Modeling and Optimization of Production Processes: Lessons from Traffic Dynamics

Dirk Helbing

Institute for Economics and Traffic, Dresden University of Technology, Andreas-Schubert-Str. 23, D-01062 Dresden
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Abstract. We will develop and study models of supply networks and how they relate to vehicular traffic. These models allow to take into account the non-linear, dynamical interactions of different production units and to test alternative management strategies with respect to their potential impacts. In this way, one can understand the preconditions of the so-called bull-whip effect (i.e. the fact that small variations in the consumption rate can cause large variations in the production rate of companies generating the requested product). Moreover, we will show how the non-linear dynamics of a particular supply chain in semiconductor production has been optimized by means of the “slower-is-faster effect” known from panicking pedestrian crowds. Driven many-particle models of pedestrian motion also offer solutions for other typical problems of non-linear production processes such as the coordination of robots, the efficient segregation of different kinds of objects, or the frictionless merging of object flows at bottlenecks. Finally, from the simulation of pedestrian behavior one can learn how fluctuations could be used to increase the order in the system, how to speed up certain production processes, or how to compensate for delays in a series of production steps.

1 Modeling the Dynamics of Supply Networks

Concepts from statistical physics and non-linear dynamics have been very successful in discovering and explaining dynamical phenomena in traffic flows [1,2]. Many of these phenomena are based on mechanisms such as delayed adaptation to changing conditions and competition for limited resources, which are relevant for production systems as well. Therefore, economists [3], traffic scientists [4], mathematicians [5], and physicists [6] have recently pointed out that methods used for the investigation of traffic dynamics are also of potential use for the study of supply networks. In this contribution, our primary attention will be directed towards the non-linear interaction between different production units or production processes and the resulting dynamics, while classical queuing theory mainly focusses on stochastic fluctuations of production processes in a stationary state.

1.1 Modelling One-Dimensional Supply Chains

For simplicity, let us start with a model of one-dimensional supply chains. The assumed model consists of a series of u suppliers b, which receive products from the next “upstream” supplier b−1 and generate products for the next “downstream” supplier b+1 [7,8]. The final products are delivered to the consumers u +1 (see Fig. 1). The consumption and delivery rates are typically subject to perturbations, which may cause variations in the stock levels and deliveries of upstream suppliers. This is due to delays in the adaptation of their delivery rates.

To study the resulting dynamics, let us denote the stock level (“inventory”) at supplier b by $N_b$. It changes in time t according to the equation

$$\frac{dN_b}{dt} = \lambda_b(t) - \lambda_{b+1}(t)$$

(1)
Here, $\lambda_b$ has the meaning of the rate at which supplier $b$ receives ordered products from supplier $b-1$, while $\lambda_{b+1}$ is the rate at which he delivers products to the next downstream supplier $b+1$. Therefore, Eq. (1) is just a continuity equation which reflects the conservation of the quantity of products. (It is easy to generalize this equation to cases where products are lost. One would just have to add a term of the form $-\gamma_b N_b(t).$) Boundary conditions must be formulated for $b=0$, which corresponds to the supplier of the raw materials (fundamental resources), and for $b=u+1$, which corresponds to the consumers. That is, $\lambda_0$ is the supply or production rate of the basic product, while $\lambda_{u+1}$ is the consumption rate.

The question remains, how the delivery rates $\lambda_b$ evolve in time. It is reasonable to assume that the temporal change of the delivery rate is proportional to the deviation of the actual delivery rate from the desired one $W_b(t)$ and its adaptation takes on average some time interval $\tau$. According to this, we have the equation

$$
\frac{d\lambda_b}{dt} = \frac{1}{\tau} [W_b(t) - \lambda_b(t)].
$$

The order rate $W_b$ will usually be reduced with increasing stock levels $N_a$, but their temporal changes $dN_a/dt$ may be taken into account as well, e.g. when the stock levels are forecasted. Therefore, it is natural to assume a general dependence of the form

$$
W_b(t) = W_b(\{N_a(t)\}, \{dN_a(t)/dt\}).
$$

The function $W_b$ reflects the management strategy, i.e. the order policy regarding the desired delivery rate as a function of the actual stock levels $N_a(t)$ or anticipated stock levels $N_a(t) + \Delta t \frac{dN_a(t)}{dt} \approx N_a(t + \Delta t)$ (in first order Taylor approximation). The simplest strategy of supplier $b$ would be to react to the own stock level $N_b(t)$. However, it may be useful to consider also the stock levels of the next downstream suppliers $a > b$, as these determine the future demand, and the stock levels of the next upstream suppliers $a < b$, as they determine future deliveries or shortages of the product (“out of stock” situations). We will, therefore, assume

$$
W_b(t) = W_b(\{N_a(t)\}, \{dN_a(t)/dt\}) = W(N_{(b)}(t)),
$$

where $W$ denotes a supplier-independent management strategy and

$$
N_{(b)}(t) = \sum_{c=-n}^{n} w_c \left( N_{b+c} + \Delta t \frac{dN_{b+c}}{dt} \right)
$$

is a weighted mean value of the own stock level and the ones of the next $n$ upstream and $n$ downstream suppliers (with $2n + 1 \leq u$). For $a = b + c < 0$ and $a = b + c > u$, the weights $w_c$ are always set to zero, and they are normalized to one:

$$
\sum_{c=-n}^{n} w_c = 1.
$$
For $\Delta t = 0$, the management adapts the delivery rate $\lambda_b$ to the actual weighted stock level $N_b(t)$, while for $\Delta t > 0$, the management orients at the anticipated weighted stock level. The parameter $\Delta t$ has the meaning of the forecast time horizon, and the specification of the parameters $w_c$ and $\Delta t$ reflects the management strategy.

1.2 “Bull-Whip Effect” and Stop-and-Go Traffic

It is possible to study the stability of the steady-state solution of Eqs. (1) and (2) analytically. This stationary solution is given by

$$N_b = N_0$$

$$\lambda_b = \lambda_0 = W(N_0).$$

Small deviations from this stationary solution will fade away, if the stability condition

$$\tau < \Delta t + \frac{1}{|W'(N_0)|} \left( \frac{1}{2} + \sum_{c=-n}^{n} cw_c \right)$$

is fulfilled, where $W'$ means the derivative of the management function $W[9]$. That is, the supply chain behaves stable if the adaptation time $\tau$ is small, the forecast time horizon $\Delta t$ is large, or the change of the management function $W$ with changes in the stock levels $N_b$ are small. However, if this condition is invalid, perturbation will grow and generate oscillations in the inventories $N_b(t)$. This so-called “bull-whip effect” has, for example, been reported for beer distribution [10]. Similar dynamical effects are known for other distribution or transportation chains with significant adaptation times. Series of production processes have similar features as well. In this case, the index $b$ represents the different successive production steps or machines, $\lambda_b$ describes the corresponding production rate, and the management function $W_b$ reflects the desired production rate as a function of the stock levels $N_b$ in the respective output buffers.

Note that, in the case $N_b = N_b$ (i.e. $w_0 = 1$ and $w_c = 0$ for $c \neq 0$), the stability condition (7) agrees exactly with the one of the optimal velocity model [11], which is a particular microscopic traffic model. This car-following model assumes an acceleration equation of the form

$$\frac{dv_b(t)}{dt} = \frac{V_{opt}(d_b(t)) - v_b(t)}{\tau}$$

and the complementary equation

$$\frac{dd_b(t)}{dt} = -[v_b(t) - v_{b+1}(t)].$$

In contrast to the above supply chain model, the index $b$ represents single vehicles, $v_b(t)$ is their actual velocity of motion, $V_{opt}$ the so-called optimal (safe) velocity, which depends on the distance $d_b(t)$ to the next vehicle ahead, and $\tau$ is an adaptation time. Comparing this equation with Eq. (2), the velocities $v_b$ would correspond to the delivery rates $\lambda_b$, the optimal velocity $V_{opt}$ to the desired delivery rate $W_b$, and the inverse vehicle distance $1/d_b$ would approximately correspond to the stock level $N_b$ (apart from a proportionality factor). This shows, that the analogy between supply chain and traffic models concerns only their mathematical structure, but not their interpretation, although both relate to transport processes. Nevertheless, this mathematical relationship can give us hints, how methods, which have been successfully applied to the investigation of traffic models before, can be generalized for the study of supply networks.

Compared to traffic dynamics, supply networks and production systems have some interesting new features: Instead of a continuous space, we have discrete production units $b$, and the control function $W_b(\ldots)$ is different from the empirical velocity-density relation in traffic. While in traffic flow, the velocity-density relation is empirically given, the new feature of supply chains is that the management has a large degree of freedom how to specify
$W_b(\ldots)$, for example as a function of the own stock level and of the stock levels of other suppliers, if this information is available. With suitable strategies, the oscillations can be mitigated or even suppressed (see Sec. 1.4). Moreover, production systems may operate in different regimes, and small changes of parameters may have tremendous effects (compare Fig. 3d to 3e, and this with Figs. 3f, g). Finally, production systems are frequently supply networks with complex topologies rather than one-dimensional supply chains, i.e. they have additional features compared to (more or less) one-dimensional freeway traffic (see Sec. 1.8). They are more comparable to street networks of cities [12].

1.3 Dynamical Solution and Resonance Effects

In the vicinity of the stationary state, it is possible to calculate the dynamical solution of the one-dimensional supply chain model with $w_0 = 1$ and $w_c = 0$ for $c \neq 0$ [13]. For this, let $\delta N_b(t) = N_b(t) - N_0$ be the deviation of the inventory from the stationary one, and $\delta \lambda_b(t) = \lambda_b(t) - W(N_0)$ the deviation of the delivery rate. The linearized model equations read

$$\frac{d\delta N_b}{dt} = \delta \lambda_b(t) - \delta \lambda_{b+1}(t) \quad (10)$$

and

$$\frac{d\delta \lambda_b}{dt} = \frac{1}{\tau} \left[ W'(N_0) \left( \delta N_b + \Delta t \frac{d\delta N_b}{dt} \right) - \delta \lambda_b(t) \right]. \quad (11)$$

Deriving Eq. (11) with respect to the time $t$ and inserting Eq. (10) results, with $B = -W'(N_0) = |W'(N_0)| > 0$, in the following set of second-order differential equations:

$$\frac{d^2 \delta \lambda_b}{dt^2} + \frac{1 + B \Delta t}{\tau} \frac{d\delta \lambda_b}{dt} + \frac{B}{\tau} \delta \lambda_b(t) = \frac{B}{\tau} \left( \delta \lambda_{b+1}(t) + \Delta t \frac{d\delta \lambda_{b+1}}{dt} \right). \quad (12)$$

This corresponds to the differential equation for the damped harmonical oscillator with damping constant $\gamma$, eigenfrequency $\omega_0$, and driving term $f_0(t)$. The eigenvalues of this (system of) equation(s) are

$$\omega_{1,2} = -\gamma \pm \sqrt{\gamma^2 - \omega_0^2} = -\frac{(1 + B \Delta t) \pm \sqrt{(1 + B \Delta t)^2 - 4B^2}}{2\tau}. \quad (13)$$

The set of equations (12) can be solved successively, starting with $b = u$ and progressing to lower values of $b$. For example, assuming periodic oscillations of the form $f_u(t) = f_0^u \cos(\alpha t)$, after a transient time of about $3/\gamma$ we find

$$\delta \lambda_u(t) = f_0^u F \cos(\alpha t - \varphi) \quad (14)$$

with

$$\tan \varphi = \frac{-2\gamma \alpha}{\alpha^2 - \omega_0^2} = \frac{-(1 + B \Delta t)\alpha}{\alpha^2\tau - B} \quad (15)$$

and

$$F = \frac{1}{\sqrt{(\alpha^2 - \omega_0^2)^2 + 4\gamma^2\alpha^2}} = \frac{1}{\sqrt{(\alpha^2 - B/\tau)^2 + (1 + B \Delta t)^2\alpha^2/\tau^2}}, \quad (16)$$

where the dependence on the frequency $\omega_0$ is important to understand the resonance effect mentioned in the next section. Equations (12) and (14) imply

$$f_{u-1}(t) = \frac{B}{\tau} \left( \delta \lambda_u(t) + \Delta t \frac{d\delta \lambda_u}{dt} \right) = f_0^{u-1} \cos(\alpha t - \delta_{u-1}) \quad (17)$$
with
\[
\tan \delta_{u-1} = -\alpha \Delta t \quad \text{and} \quad f_{u-1}^0 = \frac{B}{\tau} f_0^0 F \sqrt{1 + (\alpha \Delta t)^2}.
\]

The oscillation amplitude increases, if \( f_{u-1}^0 / f_u^0 > 1 \). One can show that this can happen for \( B > 1/(2\tau) \), which corresponds to Eq. (7) for \( w_0 = 1 \) [13]. That is, supply chains behave unstable, if the adaptation time \( \tau \) is too large or if the management reacts too strong to changes in the stock level (corresponding to a large value of \( B = |W'(N_0)| \)).

### 1.4 Discussion of Some Control Strategies

According to the stability condition (7), a supply chain can be stabilized (i.e. oscillations in the delivery rates and stock levels can be reduced) by several strategies: (1) by reduction of the adaptation time \( \tau \), (2) by anticipation of the temporal evolution of the inventories \( (\Delta t > 0) \), (3) by taking into account the inventories \( N_a \) of other suppliers \( a = b + c \) with \( w_c > 0 \) for \( c > 0 \), and (4) by modification of the functional form of the management function \( W(...) \). In the case of perturbations in the consumption rate, numerical simulation results are as follows [9]:

- Anticipation of the own future inventory is an efficient means to stabilize the production system. Even anticipation time horizons \( \Delta t \) considerably smaller than the adaptation time \( \tau \) are sufficient to reach complete stability.
- The adaptation to a variation in the consumption rate tends to be better, if not only the own inventory, but also the inventories of downstream suppliers or the consumer sector itself are taken into account by so-called “pull strategies”. In contrast, considering the inventories of upstream suppliers corresponding to “push strategies” tends to destabilize the system (cf. Ref. [14]). It is the direction of the information flow in the system which is responsible for this: The oscillations in the consumption rate travel upstream, as in stop-and-go traffic [1].
- Although the linear stability analysis gives a good idea under which conditions the oscillation amplitude in the system becomes zero, further implications are limited, because non-linear effects dominate when the evolving oscillation amplitudes become large. For example, the emerging oscillation frequency \( \omega \) in the system does often neither correspond to the frequency \( \omega \) of the external perturbation nor to the frequency, which is most unstable according to the linear stability analysis. Instead, it is often much smaller than expected [6] and must be determined by simulations [9]. When in the weighted stock level \( N(b)(t) = w_0 N_b(t) + (1 - w_0) N_a(t) \), the weight \( (1 - w_0) \) of another inventory \( N_a(t) \) is increased in the management strategy, there is a surprise: As expected, the oscillation amplitudes are significantly reduced, when the second next downstream supplier is taken into account with \( N_a(t) = N_b+2(t) \) instead of the next downstream one with \( N_a(t) = N_b+1(t) \). However, considering the variation in the consumption rate itself with \( N_a(t) = N_{a+1}(t) \) has a very weak stabilization effect, although the consumer sector is located even further downstream [9]. This point is related with resonance effects: According to Sec. 1.3, there is a frequency dependence of the emerging oscillation amplitudes.

In conclusion, there are non-trivial and unexpected effects in the behavior of one-dimensional supply chains. Therefore, simulation models describing the non-linear interactions and dynamics of supply chains and production processes are relevant for their optimization. From the practical point of view it is, for example, useful that Eq. (7) allows one to estimate the maximum adaptation time \( \tau \) or the minimum forecast time horizon \( \Delta t \) supporting a stable supply chain. Moreover, the stabilizing effect of a reaction to inventories of downstream
suppliers suggests to exchange these data on-line. Note that our conclusions regarding the stabilization by forecasts and the consideration of downstream stock levels are expected to be transferable to more complex systems than the one-dimensional supply chains treated here. They should be also applicable to cases where suppliers are characterized by different parameters, to situations with limited buffers and transport capacities [15], or to supply networks [13]. Some of these aspects will be included in the more general model of production networks discussed in the following.

1.5 Production Units in Terms of Queueing Theoretical Quantities

We will investigate a system with $u$ production units (machines or factories) $b \in \{1, 2, \ldots, u\}$ producing or using $p$ different products $i, j \in \{1, 2, \ldots, p\}$. The respective production process is characterized by parameters $c^j_b$ and $p^i_b$; in each production step, production unit $b$ requires $c^j_b$ products (“educts”) $j \in \{1, \ldots, p\}$ and produces $p^i_b$ products $i \in \{1, \ldots, p\}$. The number of production steps of production unit $b$ per unit time is a measure of the throughput and shall be represented by $Q^\text{out}_b(t)$. It can be related to the variables used in queueing theory [16]: Let $\lambda_b$ be the feeding (arrival) rate, $C_b$ the number of parallel channels, $\mu_b$ the overall processing (departure) rate (i.e. $C_b$ times the processing rate of a single channel),

$$\rho_b(t) = \frac{\lambda_b(t)}{\mu_b(t)}$$

the utilization, and $S_b$ the storage capacity of production unit $b$ (see Fig. 2). The inflow

$$Q^\text{in}_b(l_b) = \begin{cases} \lambda_b & \text{if } l_b < S_b, \\ 0 & \text{otherwise}. \end{cases}$$

The outflow $Q^\text{out}_b$ agrees with the processing rate $\mu_b$, if all $C_b$ channels are occupied, i.e. $l_b \geq C_b$. Otherwise, only a proportion $l_b/C_b$ of the channels is active, i.e.

$$Q^\text{out}_b(l_b) = \begin{cases} \mu_b & \text{if } l_b \geq C_b, \\ \frac{\mu_b}{C_b} l_b & \text{if } l_b < C_b. \end{cases}$$

Note that the indices $i$ and $j$ are formally running over all $p$ possible products, and $c^j_b$ or $p^i_b$ are typically nonzero only for a few products $i$ or educts $j$, which depends on the production unit $b$. Moreover, the products $i$ are normally different from the educts $j$, i.e. $c^j_b > 0$ normally implies $p^i_b = 0$ and $p^i_b > 0$ normally implies $c^j_b = 0$. 

![Fig. 2. Schematic illustration of a production unit b as a queueing system with a limited storage capacity S_b and C_b parallel production channels (after [6]). The arrival rate \( \lambda_b \), the departure rate \( \mu_b \), as well as the inflow \( Q^\text{in}_b \) and the outflow \( Q^\text{out}_b \) are indicated.](image-url)
If $P_b(l_b)$ denotes the probability of finding a queue of length $l_b$, the average inflow is given by
\[ \langle Q_b^{\text{in}} \rangle = \sum_{l_b} Q_b^{\text{in}}(l_b) P_b(l_b) = \lambda_b \left( 1 - \sum_{l_b=S_b}^{\infty} P_b(l_b) \right), \]  
(22) 
while the average outflow is given by
\[ \langle Q_b^{\text{out}} \rangle = \sum_{l_b} Q_b^{\text{out}}(l_b) P_b(l_b) = \mu_b \left( 1 - \sum_{l_b=0}^{C_b-1} \frac{C_b-l_b}{C_b} P_b(l_b) \right). \]  
(23)

Generally speaking, the fractions $p_b^\text{in}$ and $p_b^\text{out}$ are measures of inefficiency in production due to full or empty queues (or other reasons). In the stationary case, the average inflow and outflow agree with each other and determine the (average) throughput $\langle Q_b \rangle$:
\[ \langle Q_b \rangle = [1-p_b^\text{in}(\rho_b, C_b, S_b)] \lambda_b = [1-p_b^\text{out}(\rho_b, C_b, S_b)] \mu_b = \frac{\langle l_b \rangle}{\langle T_b \rangle}. \]  
(24)

As indicated, it is also given as the quotient of the average queue length $\langle l_b \rangle$ and the average waiting time $\langle T_b \rangle$ (which is known as Little’s law). Both are functions of the utilization $\rho_b$, the number $C_b$ of parallel channels, the storage capacity $S_b$ (and possibly other variables as well). For example, for a $M/M/1$ : $(S_b/FIFO)$ process (one channel with first-in-first-out serving, storage capacity $S_b$, Poisson-distributed arrival times and exponentially distributed service intervals), one finds for $\lambda_b \leq \mu_b$:
\[ \langle Q_b \rangle = \lambda_b \frac{1 - m_{S_b}^{-1}}{1 - \rho_b m_{S_b}^{-1}} \mu_b \frac{S_b}{S_b + 1}. \]  
(25)

Note that not only the expected value, but also the standard deviation of the queue length and the waiting time diverge for $\rho_b \to 1$ [16]. Therefore, efficient production is often related to a utilization $\rho_b \leq 0.7$.

### 1.6 Calculation of the Cycle Times

Apart from the productivity or throughput $Q_b$ of a production unit, production managers are highly interested in the cycle time, i.e. the time interval between the beginning of the generation of a product and its completion. The problem is similar to determining the travel times of vehicles entering a traffic jam [12]. Let $T_b$ denote the process cycle time between entering the queue of production unit $b$ and leaving it, assuming that all $c_i$ required educts $i$ for one production cycle are transported together and located at the same place in the queue. The change of the queue length $l_b$ in time is then given by the difference between the inflow and the outflow at time $t$:
\[ \frac{dl_b}{dt} = Q_b^{\text{in}}(l_b(t)) - Q_b^{\text{out}}(l_b(t)). \]  
(26)

On the other hand, the waiting educts move forward $Q_b^{\text{out}}$ steps per unit time. For this reason, the waiting time $t_b(t)$ until one of the channels is reached is given by the implicit equation
\[ l_b(t) = C_b = \int_t^{t+t_b(t)} dt' Q_b^{\text{out}}(l_b(t')) = \int_t^{t+t_b(t)} dt' Q_b^{\text{out}}(l_b(t')) - \int_{-\infty}^{t} dt' Q_b^{\text{out}}(l_b(t')), \]  
(27)
and the overall time $T_b$ required for the processing of the product corresponds to the sum of the waiting time $t_b$ and the treatment time by one of the channels:

$$T_b(t) = t_b(t) + \frac{C_b}{\mu_b(t + t_b(t))}.$$  

(28)

From Eqs. (26) and (27), one can finally derive a delay-differential equation for the waiting time under varying production conditions [6]:

$$\frac{dt_b}{dt} = \frac{Q_b^{in}(t_b(t))}{Q_b^{out}(t_b(t) + t(t))} - 1.$$  

(29)

As the production initially starts with a waiting time of $t_b(0) = 0$ (when the factory or production unit $b$ is opened), this equation can be solved numerically as a function of the outflow $Q_b^{out}(t')$. In this way, it is possible to determine the waiting time $t_b$ and process cycle time $T_b$.

### 1.7 Feeding Rates, Production Speeds, and Inventories

In the absence of capacity constraints, we just have the relation $\lambda_b = \rho_b^0 \mu_b$ for the feeding rate, where the actual utilization $\rho_b$ agrees with the desired utilization $\rho_b^0$, e.g. $\rho_b^0 = 0.7$. However, the production of a product requires the presence of all required parts (educts). Therefore, the actual feeding rate $\lambda_b$ is determined by the minimum of the desired production speed $\rho_b^0 \mu_b$ and the delivery rates $\lambda_b^j$ of the required educts $j$:

$$\lambda_b(t) = \min_j(\rho_b^0 \mu_b, \{\lambda_b^j(t)\}),$$  

(30)

In the following, we will assume that the delivery rates $\lambda_b^j$ of educts $j$ are proportional to the desired production speed $\rho_b^0 \mu_b$ and to the number $N_j$ of available educts $j$, divided by the quantity $c_j^0$ of educts needed for one production step: $\lambda_b^j(t) = V_b^j(\rho_b^0 \mu_b N_j / c_j^0)$. Due to transport constraints $V_b^j$, $V_b^j \rho_b^0 \mu_b$ is the maximum transport rate for getting educt $j$ into the production unit $b$.

As in Eq. (2), we will assume that the adaptation of the production speed $\rho_b^0 \mu_b$ to changing demand is again delayed by some adaptation time $\tau_b$:

$$\frac{d(\rho_b^0 \mu_b)}{dt} = \frac{1}{\tau_b} [W_b(N_b, \ldots) - \rho_b^0 \mu_b].$$  

(31)

In the case $\lambda_b = \rho_b^0 \mu_b$ when transport constraints do not matter, this equation exactly agrees with our previous formula (2), while deviations may occur, when required educts are not delivered at the desired rate. This makes production networks considerably more sensitive and complex than one-dimensional supply chains. It is reasonable to assume that the management function $W_b$ increases with decreasing stock levels $N_i$ of the products $i$ production unit $b$ produces, but that is saturates due to financial, spatial, or technological limitations and inefficiencies in the processing of high order flows. In the following, we will therefore use a function of the form

$$W_b(N_b, \ldots) = \max \left( A_b \frac{1 + B_b N_b}{1 + B_b N_b + D_b N_b T}, 0 \right)$$  

(32)

with $1/N_b(t) = \sum_i p_i^j / N_i(t)$ and suitably chosen parameters $A_b$, $B_b$, and $D_b$. $N_b$ is something like a weighted inventory of the produced products. Moreover, if no products are lost, the
stock level (inventory) \( N_i \) of product \( i \) changes according to the conservation equation

\[
\frac{dN_i}{dt} = \sum_b \left[ p_b^i Q_b^{in}(l_b(t)) - c_b^i Q_b^{out}(l_b(t)) \right],
\]

(33)

as \( p_b^i Q_b^{in}(l_b(t)) \) is the number of products \( i \) finished by production unit \( b \) per unit time, while \( c_b^i Q_b^{out}(l_b(t)) \) is the number of educts entering its queue per unit time.

If the storage capacity \( S_b \) is appropriately chosen and the buffer is sufficiently filled, we may assume \( C_b < b < S_b \) and \( Q_b = Q_b^{in} = Q_b^{out} = \lambda_b \). In the following, we will focus on this particular case for simplicity, but other cases can be numerically treated as well. Moreover, we will assume something like a conservation of materials or value: First of all, the quantity of product \( i \) consumed by the production units \( b \) should be generated somewhere, i.e.

\[
\sum_{b=1}^{u+1} c_b^i = \sum_{b=1}^{u+1} p_b^i.
\]

(34)

Second, the quantity of educts consumed by some production unit \( b \) corresponds to the quantity of its generated products, i.e.

\[
\sum_{i=0}^p c_b^i = \sum_{i=0}^p p_b^i.
\]

(35)

In the following, we will discuss the case \( p_b^i = 1 \), if \( b = i \), otherwise 0. This defines the \( u \) production units \( b \) through their \( p = u \) respective main products \( i \) and implies Leontief’s classical input-output model from macroeconomics [17] as stationary solution [6]. Moreover, we can define

\[
c_b^0 = 1 - \sum_{i=1}^u c_b^i \quad \text{and} \quad c_{u+1}^i = 1 - \sum_{b=1}^u c_b^i.
\]

(36)

Values \( c_b^0 > 0 \) allow us to describe the inflow of basic resources \( i = 0 \), while \( c_{u+1}^i > 0 \) allows one to describe the depletion of products by an additional consumer sector \( b = u + 1 \). The boundary conditions are completely defined by specifying \( N_0(t) \) and \( Q_{u+1}(t) \) (see below).

1.8 Impact of the Supply Network’s Topology

Assuming the case \( Q_b(t) = Q_b^{in}(t) = Q_b^{out}(t) = \lambda_b(t) \), we will now discuss simulations based on the equations specified in Sec. 1.7 for three different supply networks sketched in Figs. 3a–c, each with five levels: (a) a one-dimensional supply chain with 5 production units, (b) a “supply ladder” with 10 production units, and (c) a hierarchical supply tree with 31 production units. By introducing random variables \( \xi_b \), which were assumed to be equally distributed in the interval \([-\eta, \eta] \), we can take into account a heterogeneity \( \eta \) in the individual parameters characterizing the different production units. Here, we have chosen \( N_0(t) = N_0 = 20 \), \( N_i(0) = 20(1 + \xi_j) \), \( \tau_b = 180(1 + \xi_b) \), \( V_b^0 = V \), \( V_b^j = V(1 + \xi_j) \) for \( j > 0 \), \( \mu_b = W_b(N_b(0)) \), \( Q_{u+1}(t) = W_{u+1}(N_0) \min(1, N_i(t)/c_{u+1})[1 + 0.1 \sin(0.04t)] \) and \( A_b = 100/V \), \( B_b = 0.01 \), \( D_b = 0.02 \). For the one-dimensional supply chain, we have \( c_b^0 = 1 \), if \( i \) delivers to \( b \), otherwise \( c_b^0 = 0 \). For the supply ladder and the hierarchical supply tree, we have \( c_b^0 = 0.5 \), if an arrow points from \( i \) to \( b \) (see Figs. 3b, c), otherwise \( c_b^0 = 0 \). \( c_b^0 \) and \( c_{u+1}^i \) are defined in accordance with Eq. (36). These specifications guarantee that, for \( \eta = 0 \), i.e. if the production units are characterized by identical parameters, the dynamics of the inventories is the same for all three discussed network topologies. However, the topology...
Fig. 3. Illustration of different supply networks and their dynamics: (a) one-dimensional supply chain with five production units, (b) “supply ladder” with five levels, and (c) hierarchical supply tree. For identical parameters and strategies ($\eta = 0$), the dynamics of the inventories is the same for all three network topologies (a)–(c): For the transport capacity $V = 0.045$, the dynamics is shown in (d), while (e) corresponds to $V = 0.047$, see the light dotted lines. The drastic change from small oscillations of high frequency to huge oscillations of low frequency above a critical threshold of the transport capacity $V$ indicates a phase transition. The topology matters a lot, when the individual parameters vary. The dark solid lines correspond to a heterogeneity of $\eta = 0.25$ in the case of (e) a one-dimensional supply chain, (f) a supply ladder, and (g) a hierarchical supply network. One can conclude that heterogeneity in supply networks can considerably decrease the undesired oscillation amplitudes in the inventories. The strongest effect is found for supply ladders, which is relevant for the design of robust supply networks.

matters a lot, if we have a heterogeneity $\eta > 0$ in the model parameters (see Figs. 3e–f). Moreover, as our dynamical model of supply networks assumes non-linear interactions, small changes of the stationary inventory $N_0$ or the relaxation time $\tau_b$ can have large effects (see Figs. 3 and 6 in Ref. [6]). We may also have a transition from small oscillations of relatively high frequency to large oscillations of low frequency, when we change the transport capacity a little (cf. Figs. 3d, e).
1.9 Advantages and Extensions

As the variables in the model are operational and measurable, the model can be tested and calibrated with empirical data. The above model of supply networks is flexible and easy to generalize. For this reason, it can be adapted to various applications. Our approach can be related to microscopic considerations such as queueing theory or event-driven (Monte-Carlo) simulations of production processes, but, as it focusses on the average dynamics, it is numerically much more efficient and, therefore, suitable for on-line control. Nevertheless, the formulas can be extended by noise terms to reflect stochastic effects. Our system of coupled differential equations would then become a coupled system of stochastic differential equations (Langevin equations), where the noise amplitudes would be determined via relationships from queueing theory.

2 Many-Particle Models of Production Processes

Instead of modelling product flows by mean value equations, as above, we may also simulate manufacturing processes as driven many-particle systems, where the discrete products and transport devices play the role of particles. Their interactions can be specified in a way, which reproduces the observed arrival and departure rates. In a recent study, for example, we have been able to increase the actual throughput of a chain of production processes in a major semiconductor factory by up to 39% [18]. In this particular supply chain, silicium wafers require a series of chemical processes similar to the development of a photographic film. Between the chemical treatments, the wafers must be washed in water basins, and in the end they need to be dried (see top of Fig. 4). The treatment times can be varied within certain time intervals. If the treatment takes longer or shorter, the wafers will be of poor quality and cannot be sold. The main problem in this system is the limited transport capacity of the handler, i.e. the device which has to move the wafers around. If several sets of wafers had to be moved at the same time, the treatment times for one of these would probably exceed the critical time threshold. Therefore, the challenge is to find a schedule that resolves conflicts in desired handler usage (as one channel, namely the handler, has to serve several products in parallel). The problem to achieve coordination among several elements is similar to the coordination of pedestrians in the merging area in front of a door (see Sec. 2.1).

In the first step, we have analyzed the time requirements of the different treatment and transport processes in detail. Generally speaking, these time requirements are (more or less) deterministic, and queueing effects in the supply chain result from dynamical interactions of the different units, namely conflicts in desired handler usage. These conflicts imply a delayed service (cf. Secs. 1.1 and 1.7). Based on a variable Gantt diagram (with variable treatment times), we have resolved these conflicts and reduced the related delays by harmonization, i.e. coordination of the different treatment times. This has usually been reached by increasing the treatment times, as is indicated in the middle of Fig. 4 by longer bars in the optimized schedule (shown on the right) compared to the original schedule (on the left). Different sets of wafers are distinguished by different shades of grey, while the treatment times belonging to the same “run” (series of treatments) are represented by the same shade of grey. One can clearly see that, in the optimized schedule with longer treatment times, the waiting times between successive runs are significantly decreased, resulting in much higher throughputs (see bottom of Fig. 4). That is, instead of stop-and-go patterns (waiting and usage periods of the chemical or water basins), we have reached a more or less continuous usage pattern (cf. Secs. 1.2 and 1.4).
Fig. 4. Top: Schematic representation of the successive processes of a wet bench, i.e. a particular supply chain in semiconductor production. Middle: The Gantt diagrams illustrate the treatment times of the first four processes, where we have used the same shades of grey for processes belonging to the same run, i.e. the same set of wafers. The left diagram shows the original schedule, while the right one shows an optimized schedule based on the “slower-is-faster effect” (see Sec. 2.3). Bottom: The increase in the throughput of a wet bench by switching from the original production schedule to the optimized one was found to be 33%, in some cases even higher. (After [18].)

2.1 Learning from Pedestrians

Note that the above described optimization of some processes in semiconductor manufacturing is an example for the application of the “slower-is-faster effect”, which had been dis-
covered for panicking pedestrian crowds [19]. However, a closer investigation shows that one could learn many more strategies from pedestrian behavior to improve production processes with non-linear dynamics, as the basic features of pedestrian streams and many production systems are the same: (1) The system consists of a large number of similar entities (individuals, particles, products, boxes, ...). (2) The entities are externally or internally driven, i.e. there is some energy input, e.g., they can move. (3) The entities interact non-linearly, i.e., under certain conditions, small variations can have large effects. In other words, the system behavior is dominated by the interactions rather than the boundary conditions (the external control). (4) There is a competition for resources such as time (slots), space, energy, etc. (5) Each entity has a certain extension in space or time, or a certain demand. (6) When entities come too close to each other, frictional and obstruction effects occur. Therefore, we are trying to transfer our knowledge of traffic dynamics to the optimization of real production processes.

Pedestrian models have, for example, been successfully applied to the coordination of robots [20]. In the following, we will give a short introduction to this model, while details can be found in the available reviews [1,21,22]. The so-called social-force model of pedestrian dynamics describes the different competing motivations of pedestrians by separate force terms similar to granular flows, and it has the following advantages: (1) The social-force model takes into account the flexible usage of space (i.e. compressibility), but also the excluded volume and friction effects which play a role at extreme densities. (2) The model assumptions are simple and plausible. (3) There are only a few model parameters to calibrate. (4) The model is robust and naturally reproduces many different observations without modifications of the model. (5) Nevertheless, it is easy to consider individual differences in the dynamic behavior, and extensions for more complex problems are possible.

The basic version of the social force model assumes that the change of the location $x_\alpha(t)$ of some pedestrian $\alpha$ in the course of time $t$ is given by the actual velocity $v_\alpha(t)$, i.e. $dx_\alpha/dt = v_\alpha(t)$. Moreover, the acceleration $dv_\alpha/dt$ is specified by a sum of “social forces”, e.g. the driving force $(v_0^\alpha e_\alpha - v_\alpha(t))/\tau_\alpha$, which describes the adaptation of the actual velocity $v_\alpha$ to the desired velocity $v_0^\alpha$ and the desired walking direction $e_\alpha(t)$ within a certain acceleration time $\tau_\alpha$, the repulsive forces $f_{\alpha\beta}(x_\alpha, v_\alpha, x_\beta, v_\beta)$ with respect to other pedestrians $\beta$, the repulsive forces $f_{\alpha k}(x_\alpha, v_\alpha, t)$ with respect to obstacles $k$, and fluctuation forces $\xi_\alpha(t)$ reflecting individual variations in behavior:

$$
\frac{dv_\alpha}{dt} = \frac{v_0^\alpha e_\alpha - v_\alpha(t)}{\tau_\alpha} + \sum_{\beta(\neq \alpha)} f_{\alpha\beta}(t) + \sum_k f_{\alpha k}(t) + \xi_\alpha(t) .
$$

For reasonable specifications of the interaction forces see Refs. [1,19,23,24].

This model describes many self-organization phenomena in pedestrian crowds in a natural and realistic way. For example, for “relaxed” pedestrians in normal situations with small fluctuation amplitudes, our microsimulations of pedestrian counterflows in corridors reproduce the empirically observed segregation of opposite flow directions into lanes [1,23–25], see Fig. 5a. However, for large fluctuation amplitudes corresponding to “nervous” pedestrians, we find a “freezing by heating effect” characterized by a breakdown of “fluid” lanes and the emergence of “solid” blockages [23], see Fig. 5b. The same model also reproduces the observed oscillations of the flow direction at bottlenecks [1,24], see Fig. 5c. Cellular automatron Java applets visualizing these phenomena are available in the internet (see www.helbing.org/Pedestrians/Corridor.html, /Door.html). These findings are obviously relevant for production processes involving granular flows or requiring the segregation of different kinds of particles or objects.

Due to mutual interactions and environmental impacts, pedestrians suffer delays and experience different travel times, even if their desired velocities $v_0^\alpha$ are the same. In order to
Fig. 5. (a) Segregation of opposite flow directions into lanes for the case of small noise amplitudes (after [1,21,23]; cf. also [24,25]). White disks represent entities (pedestrians or objects) moving from left to right, black ones move the other way round. (b) For sufficiently high densities and large fluctuations, we observe the noise-induced formation of a crystallized, “frozen” state (after [1,23]). (c) Bottlenecks are passed by clusters of entities in alternating directions, giving rise to oscillatory flows. The phenomena (a) to (c) are relevant for heterogeneous object flows, segregation processes, the coordination of robots, and production processes with competing goals.

compensate for this and arrive in time, pedestrians adapt their desired velocities. If $l_\alpha$ is the total length of the way and $s_\alpha(t) = \int_0^t dt' \, v_\alpha(t') \cdot e_\alpha(t')$ the portion of the way they have completed until time $t$, a reasonable adaptation strategy would be

$$v_\alpha^0(t) = \min \left( \frac{l_\alpha - s_\alpha(t)}{T_\alpha - t}, v_{\alpha}^{\text{max}} \right),$$

where $T_\alpha$ is the required arrival time and $v_{\alpha}^{\text{max}}$ the maximum speed. This strategy will eventually lead to different actual speeds, which will also segregate into lanes: fast lanes and slow lanes. A similar strategy may be used to compensate for delays in production processes.

2.2 Optimal Self-Organization and Noise-Induced Ordering

Lane formation (segregation) is an optimal self-organization phenomenon [25] resulting from the combined action of driving and repulsive forces: Under certain conditions, one can show that lane formation maximizes the average speed, and the resulting pattern is “fair” for both directions of motion [25]. That is, the efficiency in the overall system is optimized based on local interactions. For this reason, one can speak of a distributed control mechanism or even distributed intelligence. Such mechanisms are usually cost-effective and robust with respect to local failures, which is interesting for applications.

When the amplitude of the fluctuations $\xi_\alpha(t)$ is increased, we expect a reduced degree of order in the system. For large noise amplitudes, we find indeed a breakdown of self-organized patterns such as lanes. In this case, we face a disordered system with homogeneous distributions in space. However, for medium noise amplitudes, one can observe an increase in the level or order. For example, instead of many narrow lanes, one can find a few wide ones (see Fig. 6). This phenomenon is called noise-induced ordering [26,27]. We have learned that a careful choice of the noise strength can speed up the time-dependent increase of the order very much. Moreover, after a given, large enough time period, the system has reached a typical level of order, which depends significantly on the fluctuation strength. In conclusion, a variation of the “applied” fluctuation strength together with a proper choice
of the “treatment times” would allow one to control pattern formation in several respects: (1) the speed of ordering, (2) the typical length scale in the system, and (3) the level of ordering. A time-dependent variation of the control parameters should even facilitate to switch between the support and suppression of structure formation, e.g. between demixing and homogenization. These points are, for example, relevant for the production, properties, handling, and transport of heterogeneous materials, for flow control, and efficient separation techniques for different kinds of particles or objects.

2.3 “Slower-is-Faster Effect” in Merging Flows

“Freezing by heating” is one of the phenomena observed in panic stampedes. Another one is the “faster-is-slower effect” or “slower-is-faster effect” [19]. It is caused by arching and clogging at bottlenecks like exits, which implies irregular outflows (see Fig. 7a). The reason are frictional interactions, when the entities touch each other. This happens when the driving forces exceed a certain critical threshold and the diameter of the bottleneck is small. Below the critical threshold, outflows are regular and efficient (see Fig. 7b), i.e. entities arrange perfectly among each other in the merging area in front of a bottleneck. This mutual coordination is thanks to the non-linear repulsive interactions. It neither requires communication nor fluctuations. For simulations see http://angel.elte.hu/~panic/ or http://www.panic.org. The relevance for production processes involving merging flows (packing processes) or granular flows through hoppers (filling processes) is obvious.

2.4 Optimization of Multi-Object Flows

Due to the non-linear dynamics, driven multi-object flows decisively depend on the geometry of the boundaries. This calls for innovative solutions, which utilize the self-organization in non-linearly interacting many-particle systems. We will exemplify this for the improvement of standard elements of pedestrian facilities [22] (see Fig. 8):

1. At high pedestrian densities, the lanes of uniform walking direction tend to disturb each other, as pedestrians expand into areas of low density and try to overtake each other. This often leads to mutual obstructions of the opposite walking directions. The lanes can be stabilized by series of small obstacles in the middle of the walkway, see Fig. 8 (left) which, in the direction of motion, has a similar effect as a wall.

2. The flow at bottlenecks can be improved by a funnel-shaped construction (see center of Fig. 8). Interestingly, the optimal funnel shape resulting from an evolutionary optimization is convex [22].
Fig. 7. (a) When the desired velocities $v_0^0$ are too high (e.g. in panic situations), pedestrians come so close to each other, that their physical contacts cause the build up of pressure and obstructing friction effects, which results in temporary arching and clogging. (b) This is related with an irregular and reduced outflow, while the outflow is regular for small enough desired velocities ($v_0^0 \leq 1.5 \text{ m/s}$) [1,19,21]. It is interesting that suitable obstacles can improve the outflow by reducing the pressure (see http://angel.elte.hu/~panic/ for online Java simulations.) This has also implications for filling and packing processes with merging flows.

3. Oscillatory changes of the walking direction and periods of standstill in between do not only occur for counterflows at doors, but also when different flows cross each other. The loss of efficiency caused by this can be reduced by railings inducing roundabout traffic, see Fig. 8 (right). Roundabout traffic can already be triggered and stabilized by an obstacle in the middle of a crossing, because it suppresses the phases of “vertical” or “horizontal” motion in the intersection area. In our simulations, this increased efficiency up to 13%.

It is natural to use similar solutions for the improvement of multi-object flows in production processes as well.

Fig. 8. Improved standard elements of pedestrians facilities: corridors (left), bottlenecks (middle), and intersections (right). Empty circles represent obstacles such as columns or trees, while full circles with arrows symbolize pedestrians and their walking directions (after [22]). Similar solutions may be also applied to other multi-object flows, e.g. filling processes.
3 Summary and Outlook

In this contribution, we have shown that various concepts from traffic theory can be transferred to production processes with several implications for their optimization. We just mention optimal self-organization (regarding the segregation or merging of multi-object flows), or noise-induced ordering. Moreover, the “slower-is-faster effect” has already been successfully applied to semiconductor production. The reason for the similarities of traffic and production systems is the presence of moving entities (persons or objects), which interact in a non-linear way with obstructive and frictional effects. Therefore, a competition for limited resources (such as capacities, time, or space) takes place. The question is, how to distribute them in an efficient and fair way. Answers have, for example, been developed in the areas of game theory and of intelligent transportation systems. The theory of complex driven many-particle systems has made significant contributions to this.

We have also sketched a novel theory of supply networks. This theory is developed to help understand the dynamical phenomena, breakdowns, instabilities and inefficiencies of production and supply networks. It is well suited for an efficient simulation and on-line control of production systems, and effects of fluctuations can be studied as well. Oscillations in the inventories (“the bull-whip effect”), resulting from non-linear interactions of production units, may occur on considerably slower time scales than the variation of the consumption rate, which may explain the existence of business cycles. Moreover, we have discovered that the supply network’s topology has a significant impact on the resulting production dynamics, and that a heterogeneity in the parameters characterizing the different production units can stabilize the production considerably. Production can also be stabilized by forecasting temporal changes in the inventories, by reduction of the time required to adapt the production speed, by consideration of the stock levels of downstream suppliers, and by modification of the management function.

Acknowledgments. The author wants to thank for the partial support by SCA Packaging Ltd. and for invitations by Prof. Radons and Prof. West. He acknowledges inspiring discussions with Dieter Armbruster, Carlos Daganzo, Illés Farkas, Arne Kesting, Megan Khoshyaran, Christian Kühnert, Péter Molnár, Takashi Nagatani, Tadeusz Płatkowski, Dick Sanders, Petr Šeba, Thomas Seidel, Tamás Vicsek, Torsten Werner, and Ulrich Witt. The author is also grateful to Dominique Fasold and Tilo Grigat for preparing the schematic illustrations. Last, but not least, he has appreciated the warm hospitality and the challenging intellectual atmosphere at the Santa Fe Institute, where this paper has been contributed for discussion by the Social Scaling Working Group in August 2003.

References

18  Dirk Helbing