

A Mesoscopic Approach to the Simulation of Semiconductor Supply Chains.

Dan Marthaler, Dieter Armbruster, Christian Ringhofer,
Department of Mathematics, Arizona State University,
Tempe, AZ, 85287-1804, USA

September 15, 2002

Abstract

Production flows through factories are modeled through conservation laws leading to nonlinear hyperbolic PDEs. For a linear production line, models based on conservation laws can be derived from first principles, using methods from gas dynamics. For re-entrant manufacturing, a heuristic model is presented merging a nonlocal state equation relating TPT to WIP through Little's law to produce a nonlinear nonlocal hyperbolic PDE. These two models can serve as the building blocks of fast simulations of the dynamics of capacity limited supply chains. We present simulations for a chain consisting of a re-entrant module, followed and preceded by a linear module. The response of the system to various production scenarios is discussed.

1 Introduction

Simulations of capacitated factories are extremely important to generate a testbed that would allow the modeling, analysis and control of large scale supply chains. While discrete event simulators have been highly successful in simulating complicated individual semiconductor factories at the tool level (WWK, 1996), they are much too computationally expensive to integrate them into even moderately complicated supply chain topologies. Static models that assign a capacity and a throughput time to a factory can be analyzed much faster but at the expense of accuracy. In particular, they are not able to model the nonlinear response of a capacitated factory.

2 Models

Real factories show a strong increase in the average throughput time τ as the loading of the factory is increased. Unfortunately, large factories are rarely run in equilibrium for any amount of time and are too costly to be run as a controlled

experiment. Hence the specific nature of the nonlinearity of the throughput time is unclear. We have developed two models that can serve as extreme case models for real factories. Most likely any real factory will behave in some intermediate manner. The two models roughly correspond to a linear factory where every production step has its own machine and a re-entrant factory where product has repeated passes through the same machine, respectively. We call the former a queuing model and the latter a re-entrant model.

Both models are based on the fundamental fact of conservation of jobs: whatever enters the factory has to come out of the factory at some time (we neglect the yield issue for the moment). We define x to be a *completion* variable with $x = 0$ denoting the start of a product into the factory and $x = 1$ the release of a finished product. Writing $\rho(x, t)$ for the density of work at stage x at time t we get the total load (WIP) as a function of time

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

The density then satisfies a hyperbolic conservation law of the form

$$\rho_t + (v(\rho)\rho)_x = 0 \tag{1}$$

where $v(\rho)$ describes the velocity of product moving in the factory. The exact nature of the transport velocity $v(\rho)$ is the major modeling issue:

1. For a re-entrant factory we assume that $v(\rho)$ is described by a state equation of the form

$$v(\rho) = v_0 \left(1 - \frac{L}{L_{max}} \right). \tag{2}$$

Here v_0 is the speed for the empty factory and L_{max} is the maximal load. Equation 2 implies that the velocity is uniform in the whole factory and that, due to the re-entrant nature of the flow, the total WIP determines this velocity. Notice that v is time dependent through $L(t)$. The start rate $\lambda(t)$ into the factory enters as the boundary condition for the flux at $x = 0$:

$$\lambda = \rho(0, t)v(t) \tag{3}$$

It is easy to see that for any equilibrium solution ρ_{eq} as well as for long term averages (if they exist) equations (1-3) satisfy Little's law: Assume $\rho = \rho_{eq}$. Then $\rho(0, t) = \rho_{eq}$ and with $v = v_{eq} = \frac{1}{\tau}$, equation (3) becomes Little's law. Clearly the state equation can be chosen in a more sophisticated way to model a given exact topology of the re-entrant factory, if so desired. The full model ((1-3) in essence describes the dynamics of the factory flow as if it were always in equilibrium, following adiabatically the state equation (2).

2. In contrast to the re-entrant model, a more sophisticated model for the queuing model can be derived (Ringhofer, 2002). It allows for non-adiabatic relaxation of the velocity fields: Consider a job arriving at a queue with processing rate μ . Its throughput time depends on the number of jobs waiting before it

$$\tau = \frac{1}{\mu}(1 + L). \quad (4)$$

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

$$\rho_t + (u\rho)_x = 0 \quad (5)$$

$$u_t + uu_x = 0 \quad (6)$$

$$u(0, t) = \frac{\mu}{1 + L(t)} \quad (7)$$

$$u(0, t)\rho(0, t) = \lambda(t) \quad (8)$$

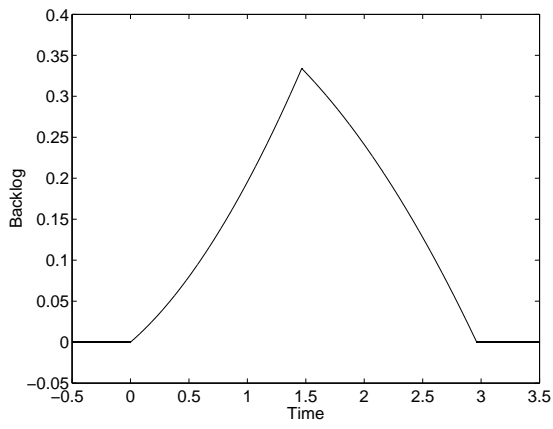
The important improvement in both models over discrete event simulators is the modeling of the flow as a transport equation for a continuous density variable. Solving these equations is independent of the number of parts moving through the factories as well as independent of the number of machines or production steps. In fact, the continuous model will improve as a model for flows with a large number of products, many steps and many machines. Similar ideas have been suggested for traffic flow [6].

3 Simulations

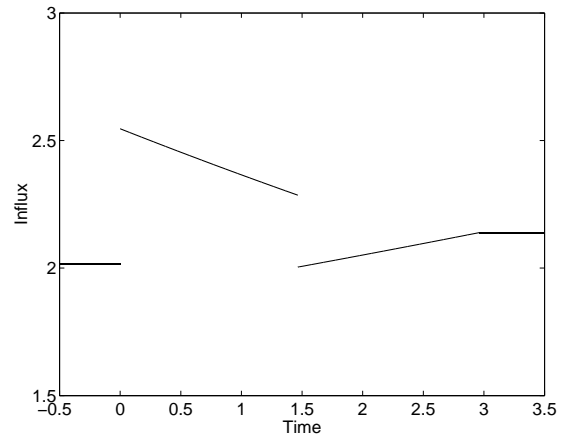
For simulations these two models have major advantages: they have Little's law built in, they can be improved to include policies and/or the topology of a factory and most importantly, they can be simulated very fast using typical codes developed for hydrodynamic flows [4], [3]. Movies of several experiments for re-entrant factories as well as for a three node supply chain can be found on the website http://math.duke.edu/~daniel/supply_chain.html

In addition, time dependent start rates to achieve outputs that are optimal relative to some chosen criteria can be derived. For instance, Figure 1(b) shows the start rate $\lambda(t)$, Figure 1(c), the WIP $L(t)$, and Figure 1(a) the backlog as a function of time for the following experiment for a re-entrant factory: We assume the factory is in equilibrium with a density of $u(x) = 2.8$ for $t < 0$. We assume that the demand changes to a value corresponding to a higher equilibrium density of $u(x) = 3.1$. Since the outflux of equations (1-3) is defined as

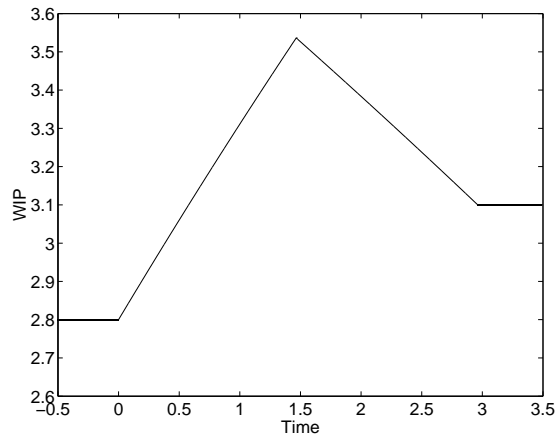
$$O(t) = \rho(1, t)v(t) \quad (9)$$



(a)



(b)



(c)

Figure 1: a) Backlog, b) influx, and c) load as a function of time for the demand jump of $O_1 = 2.01$ to $O_2 = 2.14$.

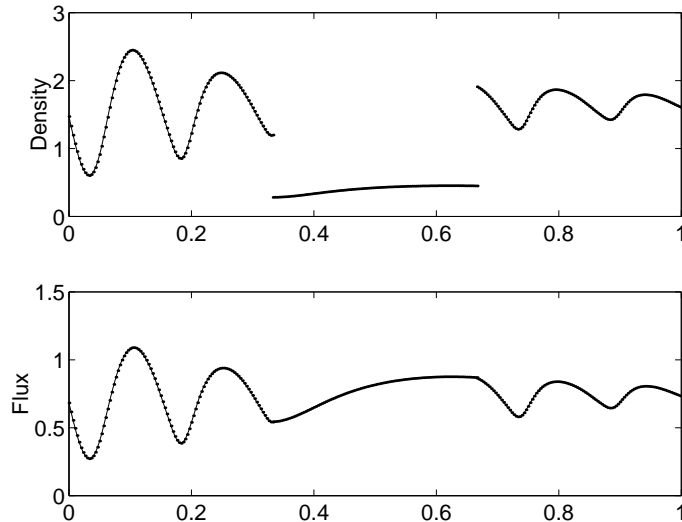


Figure 2: Snapshot of the density and local flux for a supply chain consisting of two identical queue modules surrounding a re-entrant module.

we obtain, with a maximal load of $L_{max} = 10$, an outflux of $O_1 = 2.01$ for $t < 0$. The new demand for $t > 0$ is $O_2 = 2.14$. However, any changes in the starts will show up in the output only after a delay and hence a backlog will initially accumulate. We choose to design a start rate $\lambda(t)$ such that the resulting backlog becomes zero in shortest possible time. Since we are designing an input that makes the WIP density piecewise constant, the start rate $\lambda(t)$ is discontinuous (see Figure 1b)). In addition, we are looking for a start rate that keeps WIP profiles as homogeneous as possible. Hence we choose to make up the backlog by running the factory with a constant density that is higher than necessary to satisfy the immediate demand. Optimality is determined with respect to the level of the intermediate density and the time that this density should be maintained. We chose a relatively small demand jump just to get reasonable figure sizes. Larger demand jumps can easily be accommodated in the same way but they lead to much longer control times. Details of this optimal control scheme are discussed in [1].

These continuum models for production flows open up the possibilities to generate highly efficient simulations of large networks of supply chains. As a proof of concept we have linked our models to form a 3 node linear supply chain. We simulate a queuing factory feeding into a re-entrant factory feeding into a queuing factory. All factories are flux coupled, i.e. the outflux of an upstream factory directly goes into the next downstream factory. There are no buffers between the factories. As a result, flux will be continuous along the supply chain but, due to varying velocities at the boundaries between factories, the WIP densities will have discontinuities.

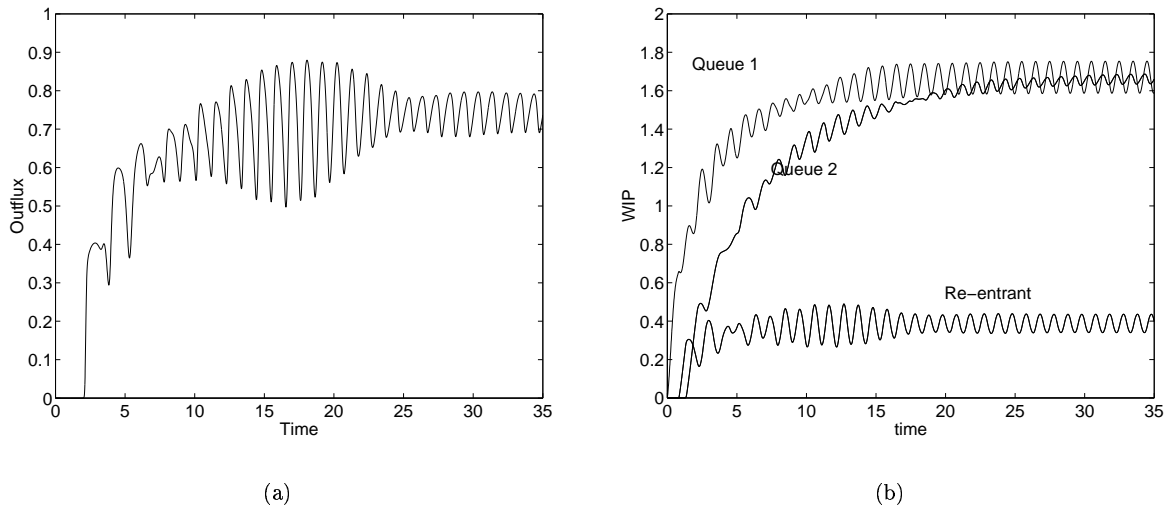
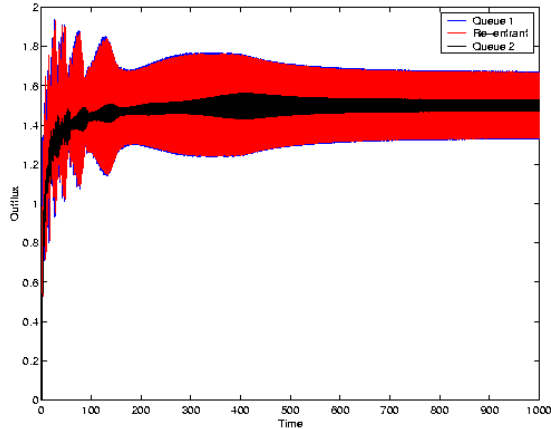
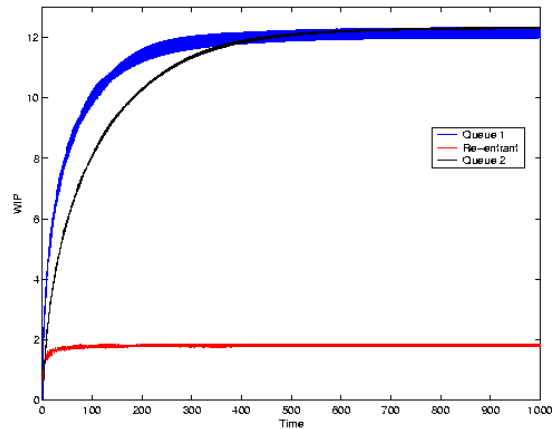


Figure 3: a) Outflux of the last factory and b) WIP for all factories for the queue→re-entrant→queue system as a function of time. The experiment corresponds to weakly overloaded linear factories and a high capacity re-entrant factory.

It is straightforward to link the two models together to form a supply chain: Experiments 6 and 7 on our webpage show transient and steady state movies for such a supply chain. Figure 2 shows a snapshot of a steady state density and flux for the chain of three factories. Each factory module makes up one-third of the total x -axis. The two queuing models are identical and have a lower velocity than the re-entrant factory. The periodic influx for this experiment is given as $\lambda(t) = 0.75 + 0.5 \sin(2\pi t)$. The queuing modules were constructed with processing rates of $\mu = 1.2$, and the raw velocity for the re-entrant module was $v_0 = 2.0$. The capacity for the re-entrant module was $L_{max} = 10$. With $\mu = 1.2$ the influx exceeds the equilibrium threshold for a short part of a period. The average influx stays below threshold and hence the supply chain equilibrates. Figure 2 represents a snapshot at $t = 32$ when the whole supply chain has settled into a steady state pattern. Figure 3(b) shows the outflux of this experiment. It clearly shows transient oscillations that are much larger than the eventual oscillations in equilibrium. This is a reflection of the fact that initially the queues are short and hence much more responsive to the influx oscillations. These simulations can be performed very fast which will allow us to simulate more realistic supply chains. We notice that the queuing factories have extremely long transients (on the order of 30 throughput times) before they get close to the equilibrium, whereas the re-entrant factory equilibrates much faster. This corresponds to the long relaxation times for queues near their capacity. In addition, the re-entrant factory acts as a damping device reducing the WIP amplitudes in



(a)



(b)

Figure 4: a) The outflux and b) WIP for all factories of the supply chain as a function of time. The experiment reflects strongly overloaded linear factories and a weakly overloaded re-entrant factory.

the downstream factory. Figure 3(a) shows WIP as a function of time for each module. Notice that the WIP is initially zero for the re-entrant module and for the second queue. This is due to the time delay and the initialization (the modules started empty). Also, we see that the amplitude for the oscillations in the second queue is less than that of the first queue; clearly demonstrating the damping effect of the re-entrant module.

With $v_0 = 1$, $L = 10$, $\mu = 1.6$ we have a critical influx for the re-entrant factory of $\lambda_c = 2.5$. Hence a periodic influx of $\lambda = 1.5 + 1.25 \sin 2\pi t$ overloads the re-entrant factory slightly and the linear factories strongly. Figure 4 shows outflux and backlog for this experiment. We notice that the outflux of the first queue and the re-entrant factory closely track each other whereas the outflux of the last queue has a significantly smaller oscillation amplitude. Again, because of the global model of the re-entrant flow, its WIP equilibrates much faster than the queueing models.

4 Conclusion

We have shown that modeling factory production via hyperbolic conservation laws can lead to a very fast and qualitatively correct simulation tool.

We can extend our approach to model multiple products and to model the influence of dispatch policies. For instance, to model production of two products,

A and B, we introduce two density variables ρ_A and ρ_B together with their state equations

$$\begin{aligned} v_A &= v_0^A(1-L) \\ v_B &= v_0^B(1-L) \\ L &= \int_0^1 \frac{u^A}{L_{max}^A} + \frac{u^B}{L_{max}^B} ds \end{aligned} \quad (10)$$

This implies that the two products A and B have their own independent production time $1/v_0^A$ or $1/v_0^B$. They will also contribute differently to the slowing down due to overcrowding ($L_{max}^A \neq L_{max}^B$) but they will slow down together based on their percentage of the total capacity of the factory. Figure 5 shows the outflux for a switching experiment: For $v_0^A = 1$, $v_0^B = 2$, $L_{max}^A = 10$, $L_{max}^B = 20$ we change the total influx of $\lambda = 3$ from $1/3$ A and $2/3$ B to $1/3$ B and $2/3$ A. Experiment 11 on the webpage shows the time evolution of the densities as a movie. The simulation confirms the intuition: Since B is moving through the factory faster and since it is less influenced by the overall load, its transient is shorter than the one for A and hence the production for B relaxes to its new equilibrium value much faster than A. Also, since A is moving slower it leads to a higher WIP for the same throughput as for B. Hence the transition leads to a buildup in WIP which is reflected in the missing outflux in Figure 5.

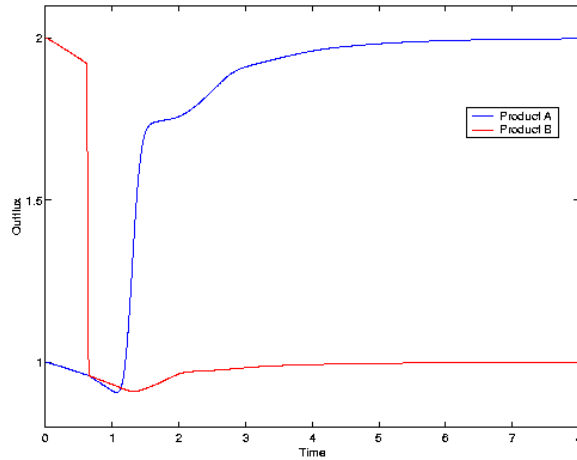


Figure 5: Outflux for a two product switch $1/3A, 2/3B$ for $t < 0$ to $2/3A, 1/3B$ for $t > 0$.

Dispatch rules and other policies are used to influence the behavior of the factory in certain critical situations. For instance, a pull policy might be used to generate a fast reaction to an increase in demand, while a push policy might be used to hedge against an upturn in demand by generating a lot of product in a semi-finished state. Such policies can be modeled by using integration kernels

$w(x, s)$ that indicate the importance of a product at location s in completion space on the speed of a product at location x :

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

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. A detailed discussion of pull vs. push policies will be published elsewhere.

These modules may also serve as mesoscale simulation models for individual factories. One may examine the topology of a factory, partitioning it corresponding to re-entrant or queuing type sections using similar weight functions as for the push or pull policies. In these sections, average statistics would be used for the parameter values of maximum load and processing time. With these parameters, accurate models representing real re-entrant factories may be inserted into supply chain models and their simulation conducted rapidly.

Acknowledgments: Support from Intel Corporation and from NSF grant DMS 0075041 is gratefully acknowledged. The authors would especially like to thank Karl Kempf for his many insightful discussions in this work.

References

- [1] Dieter Armbruster, Dan Marthaler, Christian Ringhofer, Modeling a re-entrant factory, preprint Arizona State University 9/2002
- [2] John D.C. Little, A proof for the queuing formula $L = \lambda W$, Operations Research, **9**, 383-387, 1961
- [3] R.J. LeVeque, *Finite Difference Methods for Differential Equations*. Draft Version for use in AMath 585-6 University of Washington, (1998).
- [4] R.J. LeVeque, *Numerical Methods for Conservation laws* (Birkhäuser-Verlag, 1992).
- [5] e.g. Factory Explorer, WWK products, (1996)

- [6] D. Helbing, Traffic modeling by means of physical concepts, in: *Workshop on Traffic and Granular Flow*, D.E. Wolf, M. Schreckenberg and A Bachem, eds. World Scientific, Singapore (1996)
- [7] C. Ringhofer, D. Armbruster: A kinetic model for linear supply chains, in preparation, 2002, Arizona State University.