OPERATIONS RESEARCH
Vol. 54, No. 5, September–October 2006, pp. 933–950
ISSN 0030-364X | EISSN 1526-5463 | 06 | 5405 | 0933

DOI 10.1287/opre.1060.0321

A Continuum Model for a Re-entrant Factory

Dieter Armbruster
Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804, armbruster@asu.edu

Daniel E. Marthaler
Northrup Grumman Integrated Systems, Western Region, 17066 Goldentop Road, 9V21/R3-2, San Diego, California 92127-2412, daniel.marthaler@ngc.com

Christian Ringhofer
Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804, ringhofer@asu.edu

Karl Kempf
Decision Technologies, Intel Corporation, 5000 West Chandler Boulevard, MS CH3-10, Chandler, Arizona 85226, karl.g.kempf@intel.com

Tae-Chang Jo
Mathematics Department, Inha University, 253, Yonghyun-Dong, Nam-Ku, Incheon, 402-751, South Korea, taechang@inha.ac.kr

High-volume, multistage continuous production flow through a re-entrant factory is modeled through a conservation law for a continuous-density variable on a continuous-production line augmented by a state equation for the speed of the production along the production line. The resulting nonlinear, nonlocal hyperbolic conservation law allows fast and accurate simulations. Little's law is built into the model. It is argued that the state equation for a re-entrant factory should be nonlinear. Comparisons of simulations of the partial differential equation (PDE) model and discrete-event simulations are presented. A general analysis of the model shows that for any nonlinear state equation there exist two steady states of production below a critical start rate: A high-volume, high-throughput time state and a low-volume, low-throughput time state. The stability of the low-volume state is proved. Output is controlled by adjusting the start rate to a changed demand rate. Two linear factories and a re-entrant factory, each one modeled by a hyperbolic conservation law, are linked to provide proof of concept for efficient supply chain simulations. Instantaneous density and flux through the supply chain as well as work in progress (WIP) and output as a function of time are presented. Extensions to include multiple product flows and preference rules for products and dispatch rules for re-entrant choices are discussed.

Subject classifications: production/scheduling: approximations; simulations: efficiency; mathematics.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received January 2003; revisions received October 2003, April 2004, April 2005; accepted September 2005.

1. Introduction

In recent years, fast scalable simulations of production flows in a supply chain have become an important research topic (Scalable Enterprise Initiative 1999). While the long-term goal is to optimize production across the whole supply chain, an intermediate goal is to generate simulation tools that support the exploration of business questions, and to pose "what if?" questions on these simulations. Because most production deals with individual parts and the processes that these parts undergo, the natural method of choice for accurate simulations is discrete-event simulators. While they have been very successful on the factory level (Law and Kelton 1991), e.g., to simulate semiconductor production lines, they are relatively slow, and it seems obvious that they are not scalable to a full supply chain.

This paper proposes to develop a new approach to the simulation of production flow in analogy to traffic flow. Specifically, we will model the dynamics of material flow through a factory via a hyperbolic conservation law. The main variable that is modeled will be the density of product in a factory. The specifics of the production process will enter into a state equation relating the velocity of the product to the density of the material in the factory. The focus in our model is on re-entrant factories because they represent the most important part of arguably the most efficient current production systems, the production of semiconductor devices. However, we maintain that our approach can be tailored to many other production strategies and processes. The major advantage of hyperbolic conservation law models is the existence of a large body of numerical codes to solve such equations efficiently.

The rest of this paper is organized as follows: §2 discusses the basic model and its relationship to traffic modeling. The shape of the state equation for acyclic and re-entrant flows is discussed. It shows that a hyperbolic conservation law can be written that corresponds to Little's law. Section 3 briefly discusses the numerical method to simulate our re-entrant factory model. We discuss the applicability and limitations of our model and compare it to several discrete-event simulations in §4. Section 5 discusses the existence of multiple equilibria for a nonlinear state equation with constant start rates into a factory. There exist two equilibria: an equilibrium where a high product density moves slowly, and another one where a low product density moves fast, such that the product of speed times density gives the same production rate in both cases. We will show that only the low-density equilibrium is stable under perturbations. Section 6 deals with the issue of controlling the output of the factory by adjusting the start rate to a changed demand rate. Section 7 shows proof of concept for the original goal of fast supply chain simulations: We link three nodes, two linear factories, and a re-entrant factory, each one modeled by a hyperbolic conservation law. The model of the linear factories has been discussed in Armbruster et al. (2004). Instantaneous density, as well as flux through the supply chain for several scenarios, is shown. Work in progress (WIP) and output as a function of time are presented. We conclude with a discussion of the influence of dispatch rules and an outlook on attempts to derive the continuum models from first principles.

2. The Model

There are currently three major approaches to simulate production flows: Discrete-event simulations (DES), fluid networks, and conventional queueing network models. DES has successfully been used in large simulations for semiconductor factories (for instance, Chen et al. 1988 and Law and Kelton 1991), but typically is very time consuming. Fluid models (Dai and Weiss 1996, Kumar 1993) come from traffic theory and were introduced by Newell (1965, 1973) to approximately solve queueing problems. They consider the length of a queue $q(t)$ as a continuous variable whose rate of change is given by

$$\frac{dq}{dt} = \begin{cases} \lambda(t) - \mu(t) & \text{for } q(t) \neq 0, \\ 0 & \text{for } q(t) = 0, \end{cases} \tag{1}$$

where $\lambda(t)$ is the arrival rate and $\mu(t)$ is the processing rate of the queue. This basic building block for a queue can be connected to a *work-conserving fluid model* by feeding the outflux of each queue into other queues. Dai and Weiss (1996) have analyzed the relationship between the stability of the fluid model and the stability of scheduling policies for the associated queueing networks. Here, stability of the fluid model is represented by the boundedness of the fluid variables for a given influx $\lambda(t)$, which is assumed to be less than the smallest processing rate $\mu_i(t)$ of a queue in the network. Stability in the queueing theory sense is given by a unique stationary distribution ψ for the underlying stochastic process describing the queueing network. Dai and Weiss (1996) in particular showed that a queueing discipline is stable if the corresponding fluid model is stable. Queueing network models have successfully been used, for instance, to develop input and priority sequencing policies in two-station networks (Wein 1990).

Fluid models have several distinct shortcomings:

- They do not model stochasticity very well. Either $\lambda(t)$ and $\mu(t)$ are mean rates, in which case Equation (1) is a fully deterministic system and stochasticity is not modeled at all, or $\lambda(t)$ and $\mu(t)$ are stochastic processes that allow some theoretical analysis, but drastically diminish the advantages of a continuum model as a simulation tool.
- Assuming constant production rates, fluid models are too rigid. If they are stable, there will never be any WIP waiting in a queue, and if they are unstable, the queues will explode. Similarly, fluctuations in one queue can never travel downstream (or upstream) to the next queue unless the queue actually becomes zero.
- While a fluid network models the WIP with a continuous variable, it models machines as individual discrete queues. With several hundred production steps for a typical chip, it is reasonable to also approximate the production steps along a continuum.

Systems dynamics (Forrester 1962) is related in spirit to a fluid model. It will derive ordinary differential equations (ODEs) that model the flow from one machine to another and analyze the resulting large system of ODEs. However, it is unclear how the correct throughput time will be incorporated in such a system. Either the ODEs will become delay equations (which makes them much harder to solve), or intermediate states will have to be introduced (Hines 2003). In any case, the changes in throughput time due to increased load in a factory will not be included.

Queueing network models have their own limitations:

- While they can deal with several re-entrant steps by introducing different customer classes, their complexity explodes as the number of machines increases.
- They solve control-type problems for a queueing network in steady state, i.e., all their result are valid in the limit as $t \to \infty$.
- They have to ignore all incidence in the factories that cannot be modeled by queueing type of behavior (see below for examples).

In a modern semiconductor factory, we are interested in modeling and simulating on the order of 250 production steps executed on about 100 machines, with a re-entrant part of the production line that cycles about 15–20 times. In addition, the life cycle of a product is of the order of one year, whereas the throughput time lies between 40 and 60 days. Hence, it is unlikely that the factory is ever run for any longer amount of time in steady state, and we are especially interested in *transient* behavior of such systems. A model that addresses these issues can be derived from traffic modelling: The simplest model for traffic on a one-lane highway without on and off ramps is a hyperbolic conservation law of the form

$$\frac{\partial \rho}{\partial t} + \frac{\partial (v(\rho)\rho)}{\partial x} = 0, \tag{2}$$

where subscripts denote partial differentiation. The velocity $v(\rho)$ is described by a state equation relating vehicle velocity v and vehicle density ρ through (Lighthill and Whitham 1955, Helbing 1996)

$$v(\rho(x,t)) = v_0\left(1 - \frac{\rho(x,t)}{\rho_m}\right), \tag{3}$$

with ρ_m the capacity of the highway. This is called the *Lighthill-Whitham model* and is the first step in a hierarchy of traffic models (Helbing 1996).

We propose a full-continuum model for the production flow through a re-entrant factory with a similar structure: Let x be a continuous variable representing completion of the product, i.e., product at $x = 0$ denotes a raw product and parts at $x = 1$ denote a finished product. While $x = 0$ corresponds to the entry into the factory and $x = 1$ corresponds to the exit from the factory, there is no one-to-one correspondence between factory floor space and x. However, there is a unique production process assigned to every x-value. Assuming a high-volume, many-stage factory, we model production flow with a continuum variable on a continuum domain. We write $u(x,t)$ for the density of product at stage x and at time t. Assuming a unique entry and exit for the factory, i.e., all product enters at $x = 0$ and leaves at $x = 1$, and assuming a 100% yield, the density must satisfy

$$\frac{\partial u}{\partial t} + \frac{\partial(v(u)u)}{\partial x} = 0, \tag{4}$$

where

$$v(u) = \phi(u) \tag{5}$$

is the state equation relating the speed of the product moving through the factory to the amount of product in the factory, i.e., to WIP. Note that the units of u are [parts][stage] and the units of $v(u)u$ are [parts][time]. This suggests that in conventional nomenclature of process control and performance simulation, $u(x,t)$ represents local WIP density and the flux $u(x,t)v(x,t)$ is the local throughput at stage x at time t, respectively. Recently, based on the earlier work of Graves (1985) and Karmarkar (1989), Asmundsson et al. (2002) have developed a similar idea. To model the nonlinear response of the throughput to an increase in WIP, they introduce the idea of *nonlinear clearing functions* of the form

$$\text{throughput} = \alpha(\text{WIP})\text{WIP}. \tag{6}$$

Because throughput is a flux, α(WIP) is a velocity, and hence equivalent to our state equation $\phi(u)$ (Equation (5)). Asmundsson et al. (2002) choose specific nonlinear clearing functions motivated by the theory of queueing networks, but emphasize that only empirical data will be able to decide the correct clearing function for a particular problem. They do not incorporate their clearing function idea into a differential equation. Instead, they use the clearing function to generate a linearization of the supply chain planning problem that can be solved via an LP algorithm.

The basic issue for our model is to come up with a useful state equation $\phi(u)$. As a starting point, the model has to respect the fundamental law of factory physics, Little's law (Little 1961). For a single factory, Little's law may be written as

$$N = \tau\lambda, \tag{7}$$

where N is the time-averaged load of the factory (WIP), τ is the mean cycle or throughput time over all outputs (TPT), and λ is the start rate. Little's law is fundamentally a deterministic law and results from mass conservation. However, by amending it with a description of the stochastic processes in a factory, we can generate a state equation characterizing the factory. In its simplest case, the stochastic process is represented by its means. For instance, modeling a factory as a single linear queue with Markov arrival and processing rates (an $M/M/1$ queue), the mean throughput time τ can be determined as a function of the start rate λ and processing rate μ (Gross and Harris 1985) to be

$$\tau = \frac{1}{\mu - \lambda}. \tag{8}$$

Therefore, the relationship between WIP and throughput time becomes

$$\tau = \frac{1}{\mu}(1 + N). \tag{9}$$

Equation (8) shows that μ becomes the maximal or critical start rate. With $v = 1/\tau$, we have the desired state equation. Equation (9) shows that WIP is a linear function of the throughput time with a slope μ. The linear relationship between TPT and WIP is intuitively obvious: A part entering the queue will have to wait until the queue is served (N/μ timeunits) plus its own processing time, $1/\mu$ time units.

A linear relationship as in Equation (9) is true for a queueing network that has product form (Nelson 1995), i.e., the whole network can be replaced by an effective queue. BCMP networks (Baskett et al. 1975) and Kelly networks (Nelson 1995) are of product type. Such networks have quite restrictive conditions on the stochastic processes of the network. In general, the Pollaczek Khintchine formula (Gross and Harris 1985) suggests that the mean throughput time depends on the first- and second-order moments (the means and the variances) of all the stochastic processes involved, and hence in general may lead to a nonlinear state equation $\phi(\rho)$. Specifically, any process that increases variance as the load in the factory is increased will lead to a nonlinear state equation. For instance, the detailed empirical analysis by Chen et al. (1988) of a re-entrant semiconductor factory shows that a discrete-event simulation based on a product network predicts throughput times within about 10% of the actual throughput times. It depends on your point of view whether 10% accuracy is a large or a small number for describing a factory that is a billion-dollar investment. Lu et al. (1994) show that through dispatch and scheduling rules, one can reduce the variance of throughput times inside semiconductor manufacturing plants. Hence, there are several key factors that will lead to a queueing network that cannot be approximated by a product network with constant production rates:

- Chen et al. (1988) report that 35% of the throughput time of a chip in the Hewlett Packard (HP) factory studied was used up on engineering holds. Basically, those are the queues in front of a human operator, who has to determine whether the previous process worked properly or whether reworking is in order. While Chen et al.'s work analyzes a development factory, which may result in a higher incidence of human operator interactions than in a regular production fab, it is nevertheless true for all human operators that they do not have a constant production rate. For instance, it is known that bank tellers work faster if the queues in front of them are larger. However, they often do that by cutting corners and reducing quality of service. For a model of a semiconductor factory with constant yield, this may lead to the paradoxical effect that the operators work faster, but due to errors the mean throughput time of the good chips will actually increase.
- Recent simulations of fully automated 300 mm wafer fabs (Shikalgar et al. 2002) show that increased loading may lead to crowding effects of the transportation system, in turn leading to nonlinear dependence of throughput on WIP.
- Scheduling and dispatch policies play a major role: Kumar (1993) discusses a mixed push/pull policy that leads to a blowup in inventory even though all start rates are smaller than all production rates.
- Similarly, the effects of hot lots and other priority rules for multiproduct flows will destroy the product property of a queueing network. DeJong and Wu (2002), for instance, show that priority lots have an exponential impact on the throughput of the nonpriority lots.
- Even for the restricted models of fluid networks (which do not include operator interactions and multiproduct flows), Dai and Weiss (1996) showed that dispatch and scheduling policies play a major role in the stability of the queuing regime.
- Daganzo (2003), in discussing the application of traffic flow models on controls and release strategies for a whole supply chain, analyzes situations where the state equation is less determined by any physical constraints, but determined by the reorder policies in the supply chain.
- Finally, there is anecdotal evidence from real re-entrant factories (Kempf 2001) that show a stronger than linear increase in the average throughput time τ as the loading of the factory is increased.

Unfortunately, the true state equation for a re-entrant factory is not known, as large factories are rarely (if ever) run in equilibrium, and controlled experiments are obviously impossible. It is worth noting here that standard discrete-event simulators may not be a good substitute for real experiments, specifically if human interaction is part of the cause for the increase in variance (Spier and Kempf 1995).

For future reference, we restate the full model, with $f(x)$ an initial density distribution, and taking into account that v does not depend on x:

$$\begin{aligned} &\frac{\partial u}{\partial t}+v(u)\frac{\partial u}{\partial x}=0,\\ &u(x,0)=f(x),\\ &u(0,t)v(t)=\lambda(t), \end{aligned} \tag{10}$$

with a state equation

$$v(u)=\phi(u). \tag{11}$$

As in the above examples, the velocity given by the state equation is a functional of u and typically depends on the total WIP in the factory, i.e., on $\int_0^1 u(x,t)\,dx$ or on suitably weighted WIP distributions. Note that the start rate $\lambda(t)$ into the factory enters as the boundary condition for the local throughput at $x=0$.

Our model shares many features with thermodynamic transport equations. In particular, modeling the relationship between density and velocity through a state equation is typically called an *adiabatic approximation*: Factory flow is modeled as if it were always in equilibrium, following adiabatically the state Equation (11). In the thermodynamics of gases, this corresponds to the ideal gas equations. Elementary physics tells us that the ideal gas equations are strictly valid only for infinitesimally slow perturbations. Nevertheless, they are a good approximation for many experiments. Similarly, we can expect that fast transients will be modeled poorly, but that slow transients and averages will be modeled very well. An adiabatic model is the first closure of a hierarchy of moment expansions for the time evolution of the probability distribution of the parts moving through a factory. A more elaborate model including the first-order moment leading to a dynamic equation for the evolution of the velocity is presented in Armbruster et al. (2004). Moment expansions have the advantage of finding the time evolution of average quantities of a stochastic process via the solution of deterministic equations instead of through repeated simulations.

3. Simulations

3.1. The Method

A typical numerical method to solve Equation (10) is a first-order conservative up-winding scheme. A conservative numerical method is one in which the total mass in the computational domain will be conserved, or at the very least will vary correctly, provided the boundary conditions are properly imposed (LeVeque 1998).

To implement this method, we first partition the spatial space into equal subintervals of width h. Let $0 = x_0, x_1, \ldots, x_N = 1$ be the endpoints of these subintervals. To advance the solution on these endpoints forward in time, the first requirement to be met is that the time step k satisfy the Courant-Friedrich-Levy (CFL) condition

$$0 \leqslant \frac{v(t_n)k}{h} \leqslant 1,$$

where t_n is the current time (LeVeque 1992). This condition prevents the numerical solution from traveling faster than the true solution. We compute the velocity, $v(t_n) = \phi(\int_0^1 u(x, t_n)\,dx)$, via an extended Simpson's rule quadrature.

Obtaining the time step, we may advance the solution at each grid point x_j, $j = 0, \ldots, N$, by

$$u(x_j, t_n + k) = u(x_j, t_n) - \frac{k}{h} v(t_n)[u(x_j, t_n) - u(x_{j-1}, t_n)],$$

where k is the time step, and the spatial step h is given by $h = 1/(N-1)$. To advance the solution at the left boundary, we use the boundary condition for the start rate λ in Equation (10) and the propagation scheme

$$u(x_0, t_n + k) = u(x_0, t_n) - \frac{k}{h}[v(t_n)u(x_0, t_n) - \lambda(t_n)].$$

3.2. A Basic Simulation

Utilizing the numerical method of §3.1, we run the initial boundary value problem (10) for a state equation of the Lighthill-Whitham traffic model of the form

$$v(u) = v_0\left(1 - \frac{\bar{u}(t)}{L}\right), \tag{12}$$

with

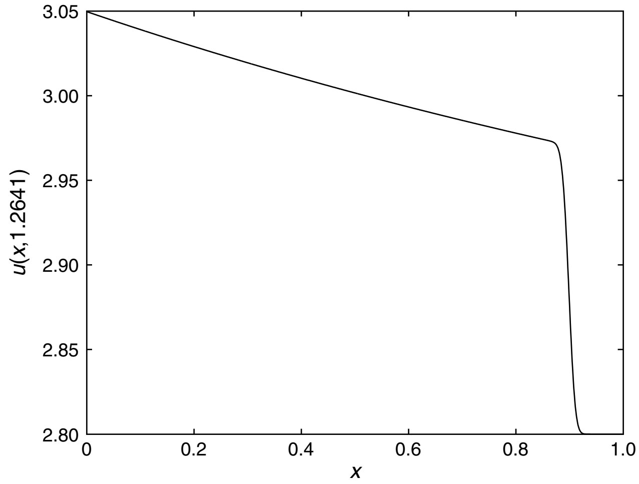
$$\bar{u}(t) = \int_0^1 u(x, t)\,dx.$$

Here, v_0 is the speed for the empty factory and L is the maximal load (capacity of the factory). The influx is given by

$$\lambda(t) = \begin{cases} 2.016, & t < 0, \\ 2.139, & t > 0. \end{cases}$$

This corresponds to a switch from steady state $u_1 = 2.8$ to a new steady state $u_2 = 3.1$ (see below). The parameter values used are $v_0 = 1$, $L = 10$, and $f(x) = 2.8$. Figures 1 and 2 show a snapshot of the solution at $t = 1.2641$ and the outflux as a function of time, respectively. We see that the density is asymptotically approaching a stable steady-state density of $u_2 = 3.1$.

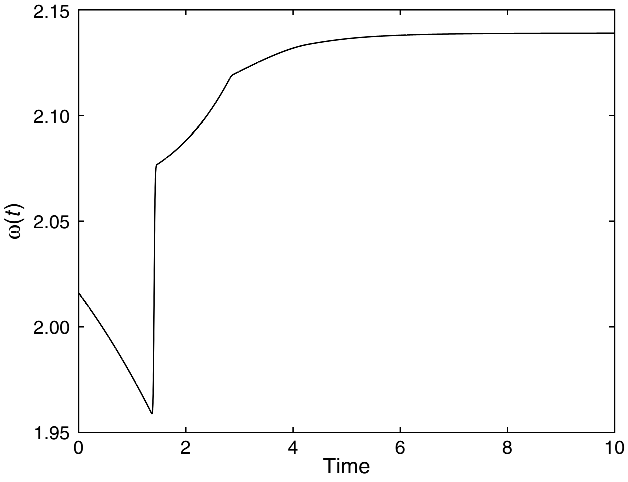
Figure 1. Snapshot of the density $u(x)$ for the solution of system (10) and (12) at time $t = 1.2641$.



Note. Note that the true solution is discontinuous, while numerical dissipation smoothes the corners.

Figure 2 shows an important concept that is typical for re-entrant manufacturing systems that has so far received little attention in the literature. The outflux $\omega(t) = u(1, t) \cdot v(t)$ initially declines when the influx increases. The intuitive reason for this in the context of re-entrant manufacturing systems is the fact that an increase in material at the beginning of the production line competes for machine availability with all the other requests. In particular, an increase at the front of the production line may slow down wafers at the end of the production line as long as they still have to go through the same re-entrant machine. This effect is known as an *inverse response* in control theory and is most pronounced for push and first-in-first-out (FIFO)

Figure 2. Outflux as a function of time for the basic simulation.



Note. Note the transient time for the solution to reach steady state.

policies. After the initial WIP profile has left the factory (which happens at about $t = 1.5$), the outflux increases drastically because the density jump entered into the factory at $t = 0$ has reached the end $x = 1$. We see that, asymptotically, a steady outflux of approximately 2.14 is reached.

The time to do these computations is not an issue here. All the figures in this paper were generated on a regular PC with a code written in C within seconds, to a couple of minutes, of processing time.

4. Discrete-Event Simulations vs. PDE Models

To compare the PDE models with standard discrete-event simulations, we have performed two sets of experiments: We have generated a discrete-event simulation in χ (Hofkamp and Rooda 2002) and studied the transient behavior of throughput and throughput time for a step-up experiment in the influx. We also simulate a temporary overload for a reverse production line. In a second experiment, we have used a large-scale model of an Intel factory and compare the discrete-event simulation of a four-level step-up experiment to the corresponding PDE simulation.

4.1. A Relatively Small Re-entrant Factory

A PDE model like Equation (10) is based on a continuum approximation for the amount of material in the factory (leading to a density u) and a continuum approximation for the number of stages in a factory (leading to an independent variable x describing the degree of completion of the product). Hence, although it is tempting to compare the PDE simulations with queuing models like tandem queues (Newell 1979) or other simple queueing networks (Dai and Vande Vate 2000), such comparisons are a new research project altogether: Typically, a queueing problem that is solvable is based on a small number of machines and a small number of steps. That is exactly the limit in which we do not expect the PDE to behave very well. It would be interesting to study how the queueing model approaches a PDE model in the limit of many machines and many steps, but this is clearly beyond the current state of our understanding (but see Lefeber 2004). In the meantime, what can be done is to compare the PDE model with discrete-event simulations. To do so, we first need to reflect on the parameterization of the PDE model: Much like the clearing function approach, the basic PDE model treats the whole factory as a black box whose input-output characteristics are given by the state equation for the velocity (Equation (11)). Hence, for a given dispatch policy and a given product mix inside the factory, we can either experimentally or via simulations find a state equation characterizing the steady-state behavior of the factory under such circumstances. The PDE will then allow us to simulate the behavior of the factory for non steady behavior such as start-up ramps and increases or decreases of the start rates. We do not at present study questions of optimizing dispatch policies or product mixes.

For our first experiment, the simulated network consists of five machines and is re-entrant, i.e., the production recipe requires that each lot will have to go through the five machines four times before it exits. We model only two types of stochastic processes: a stochastic arrival process into the factory and a stochastic exit process for every machine. Both processes are represented by exponential distributions: The arrival process has an exponentially distributed interarrival time whose mean we vary; the machine processes have exponentially distributed processing times with a mean of 0.12 for the first machine and a mean of 0.10 for the other four machines. We typically start up with an empty factory. To determine the state equation representing the steady-state behavior, we simulate a discrete-event simulation run for between 500 and 4,000 time units—long enough such that any trace of the initial transient has disappeared by about $1/3$ of the simulation time interval. We then use the last half of the simulation time interval to determine average throughput, average throughput time, and average WIP. Averaging over about 50 runs constitutes an experiment. We then repeat the experiments until we reach a predetermined confidence interval for the calculated averages (typically 5%). Figure 3 shows seven data points for throughput time versus WIP for the above experiment with a push policy. The data look almost linear and suggest a state equation for an equivalent single $M/M/1$ queue, i.e., we use a least-squares approximation of the form $\phi(u) = a\bar{u} + b$, where $\bar{u}$ is the total WIP in the factory. The fact that the factory seems to be represented by an equivalent queue is presumably a result of the fact that

Figure 3. Seven data points for a state equation describing the relationship between throughput time and WIP.

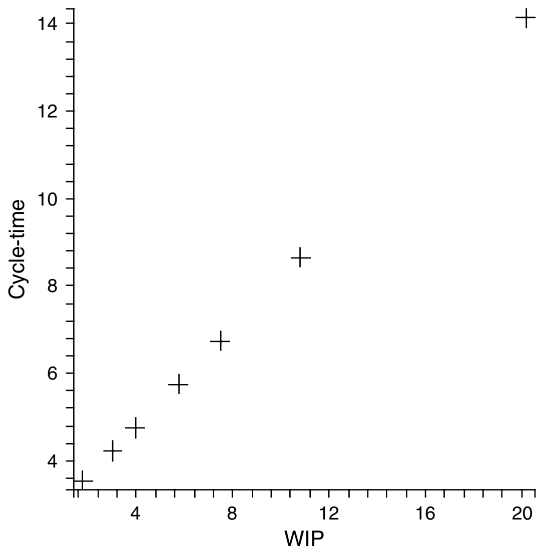


Figure 4. A ramp-up experiment showing a discrete-event simulation and PDE model: (a) Outflux and (b) TPT as a function of time.

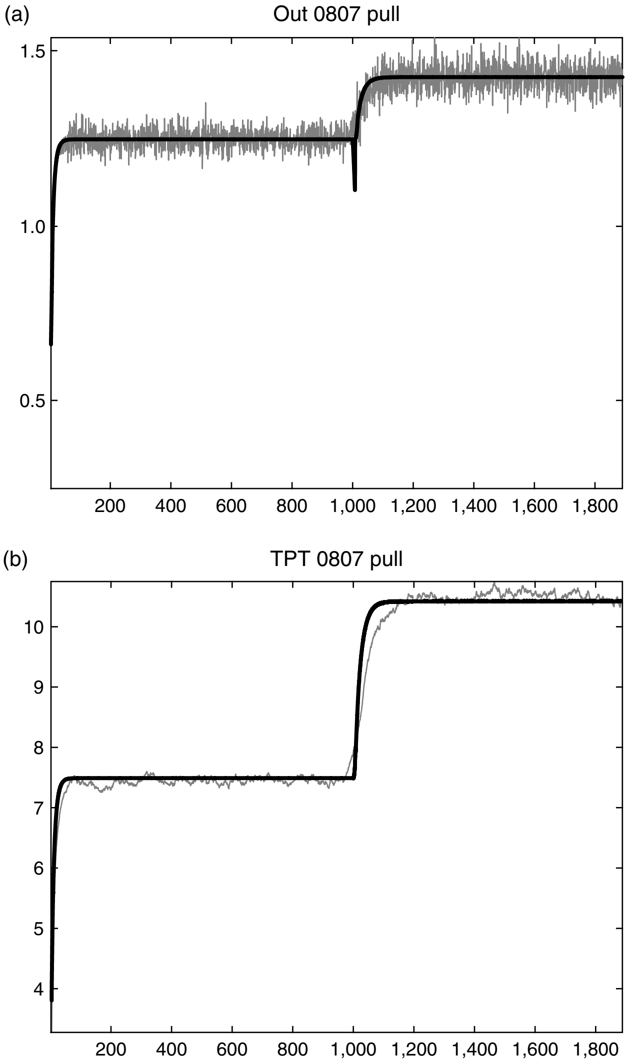
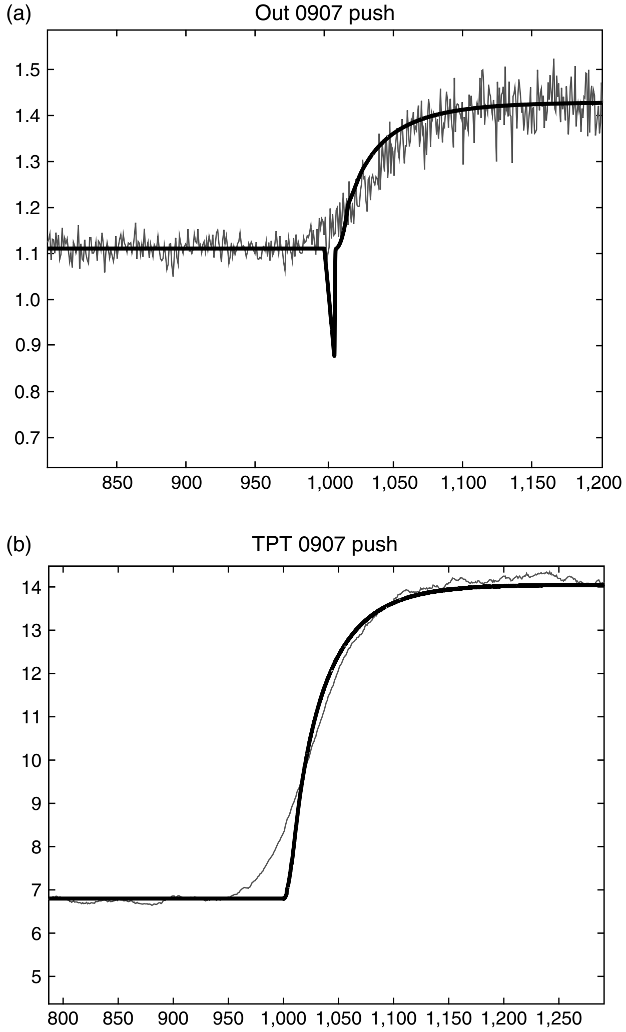


Figure 5. Like Figure 4, with a push policy and a zoom into the transition: (a) Output and (b) Throughput time.



we did not include any stochastic processes in the discrete-event simulation that may generate crowding behavior, as discussed in §2.

We use this state equation to generate PDE simulations. Trivially, the PDE simulation is very good for predicting the steady-state throughput and throughput time for a discrete-event simulation experiment with an arbitrary steady influx, confirming the linear least-square fit for the data. Comparing the PDE simulation to step-up experiments shows some nice agreement and some opportunities for further improvement: We increase the arrival rate from 1/0.9 and 1/0.8, respectively, to 1/0.7. With a critical capacity of about $\lambda = 1/0.6$, this corresponds to a step from 65% and 75% capacity to 85% capacity, respectively. We present four figures: Figure 4 shows time series for the throughput and the throughput time as a function of time for a ramp-up experiment from 75% capacity to 85% capacity with pull policy. The noisy curve is the average of the discrete-event simulations; the continuous curve is the PDE simulation with the state equation generated from Figure 3.

Figure 5 zooms into the transition time series for the throughput and the throughput time of a ramp-up experiment from 65% capacity to 85% capacity with push policy. Note that the throughput times for push and pull policies are very different, and hence we have to use a different state equation for each policy. While the transients in both cases are not perfectly resolved, the agreement is not bad either. The large downward spike in the throughput for the PDE simulation results from the fact that in our model the velocity is spatially uniform and depends on the total WIP in the factory. Hence, any increase in WIP (through, e.g., an increase in influx) will lead to an instantaneous inverse response, i.e., a reduction in velocity and hence to an instantaneous reduction in outflux. Obviously, our simulation factory is not re-entrant enough for such a strong reaction.

We have also run the same discrete-event simulation with a reverse routing: Lots visit the machines in the order $S_1, S_2, ..S_5, S_5, S_4..S_1$, repeating this loop four times. Again, we simulate about seven steady-state influx values to generate a state equation for a push and pull policy, respectively. To challenge the model, we are modeling a temporary overload influx going from about 75% of capacity to 110% of the capacity for $2{,}000 < t < 3{,}000$ before it goes back to 75%. Figure 6 shows the comparison for the throughput of the discrete-event simulation under a pull policy with the PDE simulations. The horizontal line shows the capacity limit in steady state. The PDE simulation is not perfect, but has, qualitatively, the correct behavior: It predicts quite well the throughput during the overload interval; it also has a reasonably good representation of the upswing transient. The PDE simulation again has a large inverse response that is not in the discrete-event simulation, and it also relaxes much faster than the discrete-event simulation in the downswing phase.

Trying the same comparison for the push policy fails (see Figure 7): The discrete-event simulation shows a huge inverse response spread over a very long time such that the outflux is reduced for much of the overload interval. In addition, the outflux never becomes steady. In comparison, the PDE simulation restricts the inverse response to the immediate vicinity of the increase in influx and shows a constant outflux within about 250 time units. Note, however, that the time to return to steady state after the downshift of the input seems to agree quite well for the PDE simulation and the discrete-event simulation. It seems that a push policy for such a topology leads to extreme variations in the locations of the WIP, and hence a description based on an average WIP level in the factory will have to fail badly. This is corroborated by the fact that the discrete-event simulations are much more noisy than the previous simulations, as can be seen in Figure 7. This suggests that a model based on higher moments and several stages might do much better (see §8).

Figure 6. Throughput for the reverse production line with pull policy.

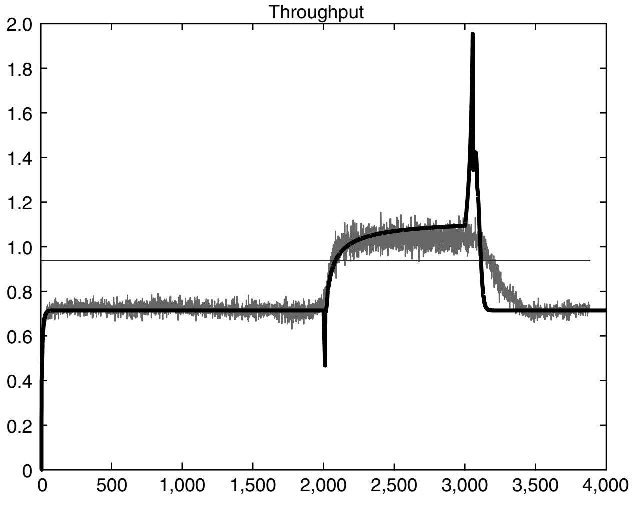
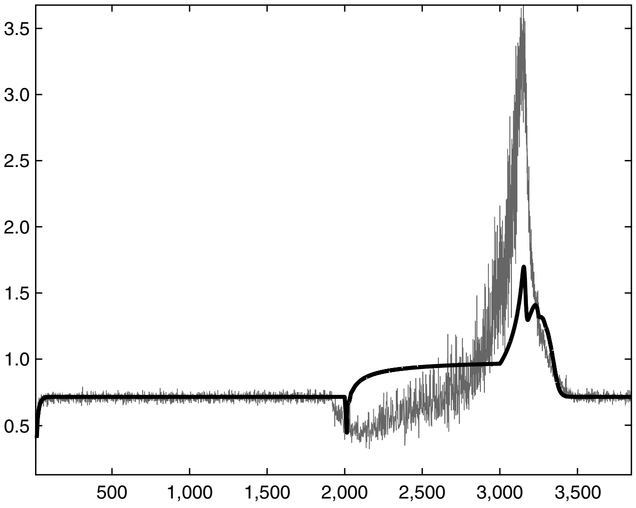


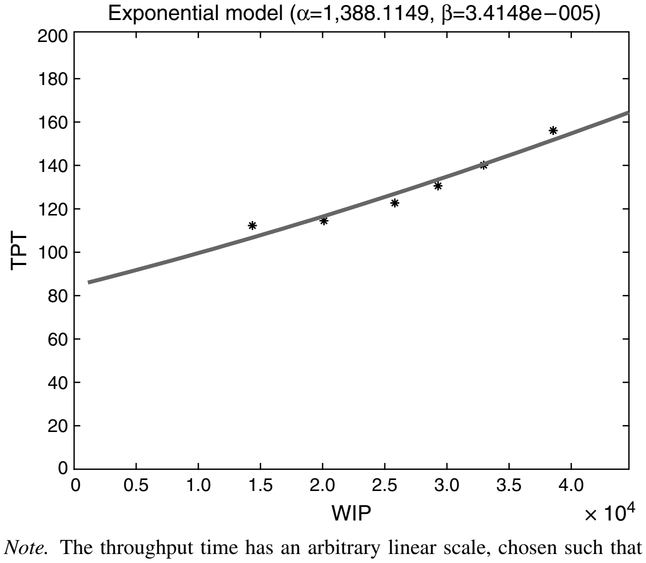
Figure 7. Throughput for the reverse production line with push policy.



4.2. A Large Discrete-Event Simulation—An Intel Factory

We compare the simulation of the PDE with a large-scale discrete-event simulation of a real Intel Corporation factory. For obvious reasons, we have changed throughput times as well as WIP numbers. The model contains approximately 100 machines and simulates 250 steps for a product mix of 10 products. We are running this model for six different start rates and determine the associated average WIP levels

Figure 8. Throughput time versus WIP levels in steady state for a large-scale discrete-event simulation.



Note. The throughput time has an arbitrary linear scale, chosen such that the intersection of the fit is at approximately 100%.

and throughput times in steady state. This results in Figure 8. The interpolating curve in the figure follows a suggestion by Asmundsson et al. (2002) and is a least-square fit of the six data points to

$$\tau = \frac{W}{\alpha(1 - e^{-\beta W})}, \tag{13}$$

with W representing the WIP in the factory; α and β are determined by the fit. With $v = 1/\tau$ as a steady-state equation, we now study the behavior of the factory to a successive ramp-up of the start rate over about a 1,000 days. Figure 9 shows the start rate increasing from 500 per day to a 1,000 per day in four different plateaus. We simulate the PDE with a deterministic start rate that is constant on a plateau, but follows the average start rates of the discrete-event simulation. We compare the output of the discrete-event simulation and the PDE in Figure 10(a). While a single discrete-event simulation run takes one hour per year of simulation time on a standard desktop computer, the PDE simulation takes seconds. While the output of the discrete-event simulation varies dramatically day by day; the PDE simulation is deterministic and seems to be more or less centered on the average of the output. To illustrate further how the discrete-event simulation and the PDE simulation are related, we smooth the output by averaging the output over three days in the future and three days in the past, i.e., a moving seven-day average. Figure 10(b) shows the averaged outs and the PDE simulation again. Figure 10(c) shows the outs averaged over a moving window of 21 days. In Armbruster and Ringhofer (2005), we discuss that closing the hierarchy of moment equations at the zero order level via a state equation implicitly defines a closure for the variance of the throughput time. It implies that the coefficient of variation stays constant. The dotted lines in Figure 10 are calculated by determining the variance for a

Figure 9. Start rate ramp-up for the discrete-event simulations.

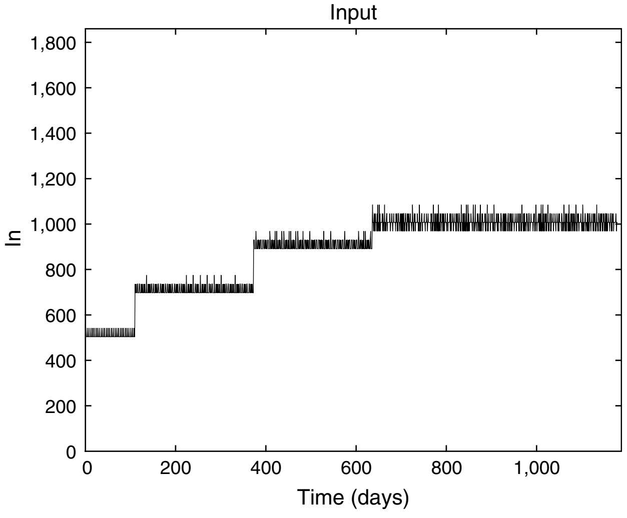
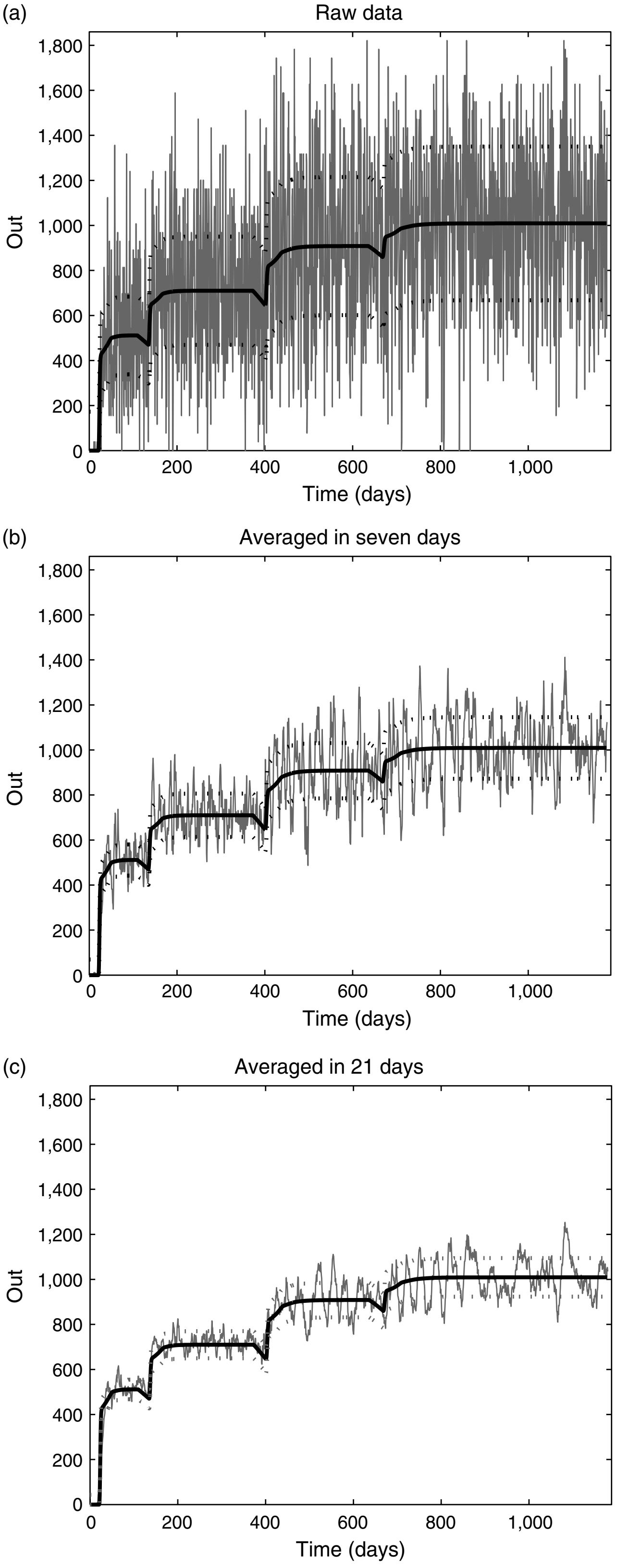


Figure 10. Discrete-event simulation and PDE simulation: (a) Raw outputs, (b) Outputs smoothed over seven days, and (c) Outputs smoothed over 21 days.



given plateau of the outputs. Assuming then that the coefficient of variation stays constant, the dotted lines are the mean ± the standard deviation. It seems that the PDE and its associated standard deviation are a good description of a smoothed transient process.

This comparison allows us to make an interesting observation that we most likely would not expect to find if we are doing only a pure discrete-event simulation. As discussed previously, for a re-entrant factory, an increase in the start rate leads to an inverse response in the output, i.e., the output initially drops before it increases to the new level. The direct discrete-event simulation Figure 10(a) shows no such inverse response; if it is there it is masked in the raw outputs by the daily variation in the outputs. It is also not commonly found in the reality of the factory, due in part to the fact that operators are paid to maintain or increase the outputs, and hence they will work hard to speed up WIP at the end of the production line (Kempf 2001). However, the smoothed outputs in Figures 10(b) and 10(c) indicate that, without the change of output policies resulting from operator interference, the inverse response can be found in the discrete-event simulation too, and it follows quite well the PDE simulation. The reason why the inverse response for the discrete-event simulations and the PDE for the Intel factory agree much better than for the simulations in §4.1 presumably depends on the fact that the Intel factory is much more re-entrant and that the relative change in its load is much lower than in the other simulations.

5. Equilibria

For the rest of this paper, we assume a nonlinear state equation analogous to the Lighthill-Whitham traffic model (Equation (12)), repeated here for convenience:

$$v(u) = v_0\left(1 - \frac{\bar{u}(t)}{L}\right). \tag{14}$$

Note, however, that all the analysis in the following chapters can easily be adapted to fit any other state equation. We begin by determining the steady-state solutions.

Proposition 1. *For a given (constant) influx* $\lambda(t) = \lambda$, *the model* (10) *and* (14) *has two steady states*:

$$u_+ = \frac{L}{2} + \sqrt{\frac{L^2}{4} - \frac{\lambda L}{v_0}} \quad \text{and} \quad u_- = \frac{L}{2} - \sqrt{\frac{L^2}{4} - \frac{\lambda L}{v_0}},$$

provided $4\lambda/v_0 < L$.

Proof. Let u_{ss} represent a constant solution to Equation (10). Then, the boundary condition becomes a quadratic equation

$$u_{ss}^2 - Lu_{ss} + \frac{\lambda L}{v_0} = 0, \tag{15}$$

and the solutions to this quadratic equation are u_+ and u_-. □

This implies that a re-entrant factory has two modes of steady operations that generate the same throughput: a high-WIP, high-TPT state and a low-WIP, low-TPT state. Note that any state equation that grows stronger than linear will show two equilibria.

Proposition 2. *The high-WIP, high-TPT state,* u_+, *is unstable, while the low-WIP, low-TPT,* u_-, *is stable to perturbations.*

Proof. Perturbing the solution near a steady state u_{ss}, we write $u(x, t) = u_{ss} + p(x, t)$, and insert into system (10),

$$\begin{aligned}
&\frac{\partial p}{\partial t}(x,t) + v_0\left(1 - \frac{1}{L}\int_0^1 (u_{ss} + p(x,t))\,dx\right)\frac{\partial p}{\partial x}(x,t) = 0,\\
&u(x,0) = u_{ss} + p(x,0), \qquad\qquad (16)\\
&(u_{ss} + p(x,0))v_0\left(1 - \frac{1}{L}\int_0^1 (u_{ss} + p(x,t))\,dx\right) = \lambda.
\end{aligned}$$

Linearizing the PDE and the boundary condition by discarding higher-order terms of $p(x, t)$, we get a linear unidirectional wave equation with constant speed $v_{ss} := v(u_{ss})$ and an integral boundary condition:

$$\begin{aligned}
&\frac{\partial p}{\partial t}(x,t) + v(u_{ss})\frac{\partial p}{\partial x}(x,y) = 0,\\
&u(x,0) = u_{ss} + p(x,0), \qquad\qquad (17)\\
&-\frac{u_{ss}v_0}{L}\int_0^1 p(x,t)\,dx + v_{ss}p(0,t) = 0.
\end{aligned}$$

Taking the time derivative of the boundary condition and solving for $(d/dt)p(0, t)$, we find

$$\begin{aligned}
\frac{d}{dt}p(0,t) &= \frac{u_{ss}v_0}{Lv_{ss}}\int_0^1 \frac{\partial p}{\partial t}(x,t)\,dx\\
&= -\frac{u_{ss}v_0}{Lv_{ss}}\int_0^1 v_{ss}\frac{\partial p}{\partial x}(x,t)\,dx\\
&= \frac{u_{ss}v_0}{L}(p(0,t) - p(1,t)).
\end{aligned}$$

To go from the second to the third equation, the wave equation for $(\partial p/\partial t)(x, t)$ (17) was used to replace $\partial p/\partial t$ with $-v(u_{ss})(\partial p/\partial x)$. Because Equation (17) has a constant wavespeed v_{ss}, the solution at the right boundary is $p(1, t) = p(0, t - 1/v_{ss})$. Hence, with $\tau = 1/v_{ss}$, $\gamma = u_{ss}v_0/L$, and $z(t) = p(0, t)$, stability of the steady state u_{ss} is determined by the stability of the trivial solution of the delay differential equation (Kuang 1993)

$$\frac{d}{dt}z(t) - \gamma z(t) + \gamma z(t - \tau) = 0. \tag{18}$$

Its characteristic equation is

$$\xi - \gamma + \gamma e^{-\xi\tau} = 0.$$

Obviously, $\xi = 0$ is a solution. This corresponds to the fact that the system is a conservation law, and hence is neutrally stable towards all perturbations that preserve the load (this reflects the fact that any factory can be run at any load with a constant WIP policy (CONWIP)) (Spearman et al. 1989).

Define $h(\xi,\tau)=(\xi-\gamma)e^{\xi\tau}+\gamma$ and consider $h(\xi,\tau)$ as a function of real ξ. Then,

$$h(0,\tau)=0, \qquad h(\gamma,\tau)=\gamma>0,$$

and

$$\frac{\partial h}{\partial \xi}(\xi,\tau)=e^{\xi\tau}+(\xi-\gamma)\tau e^{\xi\tau},$$
$$\frac{\partial h}{\partial \xi}(0,\tau)=1-\tau\gamma.$$

Hence, if $\tau>\gamma^{-1}$, then $(\partial h/\partial\xi)(0,\tau)<0$, which implies that there is a $\delta>0$ such that when $0<\xi\leqslant\delta$, $h(\xi,\tau)<0$. Because $h(\gamma,\tau)>0$, there is at least one $\bar{\xi}$, $\delta\leqslant\bar{\xi}\gamma$, with $h(\bar{\xi},\tau)=0$. Hence, the characteristic equation always has a positive root, signifying an unstable trivial solution. Conversely, if $\tau<\gamma^{-1}$, then $h(\xi,\tau)>0$ everywhere. Hence, there are no $\lambda>0$ that solve the characteristic equation, and the trivial solution is stable.

It is easy to see that u_+ implies that $\tau>1/\gamma$, while u_- implies $\tau<1/\gamma$, which proves the claim. □

Note that for a state equation $v(u)=v_0/(1+\bar{u})$ (Equation (9)), which describes a simple linear queue (or a product network), there exists only one stable equilibrium. Furthermore, it is easy to show that for any state equation $v(u)=\phi(u)$ that slows down faster with increasing load than a simple queue, there exist two equilibria and that the high-speed, low-WIP equilibrium is stable, while the low-speed, high-WIP equilibrium is unstable. In particular, the singularity at $\bar{u}=L$ is not a necessary condition for the existence of the two equilibria.

The relationship of these results to standard queueing theory is the following: Consider a small WIP perturbation of a steady state—for instance, a localized WIP increases, but keeps the start rate the same throughout. The perturbation will travel downstream and will eventually leave the factory. In the stable steady state, the perturbation will have generated secondary perturbations that are smaller than the initial perturbation and, hence, eventually the system will return to equilibrium. In the unstable steady state, an increase in WIP will slow down the system even more, leading to an even bigger increase in WIP—i.e., the secondary perturbations are bigger than the original one and, hence, eventually WIP will increase to infinity. For traditional queueing systems this corresponds to the examples of scheduling policies for re-entrant flows (Dai and Weiss 1996, Lu et al. 1994) that will lead to the growth of WIP without bound, although the utilization (measured as the start rate relative to the processing rate of the bottleneck) stays below one.

6. Control

The simplest control problem associated with running a factory is to change the production flow of the factory from one steady state, corresponding to outflux that meets a specific constant demand, to another one. Let us assume that the demand $d(t)$ changes in the following way:

$$d(t)=\begin{cases} d_1, & t<0,\\ d_2, & t>0.\end{cases}$$

The control problem is to design an influx $\lambda(t)$ that would move the system from the equilibrium u_1 corresponding to a production rate of d_1, to a new equilibrium u_2 corresponding to a production rate d_2. The outflux and the equilibrium densities u_i are related as

$$d_i=u_iv_0\left(1-\frac{u_i}{L}\right). \tag{19}$$

The influx $\lambda(t)$ is uniquely determined by the boundary condition at the left: $\lambda(t)=u(0,t)v(t)$, where the velocity depends on the global load. Hence, to generate a WIP profile that has a discontinuous step from an equilibrium $u(x)=u_1$ to new value u_2 at time $t=0$, we choose the following influx:

$$\lambda(t)=\begin{cases} u_1v_0\left(1-\dfrac{u_1}{L}\right), & t<0,\\ u_2v_0\left(1-\dfrac{1}{L}\left[\displaystyle\int_0^{x(t)}u_2\,ds+\int_{x(t)}^1 u_1\,ds\right]\right), & \\ & 0<t<\tau,\end{cases} \tag{20}$$

where $x(t)$ is the characteristic for the discontinuity emanating from $x=0$, $t=0$ in the xt-plane. At $t=\tau$, the discontinuity is at $x(\tau)=1$, and we have a new equilibrium profile $u(x)=u_2$. Because u_1 and u_2 are constants, Equation (20) reduces to

$$\lambda(t)=\begin{cases} u_1v_0\left(1-\dfrac{u_1}{L}\right), & t<0,\\ u_2v_0\left(1-\dfrac{1}{L}[u_1+(u_2-u_1)x(t)]\right), & 0<t<\tau,\\ u_2v_0\left(1-\dfrac{u_2}{L}\right), & \tau<t.\end{cases} \tag{21}$$

The characteristic $x(t)$ is the solution of the initial value problem $\dot{x}(t)=v(u(x(t)),t)$, i.e.,

$$\dot{x}(t)=v_0\left(1-\frac{1}{L}\left[\int_0^{x(t)}u_2\,ds+\int_{x(t)}^1 u_1\,ds\right]\right), \tag{22}$$
$$x(0)=0.$$

Because u_1 and u_2 are constants, the integrals can be evaluated and Equation (22) reduces to a simple linear ODE. Its solution is

$$x(t) = \frac{L-u_1}{u_2-u_1}\left(1-\exp\left(\frac{v_0(u_1-u_2)t}{L}\right)\right), \tag{23}$$

and hence $\lambda(t)$ yields

$$\lambda(t) = \begin{cases} u_1 v_0\left(1-\dfrac{u_1}{L}\right), & t<0, \\[2ex] \dfrac{v_0 u_2(L-u_1)}{L}\exp\left(\dfrac{v_0(u_1-u_2)}{L}t\right), & 0<t<\tau, \\[2ex] u_2 v_0\left(1-\dfrac{u_2}{L}\right), & \tau<t. \end{cases} \tag{24}$$

Solving $x(t)=1$ for t gives

$$\tau = \frac{L}{u_1-u_2}\ln\left(\frac{L-u_2}{L-u_1}\right). \tag{25}$$

Using the numerical method described in §3.1, the system was run with the parameter values $v_0 = 1$, $u_1 = 2.8$, $u_2 = 3.1$, and $L = 10$. With these values, we get $\tau = 1.42$. For $t > \tau$, we continue the steady influx $\lambda = u_2 v_0(1-u_2/L)$. Figure 11(a) shows a WIP profile and Figure 11(b) shows the resulting outflux: Upon the introduction of u_2 into the factory, the total load increases, and hence the velocity decreases. The velocity continues to decrease for $t < \tau$. Therefore, the outflux given by $\omega(t) = u_1 v(t)$ decreases as well, until the new density $u = u_2$ reaches the end of the production line.

If the change in demand happens without sufficient notice, then there exists a time lag during which the factory will not produce according to the new demand, as can be seen in Figure 11(b). During the transition between the equilibria we produce an outflux $\omega(t) = u(1,t)v(t)$ while we wished to produce an outflux of d_2. The difference between actual and desired production up to time t is called the *backlog*. It may be written as

$$b(t) = u_2\left(1-\frac{u_2}{L}\right)t - \int_0^t \omega(s)\,ds. \tag{26}$$

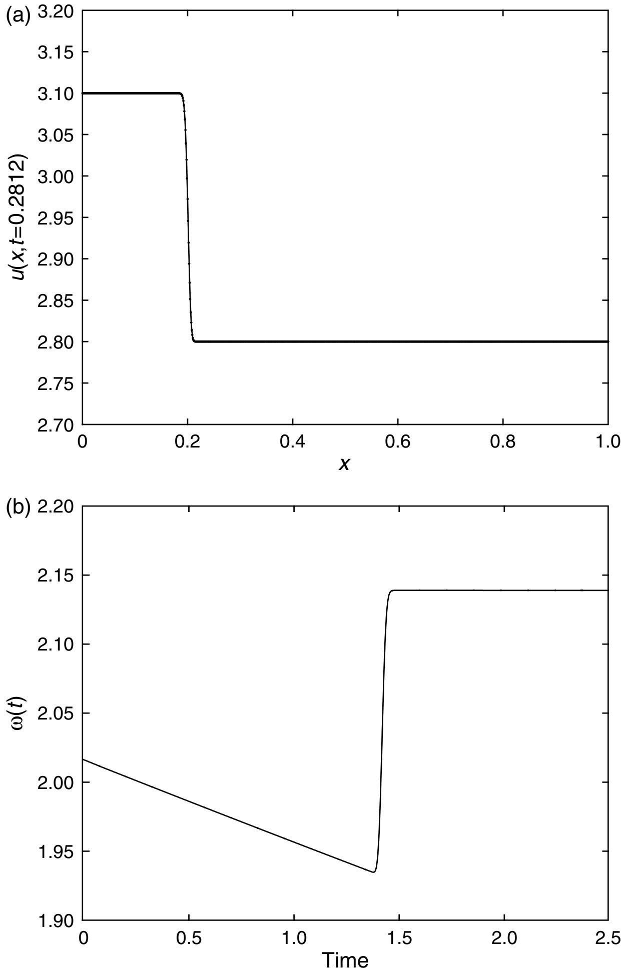
An optimal control problem can now be defined that asks for the input function $\lambda(t)$ that moves the factory from one steady state to another in shortest time, subject to the constraint of a zero backlog. Formally, we write

$$\min T$$

subject to

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial(v(u)u)}{\partial x} = 0, \\ &v(u) = v_0\left(1-\frac{\bar{u}(t)}{L}\right) \quad \text{for } t \leqslant T, \\ &u(x,0) = u_1, \\ &v(u)u(0,t) = \lambda(t), \end{aligned} \tag{27}$$

Figure 11. (a) Snapshot of the density $u(x)$ for the solution of Equation (10). The influx is given by Equation (24). Note that with this start rate, the density is piecewise constant. (b) Outflux; note that the flux is constant after the transient time of moving between the steady states.



and

$$b(t) = \int_0^t (d_2 - v(s)u(1,s))\,ds = 0 \quad \text{for } t \geqslant T. \tag{28}$$

Standard optimal control approaches in the production and inventory modeling context (e.g., Gershwin et al. 1985) cannot solve our problem here: They are based on ODEs that cannot take into account the slowdown of the factory, as the load in the factory increases due to the control actions. Lefeber (2004) suggested an approach based on control theory of delay systems. Whether such an approach will work for the re-entrant factory with its large delays remains to be proven.

As we currently cannot solve the optimal control problem in its general setting, we are going to solve the following simplified problem: We assume that we want to go from an equilibrium solution u_1 to an equilibrium solution

u_2 through one intermediate constant density $u = u_3$. We will find the optimal level u_3 and the optimal time that the system will stay in that intermediate equilibrium. While this constraint mostly is dictated by the fact that we cannot solve optimal control problems for nonlinear hyperbolic equations, it is not entirely unreasonable to have this additional constraint: We are trying to find the input sequence that solves the optimal control problem and, in doing so, limits the disruption of the WIP profile to two-step functions. Clearly, from a resources and management point of view it is desirable to have a product density in the factory that is as homogeneous as possible. Furthermore, we hope that this approach can be the basis for a numerical algorithm that solves the original optimal control problem (WIP).

We therefore assume that the solution at $x = 0$ takes the following form:

$$u(0,t) = \begin{cases} u_1, & t < 0, \\ u_3, & 0 < t < T, \\ u_2, & T < t, \end{cases} \tag{29}$$

i.e., we enter the intermediate density u_3 for T time units. Note that once we enter the desired density u_2, there is still the transition time τ before the system is completely in equilibrium. Under the assumption that v_0, u_1, u_2, and L are known, there are two parameters, u_3 and T, to choose. Therefore, the requirements for a (u_3, T) pair to satisfy are

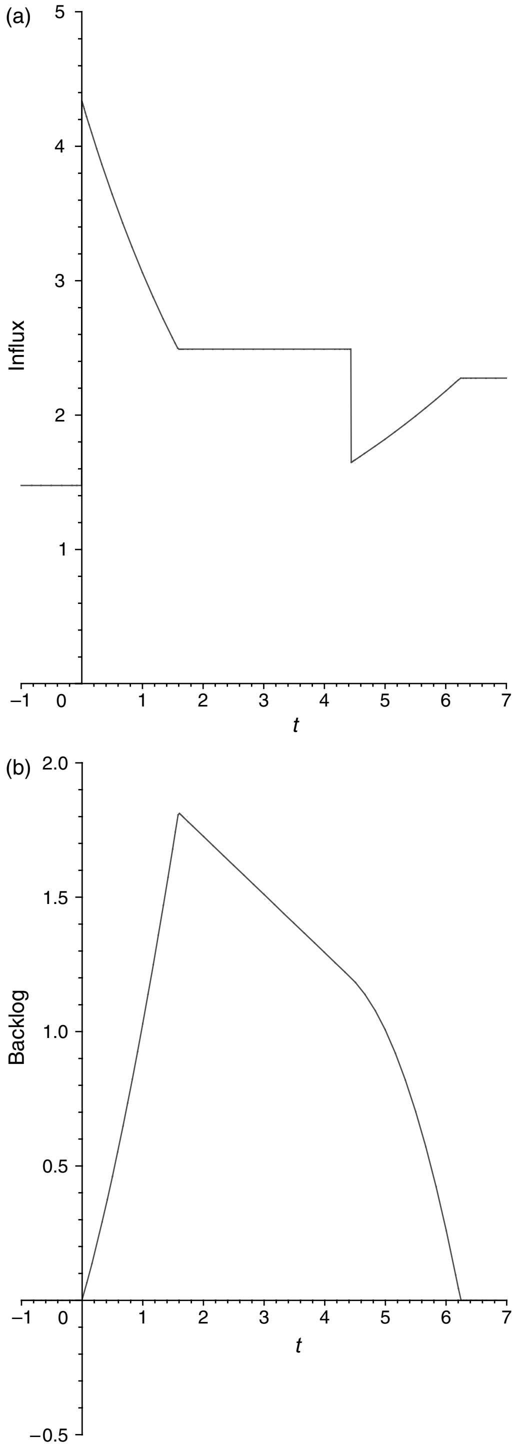
$$u_2\left(1 - \frac{u_2}{L}\right)(T+\tau) - \int_0^{T+\tau} \omega(s)\, ds = 0. \tag{30}$$

There are two cases to consider.

6.1. u_3 Saturates the Factory

Here we seek to move to an intermediate equilibrium solution, u_3, and remain on it for some time before moving again to the final equilibrium solution u_2. The method of Lagrange multipliers and simple algebra allows us to find the optimal values for the intermediate level u_3 and the time T to stay on that intermediate level. As an example, for $v_0 = 1$, $u_1 = 1.8$, $u_2 = 3.5$, and $L = 10$, we find a minimum time for $u_3 = 5.29321$ and $T = 2.85286$. Figures 12(a) and 12(b) show the start rate and backlog for the optimal solution with these parameter values. The drop at around $t = 4$ is due to the inverse response again. As we lower the start rate, we reduce WIP in the factory, and hence speed up production and subsequently increase the output. In order not to overproduce, we need to start less for a short amount of time.

Figure 12. (a) Start rate $\lambda(t)$ and (b) Backlog $b(t)$ for the optimal parameters discussed in §6.1.



Note. Note that after $t \approx 6.2421$, the influx is at the new steady state and the backlog stays zero.

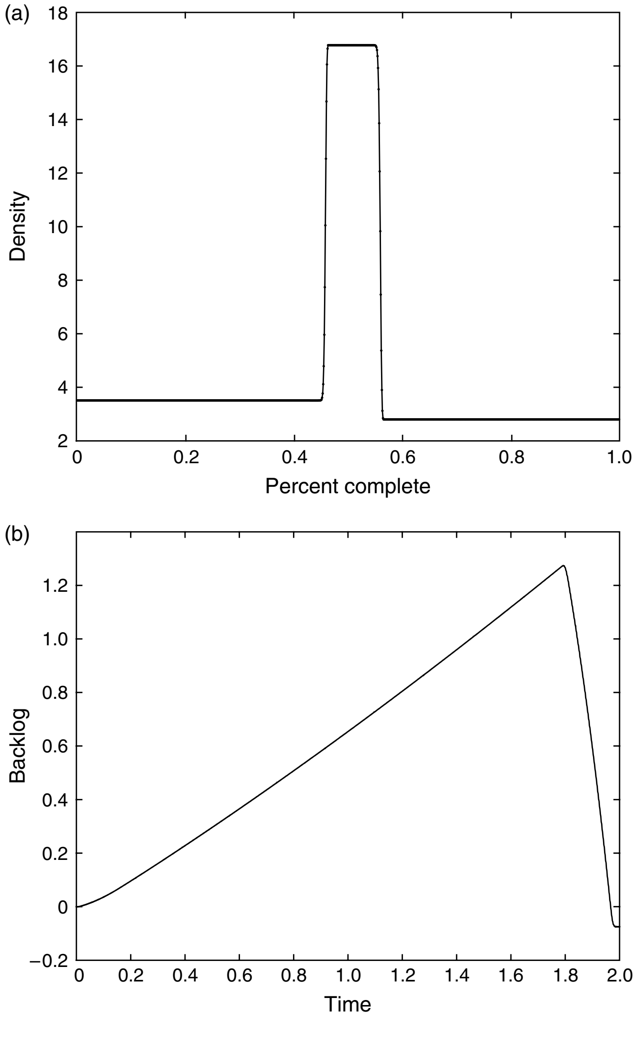
6.2. Small Backlog—u_3 Does Not Saturate

Here we input the intermediate state u_3 only for a time $t < \tau_{u_3}$, such that the factory never saturates on u_3. There will be three stages to the transient in moving from u_1 to u_2: The first stage will be introducing the density u_3 into the factory. The discontinuity from u_1 to u_3 moves through the factory along the characteristic $x_1(t)$. This will yield a decreasing velocity because the load keeps increasing. This will last until $t = T$, at which time the introduction of u_2

will begin. The discontinuity from u_3 to u_2 will move along a characteristic called $x_2(t)$. This second stage will consist of the time needed to purge the factory of u_1, so that only u_2 and u_3 are left. This time will be referred to as $T+T_1$. The third and final stage will be the expulsion of u_3 from the factory at which time the new equilibrium is reached. We call this time $t = T+T_1+\tau$ and we are seeking a minimum time subject to the backlog constraint (30) by varying the two free parameters u_3 and T_1. Again, straightforward algebra and the method of Lagrange multipliers allows us to find the minima.

We find that for every choice of u_1, u_2, and L that we checked numerically, for which there exists a solution for the backlog constraint, the shortest possible solution was to make up the backlog through an instantaneous δ-impulse with a strength of the maximal backlog. While we can prove the optimality of the δ-impulse for special cases analytically, we have not been able to do so for arbitrary u_1, u_2, and L. We have, however, confirmed it for many cases numerically. Figure 13(a) shows a WIP profile for a jump from $u_1 = 2.8$ to $u_2 = 3.5$ via a maximal density value of $u_{3_{\max}} = 17$. Figure 13(b) shows the associated temporal evolution of the backlog.

Figure 13. (a) The WIP profile for an approximation of a δ-pulse and (b) Backlog as a function of time.



There is an easy-to-understand limit for this case: Assume that a backlog of h needs to be made up. If we enter a δ-impulse with strength h into the factory, then the factory will have its maximal load just before the δ-impulse leaves the factory. At that time the load will be $u_2 + h$, assuming without loss of generality (wlog) that $u_2 > u_1$. Hence, $h < L - u_2$ because otherwise the factory is overloaded and the PDE model breaks down. There is also the practical issue of a δ-impulse in the density: Clearly, this is not a realistic model for any factory. However, if we assume that our factory is characterized by a maximal density $u_{3_{\max}}$, then our analysis provides a useful heuristic:

If the jump in demand is small (*leading to a small enough backlog*), *then the optimal strategy will be to get the factory to its maximal density instantly and keep it there for the right amount of time*, such that when this extra product leaves the factory, the backlog is zero.

We conjecture that the optimal solution for the general optimal control problem (27) for a small enough demand jump is a δ-impulse.

7. Supply Chains

The major advantage of the PDE model for a re-entrant manufacturing flow is speed and scalability. Specifically, using a model like (10) as a node, we can easily link many nodes together to form a supply chain that would be impossible to simulate via a discrete-event simulator within a reasonable time frame. For a network of conservative flows, the challenge for a simulation is not computational time, but rather display, analysis, and control of such a network. The recent book by Daganzo (2003) advocates similar ideas. As a prototype example, we are studying a three-node chain: an acyclic factory feeding into a re-entrant factory feeding into another acyclic factory.

The model for the acyclic (or queueing) factory has been discussed in Armbruster et al. (2004). It is based on a generalization of ideas from gas dynamics to queueing networks. Consider a job arriving at a queue with processing rate μ. Its throughput time depends on the number of jobs waiting before it,

$$\tau = \frac{1}{\mu}(1+N). \tag{31}$$

Using this relationship as a boundary condition, we derive a set of coupled hyperbolic conservation laws for the WIP density $u(x,t)$ and the velocity $v(x,t)$ of the form

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial (vu)}{\partial x} = 0, \\ &\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = 0, \\ &v(0,t) = \frac{\mu}{1+\bar{u}(t)}, \\ &v(0,t)u(0,t) = \lambda(t), \end{aligned} \tag{32}$$

where λ is again the arrival rate and $\bar{u}(t) = \int_0^1 u(x,t)\,dx$ is again the total load (WIP). The first equation in system (32) again has the form of a hyperbolic conservation law. However, instead of the heuristic state equation used in the re-entrant model (10) that relates the velocity to the density, we have a Burger's equation for the time evolution of the velocity. Note that the boundary condition for v provides a nonlocal feedback of the history of the production on the current velocity through the length of the queue in front of an incoming part. This model for the queueing system is more sophisticated than the model for the re-entrant system because it allows for nonadiabatic relaxation of the velocity fields, and hence a better modeling of transient phenomena. In addition, in contrast to the heuristics of §2, it has been rigorously derived in Armbruster et al. (2004).

It is straightforward to link the two models together to form a supply chain. We assume no buffers in between factories, and hence the outflux of one factory becomes the influx of the next. This implies, if we extend our completion variable x to cover the whole chain, that the flux will be a continuous variable of x while the density will show discontinuous jumps between factories.

In all experiments, because the PDE models are deterministic moment equations, steady-state modeling is not interesting: Either on the average the influx is below capacity, in which case we find a stable equilibrium given by the state equation; or the influx is above capacity, in which case WIP will explode. However, the PDE models allow us to focus on simulations driven by a time-dependent influx $\lambda(t)$.

7.1. Experiment 1: Short-Time Overload of the Queuing Factory

Figure 14 shows snapshots of a steady-state density and flux for a linear chain of three factories—two identical queue modules surrounding a re-entrant module. Each module makes up one-third of the total x-axis. The raw velocity $(1/\mu)$ for each queue is 1.2, and for the re-entrant module (v_0), it is 2.0. The capacity for the re-entrant module is $L = 12$. We start the experiment from an empty supply chain. The snapshot is taken at approximately $t = 32$ at which time the transients have disappeared. The influx is periodic, given by $\lambda(t) = 0.75 + 0.5\sin(2\pi t)$. The starts are constructed so that they exceed the threshold allowable for an equilibrium to exist in the queue. This occurs for only a short part of a period—on average, they stay below. Figure 15 shows the outflux of this experiment, while Figure 16 shows WIP profiles for each module. Interesting observations are:

- transient oscillations in the output that are much larger than the oscillations in equilibrium—notice the "bulge" in the outs between $t = 10$ and $t = 25$;
- extremely long transients (on the order of 30 throughput times) for the queueing factories;

Figure 14. Snapshot of the steady-state density and flux in the three-node chain.

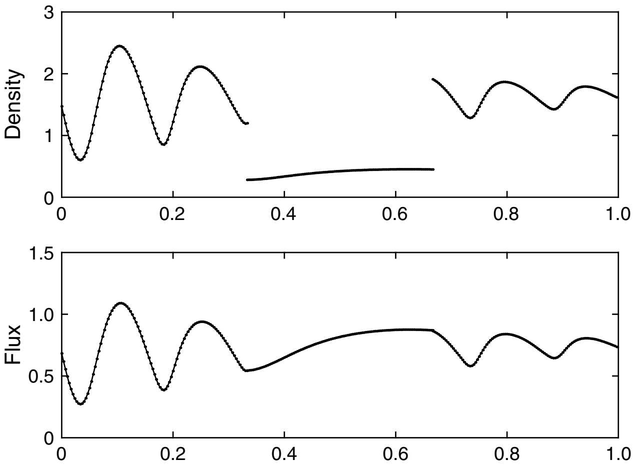


Figure 15. The outflux of the experiment of ramping up the three-node chain from zero.

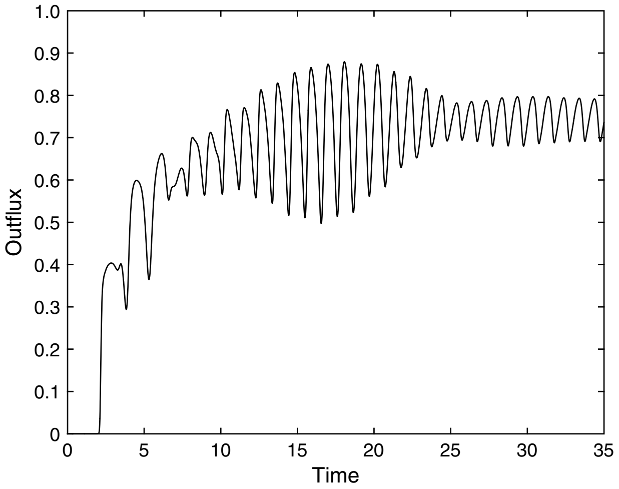
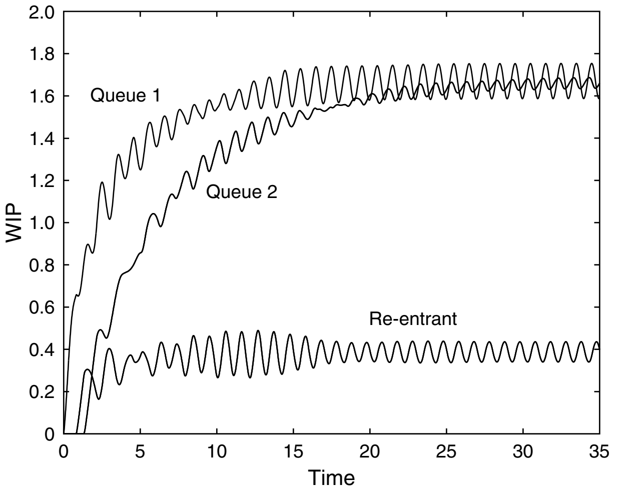


Figure 16. WIP profiles from 7.1 for each module.



Notes. Each module began empty. Notice the length of the transient to equilibria: the queues took far longer than the re-entrant module to reach steady state.

- the re-entrant factory equilibrates much faster than the queueing factories; and
- the re-entrant factory acts as a damping device reducing the amplitude of the oscillations in WIP in the downstream factory.

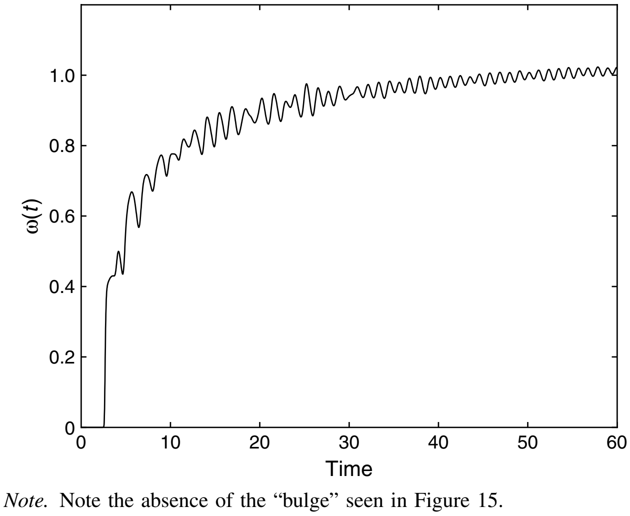
7.2. Experiment 2: Stronger Overload of the Queuing Factory

We increase the starts to $\lambda(t) = 1.1 + 0.5\sin(2\pi t)$, i.e., the mean influx of 1.1 is just below the threshold of $\mu = 1.2$ allowable for a constant steady state to exist in the queues. Figure 17 shows the outflux and Figure 18 shows the WIP for all three factories for this experiment. The major differences to the previous experiment are the much longer transients and the absence of the bulge in the outflux.

7.3. Experiment 3: Short-Time Overload of the Re-entrant Part

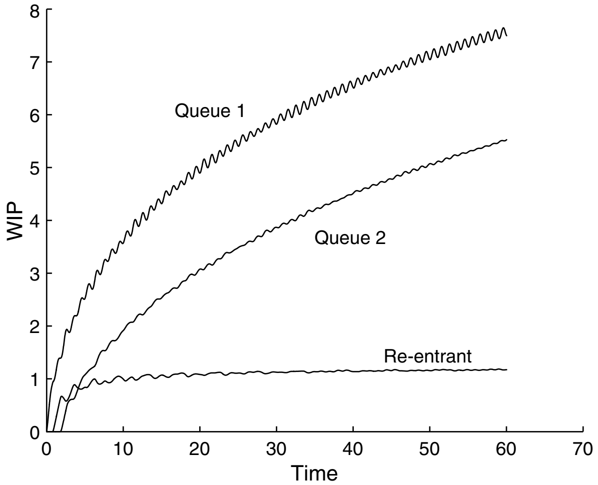
We return to the starts profile of Experiment 1: $\lambda(t) = 0.75 + 0.5\sin(2\pi t)$. However, we now reduce the capacity for the re-entrant module to $L = 4$, leading to a critical influx for the re-entrant part of $\lambda_c = 1$. This influx profile violates the threshold for constant steady states to exist for the queues (for a very short time) and for the re-entrant module (for a longer time). The time series for outflux and WIP for this experiment are very similar to those in Experiment 1. In particular, the "bulge" in the transient of the outflux reappeared. The most striking difference can be seen in the snapshot of the density and flux of the re-entrant factory: Because the re-entrant part is partially overloaded, its velocity will decrease, and hence the overall WIP will increase from an average of $u \approx 0.5$ in Figure 14 to an average of $u(x) = 1$ in Figure 19. In addition, we find that the flux and density waves travelling through the factory have a much shorter wavelength Λ than in the previous experiments, suggesting a dispersion relation $\Lambda \propto v$.

Figure 17. The outflux for Experiment 2.



Note. Note the absence of the "bulge" seen in Figure 15.

Figure 18. WIP profiles for Experiment 2.



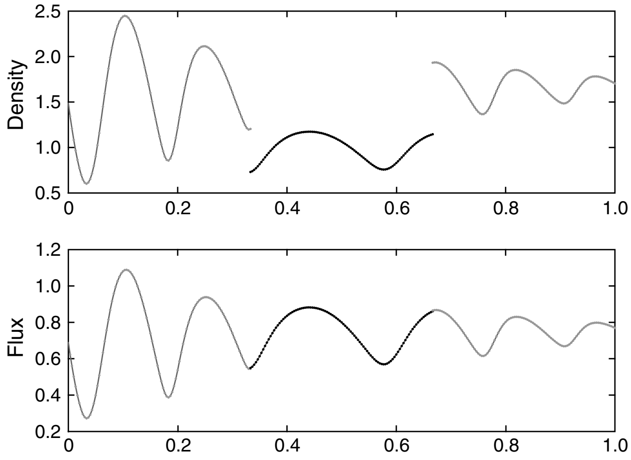
Note. The queues are clearly not yet in equilibrium.

8. Conclusions and Further Work

This study can be viewed as a response to the call for scalable simulation of supply chains (Scalable Enterprise Initiative 1999). We have shown that high-volume production in linear and re-entrant factories can successfully be modeled via hyperbolic conservation laws. This idea opens up the vast literature on traffic modeling and hydrodynamic transport equations to be adjusted for the details of production flow. In particular, standard numerical schemes allow fast and accurate simulation of production networks. Obviously, a lot of future work needs to be done to make such an approach a real competitor to the standard discrete-event simulators.

An issue that needs clarification is the issue of network topology and the related issue of prioritization or dispatch rules. The appeal of the current formulation is that it treats the dynamics of a full factory as one unit whose parameterization is completely determined by its state equation. While this allows us to discuss experiments changing the

Figure 19. Snapshot density and flux profiles for Experiment 3 at $t \approx 58$ seconds.



input of the factory and incorporating such PDEs as building blocks in supply chains (see §7), it does not allow us to easily study the changes due to changing dispatch rules or changing product routing. If we want to do this, we need to give up the concept of a homogeneous model (in completion space) of the factory and introduce completion-dependent parameters. In the simplest case, this can be done by coupling two PDE models together where one feeds into the other. For instance, the transistor and the metalization segments in chip processing would naturally be separated like that. Each part is re-entrant, but not across the transistor/metalization boundary.

A single state equation, but a more accurate model, uses appropriate weight functions $w(x, s)$ to represent the influence of WIP at stage s on the velocity of WIP at stage x:

$$N(x, t) = \int_0^1 w(x, s)u(s, t)\, ds.$$

For instance,

$$N_{\mathrm{pull}}(x, t) = \int_x^1 u(s, t)\, ds$$

restricts the WIP relevant for the velocity at point x to the WIP in front of x, whereas

$$N_{\mathrm{push}}(x, t) = \int_0^x u(s, t)\, ds$$

restricts it to the WIP behind x—mimicking pull and push policies, respectively. The experience with the failed PDE model discussed in §4.1 (Figure 7) suggests that the current model based on a global velocity will be quite good if the dependence of the velocity at stage x on the WIP at stage s is constant; and it will not be good if that dependence varies significantly, as it does, e.g., for a push policy in a reverse production line.

We are currently pursuing several first-principle models and heuristic extensions of these ideas:

- While it is extremely hard to characterize in any meaningful way the actual stochastic processes involved, we know that there exist some physical limits. In particular, a fixed production line has a maximal production capacity at every machine. No matter what policies, the recipe for a certain chip expects it to stay x hours in a diffusion oven. Hence, we can define a maximal capacity function $C(\xi)$ that describes the capacities of all machines that are involved in the production process, where ξ now is a variable that does not describe the stages, but the sequence of machines. The resulting quasi-static model then becomes

$$\begin{aligned}&\frac{\partial u}{\partial t} + \frac{\partial F}{\partial x} = 0,\\ &F(x) = \min\{uv_{eq}, \mu(x)\},\\ &v_{eq} = \Phi(u),\end{aligned} \tag{33}$$

where the maximally available capacity at stage $\mu(x)$ depends on the dispatch policy: Assume that there are n layers that are produced on the same machines, leading to n loops through those machines. The map between machine position and production stage is then given by modular division: The stage variable x acquires a layer index i such that $x_i = (i-1)/n + \xi$ for $i = 0, \ldots, n-1$ describes a production stage in the $i+1$th loop at the machine position ξ. At any particular machine, flux requests from all n loops may compete for the maximally available capacity $C(\xi)$ leading to a distribution of the maximally available capacities at stage i, $\mu(x_i)$, depending on the fluxes $F(x_j)$ in the other loops. For instance, for a "push" policy, capacities are allocated from front to end. Hence, we can iterate, for $i = 0, \ldots, n-1$, the following scheme to find the capacity distribution $\mu(x)$:

$$\begin{aligned}&\mu(x_0) = C(\xi),\\ &F(x_i) = \min\{\mu(x_i), uv_{eq}|_{x_i}\},\\ &\mu(x_i) = \max\{0, \mu(x_{i-1}) - F(x_{i-1})\}.\end{aligned}$$

Such a model will lead to the formation of bottlenecks, and hence δ-distributions in the density variable for any influx that temporarily exceeds the total capacity.

- For a completely deterministic flow network, we have derived the system (33) from first principles using models from gas dynamics (Armbruster et al. 2006a). In Armbruster et al. (2006b) we extend this to derive a general model supporting arbitrary policies.
- The fundamental PDE model (Equation (10)) with a state equation (Equation (5)) based on mass conservation can be shown to be the zero order moment expansion of a Boltzmann equation with appropriate closure assumptions (Armbruster et al. 2004). Specifying the stochastic process that generates the randomness in the production process as a stochastically varying throughput time adds a diffusion term to the mass conservation (Armbruster and Ringhofer 2005).

We are currently working on the following issues:

- A major source of stochasticity in factory production is machine breakdowns. Usually, the distributions of time to failure and time to repair can be approximated. This will lead to a stochastic version of the capacity limited flow network discussed above.
- The optimal control problem (27) should be solved in its most general form. Specifically, extending the analysis of §§6.1 and 6.2 to allow for more than one intermediate level should lead to a numerical algorithm for the optimal control problem.
- We plan to parametrize and validate PDE-based models for complicated discrete-event models and for real factories. The goal is to determine modifications of the simple state equation (12) model to describe more accurately the inverse response of a production system and the general form of transients and their dependence on policies. This will involve diffusion terms (Armbruster and Ringhofer 2005) and/or terms representing a relaxation time. Another

goal will be to determine the discretization error associated with treating a fundamentally discrete problem as a continuous flow.

- While we showed the feasibility of linking hyperbolic conservation laws together to make a simple linear chain, this is really just a proof of concept. The real test for the usefulness of our approach will be whether relevant business questions can be answered for a supply network by linking our nodes together and performing simulations. This may involve the development of an object-oriented simulation interface. A related project uses the PDE-based models as the predictive model for attempts to optimize the behavior of a whole supply chain via model predictive control (MPC) algorithms (Wang et al. 2004).

Acknowledgments

The authors thank Edward Yellig of Intel Corporation for executing the simulation runs. John Fowler, Daniel Rivera, and Matthias Kawski and their students helped to focus the authors' ideas through many constructive discussions. The χ simulations are extensions of codes written by R. van den Berg and E. Lefeber. Dieter Armbruster and Christian Ringhofer were supported by NSF grant DMS-0204543. Support from Intel Corporation is also gratefully acknowledged.

References

Armbruster, D., C. Ringhofer. 2005. Thermalized kinetic and fluid models for re-entrant supply chains. *SIAM J. Multiscale Modeling Simulation* **3**(4) 782–800.

Armbruster, D., P. Degond, C. Ringhofer. 2006a. A model for the dynamics of large queuing networks and supply chains. *SIAM J. Appl. Math.* **66**(3) 896–920.

Armbruster, D., P. Degond, C. Ringhofer. 2006b. Kinetic and fluid models for supply chains supporting policy attributes. *Transportation Theory Statist. Phys.* Forthcoming.

Armbruster, D., D. Marthaler, C. Ringhofer. 2004. Kinetic and fluid model hierarchies for supply chains. *SIAM J. Multiscale Modeling* **2** 43–61.

Asmundsson, J., R. Uzsoy, R. L. Rardin. 2002. Compact nonlinear capacity models for supply chains: Methodology. Preprint, Purdue University, West Lafayette, IN.

Baskett, F., K. M. Chandy, R. R. Muntz, F. G. Palacios. 1975. Open, closed and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Machinery* **22** 248–260.

Chen, H., J. M. Harrison, A. Mandelbaum, A. Van Ackere, L. M. Wein. 1988. Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Oper. Res.* **36** 202–215.

Daganzo, C. F. 2003. *A Theory of Supply Chains*. Springer Verlag, Heidelberg, Germany.

Dai, J. G., J. H. Vande Vate. 2000. The stability of two-station multitype fluid networks. *Oper. Res.* **48**(5) 721–744.

Dai, J. G., G. Weiss. 1996. Stability and instability of fluid models for certain re-entrant lines. *Math. Oper. Res.* **21** 115–134.

DeJong, C. D., S. P. Wu. 2002. Simulating the transport and scheduling of priority lots in semiconductor factories. J. L. Snowdon, J. M. Charnes, eds. *Proc. 34th Winter Simulation Conf.* ACM, San Diego, CA, 1387–1391.

Forrester, J. W. 1962. *Industrial Dynamics*. MIT Press, Cambridge, MA.

Gershwin, S. B., R. Akella, Y. F. Choong. 1985. Short-term production scheduling of an automated manufacturing facility. *IBM J. Res. Develop.* **29**(4) 392–400.

Graves, S. C. 1985. A tactical planning model for a job shop. *Oper. Res.* **34** 525–533.

Gross, D., C. M. Harris. 1985. *Fundamentals of Queueing Theory*. Wiley, New York.

Helbing, D. 1996. Traffic modeling by means of physical concepts. D. E. Wolf, M. Schreckenberg, A. Bachem, eds. *Workshop on Traffic and Granular Flow*. World Scientific, Singapore, 87–104.

Hines, J. 2003. Personal communication. Massachusetts Institute of Technology, Cambridge, MA.

Hofkamp, A. T., J. E. Rooda. 2002. Chi reference manual. Technische Universiteit Eindhoven, Eindhoven, The Netherlands. Retrieved November 2002 http://se.wtb.tue.nl/documentation/.

Karmarkar, U. S. 1989. Capacity loading and release planning in work-in-progess (WIP) and lead-times. *J. Manufacturing Oper. Management* **2** 105–123.

Kempf, K. 2001. Personal communications. Intel Corporation, Chandler, AZ.

Kuang, Y. 1993. *Delay Differential Equations*. Academic Press, Boston, MA.

Kumar, P. R. 1993. Re-entrant lines. *Queueing Systems* **13** 87–110.

Law, A. M., W. D. Kelton. 1991. *Simulation Design and Analysis*. McGraw-Hill, New York.

Lefeber, E. 2004. Nonlinear models for control of manufacturing systems. G. Radons, R. Neugebauer, eds. *Nonlinear Dynamics of Production Systems*. Wiley-VCH, Berlin, Germany, 71–84.

LeVeque, R. J. 1992. *Numerical Methods for Conservation Laws*. Birkhäuser-Verlag, Basel, Switzerland.

LeVeque, R. J. 1998. Finite difference methods for differential equations. Draft version for use in AMath 585-6, University of Washington, Seattle, WA.

Lighthill, M. J., G. B. Whitham. 1955. On kinematic waves II. A theory of traffic flow on long crowded roads. *Proc. Roy. Soc. Ser. A* **229** 317–345.

Little, J. D. C. 1961. A proof for the queuing formula $L = \lambda W$. *Oper. Res.* **9** 383–387.

Lu, S. C. H., D. Ramaswamy, P. R. Kumar. 1994. Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. *IEEE Trans. Semiconductor Manufacturing* **7**(3) 374–385.

Nelson, R. 1995. *Probability, Stochastic Processes, and Queueing Theory*. Springer Verlag, New York.

Newell, G. F. 1965. Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Rev.* **7**(2) 223–240.

Newell, G. F. 1973. Scheduling, location, transportation and continuum mechanics; some simple approximations to optimization problems. *SIAM J. Appl. Math.* **25**(3) 346–360.

Newell, G. F. 1979. *Approximate Behavior of Tandem Queues*. Springer Verlag, Berlin, Germany.

Scalable Enterprise Initiative. 1999. Document 99-149, National Science Foundation.

Shikalgar, S. T., D. Fronckowiak, E. A. MacNair. 2002. 300 mm wafer fabrication line simulation model. J. L. Snowdon, J. M. Charnes, eds. *Proc. 34th Winter Simulation Conf.* ACM, San Diego, CA, 1365–1368.

Spearman, M. L., D. L. Woodruff, W. J. Hopp. 1989. CONWIP: A pull alternative to Kanban. *Internat. J. Production Res.* **28**(5) 879–894.

Spier, J., K. Kempf. 1995. Simulation of emergent behavior in manufacturing systems. *Proc. 6th SEMI/IEEE Adv. SemiCond. Manufacturing Conf.* IEEE, New York, 90–94.

Wang, W., D. E. Rivera, K. G. Kempf, K. D. Smith. 2004. A model predictive control strategy for supply chain management in semiconductor manufacturing under uncertainty. *Proc. 2004 Amer. Control Conf.*, Boston, MA, 4577–4582.

Wein, L. M. 1990. Scheduling networks of queues: Heavy traffic analysis of a two-station network with controllable inputs. *Oper. Res.* **38**(6) 1065–1078.