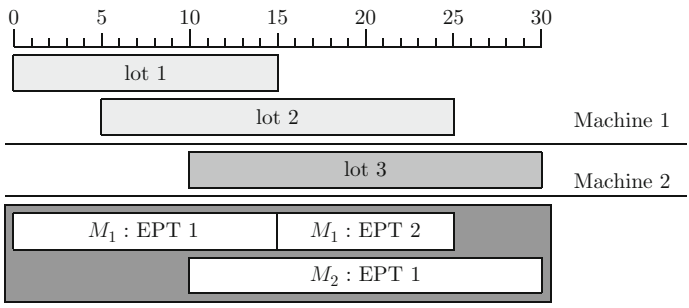

Fig. 17.5 Gantt chart of three lots at a workstation with two machines in parallel

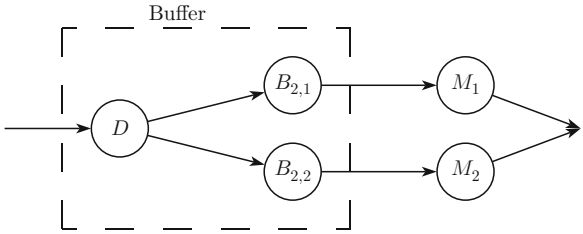

Fig. 17.6 Model of dispatching and queuing at a multimachine station

Using this abstraction, the EPTs as depicted in Fig. 17.5 follow straightforwardly for each separate machine. Notice again that the only data we need for determining the EPTs are arrival and departure event of lots. Also, we do not only need to determine the EPTs, but we also need to know the dispatching strategy. Either this policy is known from reality and can be implemented in the discrete event model, or an "effective dispatching policy" needs to be derived from manufacturing data. The latter is still a subject of current research. Furthermore, multimachine workstations with equipment that serves batches can easily be dealt with combining the results presented so far.

17.2.4 Finite buffers

In the preceding sections, we assumed infinite buffers or at least buffers that are large enough. This enabled us to analyze workstations in isolation. If buffer sizes are small and cannot be neglected, as for example in automotive industry, buffer sizes will explicitly be taken into account in the aggregate discrete event model. Therefore, the effect of blocking, will be explicitly taken into account by means of the discrete event model. Therefore, this disturbance should *not* be included in the EPT. To take into account the effect of blocking, a third event is needed. So far, we only needed arrival and departure events from lots. Or to be more precise: we needed *actual arrival* (AA) and *actual departure* (AD) events. For properly dealing with blocking we also need *possible departure* (PD) events, see also Kock et al. (2005, 2006a,c).

Consider a line of two machines in series, machine M_{j-1} and machine M_j, and assume that there is no buffer between these two machines. Let the Gantt chart of Fig. 17.7 describes a given time period, where we again abstracted from most disturbances.

- At $t = 0$, the first lot arrives at Machine M_{j-1}. At $t = 9$, this lot has been completed and moves to Machine M_j. Both the possible and actual departure at Machine M_{j-1} are at $t = 9$. Processing of the first lot at Machine M_j completes at $t = 22$.

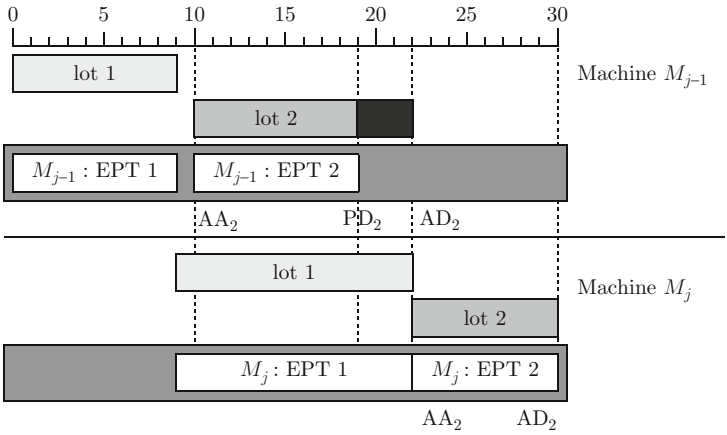

Fig. 17.7 Gantt chart of 2 lots at two sequential, unbuffered machines

- At $t = 10$, the second lot arrives at Machine M_{j-1}. At $t = 19$, this lot has been completed, but cannot yet move to Machine M_j. The possible departure for this lot is at $t = 19$. As Machine M_j only becomes available at $t = 22$, the actual departure at Machine M_{j-1} is at $t = 22$. The actual arrival at Machine M_j is at $t = 22$ for the second lot, and the actual departure at Machine M_j is at $t = 30$.

From the measured events, the EPTs follow readily. Since Machine M_{j-1} cannot help it to become blocked, the EPT for the second lot stops at $t = 19$, i.e., at the *possible departure* event. If we denote the jth EPT realization at Machine i as $\mathrm{EPT}_{i,j}$ we obtain

$$\mathrm{EPT}_{i,j} = \mathrm{PD}_{i,j} - \max\left(\mathrm{AA}_{i,j}, \mathrm{AD}_{i-1,j}\right), \tag{17.1}$$

where $\mathrm{AA}_{i,j} < \mathrm{PD}_{i,j} \leq \mathrm{AD}_{i,j}$ denote, respectively, the actual arrival, possible departure, and actual departure event at Machine i for lot j. By measuring only these three events at each machine, one is able to derive EPTs for each single job workstation in the manufacturing system.

17.2.5 Multilot Machines

By means of the results presented above, one is able to deal with both finite and infinite buffered multimachine workstations serving batches of jobs. In particular, multi might be one and batch sizes can also be one, so any kind of equipment can be dealt with which processes a single job at the time.

However, certain machines can start serving the next job before the previous one has left the machine. Typically these machines are some minifactories themselves. For these machines, we cannot use a simple queuing model. Therefore, for those machines, we cannot use the relation (17.1) to derive EPTs. A different aggregate

model is needed for those kind of machines. First attempts for an aggregate model for multiple lot machines have been made in Eerden et al. (2006) and Kock et al. (2006b). In particular, these models can also be used for aggregating parts of a manufacturing system. For the most recent results in this area, the interested reader may refer to Veeger et al. (2009) and http://se.wtb.tue.nl/~sereports.

17.3 Clearing Function Models

In the previous section, we derived how less detailed discrete event models can be build by abstracting from all kinds of disturbances like machine failure, setups, operator behavior, etc. By aggregating all disturbances into one EPT, a complex manufacturing system can be modeled as a relatively simple queueing network. Furthermore, the data required for this model can easily be measured from manufacturing data.

Even though this approach considerably reduces the complexity of discrete event models for manufacturing systems, this aggregate model is still unsuitable for manufacturing planning and control. Therefore, in this section, we introduce a next level of aggregation, by abstracting from events. Using the abstraction presented in the previous section, we can view a workstation as a node in a queueing network. In this section, we assume that such a node processes a deterministic continuous stream of fluid. That is, we consider this queue as a so-called fluid queue. In order not to loose the steady-state queueing relation between throughput and queue length, we impose this relation as a system constraint, the clearing function as introduced in Graves (1986), see also Chap. 20 of this book.

As an example, consider a manufacturing system consisting of two infinitely buffered workstations. Assume that Machine i has a mean EPT $t_{e,i}$ with a coefficient of variation $c_{e,i}$, i.e., a standard deviation of $c_{e,i} \cdot t_{e,i}$ for $i \in \{1, 2\}$. Let $u_0(k)$ denote the number of jobs started during the kth time period. Let $u_1(k)$ and $u_2(k)$ denote the utilization of Machines 1 and 2, respectively, during the kth time period. Furthermore, let $x_1(k)$ and $x_2(k)$ denote the buffer contents in workstations 1 and 2, respectively, at the beginning of the kth time period (i.e., the jobs in both buffer and machine), and let $x_3(k)$ denote the stored completed jobs or backlog at the beginning of the kth time period. Finally, let $d(k)$ denote the demand during the kth time period. Then we can write down the following discrete time fluid queue dynamics for this system

$$\begin{aligned} x_1(k+1) &= x_1(k) + u_0(k) - \frac{1}{t_{e,1}} u_1(k), \\ x_2(k+1) &= x_2(k) + \frac{1}{t_{e,1}} u_1(k) - \frac{1}{t_{e,2}} u_2(k), \\ x_3(k+1) &= x_3(k) + \frac{1}{t_{e,2}} u_2(k) - d(k). \end{aligned} \tag{17.2}$$

Consider a workstation that consists of m identical servers in parallel which all have a mean effective processing time t_e and coefficient of variation c_e. Furthermore, assume that the coefficient of variation of the interarrival times is c_a and that the utilization of this workstation is $u < 1$. Then we know from the queuing theory of Takahasi and Sakasegawa (1977) that in steady state the mean number of jobs in this workstation is approximately given by

$$x = \frac{c_a^2 + c_e^2}{2} \cdot \frac{u^{\sqrt{2(m+1)}}}{m(1-u)} + u. \tag{17.3}$$

In Fig. 17.8, this relation has been depicted graphically. In the left-hand side of this figure, one can clearly see that for an increasing utilization, the number of jobs in this workstation increases nonlinearly. By swapping axes, this relation can be understood differently. Depending on the number of jobs in the workstation, a certain utilization can be achieved, or a certain throughput. This has been depicted in the right-hand side of Fig. 17.8. For the purpose of production planning, this effective clearing function provides an upper bound for the utilization of the workstation depending on the number of jobs in this workstation. Therefore, in addition to the model (17.2) we also have the constraints

$$\begin{aligned} \frac{c_{a,1}^2 + c_{e,1}^2}{2} \cdot \frac{u_1(k)^2}{1-u_1(k)} + u_1(k) &\le x_1(k), \\ \frac{c_{a,2}^2 + c_{e,2}^2}{2} \cdot \frac{u_2(k)^2}{1-u_2(k)} + u_2(k) &\le x_2(k). \end{aligned} \tag{17.4}$$

The clearing function model for production planning consists of the model (17.2) together with the constraints (17.4). When we want to use this clearing function model for production planning, we need the parameters c_e and c_a. In the previous section, we explained how EPTs can be determined for each workstation, which provides the parameter c_e for each workstation. In addition, for each workstation,

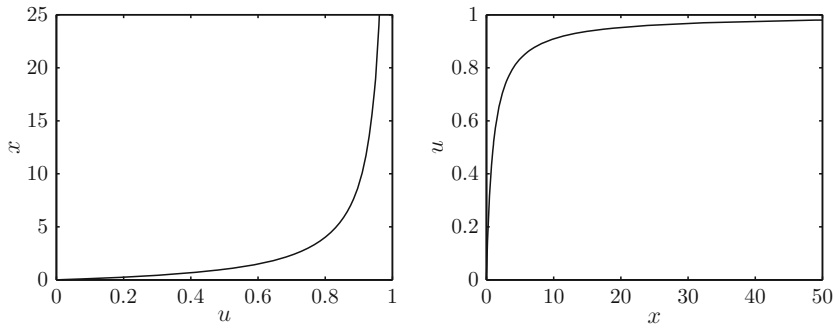

Fig. 17.8 Effective clearing function of (17.3) with $c_a = c_e = m = 1$

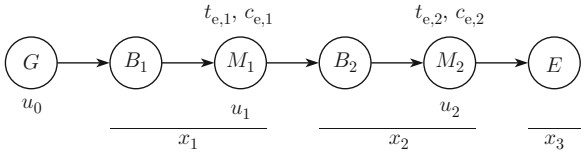

Fig. 17.9 Manufacturing system consisting of two workstations

the interarrival times of jobs can also be determined from arrival events, which provides the parameter c_a for each workstation. Therefore, both parameters can easily be determined from manufacturing data.

However, when applying this approach for production planning, one should carefully derive the EPTs. In particular, if the manufacturing execution system authorizes jobs for processing. In that case, the EPT of a lot cannot start before it has been authorized.

To illustrate this, consider the case depicted in Fig. 17.9. Assume that processing times of the workstations are exponentially distributed with means of, respectively, 0.21 and 0.23 h. Let an MPC production planning scheme be applied with time steps of 1 day (24 h) and a prediction horizon of 5 days. That is, consider a production planning scheme where each day a planning for the next 5 days is generated of which only the desired production levels for the first day are provided as targets (since the planning will be adjusted for the modified circumstances the next day). For this planning, the model (17.2) is used together with the constraints (17.4) and the obvious constraints that buffer contents and utilizations have to be nonnegative for each time period. We do allow for backlog, so x_3 is allowed to become negative. Assume that the goal is to minimize a linear cost function of the jobs in the system where the following customer demand is given:

$$d(k) = 90 + 10 \sin \frac{k\pi}{25}.$$

That is, a periodic demand with a period of 50 days (1,200 h) where demand varies between 80 and 100 jobs per day. This means that the bottleneck requires a utilization between 77 and 96%. Finally, assume that the shop floor implementation of meeting the required targets is by authorizing jobs equally distributed over time. So, if for a certain day a target of 96 jobs is set, every 15 min a new job is authorized.

Next, we consider two ways of determining EPTs. For the first (incorrect) method, we use (17.1) where the actual arrival event AA is the event of the arrival of a lot in the buffer. For the second (correct) method, we also use (17.1) for determining the EPTs, but in this case we use for the actual arrival event AA the latest of the following two events: the arrival of a lot in the buffer or the authorization of that lot for processing. In the latter case, we say that even when a lot has completed the service at the previous workstation, if it has not yet been authorized for processing, it cannot join the queue for processing and therefore actually has not yet arrived to that queue.

The difference in performance between these two ways of determining the actual arrival event AA is depicted in Fig. 17.10, where we see the evolution of the amount of jobs in the buffers and of the backlog. At the left-hand side of this figure, we see that every now and then wip levels explode. For example, around $t = 40{,}000$, we first see a backlog of about 400 lots and a little later the buffer contents in the first workstation reaches almost 5,000 lots. However, at the right-hand side of this figure, we see that the wip in the first workstation remains between 1 and 3 lots, the wip in the second workstation stays even between 2 and 3 lots, and no backlog occurs.

An explanation for this large difference in behavior can be understood if one looks at the EPT realizations. For the first method, the derived EPTs are presented in Fig. 17.11. Since we did not include any disturbances in our model, we know that the (mean) EPTs of the workstations should be 0.21 and 0.23, respectively. However, this is not what we see in Fig. 17.11. In the left-hand side of this figure, we see large EPT realizations every now and then. Also, we see periodic fluctuations in the EPT, implying that the realizations are utilization dependent, which they should not be. Recall that EPTs should be utilization independent. This periodic behavior becomes

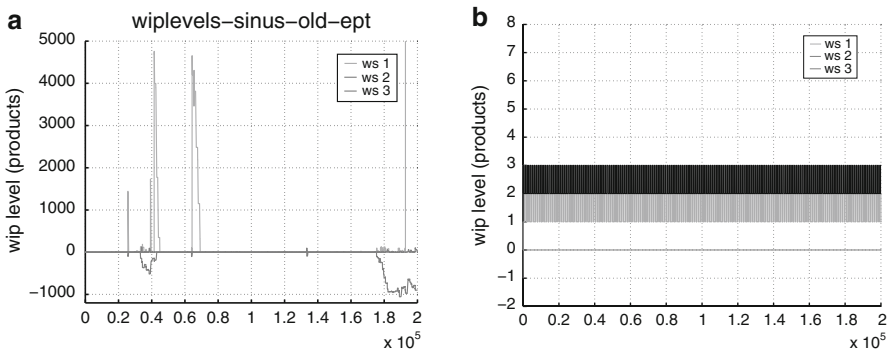

Fig. 17.10 (**a**) Resulting wip levels using incorrect EPT measurements. (**b**) Resulting wip levels using correct EPT measurements

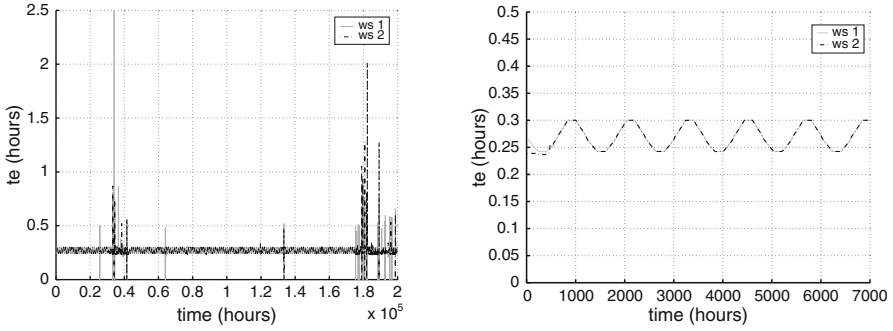

Fig. 17.11 (**a**) Incorrect EPT measurements (complete time horizon). (**b**) Incorrect EPT measurements (zoomed area)

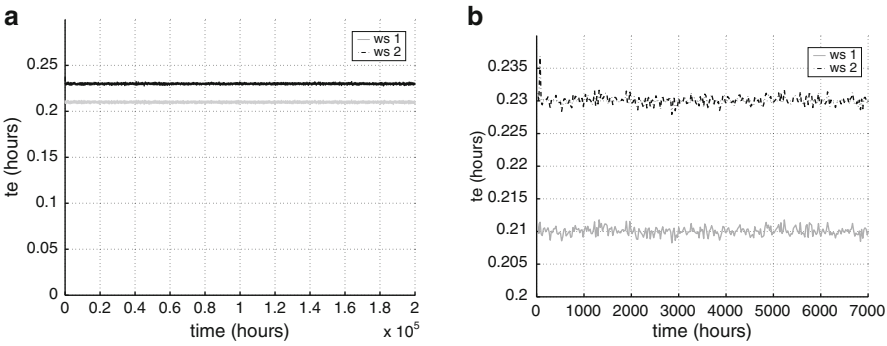

Fig. 17.12 (**a**) Correct EPT measurements (complete time horizon). (**b**) Correct EPT measurements (zoomed area)

even more clear when we zoom in on the first 7,000 time units, as depicted in the right-hand side. Furthermore, we see that the EPT realizations are also a little bit too large.

The explanation of these results is in the way EPTs are determined and the effect that this has on the production planning system. Assume that lots are waiting in the buffer and have not yet been authorized for production. Then they have to wait, even when the machine is idle. As a result, the EPT realization becomes larger. But larger EPT realizations imply that apparently less capacity is available at this machine. Therefore, for the next period, less jobs can be authorized for production. In this way, the planning system enters a viscous circle resulting in large excursions.

Indeed, if one uses as AA-event the moment when the lot has both arrived in the buffer and been authorized for production, better results are obtained, as seen in Fig. 17.12, In this figure, we see correct estimation of the EPT, where small fluctuations are only due to stochasticity. Also when we zoom in on the first 7,000 time units, no utilization dependency of EPT realizations can be found anymore.

17.4 Continuum Models

17.4.1 A Continuum of Production Stages

EPT and clearing function models can be developed for any arbitrary part of the production line. In particular, they can also be used to describe the aggregate behavior of a whole factory, replacing all the details of its production by, e.g., a clearing function relation that determines the outflux as a function of the current work in progress (WIP) in the factory. This will work well, if the associated cycle times through the factory are small and hence the change in WIP during a cycle time is also small (see Chap. 16 for a discussion for the proper timing of a clearing function). However,

if the changes in influx are on a shorter timescale than the cycle time, we need to keep track of the time already spent in the factory by a given lot at a particular place in the production line. This can be done by adding delays into ordinary differential equation models or by modeling the flow of WIP through a factory explicitly via a transport equation.

Specifically, the fluid models that use EPT and clearing functions' approaches discussed in the previous sections are really a misnomer. While individual lots are aggregated into a continuum of products, we still consider individual machines or individual machine groups where a true fluid is characterized by two continuous independent variables, a time variable and a space variable. The appropriate spatial variable for a production flow characterizes the production stages or the degree of completion. We denote this variable with x and arbitrarily restrict it to the interval $[0, 1]$. Hence the fundamental variable that we consider is the product density (lot density) $\rho(x,t)$. Note that $\mathrm{d}W(0,t) = \rho(0,t)\mathrm{d}x$ is the WIP at the beginning of the factory, while $\mathrm{d}W(1,t) = \rho(1,t)\mathrm{d}x$ is the WIP at the end of the production line. For almost all manufacturing processes, especially for semiconductor fabs where lots leaving the factory have yet to be tested for their functionality, the fundamental equation describing the transport of a continuum of product through a continuum of production stages is given by a conservation equation for the product ρ.

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial F(\rho(x,t),x,t)}{\partial x} = 0 \tag{17.5}$$

where $F(\rho(x,t),x,t)$ is the flux at position x and time t which depends in a functional manner on ρ and possibly on the exact location x and time t. The influx is then given by

$$F(\rho(0,t),0,t) = \lambda(t) \tag{17.6}$$

the outflux is given by

$$F(\rho(1,t),1,t) = \mu(t) \tag{17.7}$$

and an initial WIP distribution is characterized as

$$\rho(x,0) = \rho_0(x) \tag{17.8}$$

Note that (17.5), (17.6) and (17.8) form an initial boundary value problem for a partial differential equation. If we are defining the flux as $F(x,t) = \rho(x,t)v(x,t)$ with v the fluid velocity then (17.6) is Little's law (Little 1961) averaged on timescales t and lengthscales x where $\lambda(t)$ is the average influx rate, $\rho(x,t)$ is the average WIP, and $v(x,t)$ is the inverse of the average cycle time.

Equations (17.5), (17.6) and (17.8) are a deterministic description of the flow of products through a factory. The resulting PDE is typically nonlinear and possibly nonlocal, however it is defined just on one spatial dimension. The computational effort to solve such a PDE is minimal. Hence this description is a candidate for a real-time decision tool simulating, e.g., the network of factories that make up a complicated supply chain or that describe the possible production options for a large

company. The PDE models allow a user to explore different scenarios by varying the parameters that define the network of PDEs in real time. In addition, the PDE models are inherently time dependent allowing the study of non-equilibrium or transient behavior. The price paid for the convenience of fast time-dependent simulations is that the PDE solutions describe the average behavior of a certain factory under the conditions that define the simulation. Many production scenarios are highly volatile and the variances of output of WIP are as big or bigger than the means of the processes. In that case, a tool that predicts the mean behavior is not very useful but one can argue that such production processes are inherently unpredictable and that individual sample paths generated by a discrete event simulation are just as meaningless as the time evolution of the mean behavior. However, any process where the time dependence of the mean by itself provides useful information is a candidate for a successful description by partial differential equations. In the following, we will present short descriptions of the basic model and its refinements to capture more and more of the stochasticity of the process and of the detailed decision issues in production systems. References to more in-depth discussions are given. In Sect. 17.5, we will present open problems and directions for further improvements.

The fundamental reference for the idea of modeling production flow as a fluid is in Armbruster et al. (2006a). Daganzo (2003) uses the idea of discrete kinematic waves to describe the inventory replenishment process in a supply chain. A recent paper (Göttlich et al. 2005) extends the idea to supply chain networks.

17.4.2 Flux Models

The fundamental modeling effort has been to find the right flux function F as a function of the WIP $\rho(x)$. Several first principle, heuristic and experimental attempts to find a good flux model have been discussed. Almost all of them are quasistatic or adiabatic models in the sense that the flux is not evolving in time but has a fixed functional relation to the WIP in the factory (a state equation) usually describing the functional dependence of outflux as a function of WIP in steady state. Hence any disturbance away from the state equation through, e.g., an increase in WIP caused by an increase in influx will lead to an instantaneous relaxation to the new throughput given by the state equation. The flux is written as $F = \rho v_{\text{eq}}$, $v_{\text{eq}} = v_{\text{eq}}(\rho) = 1/\tau(\rho)$ with v_{eq} the steady-state velocity and τ the average cycle time in steady state. Typical models are

- A traffic flow model (Greenshields 1935) with the equilibrium velocity

$$v_{\text{eq}}^{\text{LW}} = v_0 \left(1 - \frac{\rho}{\rho_{\max}}\right).$$

  Here v_0 is the "raw" velocity describing the flow through an empty factory, $\rho_{\max}$ is the density at which nothing moves any more in steady state and hence the density will increase without bounds (cf. a traffic jam). Note that the velocity

at stage x depends only on the WIP at stage x. Such a property is valid for traffic models and for a-cyclic production systems where every production step is performed on a single dedicated machine set.

- A model describing the whole factory as an equivalent M/M/1 queue. In that case, we have the PASTA property and the cycle time becomes $\tau = 1/v_0(1+W)$ with W the length of the queue which here is $W = \int_0^1 \rho(x)\mathrm{d}x$, i.e., total WIP. The equilibrium velocity therefore becomes

$$v_{\mathrm{eq}}^{Q1} = \frac{v_0}{1+W} A.$$

  Notice that the M/M/1 model describes a re-entrant factory: since the equilibrium velocity is the same for all parts in the queue, any change in the length of the queue will affect all WIP in the factory uniformly. This is a crude model of a highly re-entrant factory where any increase in starts will lead to a slowdown everywhere inside the factory.
- A more sophisticated re-entrant factory model is given through the use of integration kernels $w(x,\xi)$

$$v_{\mathrm{eq}}^{Q2}(x,t) = \frac{v_0}{1+\int_0^1 w(x,\xi)\rho(\xi,t)\mathrm{d}\xi}$$

  The kernels $w(x,\xi)$ describe the influence of the competition for capacity from the product located at stage ξ on the product located at position x. E.g., assuming a re-entrant production with two passes through the same machines, then for $x \in [0, 0.5]$

$$\begin{aligned} w(x,\xi) &= 0.5\delta(\xi - x) + 0.5\delta(\xi - (x+0.5)) \text{ and} \\ v_{\mathrm{eq}}^{Q2}(x,t) &= \frac{v_0}{1+0.5\rho(x,t)+0.5\rho(x+0.5,t)}, \end{aligned}$$

  with $v_{\mathrm{eq}}^{Q2}(x,t) = v_{\mathrm{eq}}^{Q2}(x+0.5,t)$.
- Detailed discrete event simulations can be used to determine the state equation through simulation. Given a DES model, we can determine average WIP in steady state for different throughputs. Assuming a clearing function model or a queuing model, we can then use least squares fits to parameterize the equilibrium throughput or the equilibrium velocity as $v_{\mathrm{eq}} = \Phi(\mathrm{WIP})$.

Figure 17.13 shows three different clearing functions for a line of 100 identical machines and an arrival process that is identical to the first machine process. The difference between the three different curves is due to different levels of variances. Notice that the capacity of the line, i.e., the horizontal asymptote for the clearing function as well as its curvature depends crucially on the stochasticity of the line. The interpolation is a least squares fit to an exponential model for the throughput μ as a function of the WIP W, $\mu = \mu_\infty(1 + \exp(-kW))$ (Asmundsson et al. 2002).

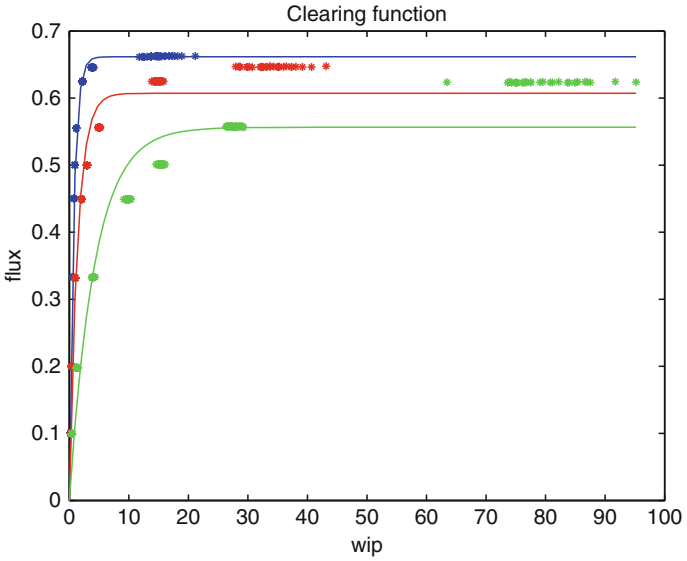

Fig. 17.13 Throughput as a function of WIP in steady state. From top to bottom, the three datasets represent coefficients of variations $c^2 = 0.1, 1$, and 6. Least squares interpolations are made for an exponential clearing function

It is obvious that the exponential decay is not a very good fit for moderate and high variances, suggesting that a low-order polynomial fit or a Pade approximation might work better. Nevertheless, only a few sets of discrete event simulations are necessary to get a general outline of the graph of the clearing function, allowing us to predict WIP and throughput times for arbitrary influxes. However, it is worth noting here that a clearing function characterizes the full state of a system—any change of the system may lead to a different clearing function. While this is obvious for the addition or removal of machines in the factory, the state is also characterized by the variances of the machines and the policies in the factory, in particular, by dispatch policies.

The major advantage of partial differential equation models is the fact that they are able to model time-dependent processes, e.g., transients. Figure 17.14a shows the average throughput for a seasonally varying input (sinusoidal) with a period of about 1 year. The noisy line comes from averaging 1,000 discrete event simulations of a model of a semiconductor factory (Perdaen et al. 2006). The continuous line shows the PDE simulation for the same experiment, where the PDE simulation is generated through a quasistatic model. The PDE simulation is quite good due to the fact that the influx varies slowly. Figure 17.14b shows the same experiment for a sinusoidal input that varies ten times faster. Now the PDE simulation seems to lag a bit relative to the discrete event simulation.

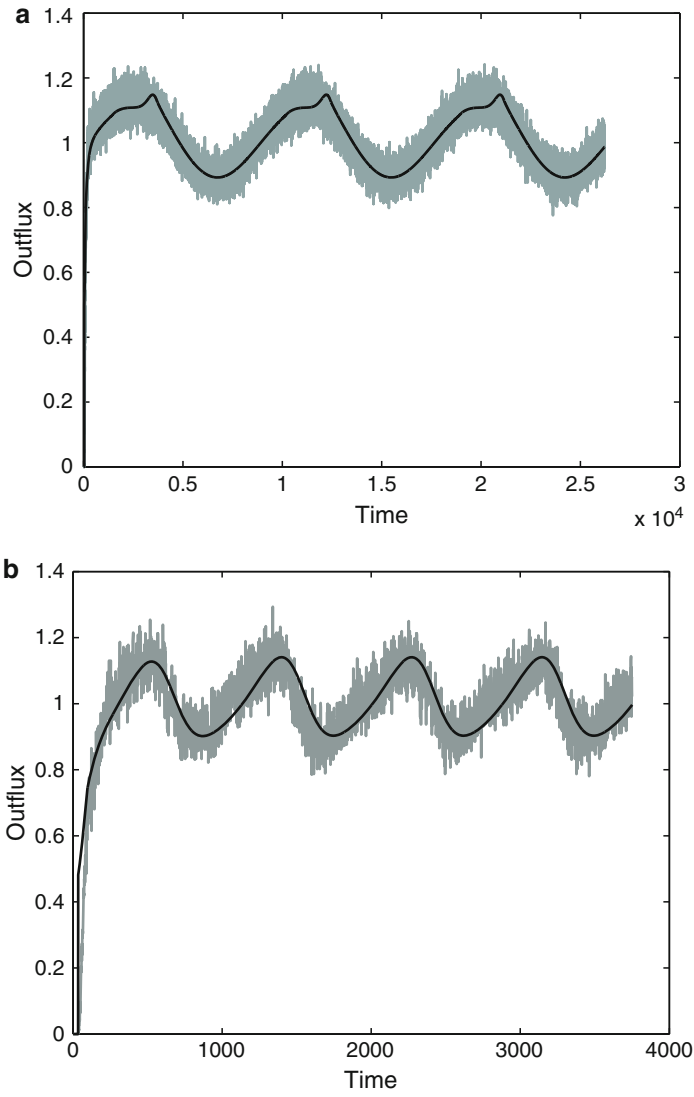

Fig. 17.14 Throughput as a function of time for a sinusoidally varying input (**a**) period of about one year, (**b**) period of about 1/10 of a year

17.4.3 Higher-Order Models and Extensions

Moment expansions. The quasistatic or adiabatic model is the zero order equation of a hierarchy of moment expansion models (Armbruster et al. 2004a). Moment expansions follow the approach of turbulence modeling or gas-dynamic modeling

of transport processes (Cercignani 1988). Here, the fundamental quantity is a probability density distribution $f(x, v, t)$ where

$$f(x, v, t)\mathrm{d}x\,\mathrm{d}v\,\mathrm{d}t = \Pr\{\xi \in [x, x + \mathrm{d}x], \eta \in [v, v + \mathrm{d}v], \tau \in [t, t + \mathrm{d}t]\}$$

describes the probability to find a particle in an x-interval with a speed in a particular v-interval in a certain time interval. The time evolution of this probability density leads to a Boltzmann equation. That Boltzmann equation is equivalent to an infinite set of equations for the time evolution of the moments of the probability distribution with respect to the velocity v. As usual a heuristic cutoff is used to reduce the infinite set to a finite set. A two moment expansion is given as

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho v}{\partial x} = 0,$$
$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = 0.$$

Boundary conditions

$$\lambda(t) = \rho(0, t)v(0, t),$$
$$v(0, t) = \frac{v_0}{1 + W(t)},$$

reflect the idea that a lot that arrives at the end of the queue has an initial expectation of a cycle time given by the length of the queue in front of it. Assuming that the velocity is constant over the whole space interval we get that $\partial v/\partial t = 0$ and hence $v = v_{\mathrm{eq}}(\rho) = v_0/1 + W$, i.e., we have the explicit closure that leads to the quasistatic approach.

Diffusion. The quasistatic approach incorporates the influence of the stochasticity, in particular, the variances of the stochastic processes, only through a shift of the means (e.g., mean capacity, mean cycle time, etc.). A typical model that includes the variances explicitly is given through an advection diffusion equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial F}{\partial x} = 0, \tag{17.9}$$
$$F = v_{\mathrm{eq}}\rho - D\frac{\partial \rho}{\partial x}, \tag{17.10}$$

where the advection process describes the deterministic evolution of the means and the diffusion process, parameterized through the diffusion coefficient D, models the behavior of a Brownian motion superimposed on these means. Armbruster and Ringhofer (2005) derive such an equation from first principles, based on a transport process that randomly updates the transport velocity from a density-dependent probability distribution. To model the re-entrant influence, the velocity is random

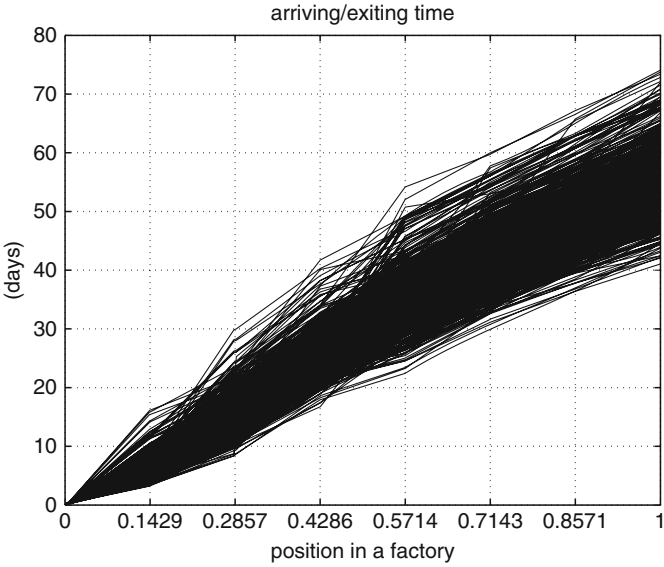

Fig. 17.15 Paths of 920 lots through an INTEL factory

in time but constant over all stages. A expansion based on an infinite number of machines and an infinite number of velocity updates of the associated Boltzmann equation leads to (17.10).

It is easy to show the presence of diffusion in real factory data as well as in discrete event simulations. Any state of the art production facility will be able to determine the exact location of any lot that goes through the factory at any given time. Figure 17.15 shows a crude approximation to the paths of 920 lots through a real INTEL factory. By starting all lots at the same place and time, the resulting fan in Fig. 17.15 is an indication of the diffusion process. Slicing the data in Fig. 17.15 at fixed times, we can generate histograms of the number of lots as a function of position in the factory. Figure 17.16 shows that, as expected from the central limit theorem, the distribution of WIP toward the end of the factory is reasonably well approximated by a normal distribution. Standard fitting procedures will allow us to determine the state equation v_{eq} and the diffusion coefficient D in (17.10) (Armbruster et al. 2004b).

17.4.4 Control of Production Lines

Having a differential equation model for a production line opens up the field of continuous control (see also Lefeber 2004; Göttlich et al. 2006). While there are still many open questions, two initial attempts have been successful.

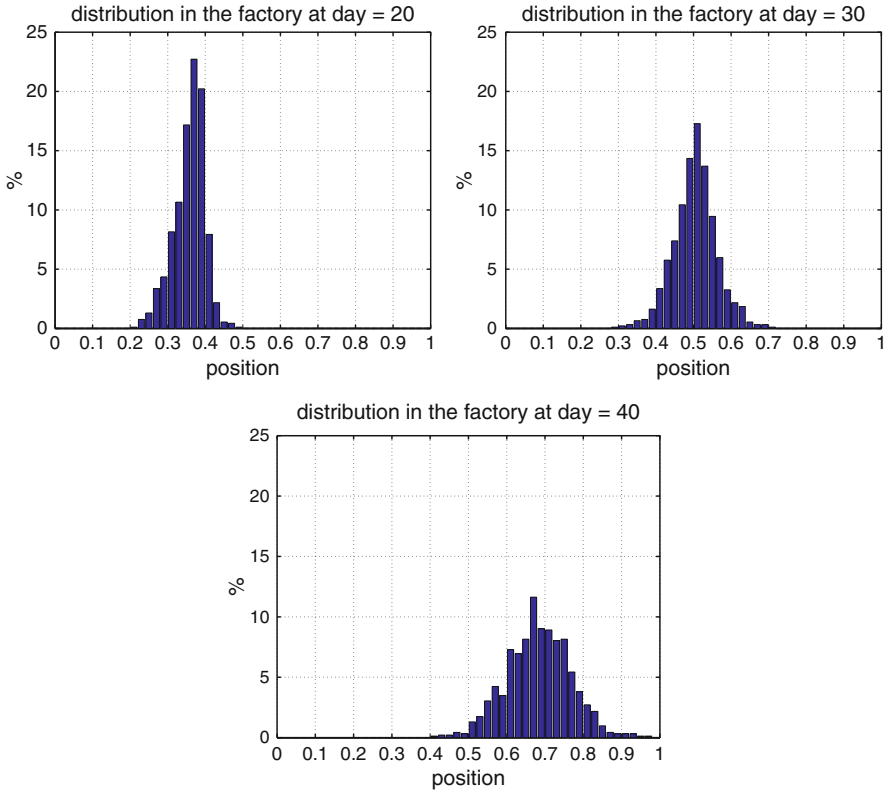

Fig. 17.16 Histograms of positions of the lots in the factory at time $t = 20$, $t = 30$, and $t = 40$

17.4.4.1 Control Via the Push-Pull Point

The cycle time through a semiconductor fab is several weeks. Hence, typically the starts into the fab are done "to plan" while the delivery out of the fab is "to orders." This reflects itself in the dispatch policies at the re-entrant machines. At the beginning of the factory, we have a push policy, favoring lots requiring early production stages over lots waiting for high production stages, whereas at the end of the factory we have a pull policy which tries to affect output by favoring the final steps over earlier steps. Somewhere in the middle of the factory there is a production stage where the push policy changes into a pull policy. That stage is called the push-pull point and it is one of the few possible control actuators inside the factory that might influence the output of the factory. In Perdaen et al. (2006), we have studied the use of changing the push-pull point to affect the tracking of a demand signal in a discrete event simulation of a semiconductor fab. We assume that we have a demand curve as a function of time, and a time interval in which the demand is of the order of magnitude of half of the total WIP of the factory. We then place the push-pull point in such a way that the demand over that time interval matches the total WIP downstream from the push-pull point.

The final result is that a push-pull control algorithm will not significantly improve the factory output for an open system where the WIP is uncontrolled. If we are using the push-pull algorithm together with a CONWIP policy, then the demand-outflux mismatch over a fixed time interval is reduced by a factor of 5–6, for a demand signal with a coefficient of variation $c = \sigma/\mu = 0.4$.

This control algorithm and its implementation have nothing to do per se with a continuum model of the factory. However, a continuum description provides a framework to understand the DES result: since the average cycle time for a lot under a pull policy is shorter than for a lot produced under a push policy, the associated average velocity for a pull policy is higher than for a push policy. Assuming for this argument a uniform velocity in the factory in steady state, the WIP profile $\rho(x) = \lambda/v$ will be constant, independent of x and t. We consider the upstream part of the production line as a homogeneous push line and the downstream part as a homogeneous pull line, each with its own constant velocity with $v_{\text{push}} < v_{\text{pull}}$. Since the throughput is the same everywhere and since $\rho v = \lambda$ has to hold, we get a jump in the WIP profile at the push-pull point by the amount

$$\frac{\rho_{\text{push}}}{\rho_{\text{pull}}} = \frac{v_{\text{pull}}}{v_{\text{push}}}. \tag{17.11}$$

Figure 17.17a shows the constant throughput and the discontinuous WIP profile.

When we now instantaneously move the PPP upstream by an amount Δx then the queues that were just upstream of the PPP and hence had the lowest priority on the line move up in priority and therefore speed up. Hence the product of $\rho_{\text{push}} v_{\text{pull}} > \lambda$, i.e., we create a flux bump. Similarly we create a flux dip by moving the PPP downstream. Keeping the PPP at its new location, the flux bump is downstream from the PPP and hence moves downstream with the constant speed v_{pull} pulling a WIP bump with it until they both exit the factory. During the time they exit, they will increase the outflux. Figure 17.17b, c shows this time evolution. After the WIP/flux bump has exited, the total WIP in the factory is lower and hence in order to satisfy the same demand, the push-pull point will have to move yet further upstream driving it toward the beginning of the factory.

In contrast, the time evolution of the flux bump for the PPP-CONWIP policy is illustrated in Fig. 17.18.

As the CONWIP policy is implemented by matching the starts to the outflux, once the WIP bump moves out of the factory, the starts will be increased to create a new WIP bump. In that way, the total throughput will stay high until the PPP point is moved downstream again. That will happen when the backlog has moved to zero and the sum of actual backlog and actual demand has decreased. In that way we have a policy that reverts all the time to a match between demand and outflux.

17.4.4.2 Creating an Arbitrary WIP Profile

One problem that represents a step to the practically more interesting problems (see Sect. 17.5) is the following: given a WIP profile $\rho_1(x), 0 \le x \le 1$ and a quasistatic model of a production system determined by $v_{\text{eq}} = \Phi(\text{WIP})$, what is the influx $\lambda(t)$

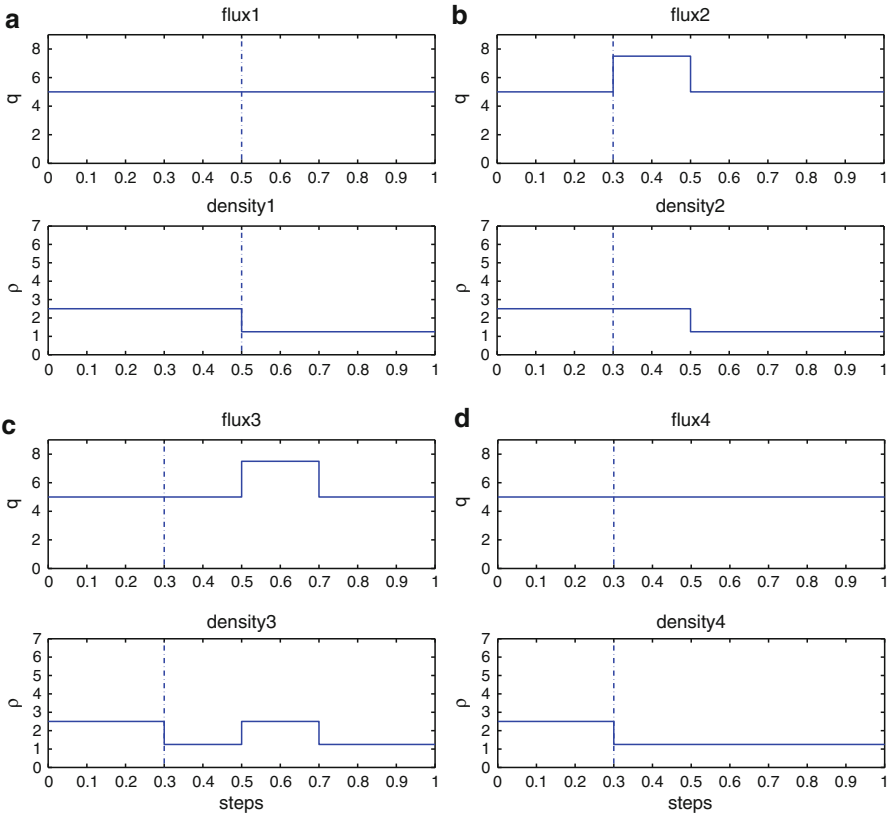

Fig. 17.17 Stages of creating a flux bump (**a**)–(**d**) show subsequent snapshots of the flux and WIP profile. For details and interpretation, see text

to generate a desired new WIP profile $\rho_2(x)$, subject to a time evolution determined by the PDE

$$\begin{aligned} &\rho_t + v_{\text{eq}}\rho_x = 0, && x \in (0,1)\,,\ t > 0. \\ &\lambda(t) = v(t)\rho(0,t), && t > 0. \end{aligned}$$

An implicit analytical solution involves the simple idea of letting the initial profile travel out through the right boundary while the new profile travels in through the left boundary.

$$\rho(x,t) = \begin{cases} \rho_1(x - \int_0^t v(s)\mathrm{d}s) & \text{if } \int_0^t v(s)\mathrm{d}s \le x \le 1 \\ \rho_2(1 + x - \int_0^t v(s)\mathrm{d}s) & \text{if } 0 \le x < \int_0^t v(s)\mathrm{d}s \le 1. \end{cases} \tag{17.12}$$

From (17.12), we can determine the influx $\lambda(t) = v(t)\rho_2(1 - \int_0^t v(s)\mathrm{d}s)$. The transit time T for the initial profile $\rho_1(x)$ is defined by $1 = \int_0^T v(s)\mathrm{d}s$. Note that (17.12) is

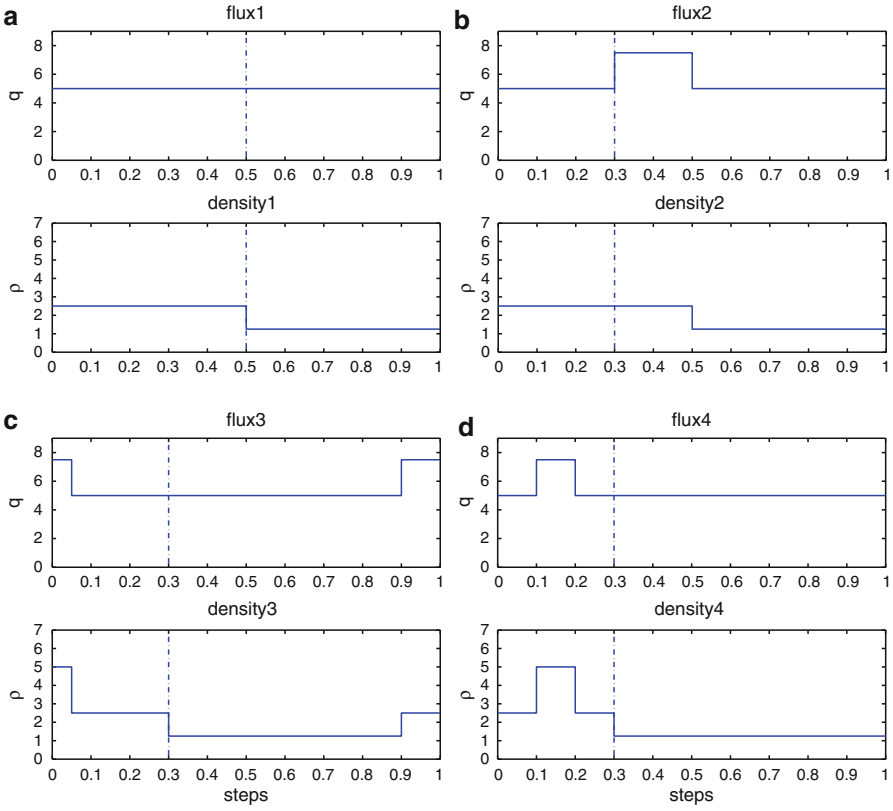

Fig. 17.18 Stages of creating a flux bump for a PPP-CONWIP policy (**a**)–(**d**) show subsequent snapshots of the flux and WIP profile. For details and interpretation, see text

a general solution for all time-dependent functions of velocity, especially including those based on the load $\int_0^1 \rho(x,t)\mathrm{d}x$. Furthermore, it is an implicit solution as the density $\rho(x,t)$ and hence the influx $\lambda(t)$ depend on the velocity $v(\rho(x,t),x,t)$ and its history.

A feasible numerical method to find an explicit solution for $\rho(x,t)$ and $\lambda(t)$ consists of the following steps:

1. Discretize in space and initialize $\rho(x_j,0)$ to $\rho_1(x_j)$ for all space points $j = 1\ldots N$.
2. Determine $\rho(x_j,t_n+\delta t)$ by using a hyperbolic PDE solver and evaluate $v = v(t_n+\delta t)$. Integrate $\int_0^{t_n+\delta t} v(s)\mathrm{d}s$ and set $\rho(0,t_n+\delta t) = \rho_2(1-\int_0^{t_n+\delta t} v(s)\mathrm{d}s)$. Set $\lambda(t_n+\delta t) = v(t_n+\delta t)\rho(0,t_n+\delta t)$. Repeat until $\int_0^{t_n+\delta t} v(s)\mathrm{d}s = 1$

Figure 17.19a shows a starting profile $\rho_1(x)$ and an end profile $\rho_2(x)$. Figure 17.19b shows the influx $\lambda(t)$ that generates the new WIP profile for the state equation $v(t) = v_0/(1+\int_0^1 \rho(x,t)\mathrm{d}x)$.

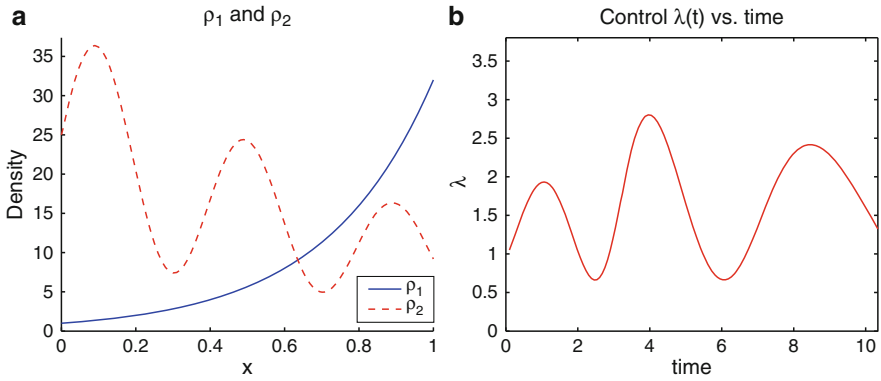

Fig. 17.19 (**a**) Two WIP profiles $\rho_1(x)$ and $\rho_2(x)$ and (**b**) the influx $\lambda(t)$ that transforms ρ_1 into ρ_2

17.5 Conclusions and Open Problems

We have presented three approaches to aggregate modeling of production lines: EPTs, clearing functions, and continuum models (PDEs). EPT is a tool to separate waiting for the availability of a machine from all other sources of variability that extend the processing time. EPTs are easy to measure and allow the development of discrete event simulations that aggregate many different and hard to characterize stochastic processes into one processing time. Alternatively, we can use EPTs to develop relatively simple queueing networks. We have shown that EPTs are utilization independent and that they can be defined for machines that work in parallel, for production lines with finite buffers, and for batch processes.

The next level of aggregation treats the products as a continuum and in that way loses the concept of an event. The resulting model consists of ordinary differential equations that reflect the queues in front of machines and their dynamics driven by the balance of influx and outflux. Together with the loss of the event, clearing function models also lose the stochastic behavior – a clearing function is a input-output relation that reflects the *average* behavior of the system that it is modeling. Simple queues allow an exact determination of the clearing function relationship but most networks require either off-line simulations or queueing approximations to determine the shape of the clearing function numerically.

Continuum models treat the whole production process as a continuum in products and a continuum in production steps. The resulting partial differential equations are typically hyperbolic and describe the movement of products through a factory as a WIP-wave. Different levels of scale and accuracy have been presented. The lowest level of accuracy is represented by a quasistatic approach that connects the PDE models to the clearing function models by using the clearing function as a state equation. The major advantage of continuum models is that they are scale independent, i.e., their simulation does not depend on the number of lots produced nor the number of stages that the lot is going through. A second advantage is

that they allow the study of nonequilibrium and transient effects, something that can rarely be done in queueing models. Like the clearing function approach they are deterministic and typically represent the mean transport behavior, although the time evolution of higher-order moments can in principle be studied. PDE models can be extended to networks of factories (supply chains) (Armbruster et al. 2006a; Göttlich et al. 2005) and they can be set up to include policies (dispatch or global) (Armbruster et al. 2006b).

An interesting study for further research would be to compare the computational efforts as well as the performance of the four modeling approaches.

A major open problem for the continuum model approach is the following:

- In Armbruster and Ringhofer (2005) we have derived an advection diffusion equation from first principles that describe the mean time evolution of a certain stochastic production process. However, the process we used involved stochastically varying spatially homogeneous velocities which are not easily related to the usual characterization of the stochasticity of production. The latter is typically described through stochastically varying capacity reflecting the tool manufacturer's characterization of a machine through its time distribution for failure and its time distribution for repair. We are working on developing PDEs whose parameters are determined by a priori given distributions for those times.

Other open problems involve control and optimization of production:

- What is the influx $\lambda(t)$ that moves a production line from an equilibrium state with throughput d_1 to a new equilibrium state with throughput d_2 in shortest possible time.
- Given an initial WIP profile $\rho_0(x, t_0)$ and a demand signal $d(t)$ for $t_0 \leq t \leq t_0 + T$ for some time interval T. What is the input $\lambda(t)$ that minimizes the difference between the output and the demand over that time interval.

We are currently exploring variational methods analogous to optimal control problems for parabolic equations (Göttlich et al. 2006) to solve these optimal control problems.

Acknowledgments This work was supported in parts by NSF grant DMS 0604986. We thank Pascal Etman, Ad Kock, and Christian Ringhofer for many insightful discussions and Bas Coenen, Ton Geubbels, Michael Lamarca, Martijn Peeters, Dominique Perdaen, and William van den Bremer for computational assistance.

References

Armbruster D, Ringhofer C (2005) Thermalized kinetic and fluid models for reentrant supply chains. SIAM J Multiscale model Simul, 3(4):782–800

Asmundsson J, Uzsoy R, Rardin RL (2002) Compact nonlinear capacitymodelsfor supply chains: methodology. Technical report, Purdue University. preprint

Armbruster D, Marthaler D, Ringhofer C (2004a) Kinetic and fluid model hierarchies for supply chains. SIAM J Multiscale model and Simul 2(1):43–61

Armbruster D, Ringhofer C, Jo T-J (2004b) Continuous models for production flows. In: Proceedings of the 2004 American control conference, Boston, MA, pp 4589 – 4594

Armbruster D, Marthaler D, Ringhofer C, Kempf K, Jo T-J (2006a). A continuum model for a re-entrant factory. Oper Res 54(5):933–950

Armbruster D, Degond P, Ringhofer C (2006b) Kinetic and fluid models for supply chains supporting policy attributes. Technical report, Arizona State University, Department of Mathematics and Statistics. Accepted for publication in Transp Theory Stat Phys

Banks J (1998) Handbook of simulation: principles, methodology, advances, applications, and practice. John Wiley & Sons, Inc., USA

Cassandras CG, Lafortune S (1999) Introduction to discrete event systems. Kluwer Academic Publishers, Norwell, MA

Cercignani C (1988) The Boltzmann Equation and its applications. Springer Verlag, New York, NY

Daganzo CF (2003) A theory of supply chains. Springer Verlag, New York, NY

van der Eerden J, Saenger T, Walbrick W, Niesing H, Schuurhuis R (2006) Litho area cycle time reduction in an advanced semiconductor manufacturing line. In: Proceedings of the 2006 IEEE/SEMI advanced semiconductor manufacturing conference, Boston, MA, pp 114–119

Göttlich S, Herty M, Klar A (2005) Network models for supply chains. *Commun Math Sci* 3(4):545–559

Göttlich S, Herty M, Klar A (2006) Modelling and optimization of supply chains on complex networks. *Commun Math Sci* 4(2):315–330

Graves SC (1986) A tactical planning model for a job shop. *Oper Res* 34(4):522–533

Greenshields BD, Bibbins JR, Channing WS, Miller HH (1935) A study of highway capacity. *Highway research board proceedings*, vol 14, Part 1, Washington DC, pp 448–477

Hopp WJ, Spearman ML (2000) *Factory Physics*. 2nd edn. Irwin/McGraw-Hill, New York, NY

Jacobs JH, Etman LFP, van Campen EJJ, Rooda JE (2001) Quantifying operational time variability: the missing parameter for cycle time reduction. In: Proceedings of the 2001 IEEE/SEMI advanced semiconductor manufacturing conference, pp 1–10

Jacobs JH, Etman LFP, van Campen EJJ, Rooda JE (2003) Characterization of the operational time variability using effective processing times. *IEEE Trans Semicond Manuf*, 16(3):511–520

Jacobs JH (2004) Performance quantification and simulation optimization of manufacturing flow lines. Phd thesis, Eindhoven University of Technology, Eindhoven, The Netherlands

Jacobs JH, van Bakel PP, Etman LFP, Rooda JE (2006) Quantifying variability of batching equipment using effective process times. *IEEE Trans Semicond Manuf*, 19(2):269–275

Kock AAA, Wullems FJJ, Etman LFP, Adan IJBF, Rooda JE (2005) Performance evaluation and lumped parameter modeling of single server flowlines subject to blocking: an effective process time approach. In: Proceedings of the 5th international conference on analysis of manufacturing systems and production management, Zakynthos Island, Greece, pp 137–144

Kock AAA, Etman LFP, Rooda JE (2006a) Effective process time for multi-server tandem queues with finite buffers. SE Report 2006–08, Eindhoven University of Technology, Systems Engineering Group, Department of Mechanical Engineering, Eindhoven, The Netherlands. Submitted for publication. http://se.wtb.tue.nl/sereports/

Kock AAA, Etman LFP, Rooda, JE (2006b) Lumped parameter modelling of the litho cell. In: Proceedings of the INCOM06, vol 2. St. Etienne, France, pp 709–714

Kock AAA, Wullems FJJ, Etman LFP, Adan IJBF, Nijsse F, Rooda JE (2006c) Performance evaluation and lumped parameter modeling of single server flowlines subject to blocking: an effective process time approach. SE Report 2006–09, Eindhoven University of Technology, Systems Engineering Group, Department of Mechanical Engineering, Eindhoven, The Netherlands. Submitted for publication. Available via http://se.wtb.tue.nl/sereports/

Lefeber E (2004) Nonlinear models for control of manufacturing systems. In: Radons G, Neugebauer R (eds) Nonlinear dynamics of production systems, Weinheim, Germany, pp 69–81

Little JDC (1961) A proof for the queuing formula $l = \lambda w$. *Oper Res*, 9:383–387

Perdaen D, Armbruster D, Kempf K, Lefeber E (2007) Controlling a re-entrant manufacturing line via the push-pull point. Technical report, Arizona State University. preprint, in revision for the Int J Prod Res

de Ron AJ, Rooda JE (2005) Fab performance. *IEEE Tran Semicond Manuf*, 18(3):399–405

Takahasi K, Sakasegawa H (1977) A randomized response technique without making use of a randomizing device *Ann Ins Stat Math*, 29a:1–8

Sattler L (1996) Using queueing curve approximations in a fab to determine productivity improvements. In: Proceedings of the 1996 IEEE/SEMI advanced semiconductor manufacturing conference, Cambridge, MA, pp 140–145

Veeger CPL, Etman LFP, Lefeber E, Adan IJBF, van Herk J (2009) Predicting cycle time distributions for integrated processing workstations: an aggregate modeling approach. EURANDOM Report 2009-034, Eindhoven University of Technology, EURANDOM, Eindhoven, The Netherlands. Submitted for publication. Available via http://www.eurandom.tue.nl/reports/