Applied Mathematical Modelling 39 (2015) 4655–4664

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

Feedback control for priority rules in re-entrant semiconductor manufacturing

Fenglan He [a], Dieter Armbruster [b,*], Michael Herty [c], Ming Dong [d]

[a] Innovative Ocean and Energy System Group, School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, PR China
[b] School of Mathematical and Statistical Sciences, Arizona State University, Tempe, USA
[c] RWTH Aachen University, Aachen, Germany
[d] Antai College of Economics & Management, Shanghai Jiao Tong University, PR China

ARTICLE INFO

Article history:
Received 10 September 2014
Received in revised form 6 March 2015
Accepted 20 March 2015
Available online 11 May 2015

Keywords:
Re-entrant production
Feedback control
Partial differential equations
Dispatch policy control

ABSTRACT

Typical semiconductor production is re-entrant and hence requires priority decisions when parts compete for production capacity at the same machine. A standard way to run such a factory is to start to plan and to finish according to demand. Often this results in a push policy where early production steps have priority over later production steps at the beginning of the production line and a pull policy where later steps have priority at the end of the production line. The point where the policies switch is called the push–pull-point (PPP). We develop a control scheme based on moving the PPP in a continuum model of the production flow. We show that this control scheme significantly reduces the mismatch between demand and production output. The success of the control scheme as a function of the frequency of control action is analyzed and optimal times between control actions are determined.

1. Introduction

Semiconductor manufacturing is arguable one of the most advanced industrial production processes. Due to its long cycle times (order of several weeks) and its re-entrant flow topology it poses many challenges for production planning and control. Especially for commodity chip production when only one type of chip is produced and the product mix cannot be altered, there are not a lot of ways to influence the production flow beyond the starts policy.

However, re-entrant production where a machine is used to serve several steps in the production schedule of a chip, allows the use of a dispatch policy to set priorities for all the steps that need to be performed at that machine. A common dispatch policy in this case is the push policy, also known as first buffer first served policy, which gives priority to earlier production steps over later production steps. A pull policy, also known as shortest-expected-remaining-process-time policy, gives priority to later production steps over earlier production steps. Push policies are usually used at the front of a production line, whereas pull policies are used at the back of the factory. The step where push policy switches to pull policy is called the push–pull point (PPP).

In a discrete simulation (DES) model of a semiconductor factory Perdaen et al. [1] study the response of a production line to demand fluctuations that occur on a much shorter timescale than the cycle time of the factory. They show that the

* Corresponding author.
E-mail addresses: hesnow1021@163.com (F. He), armbruster@asu.edu (D. Armbruster), herty@igpm.rwth-aachen.de (M. Herty), mdong@sjtu.edu.cn (M. Dong).

http://dx.doi.org/10.1016/j.apm.2015.03.061

mismatch between a fluctuating demand and the factory production can be reduced dramatically relative to many other static production scenarios when they couple a heuristics for moving the PPP with a CONWIP starts policy [2].

While DES can be made extremely accurate and their use in practice is widespread, they are extremely time consuming, both to run and to maintain. In addition they have the usual issues of simulation studies: It is difficult to extrapolate their generality from the specific parameters that are used and insights into the simulation structure and its dependence on parameters is hard to get. Therefore, in recent years aggregate production models have been developed. Initially they were based on a continuum of product leading to queuing network models [3] which led to the concept of the clearing function, relating output in a period to workload during the period (see [4] for a state of the art review). More recently models based on transport equations considering production flow as a fluid flow in a continuum in product as well as in production steps have been discussed. The latter approach leads to partial differential equations (PDEs) describing the flow of the product density and the associated flow velocity [5].

In [6] a PDE model incorporating a PPP based dispatch policy was presented. The current paper improves this model and develops a linearized control algorithm. The current model is a feedback control model, tracking the outflux to a given known demand sequence. We will analyze the PDE model with this linearized control and compare it to fixed PPP schemes and heuristics similar to the one used in [1] that are based on giving high priority to the lots needed to satisfy the demand in the next time unit.

The general topic of production planning has a long history and the literature is large and diverse. For the more specific topic of production planning in semiconductor factories there is a comprehensive recent monograph by Mönch et al. [7] which presents an overview over real world approaches to dispatching rules, order releases and production planning at different scales from short term order releases to enterprise wide planning. A complete and exhaustive definition of production planning in its full generality is discussed by Kempf et al. [8]. Armbruster and Uzsoy's review [9] is closest to our approach focussing on the interplay between discrete event simulations and aggregate models based on clearing functions and partial differential equations. Mathematical approaches to control at the aggregate level are discussed in [10,11]. All the previous approaches dealing with aggregate models were typically controlling the order release or starts policies. To our knowledge, none considered integrating the dispatch policies into an aggregate model. An optimization approach, corresponding to an open loop control problem for production planning in aggregate models integrating release and dispatch policies has been discussed in [12].

2. PDE model

As discussed in [5] and recently summarized in [9] we consider a work-in-progress (wip) density $\rho(x,t)$ where x describes the continuum of production stages for a product. Hence $\rho(0,t)$ is the wip density at the beginning production step and $\rho(1,t)$ the wip density at the final production step. Mass conservation (known for a queuing system as Little's law) is given by the continuity equation

$$\frac{\partial \rho}{\partial t}(x,t) + \frac{\partial F}{\partial x}(x,t) = 0, \tag{1}$$

where $F(x,t) = \rho(x,t)v(x,t)$ is the associated flux of the production density and $v(x,t)$ is the velocity of a part at stage x moving through the factory at time t. Note that F is also known as the production rate, its units are parts/time. All the modeling efforts have been targeted to determine the best suited functional form of this flux. In steady state, the flux becomes constant representing the output of the production unit. The steady state flux is known as the clearing function [3] representing the relationship between the expected output of a production resource in a given planning period as a function of the workload. Karmarkar [3] proposed a nonlinear clearing function where output increases as a concave nondecreasing function of W, reaching an asymptotic maximum. Writing $F(x,t) = \Phi(W(t)),\ W(t) = \int_0^1 \rho(x,t)dx$ recovers the clearing function models for a particular choice of Φ describing the production rate at time t as a function of the total wip at time t.

Choosing

$$F = \rho(x,t)v(W(t)) \tag{2}$$

reflects a homogeneous speed through the production line, determined by the total load in the factory. Highly re-entrant flow with random dispatch policies leads to such behavior since a part at stage x competes for machine capacity with all the other parts in the factory that are going through the same machine.

In all of these models, the flux or the velocity is explicitly given as a functional of the density. This is a closure assumption: in the case of the clearing function, F represents the mean production rate for a production line in equilibrium as a function of the mean wip. In the case of $v(W)$, this represents the mean velocity $\left(\text{or } \frac{1}{\text{mean cycle time}}\right)$ as a function of the mean wip. Hence Eq. (1) is based on a quasi-steady state assumption restricting its validity to slowly varying influxes.

A typical velocity model that is based on treating the whole factory as a single M/M/1 queue [13] is given by

$$v(W) = \frac{\mu}{1+W}, \tag{3}$$