

CHAPTER 1

CONTINUUM MODELS FOR INTERACTING MACHINES

Dieter Armbruster¹, P. Degond² and Ch. Ringhofer¹

(1) *Department of Mathematics, Arizona State University
Tempe, AZ 85287-1804*

E-mail: armbruster@asu.edu, ringhofer@asu.edu

(2) *MIP, UMR 5640 (CNRS-UPS-INSA), Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse cedex, France (degond@mip.ups-tlse.fr)*

A review of continuum models for production flows involving a large number of items and a large number of production stages is presented. The basic heuristic model is based on mass conservation and state equations for the relationship between the cycle time and the amount of work in progress in a factory. Heuristic extensions lead to advection diffusion equations and to capacity limited fluxes. Comparisons between discrete event simulations and numerical solutions of the heuristic PDEs are made. First principle models based on the Boltzman equation for a probability density of a production lot, evolving in time and production stages are developed. It is shown how the basic heuristic model constitute the zero order approximation of a moment expansion of the probability density. Similarly, the advection diffusion equation can be derived as the first order Chapman-Enskog expansion assuming a stochastically varying throughput time. It is shown how dispatch policies can be modeled by including an attribute in the probability density whose time evolution is governed by the interaction between the dispatch policy and the capacity constraints of the system. The resulting zero order moment expansion reproduces the heuristic capacity constraint model whereas a first order moment will lead to multiphase solutions representing multilane fluxes and overtaking of production lots. A discussion on the similarities and differences of industrial production networks and biological networks is also presented.

1. Introduction

As large factories or other large production systems have become increasingly more complicated, their dynamic behavior over time and in response

to changes in the production environment is of utmost importance for attaining the overall goals of timely and cost efficient production. As a result, there is a substantial research endeavor worldwide to simulate, optimize and control such production systems. This paper will give an overview of our recent contributions to that endeavor - through continuous models and their simulation of networks of interacting machines.

Before we will go into details we will try to discuss our view of the theme of this book, in particular we will outline some of the dichotomies between naturally occurring networks of machines, especially biomolecular machines²⁰, and factory production.

- *Self-organizing vs. planning*: Molecular biological networks spontaneously cluster and synchronize in order to produce a desired product, say a protein. Factories production as well is highly organized and synchronized (e.g just in time production). However, rarely does this synchronization occur spontaneously through interaction with other production processes. Typically it is facilitated through production layouts, production rules and recipes and is controlled through an external management system.
- *Stochastic behavior vs. regular production*: Biological processes typically live in a highly fluctuating, unstable environment where transport may be diffusive and production efficiency is very low. Nevertheless, on average most biomolecular processes are remarkably stable. In contrast for instance, the production of semiconductor chips which is one of today's most advanced manufacturing processes is run in the most physically controlled way imaginable.
- *Evolutionary optimality vs. profit*: Evolution is a long term randomly executed continuous search process for a optimally adjusted organism. The goal for a production system is to maximize profit over usually a short timescale.
- *Reaction kinetics, diffusion and continuous transport equations vs. discrete event models*: Often the intermediate steps and the intermediate products of many biomolecular production systems are not known and neither are the transport paths. Hence often models in the form of reaction diffusion equations and continuous transport equations are developed that either aggregate the biomolecular production in space and/or split it into short and long timescales where only the time evolution of the long timescale is modeled. In contrast, in a typical model for production systems individual prod-

ucts and individual production steps can be characterized in great detail, including their stochastic behavior. The resulting models therefore tend to be discrete event simulations.

This list is certainly not exhaustive but it seems to suggest that there is not much that the biomolecular production networks and the industrial production networks have in common and that research concepts and techniques may not be successfully transferred from one realm to the other. However a more detailed inspection of these dichotomies reveals that both areas have much more in common than meets the eye initially.

- The stability of self-organizing systems under perturbations resulting in redundancy and self-correcting behavior has been discovered in several production systems. A typical example is the case of a bucket brigade production system: The term “bucket brigade” was coined by Bartholdi and Eisenstein⁹, for production lines in which the workers hand over their workpiece to the next worker down the line, whenever the last worker has finished his job. If workers are sequenced from slowest to fastest, then there is a stable fixed point that the system will converge to. Not only is this self-organized production optimal, it also selfadjust to the loss of a worker by distributing the work over the remaining workers in an optimal way. Another example has been discussed in²² where a network of machines leads to synchronization at least for the average production rates in different production levels. Agent based models¹⁰ have also been used to study emergent dynamic behavior in supply chain modeling. The overall concept of self-organizing production has lead to at least one large research unit that is completely dedicated to this theme²³.
- Stochasticity is far from under control in industrial production systems. Although every effort is made to have a stable and completely controlled production environment for a semiconductor chip factory the actual production varies dramatically: For instance, figure 1 shows the actual time dependent path of 200 lots in an actual INTEL factory (the data have been rescaled to protect proprietary information). The resulting throughput times vary by about a hundred percent. Given that the raw throughput time through such a factory is of the order of weeks, it is clear that prediction for the time when a particular chip is leaving the factory is next to impossible, or at least has to reflect a very high uncertainty. Similarly,

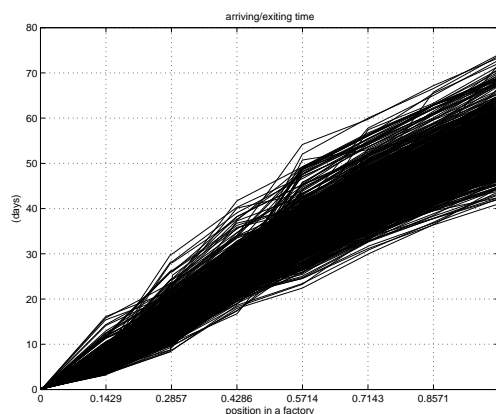


Fig. 1. Paths of 200 lots through an actual factory

for a typical new production process for the latest and fastest chip the resulting yields are quite small and vary dramatically between different production lots. As a result in order to produce the high value products many low value byproducts are produced that may or may not be commercially valuable. In that sense the uncertainty in both, biomolecular production and industrial production may be very high and similar questions pertaining to efficiency and stability are unanswered. This suggests that methods to understand, control or reduce the stochasticity in one area may well be applicable in the other.

- Biological evolution produces the blueprint and the production recipes for a successful species in much the same way as capital investment into machines and research produces the means for a successful industrial enterprise. However, in both cases, it is often very unclear how the individual production step relates to the overall evaluation function and whether a particular change will lead to an improvement. Industrial production typically generates meta-goals like a high utilization of a factory, low inventories, just in time delivery etc. every one of which may contribute to the success of the whole enterprise but collectively those goals are often contradictory and often have unspecified trade-offs. It is yet completely unexplored whether evolutionary optimization and the fitness of an industrial enterprise can learn from each other.

- There seems to be a clear case for discrete event simulations or at least for hybrid simulations¹² in biomolecular systems: Many biological processes are triggered in some way, suggesting that a simulation that has many concurrent processes that are synchronized by specific events may be a useful model. Hybrid processes would be applicable in cases with different timescales - a low one evolving on a continuous description of space and time interrupted by specific events that trigger a different slow evolution. On the other hand, we have spent the last several years to develop models for production systems that are continuum based. The rest of this paper will discuss them and will show a detailed description of heuristic as well as first principle models.

2. Heuristic models

2.1. Quasistatic models

Detailed modeling of complicated production networks is usually done via discrete event simulations. These are stochastic simulations and hence many experiments are performed to generate a large dataset. Subsequently this ensemble of experiments is postprocessed to extract the desired characteristic quantities like average cycle time, average throughput and variances for these characteristic quantities. For time dependent phenomena there are two major concerns associated with that approach: i) postprocessing is not always straightforward and the detailed algorithms used may influence the results, ii) time dependent phenomena have their own timescale that typically leads to a need for larger ensembles. Hence small prototype systems can most often not be scaled up to relevant systems because the computational costs become prohibitively expensive.

One way to deal with this problem is to consider the time evolution of densities rather than individual lots. This leads to the so called fluid models which consist of networks of coupled ordinary differential equations coupled through flux conservation: The rate of change of a queue in front of a machine is given through the influx to that machine (λ_i) minus its outflux (μ_i), i.e.

$$\frac{dq_i}{dt} = \lambda_i - \mu_i \quad (1)$$

where $\mu_i = \lambda_{i+1}$. Fluid models have been studied for quite some time (e.g.¹³) and it can be shown that a queuing system is stable (has finite queue lengths) if and only if the associated fluid model has a stable steady

state. The major drawback of a fluid model is that it does not have good representation of the delay in a system. Any part of the production system that is modelled by an equation like Eq. (1) has an instantaneous change in outflux if there is a change in influx. In a sense, a fluid model stops half way to a real continuum model like a fluid flow. Hence we proposed in ² a second independent variable x which describes the degree of completion for a production process. Note that the degree of completion as the "spatial" variable x allows us to untangle any topologically complicated physical path through the machines in a factory into a simple one dimensional linear chain. Our state variable is now $\rho(x, t)$, the density of product at a stage x at a given time t . Clearly in its simplest form the time evolution of this density has to respect mass conservation. Hence we have a partial differential equation of the form

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

$$v\rho|_{x=0} = \lambda(t) \quad (3)$$

$$\rho(x, 0) = f(x) \quad (4)$$

This is a conservation law with $v\rho$ the flux, the boundary condition $\lambda(t)$ being the external influx and $f(x)$ an arbitrary initial condition. Such conservation laws typically lead to a hyperbolic wave equation where the details of the dynamics have to be represented by a state equation for the velocity v , i.e. the functional form of the dependence of the velocity on the density. Typical models are

$$v_{LW}(\rho) = v_0 \left(1 - \frac{\rho}{\rho_c}\right), \quad (5)$$

$$v_Q(\rho) = \frac{v_0}{1 + \frac{L(\rho)}{L_c}}, \quad (6)$$

$$v_{eq}(\rho) = \Phi(L), \quad (7)$$

with L the total load (Work in progress, WIP) given as

$$L(\rho) = \int_0^1 \rho(x, t) dx. \quad (8)$$

Equation (5) for instance, treats production flow like a traffic flow as in the traffic model of Lighthill and Whitham ¹⁸. The interaction of the various products in the factory here is strictly local: An increase in the local density reduces the velocity at that position until, at a critical density ρ_c the velocity goes to zero. Such a model typically has shock waves corresponding to traffic jams. Equation (6) corresponds to a queuing theory model where,

in steady state, the time $\tau = 1/v_Q$ to exit a queue becomes $\tau = \tau_0(1 + L)$ with τ_0 the time that a product spends in the machine without any waiting and L represents the total length of the queue, or the total work in progress (WIP) (Eq. 8). In general the quasi-static approach using a state equation for the velocity is closely related to the concept of a clearing function^{8,19} which is a static representation of the dependency of the throughput on the current WIP in the factory. Equation 7 is a generic model for such a clearing function.

The dependence of the velocity on the total WIP rather than on the local density in all but the Lighthill and Whitham model reflects our emphasis on semiconductor manufacturing. There, chips are produced in layers which are build up by cycling repeatedly through the same factory. Hence, since lots that are at different layers compete for machine time, fluctuations at the beginning of the production process can influence not only the production upstream but also the production downstream.

All state equation models are based on abstractions on the actual stochastic processes inside the network of machines. These abstractions can be parameterized using either real data or through a very detailed discrete event simulation. The latter not only represents the 'hardware' (i.e. the machines) of the production process but also the 'software' which are the production rules and policies. In any case, we can extract the functional relationship $\Phi(L)$ i.e. the clearing function (Eq. 7) from the available data.

Figure 2 shows such a state equation resulting from a detailed computational experiment.

The simulated network consists of 5 machines and is re-entrant, i.e. the production recipe requires that each lot will have to go through the 5 machines four times before it exits. We see that a linear fit as in a queuing model will be quite appropriate. Clearly such a state equation can be interpolated between a few detailed simulations and then used to predict the outflux or the throughput time in steady state for an influx that has not been simulated. That is the way that static clearing functions models are used. It clearly does not need any dynamics as represented in the PDE. The PDE however, allows us to study transient behavior like for instance a step up or down of the influx into a factory. The continuity equation (Eq. 2) with a steady state model for the velocity represents a quasi-static or adiabatic model: any fluctuation in the WIP-level leads to an instantaneous relaxation of the velocity to the steady state velocity given by the state equation. Figures 3 and 4 show such an experiment: For the factory model that gave us the steady state relationship in Figure 2 we performed

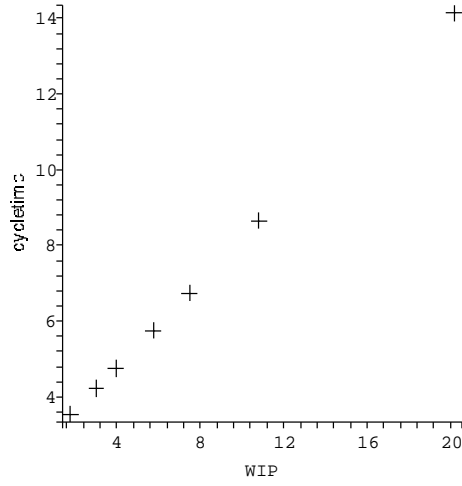


Fig. 2. Seven datapoints for a state equation describing the relationship between cycle time and WIP.

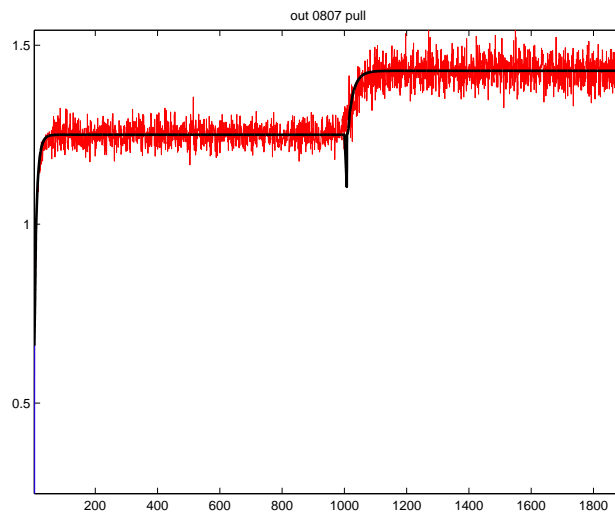


Fig. 3. Throughput as a function of time for a step up in input at $t = 1000$.

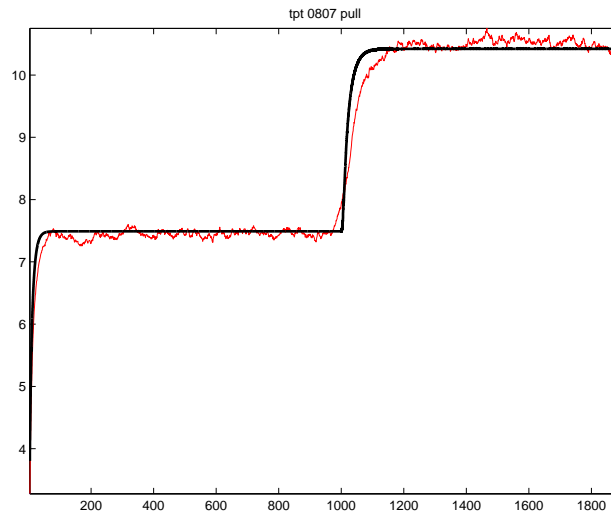


Fig. 4. Cycle time as a function of time for a step up in input at $t = 1000$.

a step-up experiment from an influx of $\lambda = 1/0.8$ to an influx of $\lambda = 1/0.7$. The noisy curves in Figs 3 and 4 are the averages for the throughput and the cycle time of a large number of discrete event simulations of that experiment. The smooth curves represent the quasi-static PDE simulation using the state equation derived from Figure 2. While the transient is 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 reduction in velocity and hence to an instantaneous reduction in outflux. Obviously, while the discrete event simulation has some of this features, it is not re-entrant enough for such a strong reaction.

2.2. Advection-diffusion equations

In analogy to stochastic particle or fluid transports one would expect that the next higher order effect beyond pure advection would involve diffusion. The resulting differential equation is an advection-diffusion equation of the

form

$$\frac{\partial \rho}{\partial t} + \frac{\partial(v\rho)}{\partial x} = D(\rho, t) \frac{\partial^2 \rho}{\partial \rho^2}, \quad (9)$$

$$v_{eq}(\rho) = \Phi(L), \quad (10)$$

where the diffusion coefficient D might depend on the density ρ and on time. The factory data displayed in Fig 1 can be used to estimate this diffusion coefficient: The fan of paths through the factory becomes wider over time. We can think of this figure as a δ -distribution that was started at $t = 0$ at the position $x = 0$. By determining the positions of the lots at a given time we can get a histogram of the widening distribution. Data matching of the Gaussian solutions to the corresponding advection diffusion equation allows us to determine the speed of the center of the distribution as well as the diffusion coefficient. In principle, with better resolution and more lots it would be possible to identify regions of high diffusion in the factory. With only 200 lots evaluated at less than 10 positions along the factory we can only get an order of magnitude for the diffusion coefficient. Details of this discussion can be found in ¹. Figure 5 shows the influence of the diffusion coefficient on the motion of a WIP-wave through the factory.

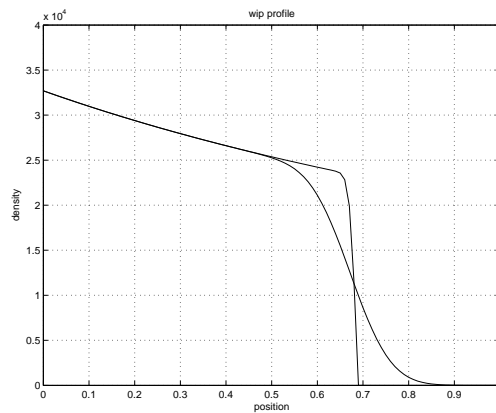


Fig. 5. WIP profile after a step up in influx, with and without diffusion

2.3. Policies and bottlenecks

So far we have treated the factory as homogeneous in the stage variable x . I.e. the velocity of a moving lot was the same at any position in the factory. While we might not want to go into the details of modeling individual stochastic behavior of machines or the production step at a certain stage (since in that case we would be back at discrete event simulations that are very time consuming), we might be interested in the influence of dispatch policies and in the behavior at or near certain important machines that represent a bottleneck. To study these issues, we can augment our heuristic models.

2.3.1. Dispatch rules

For any topologically complicated flow where lots at different stages of the production process approach the same machine, dispatch rules are needed to decide which step the machine should do next, once it becomes free. The three major dispatch rules that are used are FIFO, PUSH and PULL policies. In FIFO the sequence of lots through the machine corresponds to the sequence of arrivals. In a PUSH policy the arriving lots are ordered according to their stage number and lots with lower stage number receive preference. In a PULL policy, lots with higher stage number receive preference. Typically a PULL policy is used for "make to order" processes where production tries to satisfy a given order at a given date. A PUSH policy typically reflects a "made to plan" process where production typically fills a warehouse. While to first order FIFO does not distinguish between stages in the factory and hence a homogeneous model will be a good approximation, PUSH, PULL and any other more complicated rule can be incorporated via integration kernels in our state equation. Let $w(x, s)$ indicate the importance of a queue at location s in completion space on the speed of a lot at location x . Then the velocity at position x can be written as

$$v(x, t) = v_0 \left(1 - \frac{1}{L_{max}} \int_0^1 w(x, s) u(s, t) ds \right). \quad (11)$$

As a result, the velocity will cease to be uniform throughout the factory. For instance, a pull policy is modeled by the kernel

$$w(x, s) = \begin{cases} 0 & \text{if } s < x, \\ 1 & \text{if } x \leq s, \end{cases} \quad (12)$$

leading to

$$v(x, t) = v_0 \left(1 - \frac{1}{L_{max}} \int_x^1 u(s, t) ds \right). \quad (13)$$

Hence, $v(1, u) = v_0$, indicating that product at the end of production moves independently of the load of the factory, while $v(0, u)$ shows the full impact of the loading of the factory on the motion at the beginning of the production line.

2.3.2. Bottlenecks and maximal capacities

There are several fundamental stochastic processes that occur in a production system: Anything involving operators will lead to uncertainty as well as unscheduled machine breakdowns and repairs. A state equation model describes the influence of these processes in a very comprehensive average. We will show in the first principle models in the next section how to do more sophisticated averages. However, there is one more heuristic extension that can be done: 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

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \frac{\partial F}{\partial x} &= 0 \\ F(x) &= \min\{\rho v_{eq}, \mu(x)\} \\ v_{eq} &= \Phi(L) \end{aligned} \quad (14)$$

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 $x_i = \frac{i-1}{n} + \xi$ for $i = 0..n-1$ 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 $i = 0..n - 1$ the following scheme to find the capacity distribution $\mu(x)$.

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

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.

3. First Principle Models

3.1. Kinetic models

The models described in the previous section are more or less heuristic in nature. That is, the only physical principle used was product conservation, leading to the obvious conservation law $\partial_t \rho + \partial_x F = 0$ for the part density ρ and the flux density F . The dependence of F on ρ was assumed to be given, based on specific models. In this section we will review a more subtle approach, namely to derive equations for densities of parts based on simple rules for the movement of parts themselves. Generally, whenever particles which carry some attribute y move through space x with a certain velocity $u(x, y, t)$ dependent on position and attribute, the underlying equation for the density $f(x, y, t)$ of a particle being at position x with attribute y at time t is the Boltzmann equation

$$\partial_t f + \partial_x [u(x, y, t)f] + \partial_y (E(x, y, t)f) = Q[f], \quad (15)$$

where the term $\partial_y (Ef)$ models acceleration by a force E , i.e. a continuous change in the attribute, and the term $Q[f]$ models a random and discontinuous change in the attribute. $Q[f]$ will be an integral operator in attribute direction whose kernel is related to the probability distribution of the newly chosen attribute once the discontinuous change has occurred. Because of conservation of product (no parts are lost in this abrupt change) the total number of parts at position x has to remain the same before and after the change, resulting in the condition

$$\int Q[f](x, y, t) dy = 0, \quad \forall f. \quad (16)$$

If we define the part density ρ and the flux density F at position x and time t by

$$\rho(x, t) = \int f(x, y, t) dy, \quad F(x, t) = \int u(x, y, t) f(x, y, t) dy,$$

integrating (15) with respect to the attribute variable y gives, using (16), the conservation law

$$\partial_t \rho + \partial_x F = 0.$$

The advantage of the kinetic approach lies in the fact that the rules and underlying assumptions going in the formulation of equation (15) can be much more detailed than those involved in modeling F simply as a function of ρ and its derivatives and integrals. From an information theoretic point of view, the models in the previous section basically state that we can compute the flux density F by just knowing the value of ρ (and maybe some of its derivatives and integrals). In contrast, the relation between ρ and F is much more complex here, involving the density f , and implies that we need to know the whole history of ρ to compute the flux density F . We heavily use the methodology of classical gas dynamics, where equation (15) corresponds to the Boltzmann equation for a rarefied gas and the models of the previous section play the role of the basic equations of gas dynamics, the compressible Euler equations and the Navier - Stokes equations. We will refer to (15) as the kinetic equation underlying the macroscopic equations of the previous section. We refer the reader to ¹¹ for an overview of the underlying theory. It should be noted, that, from a theoretical as well as a numerical point of view, the kinetic theory of supply chains is simpler than the corresponding theory of gas dynamics, since the stage variable x ('space') will in general be one dimensional.

3.2. Deterministic kinetic models

We will first address the case of a purely deterministic movement of parts through the system. This will lead to gas dynamic equations of the type discussed in Section 2. This means the resulting kinetic equation (15) will be purely hyperbolic and all information about the system at any time is given deterministically by the influx. We start by defining the motion of parts through the system via trajectories. The Newton equations of motion of part number n with position $x = \xi_n(t)$ and attribute $y = \eta_n(t)$ are given by

$$\partial_t \xi_n = u(\xi_n, \eta_n, t), \quad \partial_t \eta_n = E(\xi_n, \eta_n, t), \quad (17)$$

where the functions u and E are the velocity and the accelerating force of the particle. The choice of u and E depends on the model under consideration. We assume that the part number n enters the production system at time $t = a_n$ with an attribute $y = r_n$ and therefore impose the initial conditions

$$\xi_n(a_n) = 0, \quad \eta_n(a_n) = r_n,$$

for (17). Considering a system of N parts, the density function $f(x, y, t)$ in this picture will be given by

$$f(x, y, t) = \sum_{n=1}^N \delta(x - \xi_n(t)) \delta(y - \eta_n(t)) H(t - a_n),$$

where $H(t)$ denotes the usual Heaviside function, denoting that the part does not exist in the system for $t < a_n$. Since f is a measure, equation (15) can of course only be valid in its weak form, which is obtained by multiplying (15) with a compactly supported test function $\psi(x, y, t)$ and integrating by parts.

$$\int_0^\infty dx \int_0^\infty dt \int dy f(x, y, t) [\partial_t \psi + u \partial_x \psi + E \partial_y \psi] = \quad (18)$$

$$\int_0^\infty dt \int dy f(0, y, t) \psi(0, y, t) + \int_0^\infty dx \int dy f(x, y, 0) \psi(x, y, 0).$$

On the other hand, computing the total derivative of the function $\psi(\xi_n(t), \eta_n(t), t)$ yields, using the Newton equations (17)

$$\frac{d}{dt} \psi(\xi_n, \eta_n, t) = u(\xi_n, \eta_n, t) \partial_x \psi(\xi_n, \eta_n, t) + E(\xi_n, \eta_n, t) \partial_y \psi + \partial_t \psi.$$

Integrating the above from $t = a_n$ to $t = \infty$ gives

$$\psi(0, \omega_n, a_n) = \int_0^\infty dt \int_0^\infty dx \int dy \times$$

$$H(t - a_n) \delta(x - \xi_n(t)) \delta(y - \eta_n(t)) [u(x, y, t) \partial_x \psi(x, y, t) + E(x, y, t) \partial_y \psi + \partial_t \psi],$$

and summing up over all parts gives

$$\int_0^\infty dt \int dy \psi(0, y, t) \sum_{n=1}^N \delta(t - a_n) \delta(y - r_n) = \quad (19)$$

$$\int f(x, y, t) [u(x, y, t) \partial_x \psi(x, y, t) + E(x, y, t) \partial_y \psi + \partial_t \psi] dx dy dt$$

Comparing (18) with (19), we see that the measure f satisfies the initial boundary value problem

$$(a) \partial_t f + \partial_x[uf] + \partial_y[Ef] = 0, \quad (b) f(0, y, t) = f^B(y, t) = \sum_{n=1}^N \delta(t - a_n) \delta(y - r_n), \quad (20)$$

$$(c) f(x, y, 0) = 0$$

in the weak sense. (Here we start for simplicity with an empty system. A nonzero initial condition can easily be included by assigning parts to the system at time $t = 0$ ⁴.) The advantage of the kinetic model (20) lies in the fact that the boundary density f^B can be replaced by a continuous function, thus giving a scalable model for a part density. Moments of (20) with respect to the attribute y give the macroscopic equations in section (2.2). If we c.f. postulate that all particles in (20) at a given stage are supported by a single δ -function of the form

$$f(x, y, t) = \rho(x, t) \delta(y - Y(x, t)),$$

we obtain the model (2) of the form $\partial_t \rho + \partial_x(v\rho) = 0$ with $v(x, t) = u(x, Y(x, t), t)$. To be compatible with boundary condition (20)(b), we have to choose the macroscopic attribute Y at the boundary such that $Y(0, a_n) = r_n$ holds. This, however is only the most simplistic approach to modeling. The advantage of kinetic models lies in the fact that the relations between attributes and velocities can be much more complex and better adjusted to the real situation.

Models with capacity constraints:

We now turn to a more sophisticated model where we incorporate the fact that the system will have a limited capacity. If the influx into the system exceeds this capacity bottlenecks will form. This raises the question of policies or service rules, i.e. we have to make a decision which clients / parts to serve first if they cannot all be served simultaneously. The details of the theory outlined below are given in ⁶. We consider the case of a production system with a limited capacity where the attribute y denotes the inverse priority of the part. That is, we assume that the total flux at the stage x cannot exceed a certain capacity $\mu(x)$, and we process parts in order of their priority, i.e. parts with a smaller attribute y get processed first until the capacity is reached. The service rule is therefore given by the value we assign the attribute at entry into the system and by the way the attribute

evolves in time. This is modelled by setting the velocity $u(x,y,t)$ in (17) equal to

$$u(x, y, t) = v_0(x)H(b(x, t) - y) , \quad (21)$$

where v_0 denotes the free velocity, i.e. the throughput time of stage x , and $b(x, t)$ denotes some cutoff value. This means that parts move with the free velocity v_0 for those attributes $y < b$ and they stop for $y > b$. To enforce the capacity constraint $\mu(x)$ we define the free cumulative flux $G(x, y, t)$ by

$$G(x, y, t) = v_0(x) \int_{-\infty}^y f(x, y', t) dy' ,$$

and the actual flux $F(x, t)$ by

$$F(x, t) = v_0(x) \int H(b(x, t) - y) f(x, y, t) dy = G(x, b(x, t), t) . \quad (22)$$

To enforce the constraint $F(x, t) \leq \mu(x)$ we have to therefore set $b(x, t)$ equal to $G^{-1}(x, \mu(x), t)$, where G^{-1} denotes the functional inverse of G with respect to its second variable ($G(x, y, t)$ is a monotone function of y and therefore possesses a functional inverse.) Since G is monotone, we can replace the term $H(b(x, t) - y)$ in (21) by $H(G(x, b(x, t), t) - G(x, y, t)) = H(\mu(x) - G(x, y, t))$. This implies for the total flux $F(x, t)$ in (22)

$$F(x, t) = v_0(x) \int H(\mu(x) - G(x, y, t)) f(x, y, t) dy = \quad (23)$$

$$\left(\begin{array}{ll} G(x, \infty, t) = v_0(x)\rho(x, t) & \text{for } v_0\rho < \mu \\ \mu(x) & \text{for } v_0\rho > \mu \end{array} \right) = \min\{\mu(x), v_0(x)\rho(x, t)\} .$$

This gives the macroscopic equation postulated in Section (2.3) and in ⁵. This is somewhat trivial in the case when parts or clients are processed on a first - come - first - serve basis. It becomes more interesting when the service rule is more complex. Suppose we give each part a deadline or due date and process parts in order of their remaining time to this due date. This means we set the time dependent attribute for part number n equal to $\eta_n(t) = d_n - t$, where d_n denotes the due date. This gives for the boundary density f^B in (20)

$$f^B(y, t) = \sum_{n=1}^N \delta(t - a_n) \delta(y - d_n + a_n)$$

and for the kinetic equation

$$\partial_t f + \partial_x [v_0(x)H(\mu(x) - G(x, y, t))f] - \partial_y f = 0, \quad (24)$$

$$G(x, y, t) = v_0 \int_{-\infty}^y f(x, y', t) dy' .$$

So the attribute y of a part equals the time to due date d_n at the entry time $t = a_n$ and decays continuously, making late parts more important. Now, the conservation law $\partial_t \rho + \partial_x F = 0$ will still hold with the macroscopic flux function F given by (23). However, this equation provides no information about which parts come out first, and whether we met the due dates or not. To obtain this information, we would have to solve the kinetic model (24). The solution of (24) can be quite involved, especially in the case when more than one type of attribute is considered and therefore the attribute variable y is in a higher dimensional space.

The multi - phase model

One way to obtain more information than given by the simple mass conservation law is to consider higher order moments of (24). Integrating (24) against powers of the attribute variable y gives

$$\partial_t m_j + \partial_x F_j - j m_{j-1} = 0, \quad j = 0, \dots, 2J - 1 \quad (25)$$

where the moments m_j and the moment fluxes F_j are given by

$$m_j(x, t) = \int y^j f dy, \quad (26)$$

$$F_j(x, t) = v_0(x) \int y^j [H(\mu(x) - G(x, y, t))] f(x, y, t) dy, \quad (27)$$

$$(28)$$

So $m_0 = \rho$ and $F_0 = F$ holds in the previously used notation. This gives J equations for the $2J$ unknowns $m_j, F_j, j = 0, \dots, J - 1$. As often in kinetic theory (see e.g. ¹⁷), the moment system is not closed. Some Ansatz must be made to find additional J relations among these $2J$ data. To express the various unknown fluxes in terms of the moments, we close the expressions by the Ansatz that at each stage x of the process the kinetic density $f(x, y, t)$ is given by a superposition of J δ - functions and set

$$f(x, y, t) = \sum_{k=1}^J n_k(x, t) \delta(y - Y_k(x, t)) . \quad (29)$$

Equation (29) has the following interpretation. By making the Ansatz of a superposition of J concentrations we allow for overtaking of parts, i.e. parts with a higher priority might pass lower priority part. However, we limit the passing process by assuming that no more than J parts can pass

each other at the same time, i.e. we consider essentially a traffic model with J lanes. For the case $J = 1$ (single lane traffic, no passing) this reduces to the case when the flux F_0 is given by (23) ⁶. This gives rise to what is called multi-phase fluid models, which were originally developed in ¹⁴ as an alternative to WKB methods for the Schrödinger equation of quantum mechanics. Using the Ansatz (29), the moments fluxes and acceleration terms in (26) and (27) are given in terms of the concentrations n_k and the macroscopic attributes Y_k in terms of

$$m_j(x, t) = \sum_{k=1}^J n_k Y_k^j, \quad (30)$$

$$F_j(x, t) = v_0(x) \sum_{k=1}^J n_k Y_k^j Z_k \quad (31)$$

$$Z_k = \max\{0, \min\{1, \frac{\mu(x) - v_0 \sum_{Y_s \neq Y_k} n_s H(Y_k - Y_s)}{v_0 \sum_{Y_s = Y_k} n_s}\}\}. \quad (32)$$

Inserting these relations into the moment equations (25) yields $2J$ partial differential equations for the $2J$ unknowns $n_1, \dots, n_J, Y_1, \dots, Y_J$. To illustrate the meaning of the multi-phase model we briefly discuss the case $J = 2$, i.e. we construct a flow with two phases, a high and a low priority phase. In this case, inserting (30) and (31) into the moment equations gives the following set of four equations for n_1, n_2, Y_1, Y_2 :

$$\begin{aligned} \partial_t(n_1 Y_1^j + n_2 Y_2^j) + \partial_x v_0(n_1 Z_1 Y_1^j + n_2 Z_2 Y_2^j) \\ = j[n_1 Y_1^{j-1} + n_2 Y_2^{j-1}] \end{aligned} \quad (33)$$

The issue is now the computation of Z_k , $k = 1, 2$. Let us suppose that $Y_1 < Y_2$ to fix the ideas. Then, (32) leads to the following discussion :

$$(i) \text{ if } \mu < n_1 v_0 \quad \text{then} \quad n_1 v_0 Z_1 = \mu \quad \text{and} \quad n_2 v_0 Z_2 = 0, \quad (34)$$

$$(ii) \text{ if } n_1 v_0 < \mu < v_0(n_1 + n_2) \quad \text{then} \quad n_1 v_0 Z_1 = n_1 v_0 \\ \text{and} \quad n_2 v_0 Z_2 = \mu - n_1 v_0, \quad (35)$$

$$(iii) \text{ if } v_0(n_1 + n_2) < \mu \quad \text{then} \quad n_1 v_0 Z_1 = n_1 v_0 \\ \text{and} \quad n_2 v_0 Z_2 = n_2 v_0. \quad (36)$$

Of course, the roles of 1 and 2 must be exchanged in the case $Y_1 > Y_2$. When $Y_1 = Y_2$, then

$$n_1 v_1 Z_1 = \min\{n_1 v_1, \mu \frac{n_1}{n_1 + n_2}\}, \quad n_2 v_2 Z_2 = \min\{n_2 v_2, \mu \frac{n_2}{n_1 + n_2}\}. \quad (37)$$

What formulas (34)-(36) express is very simple. $n_k v_0$ is the 'free flux' of parts in phase k and $v_0(n_1 + n_2)$ is the total 'free flux' (we call 'free fluxes' the fluxes if there would be no flux limitation). In the first case, the flux limitation μ is already below the free flux of parts 1 and therefore, the actual flux of these parts is equal to the flux constraint and parts 2 simply do not move. In the second case, the flux constraint μ is larger than the free flux of parts 1 but below the total free flux. Therefore, the flux constraint does not apply to parts 1 which move with actual flux equal to their free flux. The actual flux constraint which applies to parts 2 is the total flux constraint c diminished by the flux of parts 1 and therefore, parts 2 move under this flux constraint. In the last case, there is no flux constraint at all because the flux constraint is above the total free flux and each parts actually moves according to its own free flux. Clearly, this is consistent with the policy consisting in processing parts with lower attributes first. Again, the role of 1 and 2 must be exchanged in the case $Y_2 < Y_1$. The multiphase model with J phases can be interpreted as a model for a reentrant production system with loops if the phases are linked through the boundary conditions, i.e. for a pull policy we would reduce the phase each time a part runs through the system and set $n_k Y_k Z_k(0, t) = n_{k+1} Y_{k+1} Z_{k+1}(1, t)$, $k = 0, \dots, J-1$ with $n_J Y_J Z_J(0, t)$ equal to the total influx λ .

An example with bottlenecks:

We conclude the discussion of purely deterministic models with an example of a supply chain with two bottlenecks. We solve a 'hot lot' problem, i.e. a system with a steady flow of low priority parts ('cold lots') which is disrupted by a sudden influx of high priority parts ('hot lots'). We consider a chain of 20 stations, all with throughput time =1. So the total minimal throughput time is 20. They all have a capacity of $\mu = 160$ parts per unit time, except for number 5, which has $\mu = 80$ and number 15 which has $\mu = 40$ (two bottlenecks). We consider a constant influx of 'low priority' parts, i.e. with a due date far in the future, of 60 parts per unit time. At time $t = 40$ 'hot lots' (parts with a much closer due date) arrive at a rate of 60 parts per unit time. With these data, the first bottleneck with $\mu = 80$ can accommodate the flow of one of either parts (hot or low) but not both together. The second bottleneck ($\mu = 40$) cannot even accommodate one single flow. Within the low priority lot and the hot lot population the due dates are chosen randomly in a given interval.

The phenomena we expect to see are the following. The low priority lots pass freely through the first bottleneck but start to pile up at station 15.

This is the picture until the hot lots arrive at $t = 40$. Once the hot lots arrive, they pass freely through the first bottleneck, but constrict the flow of the low priority lots there. As soon as they reach the second bottleneck, they start to pile up and strangle the low priority flow there completely. Once the hot lots have passed through, the queues start to dissolve. The simulation runs from $t = 0$ to $t = 140$. To verify the two phase model outlined above we have directly simulated the kinetic equation (24) using a particle simulator (see ⁶ for details). Given the particle solution we generated a two phase approximation to the densities resulting from the particle solution and compared this to the actual solution of the two - phase model (33). So, given the particle solution, we first compute its first four moments m_0, \dots, m_3 and compute a corresponding phase and density according to (30). The corresponding result is compared with the solution of the 2-phase model in Figure 6 for different times. The solid and the dashed lines denote the hot and the cold phase of the 2-phase model. The triangles and \times 's denote the data points for the corresponding phases extracted from the particle model. (Note, that, numerically, there will always be two phases!). The left panel shows the values of the attributes Y_1 and Y_2 , and the right panel shows the densities n_1 and n_2 . The densities are plotted on a logarithmic scale. So, for perfect agreement, the \times symbols, the values for the 'cold' phase of the particle model, should be on top of the dashed line, the 'cold' phase of the two phase model. The triangles, the values for the 'hot' phase of the particle model, should be on top of the solid line, the 'hot' phase of the two phase model.

Figure 6 shows a reasonable agreement between the particle solution and the 2-phase model. Note, that the two phases coincide for a while to the right of the bottleneck at station 15, meaning that there is only the high priority flux there, since the low priority flux has been cut off completely. At this point both densities (in the right panel) are large since we have parts of both, the hot and the cold phase, accumulating. On the other hand, at the first bottleneck, at station number 5, only the densities of the low priority flux (the symbols \times in the right panel) become large.

3.3. Stochasticity and diffusion

The results of Section (3.2) pertain to a strictly deterministic system. In essence, we have assumed so far, that the production system works like an automaton, i.e. given a current state of the system we can determine with absolute certainty the further progress of the part traveling through the

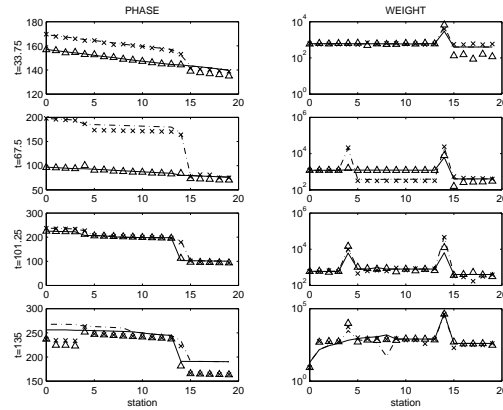


Fig. 6. 2-Phase picture, left panel=attributes Y_1, Y_2 , right panel=densities, \times, Δ =particles, $\cdot, \cdot-$ =2phase model.

system. This approach is insufficient for the following two reasons.

- Production systems are inherently random. Each individual node in the system can incur random breakdowns. The influx in the system will exhibit some statistical variations.
- The system might be too complex to model every single node in detail. Transport parameters, like the form of the velocity $u(x, y, t)$ in (17) might therefore be obtained from observation of the actual physical system over a period of time, measuring WIP's and throughput times, and, rather than fitting one particular function to these data, it will be more accurate to use them in form of statistical distributions.

We therefore introduce the element of randomness into the system. The basic picture for an individual part entering the system at the arrival time a_n with attribute r_n is the following.

Random velocity updates

We consider a production system with M stages. We divide the interval $0 < x < 1$ into M equal subintervals and attribute to each production stage and interval of length $\frac{1}{M}$. Part number n moves through this system with a velocity $\eta_n(t)$. This corresponds to identifying the attributes with velocity and setting $u(x, y, t) = y$ in (17). Each time we enter a new stage in we update the velocity η_n , i.e. the estimated throughput time for the next

stage, randomly from a given distribution which, in some way, depends on the over all state of the system. In this picture, we would update the velocity randomly each time the part has travelled a distance $\Delta x = \frac{1}{M}$. To obtain a more tractable mathematical model, we replace this picture where we update the velocity not at fixed spatial intervals Δx but, on average, at certain time intervals. So we update the velocity with a certain frequency ω or, in a probabilistic picture, we update the velocity every infinitesimal time step Δt with a probability $\omega \Delta t$. In order to update, on average, each time the part has travelled a distance $\frac{1}{M}$ we have to set $\frac{\eta}{\omega} = \frac{1}{M}$. This corresponds to the concept of a mean free path in gas dynamics⁷. Therefore the advance of a part with position $\xi(t)$ in the infinitesimal time interval Δt is governed by

$$\xi(t + \Delta t) = \xi(t) + \Delta t \eta(t), \quad \eta(t + \Delta t) = (1 - \kappa(\eta(t)))\eta(t) + \kappa\gamma(\xi, t) \quad (38)$$

$$d\mathcal{P}\{\kappa(\eta) = k\} = [\Delta t M \eta \delta(k - 1) + (1 - \Delta t M \eta) \delta(k)] dk,$$

$$d\mathcal{P}\{\gamma(\xi, t) = z\} = P(\xi, z, t) dz,$$

So each Δt we toss a weighted coin and choose $\kappa = 1$ with a probability $\Delta t M \eta$ and $\kappa = 0$ with probability $1 - \Delta t M \eta$. If $\kappa = 1$ holds we update the velocity η randomly from the distribution $P(x, z, t)$ which depends on the state of the production system at time t in some form. If $\kappa = 0$ holds we keep the velocity η at its current level. It should be pointed out here that, at this point we make a form of mean field assumption, namely that the probability distribution P is itself independent of the particle coordinates (ξ, η) . P will depend in general of some averaged quantities of the whole ensemble. Taking P independent of (ξ, η) means that we assume that there are many parts in the system, such that the influence of one individual part on the whole ensemble can be neglected. To obtain equations for densities we define the probability distribution

$$f(x, y, t) dx dy = d\mathcal{P}\{\xi(t) = x, \eta(t) = y\}.$$

At this point we encounter the usual problem defining densities for open systems. The number of parts in the system will not be constant (the part with position ξ does not exist for $t < a$), whereas the probability density $f(x, y, t) dx dy$ has to integrate to unity for all time. We remedy this situation by starting the part at some appropriate negative spatial coordinate at time $t = 0$ and move it with a constant velocity for $x < 0$, such that

it arrives at the entrance $x = 0$ of the production system precisely at the time a . Thus, we replace (38) by

$$(a) \quad \xi(t + \Delta t) = \xi(t) + \Delta t \eta(t), \quad \eta(t + \Delta t) = H(-\xi)\eta(t) + H(\xi)[(1 - \kappa(\eta(t)))\eta(t) + \kappa\gamma] \quad (39)$$

$$(b) \quad \xi(0) = -ra, \quad \eta(0) = r .$$

Therefore $\xi(t) = r(t - a)$, $\eta(t) = r$ will hold as long as $t < a$ holds and the part will arrive at $x = 0$ at time $t = a$ with velocity r . Summing up over all possible scenarios we obtain from (39)

$$f(x, y, t + \Delta t) =$$

$$\int \delta(\xi + \Delta t \eta - x) \delta(H(-\xi)\eta + H(\xi)[(1 - k)\eta + kz] - y) f(\xi, \eta, t) \times$$

$$d\mathcal{P}\{\kappa(\eta) = k\} d\mathcal{P}\{\gamma(\xi, t) = z\} d\xi \eta ,$$

which in the limit $\Delta t \rightarrow 0$ yields (see ³ for the details)

$$\partial_t f + y \partial_x f = MH(x) [P(x, y, t) \int z f(x, z, t) dz - y f(x, y, t)] . \quad (40)$$

Assuming that the arrival times a and the initial velocities r are distributed according to a distribution $A(\alpha, \beta)$ $d\alpha d\beta = d\mathcal{P}\{a = \alpha, r = \beta\}$, we obtain the initial condition

$$f(x, y, 0) = \frac{1}{y} A\left(-\frac{x}{y}, y\right) .$$

Since f satisfies the pure convection equation $\partial_t f + y \partial_x f = 0$ for $x < 0$, we can replace the above problem by an initial boundary value problem by solving (40) via the method of characteristics for $x < 0$ and obtain the initial boundary value problem

$$\partial_t f + y \partial_x f = M [P(x, y, t) \int z f(x, z, t) dz - y f(x, y, t)], \quad x > 0, \quad (41)$$

$$y f(0, y, t) = A(t, y) .$$

Asymptotics for many stage processes:

To reduce the initial boundary value problem (41) for the kinetic density function f to an equation for the part density ρ in the previous section, we take the zero order moment of (41), and define the quantities

$$\rho(x, t) = \int f(x, y, t) dy, \quad F(x, t) = \int y f(x, y, t) dy . \quad (42)$$

So, $\rho(x, t) dx$ denotes the probability of the part being at stage x of the production process at time t , and F is the corresponding flux. Integrating (41) with respect to y gives the mass conservation law

$$\partial_t \rho + \partial_x F = 0, \quad 0 < x < 1, \quad F(0, t) = \lambda(t) = \int A(t, y) dy, \quad (43)$$

and the goal is now to express the flux F in terms of the density ρ . This is done by a functional expansion of the kinetic density f in (41) for large values of M , i.e. for a system with many stages, which is the equivalent of the Chapman - Enskog expansion for the solution of the Boltzmann equation of gas dynamics (see c.f. ⁷). We formally set

$$f(x, y, t) = \phi(\rho(x, t), y, t) .$$

So we try to express the kinetic density f as a 'shape function' dependent on space only through the macroscopic density ρ . Because of the definition (42) the shape function ϕ has to satisfy

$$\int \phi(\rho, y, t) dy = \rho, \quad \forall \rho . \quad (44)$$

The macroscopic flux F is then given by

$$F(\rho, x, t) = \int y \phi(\rho(x, t), y, t) dy .$$

Inserting this Ansatz into the kinetic equation (13.9), using the macroscopic conservation law (43) gives

$$\partial_t \phi - (\partial_\rho \phi) \partial_x F + \partial_x [y \phi] = M [PF - y \phi] , \quad (45)$$

and we obtain asymptotic expressions for the shape functions ϕ and F by an expansion of (45) for large values of M , i.e. setting $\phi(\rho, y, t) = \phi_0 + \frac{1}{M} \phi_1 + \dots$ and $F(\rho, t) = F_0 + \frac{1}{M} F_1 + \dots$. For the zero'th order term of this expansion we obtain the equation

$$PF_0 - y \phi_0 = 0, \quad F_0 = \int y \phi_0 dy ,$$

which implies that ϕ_0 is given by the probability density P divided by y and multiplied by an arbitrary function of x and t . We denote this arbitrary function by ρv and obtain the macroscopic velocity $v(x, t)$ from the normalization condition (44), giving

$$\phi_0(\rho, x, y, t) = \rho v(x, t) \frac{P(x, y, t)}{y}, \quad \frac{1}{v(x, t)} = \int \frac{P(x, y, t)}{y} dy, \quad (46)$$

$$F_0(\rho, x, t) = \rho v(x, t)$$

This gives in zero'th order the heuristic model (9) with the macroscopic velocity v computed as the reciprocal value of the reciprocal microscopic velocity. For a part moving with velocity y the random throughput time $\tau(x, t)$ for the station at stage x , occupying an interval of length $\frac{1}{M}$, is given by $\tau = \frac{1}{My}$. Therefore the macroscopic velocity $v(x, t)$ is related to the random throughput time as

$$v(x, t) = \frac{1}{ME[\tau(x, t)]}, \quad (47)$$

where $E[\tau]$ denotes the expectation of τ under the probability distribution P . This fact is important if one wishes to generate the distribution P from observed data of an actual production system. Going to the next term in the expansion will give the diffusive term. The balance of order $O(\frac{1}{M})$ terms in (45) reads

$$\partial_t \phi_0 - (\partial_\rho \phi_0) \partial_x F_0 + \partial_x [y \phi_0] = P F_1 - y \phi_1$$

or using (46)

$$\rho \partial_t \left[\frac{vP}{y} \right] - \frac{vP}{y} \partial_x (v\rho) + \partial_x [\rho v P] = P F_1 - y \phi_1$$

The shape F_1 is again computed from the normalization condition (45), which in first order reads $\int \phi_1(\rho, x, y, t) dy = 0 \forall \rho$. This gives

$$\rho \partial_t \left[v \int \frac{P}{y^2} dy \right] - v \partial_x (v\rho) \int \frac{P}{y^2} dy + \partial_x [\rho v \int \frac{P}{y} dy] = F_1 \int \frac{P}{y} dy$$

or

$$F_1 = v \rho \partial_t \left[v \int \frac{P}{y^2} dy \right] - v^2 \partial_x (v\rho) \int \frac{P}{y^2} dy + v \partial_x \rho = -D \partial_x \rho + v R \rho$$

with the diffusion coefficient D and the first order correction R of the velocity being given in terms of the variation coefficient $V[\frac{1}{y}]$ of the throughput time $\frac{1}{y}$, i.e. the variance of this quantity scaled by the square of the mean. D and R can be expressed as

$$(a) \quad D(x, t) = vV\left[\frac{1}{y}\right], \quad V\left[\frac{1}{y}\right] = v^2 \int \frac{1}{y^2} P dy - 1 = \frac{E\left[\frac{1}{y^2}\right] - E\left[\frac{1}{y}\right]^2}{E\left[\frac{1}{y}\right]^2}, \quad (48)$$

$$(b) \quad R = \partial_t \left(\frac{V\left[\frac{1}{y}\right] + 1}{v} \right) - (V\left[\frac{1}{y}\right] + 1) \frac{1}{v} \partial_x v = (\partial_t + v \partial_x) \frac{V\left[\frac{1}{y}\right] + 1}{v} - \partial_x V\left[\frac{1}{y}\right]$$

$V[\frac{1}{y}]$ is a dimensionless measure of the stochasticity of the system. $V = 0$ holds for the degenerate case if the probability distribution P is a δ -function concentrated on the mean. It is important to notice that the diffusion coefficient will always be positive. Thus, the flux in the continuity equation (43) is given up to order $\frac{1}{M^2}$ terms as

$$F = \rho v \left(1 + \frac{R}{M}\right) - \frac{1}{M} v V\left[\frac{1}{y}\right] \partial_x \rho. \quad (49)$$

Note that, because of (47) the macroscopic velocity v , and therefore the whole flux F , will be of order $O(\frac{1}{M})$. The reason for this is that we measure time in units corresponding through the throughput time of one individual node in the supply chain. So the whole system will therefore evolve very slowly on this time scale. Rescaling time velocity and fluxes via

$$t \rightarrow tM, \quad \partial_t \rightarrow \frac{\partial_t}{M}, \quad v \rightarrow \frac{v}{M}, \quad F \rightarrow \frac{F}{M}, \quad \lambda(t) \rightarrow \frac{1}{M} \lambda\left(\frac{t}{M}\right)$$

would yield the same initial value problem for the conservation law (43) with the same flux function F as in (49), but with a macroscopic velocity $v(x, t)$, according to (47) which is now of order $O(1)$.

So far, $\rho(x, t)$ has denoted the probability distribution that a certain part is at stage x at time t and $\lambda(t)$ was the probability distribution that the part arrives in the system at time t . Under the mean field assumption that the density P is independent of a single part, all parts in the system will obey the same equation. Therefore the probability ρ can be identified with the part density ρ up to a constant factor and the probability density λ can be identified with the start rate (the influx into the system) up to the same factor. In zero'th order (for $M = \infty$) the flux function (49) reduces to the one given in (2) with $\frac{1}{v}$ given by the expectation of the local throughput time τ . In first order, including the $O(\frac{1}{M})$ terms in (49) we obtain the diffusion model (9) with the diffusion coefficient given in terms of the variation coefficient of the statistics. (The term $\frac{R}{M}$ can be neglected since it will always be dominated by v .) However, the above analysis gives an indication of how to match the transport coefficients v and D to a given production system in a more sophisticated way than by just fitting parameters.

The usage of the above analysis is the following. To obtain a simple model of an actual supply chain we wish to determine transport coefficients for the conservation law (43) from observations of the system for a certain period of time. To this end we break down the process into M stages and observe the time each part has arrived at each stage for a large number

of parts, i.e. we record the numbers t_n^m , $n = 1, \dots, N$, $m = 1, \dots, M + 1$ for $N \gg 1$, where t_n^{M+1} denotes the exit time. Then we proceed as follows.

- From this we compute the throughput times $\tau_n^m = t_n^{m+1} - t_n^m$, $n = 1, \dots, N$, $m = 1, \dots, M$, i.e. the time it took part number n to complete stage number m of the process.
- We also record a certain integral quantity (like the WIP) of the whole system, computing a state variable $S_m(t_n^m)$ for each stage.
- Then we fit a probability distribution $T_m(\tau, S_m(t_n^m))$ to the throughput times τ_n^m .
- After making everything continuous, we obtain a probability distribution $T(x, \tau, S) d\tau$ for the throughput time through the stage x , given a state S of the system. This probability density is related to the probability density for the velocity in the above derivation via $P(x, y, S) = \frac{1}{My^2} T(x, \frac{1}{My}, S)$.
- Then we compute the transport coefficients v, D and R from (47) and (48), which are now dependent on the state variable $S(x, t)$ as well.

After completing this process we have a relatively simple model of the form

$$(a) \quad \partial_t \rho + \partial_x F = 0, \quad 0 < x < 1, \quad F(0, t) = \lambda(t), \quad (50)$$

$$(b) \quad F = [1 + \frac{1}{M} R(x, t, S)] \rho v(x, t, S) - \frac{1}{M} v V[\frac{1}{y}](x, t, S) \partial_x \rho$$

for the supply chain. It remains to define the state variable $S_m(t_n^m)$, or in the continuous version $S(x, t)$. This is a question of how to interpret given experimental data and what integral state of the system to record in the experiment. For a linear supply chain without any reentrant processes it will be reasonable to take $S_m(t_n^m)$ equal to the number of parts in stage m at time t_n^m , or to take $S(x, t) = \rho(x, t)$. For a more complex system with reentrant processes, it will be more appropriate to take S equal to the total work in progress, i.e. $S(x, t) = S(t) = \int \rho(x, t) dx$. If the supply chain contains reentrant processes which are solely governed by pull policies, then the throughput times will depend only on the load downstream and we will set $S(x, t) = \int_x^1 \rho(x', t) dx'$.

A numerical comparison

We conclude this section with a numerical example. We consider a reentrant production system, for which we prescribe a uniform spatial velocity

distribution between a minimum and a maximum local throughput time $\tau(x)$. We simulate the system via the zeroth order model (2) taking v as the expectation. This is referred to as the queuing model. Then we include the random updates according to (39) in a particle based simulation (see ³ for details). Figure 7 shows the outflux of the system for the deterministic model (the dashed line) and a time averaged version of the randomized velocity updates (the solid line). We see the effect of randomness lowering the initial outflux significantly.

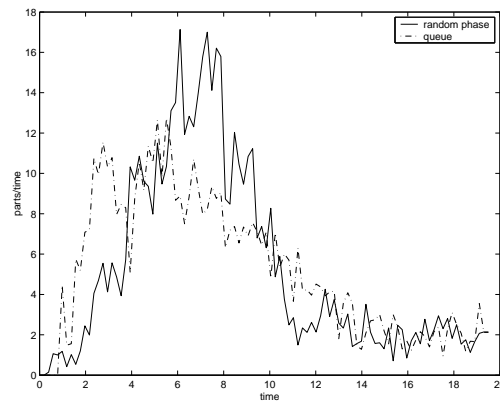


Fig. 7. Outflux (Random phase and queue) $N=400$ lots, 20 ensembles.

Figure 8 compares the outflux of the time averaged particle model with the conservation law when diffusion is added (i.e. all the terms in the flux function in (50) are considered).

4. Conclusions

We have shown how the heuristic models for production flows (Eqns 2, 9, 14) arise from different closure models of expansions of the Boltzmann equation (Eq.15). These first principle models can also be extended beyond simple heuristics to generate dynamical models for production flows with dispatch rules based on the evolution of an attribute (e.g. the due date of a particular product). The present review is a snapshot of the state of a larger research project to develop the fundamental macroscopic dynamical equations for production flows based on kinetic theory of the motion of individual production lots, and their interactions based on production rules

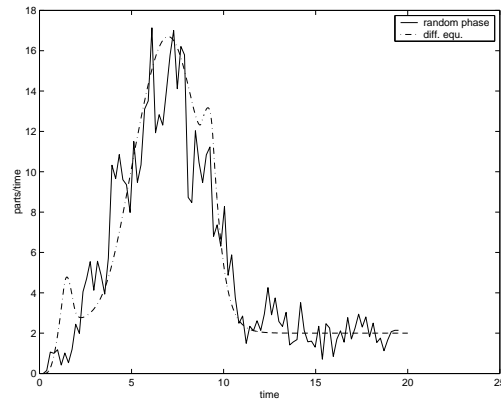


Fig. 8. Outflux for the random phase model and the diffusion equation, $N=400$ lots, 20 ensembles.

and competition for machine time and floor space. Open problems currently under investigation are:

- A macroscopic transport equation reflecting the fact that the capacity along the production line varies stochastically.
- Detailed studies comparing large scale discrete event simulations with the different macroscopic models. This involves in particular studies of time dependent flows - transient or stochastic. Open questions involve the errors generated by treating a fundamentally discrete flow as a continuum flow, the errors resulting from quasi-static and diffusive models for flows changing on a fast timescale and the ability to extract the diffusion coefficient and the state equation from discrete event data.
- The kinetic models for the flow evolution with attributes will allow us to make connections to scheduling algorithms²¹. In discrete time and for discrete production many of these algorithms are NP-hard. It remains to be seen whether a continuum approximation will lead to simpler optimization problems that are still useful as scheduling rules.

Based on the dichotomies discussed in the introduction between industrial production networks and biomolecular networks we also think that the dialog between the two fields should be continued and extended. As a minimum, thinking about the differences in the two fields sharpens the

awareness of the essential features of industrial production or biological production, respectively. However, beyond introspection, we are also quite optimistic that techniques and insights can actually be transferred between these fields.

Acknowledgments: This work was supported in parts by a grant from Intel Corporation and by NSF grant DMS 0204543. We thank Karl Kempf, Erjen Lefebber and Dan Rivera and for many insightful discussions and Tae-Chang Jo and Roel van den Berg for computational assistance.

References

1. D. Armbruster, C. Ringhofer, T-J. Jo, Continuous models for production flows, in: Proceedings of the 2004 American Control Conference, Boston, pp 4589 - 4594, 2004
2. Dieter Armbruster, Daniel Marthaler, Christian Ringhofer, Karl Kempf, Tae-Chang Jo: A continuum model for a re-entrant factory, in revision for Operations research 38 pages 9/2003
3. D. Armbruster, C. Ringhofer: "Thermalized kinetic and fluid models for re-entrant supply chains", in print, SIAM J. Multiscale Modeling and Simulation, 2005.
4. D. Armbruster, D. Marthaler, C. Ringhofer: Kinetic and Fluid Model Hierarchies for Supply Chains, *SIAM J. on Multiscale Modeling* 2, pp.43-61 (2004).
5. D. Armbruster, P. Degond, C. Ringhofer: A Model for the Dynamics of large Queuing Networks and Supply Chains, submitted, 2004
6. D. Armbruster, P. Degond, C. Ringhofer: "Kinetic and fluid models for supply chains supporting policy attributes submitted, (2004).
7. S. G. Brush: Kinetic Theory, *Pergamon Press* (1972).
8. Jakob Asmundsson, Reha Uzsoy, and Ronald L. Rardin: Compact nonlinear capacity models for supply chains: Methodology, preprint, 2002, Purdue University
9. Bartholdi, J.J. III, D. D. Eisenstein, A production line that balances itself, *Operations Research*, **44:1** (1996) 21–34.
10. H. Baumgaertel, S. Brueckner, V. Parunak, R. Vanderbok, and J. Wilke. "Agent Models of Supply Network Dynamics." in Terry Harrison et al (Eds), *The Practice of Supply Chain Management*, Kluwer, 2003
11. C. Cercignani: The Boltzmann equation and its applications, *Applied Mathematical Sciences*, vol. 67, *Springer Verlag* (1988).
12. Chi reference manual at <http://se.wtb.tue.nl/documentation/>
13. J. G. Dai, J.H. Vande Vate, The stability of two-station multitype fluid networks, *Operations Research*, **48**(5), 721-744 (2000)
14. S. Jin, X. Li, *Multi-phase Computations of the Semiclassical Limit of the Schrodinger Equation and Related Problems: Whitham vs. Wigner*, *Physica D* **182**, pp. 46-85 (2003).
15. K. H. Karlsen, N. H. Risebro, J. D. Towers, L^1 stability for entropy solutions

- of nonlinear degenerate parabolic convection-diffusion equations with discontinuous coefficients*, Skr., K. Nor. Vidensk. Selsk. **3** pp. 1–49 (2003).
16. A. Klar, R. Wegener, *Enskog-like kinetic models for vehicular traffic*, J. Stat. Phys., (1997), pp. 91–114.
 17. C. D. Levermore, *Moment closure hierarchies for kinetic theories*, J. Stat. Phys., **83** (1996), pp. 1021–1065.
 18. M.J. Lighthill, G.B. Whitham, *On kinematic waves II. A theory of traffic flow on long crowded roads*, Proceedings of the Royal Society, Series A, **229**, 317–345 (1955)
 19. U.S. Karmarkar: *Capacity loading and release planning in Work-in-Progress (WIP) and Lead-times*, J. Mfg. Oper.Mgt., 2, 105 - 123 (1989)
 20. B. Hess, A.S. Mikhailov, *Science* **264**, 223 (1994)
 21. Michael Pinedo, *Scheduling*, Prentice Hall, 1995
 22. Bart Rem, Dieter Armbruster, *Control and Synchronization in Switched Arrival Systems*, Chaos **13** (1), 128–137 (2003)
 23. Collaborative Research Centre 637 "Autonomous Cooperating Logistics Processes: A Paradigm Shift and its Limitations" Universität Bremen <http://www.sfb637.uni-bremen.de>