Submitted to *Manufacturing & Service Operations Management*
manuscript (Please, provide the manuscript number!)

Order release optimization for time-dependent and stochastic manufacturing systems

Hubert Missbauer
University Innsbruck, Department of Information Systems, Production and Logistics Management, Universitätsstrasse 15, 6020 Innsbruck, Austria, hubert.missbauer@uibk.ac.at,

Raik Stolletz
University of Mannheim, Business School, Chair of Production Management, Schloss, 68131 Mannheim, Germany, stolletz@bwl.uni-mannheim.de,

Order release planning models with load-dependent lead times must anticipate the time-dependent work-in-process and output for any given release schedule and thus require an anticipation model that approximates the transient behavior of queueing systems. We present a generic optimization model for order release planning in time-dependent and stochastic manufacturing systems that includes a well-defined interface to this anticipation model. We develop two stationary backlog carryover (SBC) approaches to approximate the time-dependent queueing behavior. ~~Its~~ consistency with the order release model is proven. The resulting nonlinear programming model is shown to be a special case of the well-known clearing function models. A numerical study shows the reliability of the method and presents structural insights.*

Key words:

History:

General comments:

- abstract: Each submitted article to MSOM should contain an abstract (no more than 300 words) that is based on the following subsections (without technical jargon):

1. Problem definition: What is your research problem?

* **more results?**

2. Academic / Practical Relevance: How is your research problem relevant to the OM research / practice community?
3. Methodology: What is the underlying research method?
4. Results: What are your key findings?
5. Managerial Implications: How can academics / managers / decision makers benefit from your study?

Please include the above five sections (highlighted in bold) in your abstract followed by the appropriate information.

- Model

— demand behind bottleneck?

— maybe place the paper for single-stage system

— WIP on server in balance equation

- Numerics

— comparison CF approaches

— backlog? $I_t, W_t \geq 0$???

— reliability for lines not shown

- Contribution

— second not that big

— third: Insights???

- SBC method: we are working with expectations

1. Motivation and Problem Description

Manufacturing planning and control systems, especially in discrete manufacturing, are mostly structured hierarchically ~~with an upper (top) level and a lower (base) level~~, ~~see Schneeweiß (2003).~~ The upper level sets the targets for the production units, that is the production orders and their release and required due dates. The lower (base) level performs detailed scheduling of the orders/operations within the production units. This definition of the top level decisions, referred to as goods flow control (Bertrand et al. 1990) or - in the context of supply chain planning - supply chain operations planning (De Kok and Fransoo 2003) implies two tasks (see Bertrand et al. (1990) for more details): (1) coordination of the production units along the material flow by setting coordinated production targets, (2) maintaining an appropriate state of the production units in order to guarantee feasibility of the production targets. In the following, we consider one production unit and limit ourselves to the second task mentioned above.

[1]

The targets for the production units are defined in terms of work orders and their release and required due dates. Feasibility can only be guaranteed if flow times are predictable~~, which~~ can be achieved by maintaining a desired level of work-in-process (WIP) at the work centers (*workload control concept*; see Bertrand et al. (1990) and Zäpfel and Missbauer (1993)). Workload control emphasizes the importance of order release as a decision problem: Order release - together with capacity planning/adjustment - largely determines the WIP level in the shop (queue lengths at the work centers). This determines the output that can be achieved via queueing relationships and leads to certain average flow times at the work centers via Little's law.

Considering order release as the essential interface between the goods flow control level and the production unit is consistent with the hierarchical manufacturing planning and control concept: ~~Multi-period models that perform~~ production planning at the goods flow control level cannot decide on the material flow within the production units since these decisions are made at the ~~lower~~ level (by schedulers or dispatchers) within the production units. Hence, these models eventually determine order releases, defined as work orders, their release (=earliest possible start) dates, and required due dates. ~~In the following, we deal with~~ multi-period models that make these decisions for one production unit for a finite planning horizon.

[1] **brief and clear statement of the decision problem first**

Seen from the perspective of goods flow control, the material flow within the production units that results from specified order releases is not known exactly; it is determined by the dispatching decisions at the shop flow level. Hence, in the context of order release planning, the production units (that usually are networks of work centers) are stochastic systems and – since they are controlled by order release – can be considered as controlled queueing systems (Kushner 2001) with order release as the relevant control. Since the order release~~s~~ usually will vary over time responding to time-varying demand, the production units mostly will be in a transient state. Consequently, the order release models must accurately describe the transient state(s) of a network of work centers, and the optimization must select the optimal order releases and trajectory of system states. This necessity to incorporate transient queueing models in order release models has been recognized in the literature (Missbauer 2009, 2011) but has only been realized with substantial limitations (see Section 2).

The aim of this paper is to perform a proof-of-concept study to realize this idea: We consider one production unit whose dynamic behaviour is largely determined by one bottleneck machine. We model this production unit by a well-defined approximation for transient single-stage queueing systems, namely the Stationary Backlog Carryover (SBC) approach (see Stolletz (2008)). We develop an SBC approximation to optimize the order releases over time using nonlinear programming. Hence, we show that the idealized procedure described above - modelling the transient behaviour of a production unit, inevitably by approximation, and solving the resulting optimization model - in principle is feasible.

The contribution of the paper is manyfold.

- First, we present a generic optimization model for order release planning in time-dependent and stochastic manufacturing systems. *Generic* means that the model includes a well-defined interface to a model that anticipates the time-dependent behavior of the production system for a given release schedule, but it does not specify this anticipation model.

- Second, two SBC approximations for $M_t/M/1$ queues are developed and their consistency with the order release model ~~are~~ proven. In addition we discuss the relationship of the resulting non-linear optimization model to ~~a~~ Clearing Function model for which a well-established theory is available. [2] [3]
- Third, the numerical study shows the ~~reliability of the method and presents structural insights.~~[4]

The paper is organized as follows: Section 2 reviews the relevant literature and motivates the method we use. The conceptional optimization model is introduced in Section 3. Section 4 is the main part and introduces and analyzes the order release model based on two SBC approximations. We also show the consistency of the resulting non-linear optimization model to clearing functions. Section 5 provides a numerical study. Conclusions and further research topics are presented in Section 6.

2. Related Literature

2.1. Workload Control and Order Release Planning

The workload control concept and order release methods within this concept have been described in conceptual publications mainly in the 1980s and 1990s (Bertrand and Wortmann 1981, Bertrand et al. 1990, Zäpfel and Missbauer 1993). As for order release algorithms, two streams of research can be distinguished (see Pürgstaller and Missbauer (2012)): Firstly, traditional order release mechanisms that determine the release times of the work orders for a short time horizon (typically one planning period) without explicit reference to an objective function, following a set of rules. Secondly, multi-period optimization models where the order releases in the periods of the planning horizon are decision variables. These models aim at optimizing the order releases for the entire planning horizon with respect to an appropriate objective function.

[2] **why not numerically shown?**

[3] **CF are not introduced before!**

[4] **really, what are our structural insights?**

Multi-period optimization models for order release planning can be classified mainly according to the way they handle lead times that are based on estimates of the flow times through the manufacturing systems (for an overview, see Missbauer and Uzsoy (2011)): Models with fixed planned lead times consider lead times as exogenous parameters and constrain the release quantities over time such that the lead times can be met. This facilitates material coordination across the manufacturing stages and leads to relatively straightforward models. However, flow times strongly depend on the utilization of the production resources, and imposing a fixed lead time constraint either partly neglects this relationship or ~~keeps~~ the utilization at a fairly constant level. It has been shown in previous work that this conceptual weakness actually degrades performance compared to models that allow load-dependent lead times (see Asmundsson et al. (2009), Kacar et al. (2013)). Therefore, we only consider order release models that allow load- (and hence time-) dependent lead times and thus must model the transient dynamics of the production units, especially the time-dependent WIP level that is necessary to realize a certain output over time.

One important and perhaps the most mature research direction in order release planning with load-dependent lead times are clearing function models. These models represent the production unit as a network of work centers. The planning horizon is divided into planning periods $t = 1, ..., T$. The material flow is represented by inventory balance equations for the WIP at each work centre and for finished stock keeping units (SKUs), usually distinguishing different SKUs or groups of SKUs with similar routing. The relationship between WIP and output at the work centers is represented by nonlinear, saturating clearing functions. A *clearing function* of a work center is the functional relationship between some measure of WIP in a period t and the expected or maximum output of the work center in period t. This allows the adaptation of WIP and planned lead times to the load situation in the shop. For an overview of clearing function models, see Missbauer and Uzsoy (2011).

Although clearing functions are based on queueing relationships, they are not explicitly based on theory on transient queueing systems (Dobson and Karmarkar 2011). Usually they are parameterized from simulation data and thus "the clearing functions employed by most researchers to date represent an average relation over a wide range of operating states, but may be quite inaccurate for a given sample path of system evolution" (Kacar and Uzsoy 2010). It can be shown by transient queueing analysis that the output estimated by clearing functions can deviate substantially from the expected output derived from transient queueing models (Missbauer 2011).

Accurately anticipating the time-dependent output and flow times that result from specified releases over time is possible by simulation or by transient queueing models. ~~So far,~~ these anticipation models cannot be incorporated into tractable optimization models. [5] ~~Known~~ order release methods iterate between an order release model with fixed lead times (usually a linear program) and a flow time and/or capacity anticipation model (simulation as in Hung and Leachman (1996) or Kim and Kim (2001), transient queueing as in Riaño (2003)). The optimal releases obtained from the release model are fed into the anticipation model that yields time-dependent flow times and outputs. These variables are used to update the planned lead times and possibly also the capacities in the release model. Its solution leads to new release volumes, etc., until convergence is achieved. However, convergence is not guaranteed, and the theory behind these mechanisms and their shortcomings are not yet understood (for numerical tests and interpretations, see Irdem et al. (2010)).

~~It is evident from theory that order release models must be based on transient queueing models.~~[6] In order to make those release models tractable and to allow the queueing approximation internal to the release model, relatively simple approximations to transient queueing models

[5] **Justus: why, see Parlar 1984**

[6] **citation?**

that are convenient for engineering applications as suggested by Abate and Whitt (1994) are necessary. Therefore, we review the available non-stationary performance evaluation methods and subsequently choose a suitable method, namely the SBC approach (Stolletz 2008), to formulate the model.

2.2. Non-Stationary Performance Evaluation

Overviews about methods to analyze non-stationary queues are given in Ingolfsson et al. (2007) and Schwarz et al. (2015). According to Schwarz et al. (2015), the methods may be classified into three categories:

1. Analytical and numerical solutions.

   Analytical solutions are rare and exist only for special settings, e.g., $M_t/M_t/\infty$ systems by Collings and Stoneman (1976). For Markovian systems, the numerical solution of the respective set of ordinary differential equations may be used (e.g., Koopman (1972) and Ingolfsson et al. (2002)). Although these methods provide (nearly) exact results, the numerical solution is rather time-consuming (Ingolfsson et al. 2007). [7]

2. Approximations based on the modification of the system characteristics.

   One class of approaches approximates discrete events by continuous processes. Deterministic fluid approaches are fast and suitable for the time-dependent analysis of overloaded systems (e.g., Newell (1971)). However, any waiting due to stochastic impacts in an underloaded system is neglected. Diffusion approximations capture these stochastic effects (see Chen and Mandelbaum (1994)). [8] Another class of approximations is based on infinite server models (e.g., Jennings et al. (1996)) and has a low approximation quality for single server systems.

3. Approximations based on models with constant parameters.

   There exist approximations which use transient models for piecewise constant parameters, for

[7] **There exist no closed-form formulas that can be used in an integrated approach.**

[8] **why not using them?**

example the approach of Upton and Tripathi (1982) or the randomization approach by Gross and Miller (1984). Another class of approximations is based upon the application of steady-state models. Ingolfsson et al. (2007) show that the stationary independent period-by-period (SIPP) approximation achieves good results within a reasonable time. This method divides the observed time horizon into multiple smaller periods and then analyzes each period independently via stationary models (Green et al. 2001). [9] In contrast, the stationary backlog-carryover (SBC) approach by Stolletz (2008) considers the dependencies between successive periods. This method builds (virtual) backlogs of non-served arrivals and carries them over to the succeeding period.

A fourth category we consider here are approximations for the time-dependent WIP level as a function of the time-dependent input at a work center that have been developed based on analyses of transient queueing systems and/or simulation results in order to be used for engineering applications. Gans et al. (2003) state that ”overloaded regimes are well approximated by fluid models”. This is supported by the simulation results in **?**, p.1799f Abate and Whitt (1987) show that for utilization < 1 and zero initial conditions the evolution of WIP can be approximated by exponential or hyperexponential functions and derive these expressions. For general initial conditions, the WIP evolution is more complex. The number in system of an $M/M/1$ queue for general initial conditions can be approximated by a linear combination of four exponentials Abate and Whitt (1988). Essential properties of the time-dependent behaviour extend to the $M/G/1$ workload process (**?**). Missbauer (2009) shows that in principle these insights can be the basis of order release models, but essential questions are still open. In particular, different load patterns (overload - underload) lead to different functional relationships.

Out of the reviewed methods the SBC is a simple approximation which yields reasonable approximation quality. [10] It is valid for arbitrary load patterns, including temporary overload, without having to switch between different mathematical formulations. Therefore, SBC is used

[9] **Hence, overloaded periods are not possible to analyze.**

[10] **Abgrenzung von anderen Verfahren, z.B. MOL**

to provide a proof-of-concept that well-defined transient queueing approximations can be used in order release models. Since the SBC method is used in the following, it is described in more detail in Section 4.1.

3. Formulation of the Generic Order Release Model

3.1. Assumptions on the Production System

- We consider a production unit that is modelled as a single-server queueing system. The conceptual model is a production unit where the material flow is largely determined by one bottleneck machine. The non-bottlenecks are not modelled explicitly. [11]
- The planning horizon is divided into (micro-)periods with (small) period length l.
- D_t is the demand for finished products in period t. Demand is deterministic and known. [12]
- Due to the stochastic material flow through the non-bottleneck work centers, the arrival to the considered bottleneck machine is stochastic. Only the arrival rate per period, denoted by λ_t, can be controlled by the order release function. The λ_t are the essential decision variables. The time-dependent releases at the entry station of the line that lead to this pattern of λ_t may differ. [13]
- The arrival process is Poisson with time-dependent arrival rate λ_t at the server.
- The service rate μ [orders per period] is constant over time and assumed to be $\mu = 1$. Assuming exponentially distributed processing times, the bottleneck station is modelled as an $M_t/M/1$ queueing system.

The assumptions on the structure of the production unit mean that orders are released and may pass one or more non-bottleneck work centers before arriving at the bottleneck machine under consideration. The release pattern and the stochasticity of the material flow at these

[11] **check BNWC: bottleneck work center model, all other workcenters are modelled as fixed time delays; Find real citations about this!**

[12] **how to argue for derivation of demand rate for bottleneck station?**

[13] **present derivation of release rates for long lines**

non-bottleneck work centers lead to a Poisson arrival process with rate $\lambda(t)$ at the bottleneck machine, and $\lambda(t)$ (represented in discrete time as λ_t for each micro-period t) is the decision variable.

This dynamic behaviour can be achieved exactly only under strong assumptions. If the order release process is Poisson with time-dependent release rate and there is a fixed delay between release and order arrival at the bottleneck, this is the case. Beyond this, it is difficult to find a queueing network structure that leads to a Poisson arrival process with time-varying arrival rate λ_t at the bottleneck where the λ_t can change arbitrarily over the periods. However, the purpose of this paper is to provide a proof-of-concept study for integrating a well-defined and tested approximation method for transient queueing systems into a model for order release planning. Refining the model in order to improve its consistency with the structure of queueing networks is a topic for future research. Relaxing the assumption of Poisson arrivals by using an SBC approximation of more general queueing systems (Stolletz and Lagershausen 2013) is a research topic as well.

3.2. Generic Optimization Model

Based on these assumptions, the decision problem at hand is to plan the time-dependent ~~release~~ rate λ_t for the single-stage stochastic [14] production systems. The expectations of queue length $Q_t(\lambda)$ at the end of t, of number of orders $W_t(\lambda)$ in the system (in queue or at the server) at the end of period t, of the final product inventory $I_t(\lambda)$ at the end of period t, and of the production rate $X_t(\lambda)$ in period t depend on all λ_τ with $\tau \leq t$.

The general optimization problem is as follows:[15]

$$Minimize \sum_{t=1}^{T} (h * W_t(\lambda) + g * I_t(\lambda)) \tag{1}$$

[14] **define λ**

[15] **parameter not defined**

subject to the constraints

$$I_t(\lambda) = I_{t-1}(\lambda) + X_t(\lambda) - D_t \qquad t \in \{1, ..., T\} \tag{2}$$

$$\lambda_t, X_t(\lambda), Q_t(\lambda), I_t(\lambda), W_t(\lambda) \geq 0 \qquad t \in \{1, ..., T\} \tag{3}$$

The main decision variable is the time-dependent arrival rate λ_t. The objective function (1) minimizes the cost for the WIP in the queue and at the server plus the cost for finished goods inventory. The expected inventory of finished goods I_t [16] ~~is~~ described via Equation (2). All variables are non-negative (3) and all undeclared variables with $0 \geq t$ are set to 0. Note that the X_t trajectory is entirely determined by the vector of the work inputs λ via the queueing characteristics of the server. Approximating this relationship is the topic of the next section.

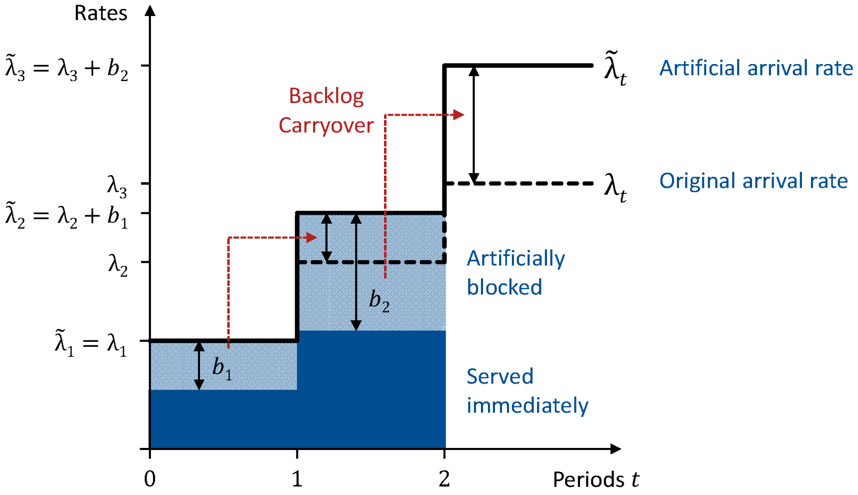
4. Order Release Optimization Using Stationary Backlog Carryover

The time-dependent behaviour of the server is determined by the SBC approximation which yields the expected utilization and the expected WIP (number in system) for each period t for a given input rate trajectory $\lambda_\tau, \tau = 1, ..., t$. The SBC approach can be considered as a clearing function model with a specified period length and analytically set parameters (**?**), hence the model is a special case of a clearing function model (for a description of clearing function models, see Missbauer and Uzsoy (2011)).

The SBC approximation yields the expected throughput per period, and the model takes this as the planned throughput. The actual throughput is a random variable and can be lower, which, in the actual operation of the system, can lead to backlogs and associated costs. In conventional clearing function models, this can be compensated by adjusting the clearing function parameters towards a more "conservative" output estimation (see Asmundsson et al. (2009), **?**, p. 216ff., for the analogous effect resulting from different functional forms of the clearing function). Our model does not provide this possibility and thus reveals the properties of a clearing function that

[16] **does the numerics allows for backlog ($I_t < 0$)?**

Figure 1 **Backlog carry-over within the SBC approach**



estimates expected values. In the case of series production, it can be complemented by a safety stock to change the compromise between inventory and due-date performance.

4.1. Performance Evaluation of $M_t/M/1$ Via SBC

Stationary Backlog Carryover (SBC), introduced by Stolletz (2008), is a method to approximate the time-dependent behaviour of queueing systems. The key idea of SBC is as follows. The time horizon is divided into (micro-)periods t with constant parameters. Two evaluation steps are performed for every period t.

Phase I In the first step, the corresponding *stationary loss system* without any waiting position is considered and the expected utilization $E[U_t^{loss}]$ is determined. Calculating $\mathrm{E}[U_t^{loss}]$ makes use of an artificial arrival rate $\tilde{\lambda}_t$ that includes the original arrival rate λ_t and the backlog rate b_{t-1} (rate of artificially rejected jobs), which is carried over from the previous period $t-1$, see Figure 1. The backlog b_t of a period t is derived based on the artificial arrival rate $\tilde{\lambda}_t$ and the resulting expected probability of blocking in the same period.

Phase II In the second step of the SBC approach, the expected number in system and the expected number in queue of the original system are approximated by a *stationary waiting system*. A modified arrival rate λ_t^{MAR} is used as an input for these calculations. This modified arrival rate is chosen so that the expected utilization of the considered waiting system equals the approximated expected utilization of the loss queueing system $E[U_t^{loss}]$ from the first step. The performance measures of the corresponding waiting model approximate the time-dependent performance of the original system.

According to Stolletz (2008), this method reaches the theoretical steady-state measures for stable $M/M/c$ systems with constant parameters.

In the following, we discuss the first and second phase of the SBC for the performance approximations for given arrival rates λ_t, as well as an alternative to the second phase via balance equations. To derive the SBC-approach for an $M_t/M/1$ queue, we assume a period length of $l = \mu^{-1} = 1$.

SBC Phase I: The artificial arrival rate $\tilde{\lambda}_t$ of period t consists of the original arrival rate λ_t and the backlog rate b_{t-1} of the preceding period. Starting with $b_0 = 0$ results in

$$\tilde{\lambda}_t = \lambda_t + b_{t-1} \qquad t \in \{1, ..., T\} \tag{4}$$

For a given artificial arrival rate $\tilde{\lambda}_t$, the stationary $M/M/1/1$ loss system is applied. This yields the probability of blocking $P(blocking_t) = \frac{\tilde{\lambda}_t}{\mu+\tilde{\lambda}_t}$, which equals the expected utilization $E[U_t^{loss}]$ due to the "PASTA" (Poisson arrivals see time averages) property:

$$E[U_t^{loss}] = \frac{\tilde{\lambda}_t}{\mu + \tilde{\lambda}_t} \qquad t \in \{1, ..., T\} \tag{5}$$

For the formulas for those $M/M/1/1$ system, see, e.g., Papadopoulos et al. (1993).

The backlog rate b_t is derived via

$$b_t = \tilde{\lambda}_t P(blocking_t) = \tilde{\lambda}_t \frac{\tilde{\lambda}_t}{\mu + \tilde{\lambda}_t} \qquad t \in \{1, ..., T\} \tag{6}$$

Based on $P(blocking_t)$ the output rate X_t is approximated via

$$X_t = \tilde{\lambda}_t(1 - P(blocking_t)) = \tilde{\lambda}_t(1 - \frac{\tilde{\lambda}_t/\mu}{1 + \tilde{\lambda}_t/\mu}) = \frac{\tilde{\lambda}_t}{1 + \tilde{\lambda}_t/\mu} \qquad t \in \{1, ..., T\} \qquad (7)$$

SBC Phase II: In the second phase of the SBC we apply the $M/M/1/\infty$ waiting model with the artificial rate λ_t^{MAR}. To reach the expected utilization $E[U_t^{loss}]$ with this system, the rate has to be defined as[17]

$$\lambda_t^{MAR} = X_t = E[U_t^{loss}]\mu = \frac{\tilde{\lambda}_t}{\mu + \tilde{\lambda}_t}\mu \approx \frac{\tilde{\lambda}_t}{1 + \tilde{\lambda}_t} \qquad t \in \{1, ..., T\} \qquad (8)$$

This results in the expected number in system of (by substituting (8) for λ_t^{MAR})

$$E[L_t^s] = \frac{\lambda_t^{MAR}/\mu}{1 + \lambda_t^{MAR}/\mu} = \frac{\frac{\tilde{\lambda}_t}{\mu + \tilde{\lambda}_t}}{1 - \frac{\tilde{\lambda}_t}{\mu + \tilde{\lambda}_t}} = \frac{\tilde{\lambda}_t}{\mu} = \frac{\lambda_t + b_{t-1}}{\mu} \approx \lambda_t + b_{t-1} \qquad t \in \{1, ..., T\} \qquad (9)$$

Thus, the expected number in system in period t is equal to the expected number of arrivals in period t plus the backlog rate b_{t-1}. In the actual shop operation, b_{t-1} is the expected number in system at the beginning of period t.

As $E[U_t^{loss}]$ is the expected number of customers in process at the server and is equal to X_t/μ (see Equation (8)), the expected number in queue is for $\mu^{-1} = 1$

$$E[L_t^q] = E[L_t^s] - E[U_t^{loss}] = \lambda_t + b_{t-1} - X_t \qquad t \in \{1, ..., T\} \qquad (10)$$

Proposition 1. *The approximated expected number in queue $E[L_t^q]$ equals the backlog rate b_t in the SBC approximation for an $M_t/M/1$ system with $l = \mu^{-1} = 1$.*

As Equation (7) can be rewritten as $\tilde{\lambda}_t P(blocking_t) = \tilde{\lambda}_t - X_t$, from Equation (6) follows $b_t = \tilde{\lambda}_t P(blocking_t) = \tilde{\lambda}_t - X_t$. Using the definition $\tilde{\lambda}_t = b_{t-1} + X_t$ from Equation (4) results in

$$b_t = b_{t-1} + \lambda_t - X_t \qquad t \in \{1, ..., T\} \qquad (11)$$

[17] To indicate when the assumption $l = \mu^{-1} = 1$ is applied we use $\approx$ instead of the equal sign.

which equals $E[L_t^q]$ according to Equation (10). Hence,

$$b_t = \tilde{\lambda}_t P(blocking_t) = \tilde{\lambda}_t - X_t = b_{t-1} + \lambda_t - X_t = E[L_t^q] \qquad t \in \{1, ..., T\} \qquad (12)$$

Note that the evolution of $E[L_t^q] = b_t$ over time is given by Equation (11) that refers to the system state at the end of the period and can be interpreted as a balance equation. Next we explore its relationship to the inventory balance equation of the release model.

4.2. WIP Estimation Via Balance Equations and Consistency

[18] The second phase of SBC calculates the average WIP (number in system) in a period t from the utilization in that period. Alternatively, we can use inventory balance equations to describe the expected WIP (number in system) at the end of the periods of the planning horizon. The expected number in system at the end of period t, denoted by W_t, is seen as the expected number in system at the end of the former period reduced by what is leaving the system plus the number of arriving jobs. For $l = \mu^{-1} = 1$, the rates equal the absolute expected numbers and the balance equation for the queue length is [19]

$$W_t = W_{t-1} + \lambda_t - X_t \qquad t \in \{1, ..., T\} \qquad (13)$$

Since Equation (13) calculates the WIP from the output calculated by SBC Phase I and Equations (9) and (10) calculate the WIP by SBC Phase II, this clarifies the consistency of the SBC Phases I and II:

Proposition 2. *For $l = \mu^{-1} = 1$,*

- *The expected number in system W_t calculated by the balance Equation (13) equals the backlog rate b_t.*

Proof: Comparing Equation (11) with the start condition $b_0 = W_0 = 0$ results directly in $W_t = b_t$.

[18] **HM: Schwer zu verstehen, wenn man nicht drin ist.**

[19] **besser rausstellen, dass balance equation den WIP auf den server ignoriert**

- *Hence, the work in process W_t based on the balance Equation (13) equals the SBC Phase II approximation of the average queue length $E[L^q_t]$ (Equation (12)).*
- *The SBC Phase II approximation of the average number in system $E[L^s_t]$ equals the expected available work based on the balance equation (termed load, defined as $W_{t-1}+\lambda_t = b_{t-1}+\lambda_t$) in period t (Equation (9)).* [20]

Therefore,[21] there is a well-defined difference between the WIP estimation by balance equations and by SBC phase II due to the approximative nature of SBC. The relative difference approaches zero as the WIP level increases to infinity for given throughput. In the following, we use the balance equations to calculate the WIP level at the end of the periods since this is consistent with the work input and output by definition and also follows the usual standards of production planning (including order release) models. [22]

4.3. Optimization Model and Mathematical Structure in Relation to Clearing Functions

Combining the general formulation of the order release planning model of Section 3 with the SBC approach to relate the performance variables with the release rate, we get the following nonlinear decision model. The model calculates the planned output over time using phase I of the SBC approximation and derives the WIP inventory from the inventory balance equations (that is, SBC Phase II is not applied). Table 1 shows the notation used.

All rates are measured in product units per periods, all inventory variables are measured in product units.

$$Minimize \sum_{t=1}^{T} (h * W_t + g * I_t) \tag{14}$$

[20] **any insights?**

[21] **why use of L^Q as WIP?**

[22] **ignore item on the server**

Table 1 Notation.

Indices	
$t = 1, ..., T$	Periods
Parameters	
μ	service rate at the bottleneck station
D_t	demand rate in period $t \in \{1, ..., T\}$
$l = \mu^{-1}$	period length
h, g	holding cost coefficient for WIP and final product inventory, respectively
Continuous non-negative decision variables:	
λ_t	arrival rate (at the bottleneck station) in period $t \in \{1, ..., T\}$
X_t	expected production rate in period $t \in \{1, ..., T\}$
$\tilde{\lambda}_t$	artificial arrival rate in period $t \in \{1, ..., T\}$ ($\tilde{\lambda}_0 = 0$)
b_t	SBC backlog rate in period $t \in \{1, ..., T\}$ ($b_0 = 0$)
Q_t	expected queue length at the end of $t \in \{1, ..., T\}(Q_0 = 0)$
W_t	expected number of orders in the system (in queue or at the server) at the end of period $t \in \{1, ..., T\}$
I_t	expected final product inventory at the end of period $t \in \{1, ..., T\}$ ($I_0 = 0$)

subject to the constraints

$$b_t = b_{t-1} + \lambda_t - X_t \qquad t \in \{1, ..., T\} \tag{15}$$

$$\tilde{\lambda}_t = \lambda_t + b_{t-1} \qquad t \in \{1, ..., T\} \tag{16}$$

$$X_t = \frac{\tilde{\lambda}_t}{1 + \tilde{\lambda}_t} \qquad t \in \{1, ..., T\} \tag{17}$$

$$W_t = b_t \qquad t \in \{1, ..., T\} \tag{18}$$

$$I_t = I_{t-1} + X_t - D_t \qquad t \in \{1, ..., T\} \tag{19}$$

$$b_t, \lambda_t, \tilde{\lambda}_t, X_t, Q_t, I_t, W_t \geq 0 \qquad t \in \{1, ..., T\} \tag{20}$$

The main decision variable is the time-dependent arrival rate λ_t. The objective function (14) minimizes the cost for the WIP in the queue and at the server plus the cost for finished good inventory. Equations (15)-(18) model the time-dependent performance using the first phase of the SBC based on equations (12), (4), (7), and Proposition 2. Here, the balance equation for W_t is used instead of the second phase of the SBC, therefore $W_t = b_t$ (see Proposition 2). The expected inventory of finished goods I_t is described via Equation (19). All variables are non-negative (20) and all undeclared variables with $0 \geq t$ are set to 0.

The model calculates the objective function (14) as a function of the decision variables λ_t. Since all other variables $(X_t, b_t, \tilde{\lambda}_t, W_t, I_t)$ are determined by equality constraints, in principle they can be omitted from the model. Calculating the X_t by resolving the recursions (15) - (17) leads to fractions that involve high polynomials which can hardly be interpreted. Therefore, we show that the model is similar to a clearing function model which helps clarify its structure.

We define:

$f^{M/M/1/1}(\lambda)$ Steady-state output of an $M/M/1/1$ loss system and a specified service rate μ, as a function of the arrival rate λ.[23]

$P^B(\lambda)$ Blocking probability of an $M/M/1/1$ loss system as a function of the arrival rate λ (=probability of rejecting an order).

The following relationship holds:

$$f^{M/M/1/1}(\lambda) = \lambda * (1 - P^B(\lambda)) \tag{21}$$

Proposition 3. *$f^{M/M/1/1}(\lambda)$ is less or equal λ for $\lambda \geq 0$, concave in λ, and $\lim_{\lambda\to\infty} f^{M/M/1/1}(\lambda) = \mu$.*

Proof: Substituting $P^B(\lambda) = \frac{\lambda}{\lambda+\mu}$ into (21) yields

$$f^{M/M/1/1}(\lambda) = \frac{\lambda\mu}{\lambda+\mu}, \tag{22}$$

which exhibits the properties stated in Proposition 3. Hence $f^{M/M/1/1}$ exhibits the properties of a nonlinear, saturating clearing function (for these properties, see Missbauer and Uzsoy (2011), p. 473f.). □

Constraints (15) and (17) now can be rewritten as follows (note that $\mu = 1$): The output in period t (17) can be rewritten by substituting (22), $\mu = 1$ and (16) for $\tilde{\lambda}_t$:

$$X_t = f^{M/M/1/1}(\tilde{\lambda}_t) = f^{M/M/1/1}(\lambda_t + b_{t-1}) \tag{23}$$

The inventory balance equations (15) for the (virtually) backlogged customers result from Equation (12):

$$b_t = \underbrace{b_{t-1} + \lambda_t}_{\tilde{\lambda}_t} - \underbrace{X_t}_{f^{M/M/1/1}(\tilde{\lambda}_t)} \tag{24}$$

The artificial backlog b_t is just a mathematical entity, but since it is the difference between the cumulative expected input to and the cumulative estimated output from the WIP at the work center, it is the estimated *actual* WIP in the system, in our model expressed as number in system (Equation (18)).

Thus, the model can be formulated as follows:

$$Minimize\ (14)$$

subject to the constraints (repeated for convenience)

$$\begin{aligned}
&I_t = I_{t-1} + X_t - D_t && t \in \{1, ..., T\}\\
&b_t = b_{t-1} + \lambda_t - X_t && t \in \{1, ..., T\}\\
&W_t = b_t && t \in \{1, ..., T\}\\
&X_t = f^{MMcc}(\lambda_t + b_{t-1}) && t \in \{1, ..., T\}\\
&b_0 = 0 &&\\
&b_t, \lambda_t, X_t, W_t, I_t \geq 0 && t \in \{1, ..., T\}
\end{aligned}$$

This is a clearing function model (for an overview, see Missbauer and Uzsoy (2011)) formulated without a linearization of the clearing function. The only difference is the equality condition in the clearing function constraints - usually the clearing function is used as an upper bound to the output. This equality condition complicates the model since it can lead to a non-convex feasible set. [24] Using an inequality (less or equal) constraint is more intuitive; at the scheduling level, this means that the schedule does not need to be semi-active. [25] However, it remains to be clarified whether the SBC approximation still holds in this case. In this paper, we use the equality constraint on the output since this maintains consistency with the SBC approach. In our numerical experiments, the model was easily solvable by NLP solvers. We did not encounter any problems due to the non-convexity issue. The extensive numerical studies that have been performed on clearing function models indicate that the clearing function formulated as a less or equal constraint holds to equality in the optimal solution except in a few cases (sudden decrease in demand; see Kefeli et al. (2011)) that can occur in our experiments.

5. Numerical Study

[26]

The starting point of our numerical experiments is the properties of SBC as a descriptive model and the behaviour of clearing function models.

Single-stage clearing function models build up only the WIP level that is necessary to provide the capacity that is used. They exhibit a production smoothing property due to the concave shape of the clearing function. However, the variations of the release quantities over time may be higher than the variations of the demand. Clearing functions only consider the load (available work) of the period as explanatory variable for estimating the output of this period. This imposes certain limitations since the history of the process is not considered precisely (see Missbauer (2011)). The clearing function tends to overestimate the impact of sudden changes in the arrival rate, and it can be assumed that this property extends to our model. Therefore, we need to explore

[24] See Carey (1987) and Merchant and Nemhauser (1978). These authors use an exit function analogous to a clearing function $X_t = f(W_{t-1})$.

[25] **semi-active?**

[26] **1st paragraph sounds like conclusion, leave it out? HM: Yes, done.**

1. the dynamic behaviour of the model that is influenced by the inaccuracies of the (descriptive) SBC model,
2. the similarities and differences to the behaviour of clearing function models due to the short (micro) periods, and[27]
3. the compromise between finished goods inventory (FGI) and backlog level that results from the queueing model without adjustable parameters.[28]

We assume that the order release model handles a smooth demand curve very well and leads to the minimum WIP level since SBC converges to the steady-state WIP values of queueing models. We must explore the reaction of the model to changes, in particular sudden changes, of the demand, distinguishing increase and decrease, zero and non-zero initial conditions as well as different utilization levels.

In the following experiments, the planning horizon is 100 periods. The time-dependent demand D_t is illustrated in the following figures.[29] The λ_t obtained from the optimization are the input into the simulation, the actual order arrivals are random due to the Poisson arrivals within a period. The simulation results given below are the averages over one million replications. The holding cost coefficients for WIP and FGI are $h = 0.8$ and $g = 1$, respectively.

Figure 2 depicts the optimization results and the simulated behaviour for a rectangular demand curve. WIP and FGI are depicted in Figure **??**.

As expected, the output is overestimated in the ramp-up phase, for constant demand the estimation is very accurate, after the decrease in demand there is no clear tendency of the approximation error. Overall, the error in the output estimation is small (mean absolute deviation

[27] **Vergleich zu anderen release policies?**

[28] **How are backlogs possible?**

[29] **change notation in Figures**

Figure 2 **Demand, arrivals and output for demand rate = 0.7 from period 6 to period 50, otherwise zero.**

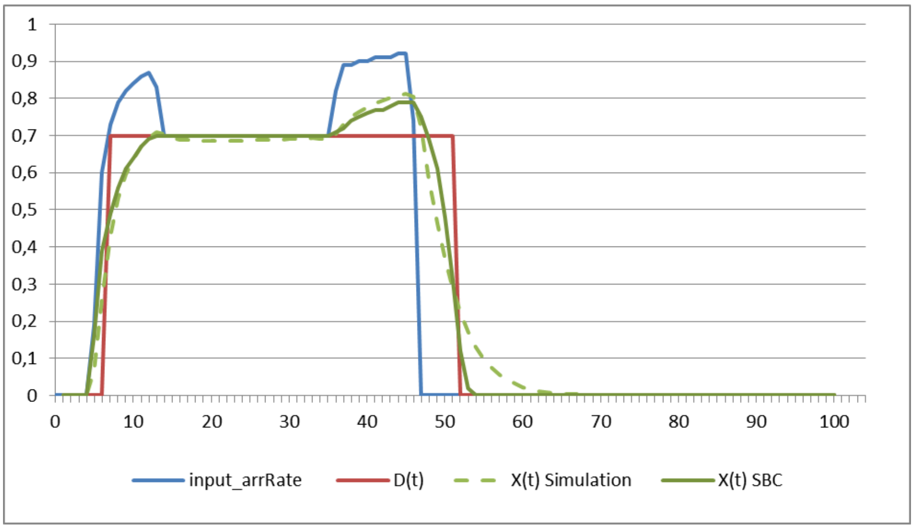
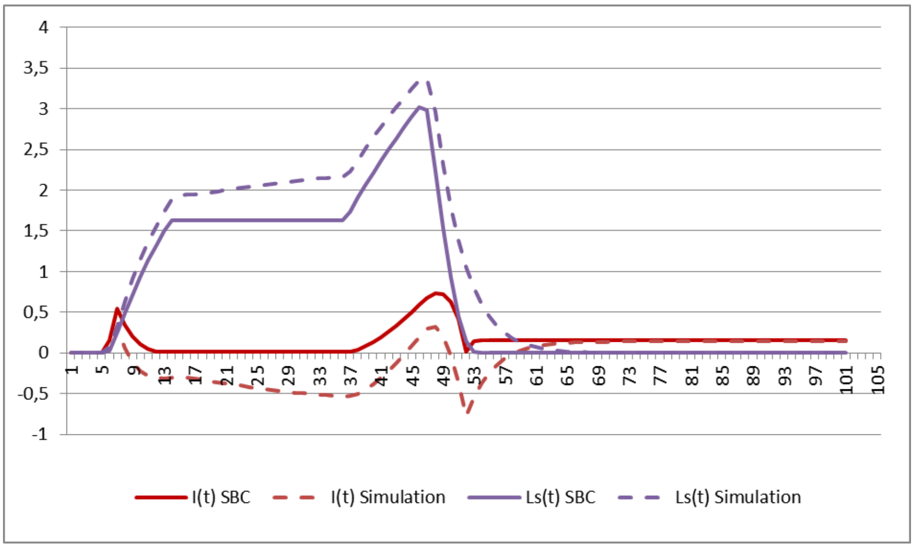


Figure 3 **Demand, WIP and FGI for the example of Figure 4.**



$\sim 2\%$). After demand drops to zero, the actual (simulated) output approaches zero slower than estimated. In this phase, SBC does not estimate the effects of the variability of WIP correctly. In most periods, there is backlogging due to the frequent overestimation of output. In clearing function models, this can be compensated by adjusting the clearing function parameters towards a more conservative output estimation. An analogous adjustment of SBC is possible by calibration of the period lengths as shown in Stolletz and Lagershausen (2013), p. 1371 f.

Figure 4 Demand, arrivals and output for demand rate = 0.90 from period 22 to period 75, otherwise zero. h=0.8, g=1

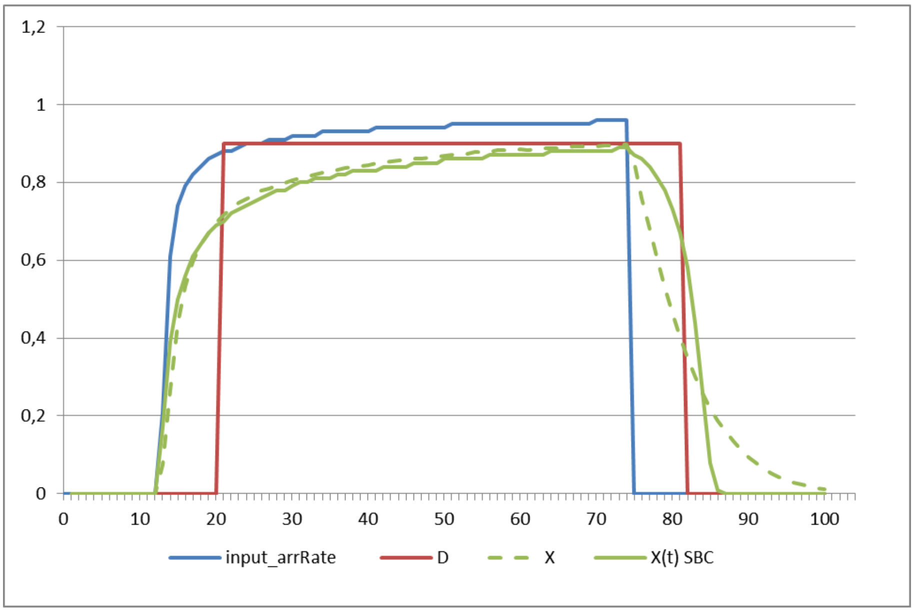
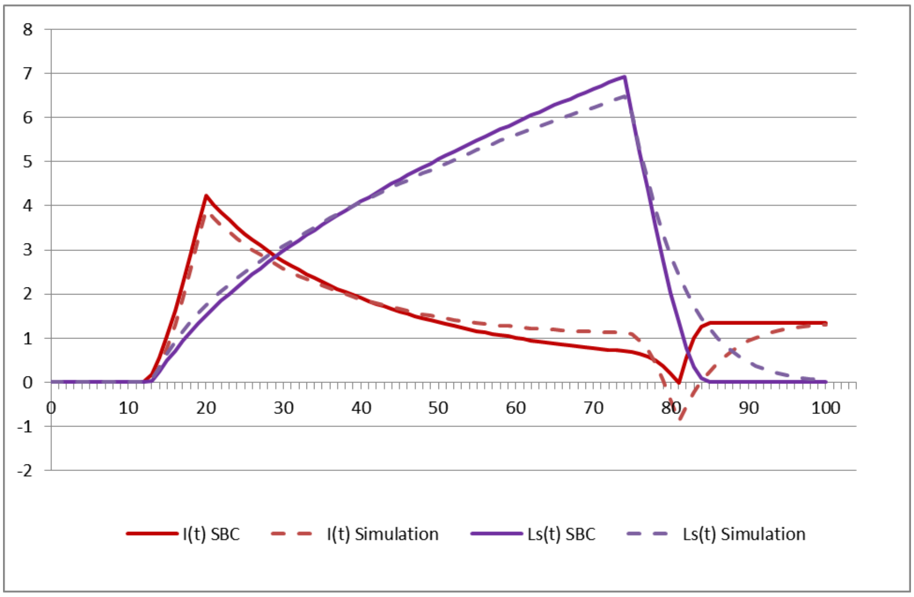


Figure 5 Demand, WIP and FGI for the example of Figure 5



We now increase the demand rate to 90% of the capacity in periods 22 to 75. Constant demand close to the capacity (90%) almost avoids backlogging and exhibits a smooth release pattern (Figures 4 and 5).

The results for a more complex demand pattern with temporary overload are shown in Figures 6 and 7.

Figure 6 **Demand, arrivals and output for two demand peaks. h=0.8, g=1**

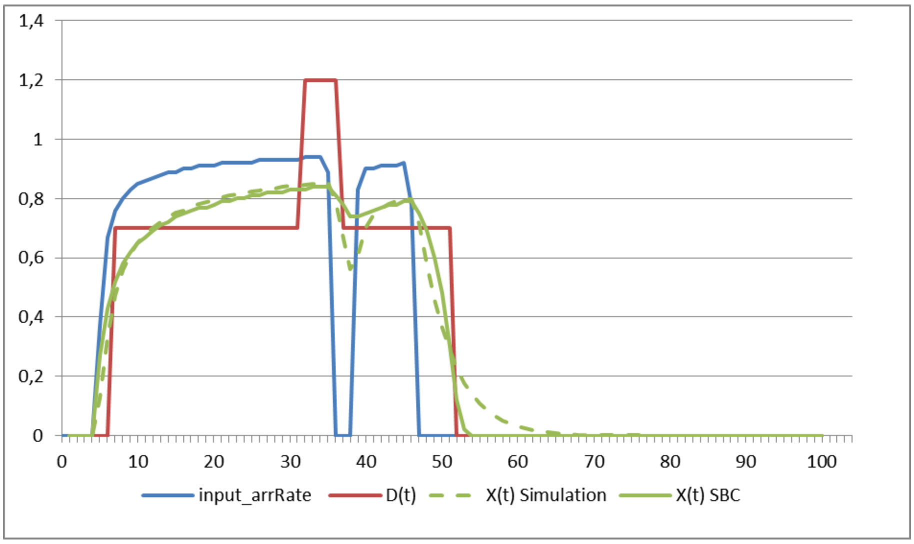
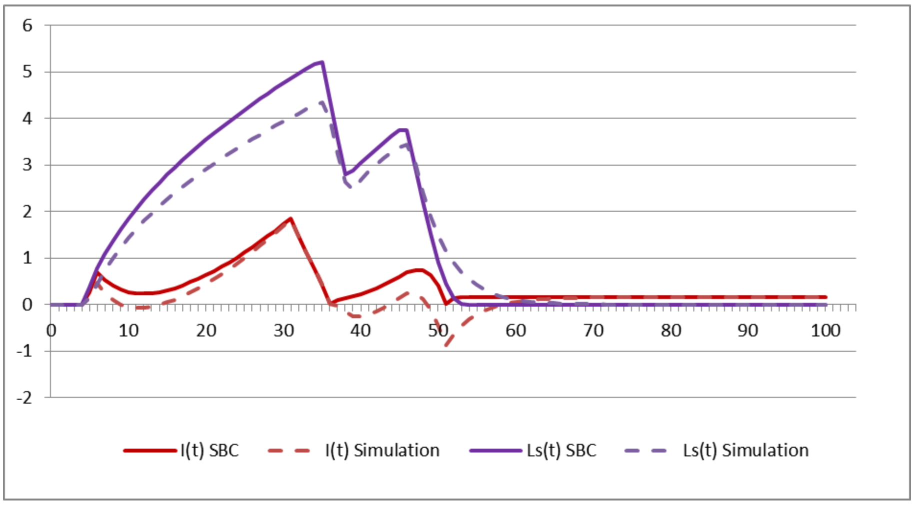


Figure 7 **Demand, WIP and FGI for the above example**



The results are similar with substantial differences between optimization and simulation results in response to sudden demand changes. Note also the sudden changes in the input rate resulting from the optimization which is in line with the dynamic behavior of clearing function models. We further explore this behavior in numerical examples where the demand per period oscillates between 65% and 90% of the capacity for three oscillation frequencies. Figures 8, 9 and 10 depict the results.

The[30] output from the optimization model (from the SBC approximation) exhibits the typical

[30] **difference in last two figures (9 and 10)?**

Figure 8 **Demand, input rate and output for oscillating demand, high frequency**

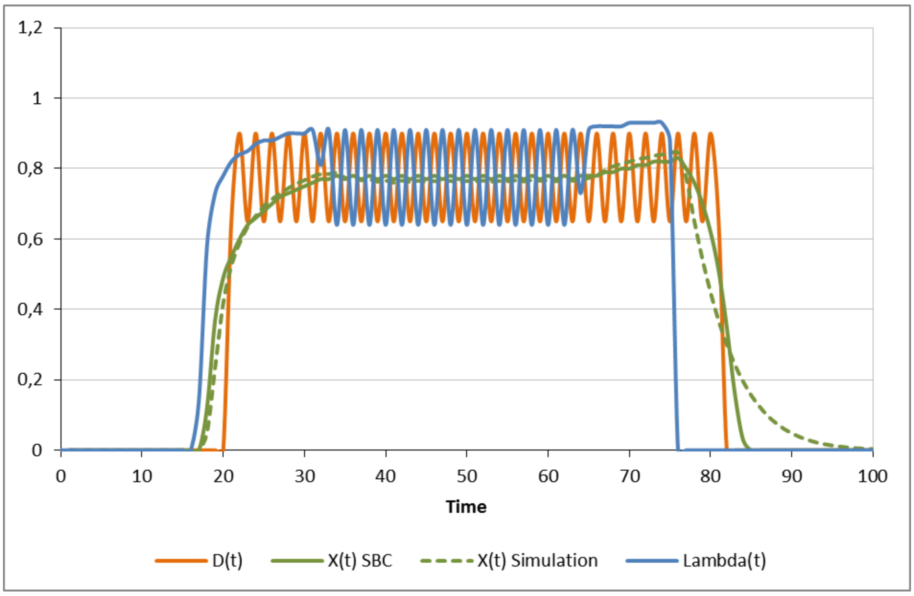
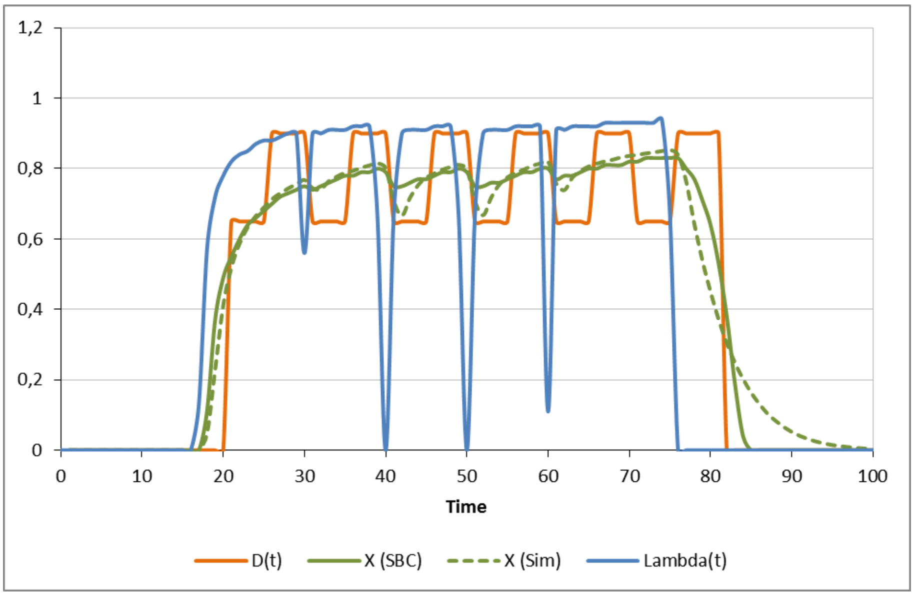
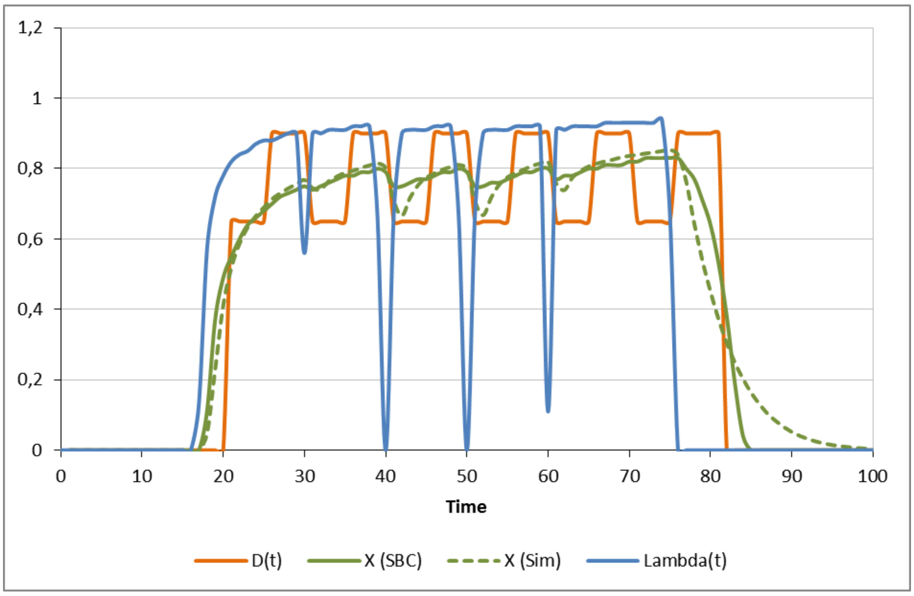


Figure 9 **Demand, input rate and output for oscillating demand, medium frequency**



smoothing behavior since the output is a concave, saturating function of the artificial arrival rate $\tilde{\lambda}_t$. The variations of the simulated output are higher than estimated especially for slower demand oscillations. The input rate λ_t exaggerates the demand variations for medium and low frequencies of the demand oscillations. This is the same behavior as observed in clearing function models with macro-periods. Whether or not this is ”truly” optimal (that is, optimal given a perfect transient queueing model that accurately models the event oriented process) is not known and an important

Figure 10 **Demand, input rate and output for oscillating demand, low frequency**



research topic (see Missbauer (2009)).

[31]

6. Conclusions and Further Research

In this paper, we developed a proof-of-concept study where an order release model is based on a well-defined and tested approximation of a time-dependent queueing system. The study was motivated by the well-known limitations of order release models that consider lead times as fixed, exogenous parameters and by the theoretical requirement to base order release models on the theory of transient queueing systems. We showed that such a theoretically consistent procedure is feasible and leads to reasonable results.

We showed that the transient queueing approximation applied here - stationary backlog carry-over (SBC) developed in Stolletz (2008) - leads to a clearing function model with micro-periods and a strictly specified [32] clearing function without adjustable parameters except the period length.

[31] **Clear managerial insights from num. study HM: To be discussed**

[32] **strictly specified CF? HM: Ok this way?**

SBC in its original form requires the output to be defined by equality constraints. According to our preliminary tests, the results are very similar to the less-or-equal formulation commonly applied in clearing function models, and although the solution space is not necessarily convex for equality output constraints, we did not encounter any difficulties in solving the model.

Our numerical results indicate that the release model performs very well for constant demand or smooth demand curves. Sudden increase or decrease [33] of demand leads to errors in the output estimation (deviation of SBC vs. simulation) without any clear pattern concerning the sign of the deviation. The deviation tends to be larger for sudden decrease in demand than for sudden increase. As we know from clearing function models, sudden demand changes lead to release patterns that seems ”nervous”, and it is unknown to what extent this is an artifact that results from the approximation errors of the descriptive (SBC or clearing function) model. In the experiments, we observed some backlogging for longer time intervals, which indicates that either introducing slack (safety stock or safety time) or a possibility to adjust the output estimation in order to achieve the target compromise between inventory and due-date performance is necessary.

Substantial research efforts are necessary to make models of this type applicable in practice. Generalizing the assumptions of the underlying queueing model (see, e.g. Stolletz and Lagershausen (2013) for the SBC approximation of a $G(t)/G/1/K$ queue) is an important research topic as well as the integration of single-stage queueing models into network structures (e.g., job shops). Since clearing functions (and SBC as a special case) do not model the impact of the history of the process precisely, improvements in this respect are highly desirable. The ultimate goal are release models that are based on the theory of transient queueing networks, use sufficiently accurate approximations to this systems and allow some parametrization in order to incorporate trade-offs between different goals and details of the dispatching decisions that are difficult to observe. It seems there is still a long way to go.

[33] **explain or leave out this sentence HM: Ok this way?**

Acknowledgments

Gregor Sim

References

Abate J, Whitt W (1987) Transient behavior of the M/M/1 queue: Starting at the origin. *Queueing Systems* 2(1):41–65.

Abate J, Whitt W (1988) Transient Behavior of the M/M/1 Queue via Laplace Transforms. *Advances in Applied Probability* 20(1):145.

Abate J, Whitt W (1994) Transient Behavior of the M/G/1 Workload Process. *Operations Research* 42(4):750–764.

Asmundsson J, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning with resources subject to congestion. *Naval Research Logistics* 56(2):142–157.

Bertrand JWM, Wortmann JC (1981) *Production control and information systems for component-manufacturing shops* (Amsterdam-Oxford-New York: Elsevier).

Bertrand JWM, Wortmann JC, Wijngaard J (1990) *Production control : a structural and design oriented approach* (Amsterdam [u.a.]: Elsevier).

Carey M (1987) Optimal Time-Varying Flows on Congested Networks. *Operations Research* 35(1):58–69.

Chen H, Mandelbaum A (1994) Hierarchical Modeling of Stochastic Networks, Part I: Fluid Models. Yao DD, ed., *Stochastic Modeling and Analysis of Manufacturing Systems*, 47–105 (New York, NY: Springer New York).

Collings T, Stoneman C (1976) The M/M/∞ in Queue with Varying Arrival and Departure Rates. *Operations Research* 24(4):760–773.

De Kok TG, Fransoo JC (2003) Planning supply chain operations: Definition and comparison of planning concepts. De Kok TG, Graves SC, eds., *Supply Chain Management: Design, Coordination and Operation*, 597–675 (Amsterdam: Elsevier).

Dobson G, Karmarkar US (2011) Production Planning Under Uncertainty with Workload-Dependent Lead Times: Lagrangean Bounds and Heuristics. Kempf GK, Keskinocak P, Uzsoy R, eds., *Planning Produc-*

tion and Inventories in the Extended Enterprise: A State-of-the-Art Handbook, Volume 2, 1–14 (New York, NY: Springer).

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: tutorial, review, and research prospects Production planning with resources subject to congestion. *Manufacturing & Service Operations Management* 5(2):79141.

Green LV, Kolesar PJ, Soares J (2001) Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands. *Operations Research* 49(4):549–564.

Gross D, Miller DR (1984) The Randomization Technique as a Modeling Tool and Solution Procedure for Transient Markov Processes. *Operations Research* 32(2):343–361.

Hung YF, Leachman R (1996) A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. *IEEE Transactions on Semiconductor Manufacturing* 9(2):257–269.

Ingolfsson A, Akhmetshina E, Budge S, Li Y, Wu X (2007) A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems with Exhaustive Discipline. *INFORMS Journal on Computing* 19(2):201–214.

Ingolfsson A, Amanul Haque M, Umnikov A (2002) Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research* 139(3):585–597.

Irdem DF, Kacar NB, Uzsoy R (2010) An Exploratory Analysis of Two Iterative Linear Programming - Simulation Approaches for Production Planning. *IEEE Transactions on Semiconductor Manufacturing* 23(3):442–455.

Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server Staffing to Meet Time-Varying Demand. *Management Science* 42(10):1383–1394.

Kacar NB, Moench L, Uzsoy R (2013) Planning Wafer Starts Using Nonlinear Clearing Functions: A Large-Scale Experiment. *IEEE Transactions on Semiconductor Manufacturing* 26(4):602612.

Kacar NB, Uzsoy R (2010) Estimating Clearing Functions from Simulation Data. *Proceedings of the Winter Simulation Conference*, 1699–1710.

Kefeli A, Uzsoy R, Fathi Y, Kay M (2011) Using a mathematical programming model to examine the marginal price of capacitated resources. *International Journal of Production Economics* 131(1):383–391.

Kim B, Kim S (2001) Extended model for a hybrid production planning approach. *International Journal of Production Economics* 73(2):165–173.

Koopman BO (1972) Air-Terminal Queues under Time-Dependent Conditions. *Operations Research* 20(6):1089–1114.

Kushner HJ (2001) *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*, volume 47 (New York, NY: Springer).

Merchant DK, Nemhauser GL (1978) A Model and an Algorithm for the Dynamic Traffic Assignment Problems. *Transportation Science* 12(3):183–199.

Missbauer H (2009) Models of the transient behaviour of production units to optimize the aggregate material flow. *International Journal of Production Economics* 118(2):387–397.

Missbauer H (2011) Order release planning with clearing functions: A queueing-theoretical analysis of the clearing function concept. *International Journal of Production Economics* 131(1):399–406.

Missbauer H, Uzsoy R (2011) Optimization Models of Production Planning Problems. Kempf GK, Keskinocak P, Uzsoy R, eds., *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook, Volume 1*, 437–507 (Boston, MA: Springer US).

Newell GF (1971) *Applications of queueing theory* (London: Chapman and Hall).

Papadopoulos HT, Heavey C, Browne J (1993) *Queueing theory in manufacturing systems analysis and design* (London: Chapman and Hall), 1 edition.

Pürgstaller P, Missbauer H (2012) Rule-based vs. optimisation-based order release in workload control: A simulation study of a MTO manufacturer. *International Journal of Production Economics* 140(2):670–680.

Riaño G (2003) *Transient Behavior of Stochastic Networks: Application to Production Planning with Load-Dependent Lead Times*. Ph.D. thesis, Georgia Institute of Technology.

Schneeweiß C (2003) *Distributed decision making* (Berlin ; Heidelberg: Springer), 2 edition.

Schwarz JA, Selinka G, Stolletz R (2015) Performance analysis of time-dependent queueing systems: Survey and classification. *Omega* 1–51.

Stolletz R (2008) Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research* 190(2):478–493.

Stolletz R, Lagershausen S (2013) Time-dependent performance evaluation for loss-waiting queues with arbitrary distributions. *International Journal of Production Research* 51(5):1366–1378.

Upton RA, Tripathi SK (1982) An approximate transient analysis of the M(t)/M/1 queue. *Performance Evaluation* 2(2):118–132.

Zäpfel G, Missbauer H (1993) New concepts for production planning and control. *European Journal of Operational Research* 67(3):297–320.