Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

AN EXPLORATORY COMPARISON OF CLEARING FUNCTION AND DATA-DRIVEN PRODUCTION PLANNING MODELS

Karthick Gopalswamy
Reha Uzsoy

Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University
111 Lampe Dr,
Raleigh, NC 27695-7906, USA

ABSTRACT

Production planning models face the fundamental problem of capturing the nonlinear relation between resource workload and cycle times. One approach to this has been the use of nonlinear clearing functions that represent the expected output of a production resource in a planning period as a function of the expected workload in that period. Recently an alternative data driven approach that represents the behavior of the system using a set of system states and their corresponding output levels has been proposed. We compare the clearing function based approach to the data driven approach using a simulation model of a scaled-down semiconductor wafer fabrication facility and discuss the strengths and weaknesses of the two approaches.

1 INTRODUCTION

A central problem in production planning is that of determining the timing and quantity of releases of work into the production system to ensure that output matches demand in an optimal or near-optimal manner. This requires modeling the cycle times, the delay between work being released into the production system and its emergence as finished product that can be used to meet demand. Both queuing theory (Buzacott and Shanthikumar 1993) and simulation models (Atherton and Atherton 1995) have demonstrated a nonlinear relation between mean cycle time and mean resource utilization. In general, the cycle time of a job through a production system is a random variable whose distribution depends on the level of resource utilization, among other factors. However, resource utilization is determined by the workload, the amount of work available to a resource in a planning period, which is, in turn, determined by the release decisions. This circularity, where the cycle time depends on release decisions that themselves require knowledge of the cycle times, has been a persistent issue in production planning for several decades.

The most common approach to this problem in the literature has been the use of exogenous, workload-independent lead times to represent the delay between work being released into the system and its completion. The widely used Material Requirements Planning (MRP) approach (Vollmann et al. 2005) and most linear and integer programming models proposed for production planning (Hackman and Leachman 1989; Missbauer and Uzsoy 2011) take this approach. These lead time parameters can be implemented in different ways; the majority of models assume that material consumes production capacity in the final planning period of the lead time, while others (Spitter et al. 2005; Pürgstaller and Missbauer 2012) allow capacity to be consumed anywhere within the lead time interval. This approach generally yields computationally tractable models, but fails to account for the nonlinear relationship between workload and cycle time. Thus these models can give poor results, especially in situations where workload fluctuates over time due to time-varying demand, or under high utilization where small variations in workload can lead to large changes in cycle times.

A number of authors (Byrne and Bakir 1999; Hung and Leachman 1996; Kim and Kim 2001; Albey, Bilge, and Uzsoy 2014) have attempted to address this difficulty by decomposing the production planning

problem into two subproblems. The first of these takes fixed lead times as an input, and produces an optimal release schedule for these lead times. The second subproblem then attempts to estimate the average cycle times that would be incurred under the given release schedule, usually using a simulation model of the production system under consideration. The observed cycle times from the simulation model are then used to derive updated lead time estimates for the release planning model, and this iteration between the two models continues until some termination criterion is satisfied. However, the convergence of these methods is not well understood (Irdem, Kacar, and Uzsoy 2010), and their need to incorporate multiple replications of a potentially complex simulation model result in high computational burden.

An alternative approach has been the use of nonlinear clearing functions (CFs) to represent production resources such as machines. A CF is essentially a metamodel of a queue, describing the expected output of a production resource in a planning period as a function of its expected workload during the planning period (Missbauer and Uzsoy 2011). Initially proposed by Karmarkar (1989), Graves (1986) and Srinivasan et al. (1988), CFs can be incorporated into mathematical programming models to plan releases for production systems. The majority of the literature uses univariate CFs that are monotonically increasing and concave in the workload, yielding tractable optimization models with convex feasible regions. Initial models encountered difficulties in modeling production systems with multiple products, which have been largely, though not completely, addressed by the Allocated Clearing Function (ACF) formulation of Asmundsson et al. (2009). This formulation, which forms the basis of the work in this paper, results in a convex optimization problem when used with a concave univariate CF. In particular, when the CF is approximated by piecewise linearization, a linear programming (LP) model is obtained. A number of computational studies (Kacar, Irdem, and Uzsoy 2012; Kacar, Monch, and Uzsoy 2013; Kacar and Uzsoy 2014; Haeussler and Missbauer 2014; Pürgstaller and Missbauer 2012) have shown that when appropriately parameterized, production planning models using CFs can outperform fixed lead time models, especially under time-varying workloads. The estimation of the CFs from empirical data is a subject of ongoing research, but is most commonly accomplished by simulating the production system of interest off-line to obtain observations of the workload and output under a range of operating conditions, and then using some form of regression to fit the CF. A number of studies (Albey, Bilge, and Uzsoy 2014; Haeussler and Missbauer 2014; Kacar and Uzsoy 2014; Kang, Albey, Hwang, and Uzsoy 2014) have examined multivariate CFs. While the issues in fitting these are on the whole similar to those for univariate CFs, although computationally more demanding, these CFs result in non-convex optimization models whose solution is computationally demanding.

The CF as widely implemented in the literature can be viewed as a parametric approach to describing the capabilities of the production resource it represents. The usual manner of fitting a CF is to postulate a specific functional form, whether univariate or multivariate, and then obtain estimates of its parameters from empirical data that is usually obtained by simulating the production resource of interest under a range of operating conditions. The CF approach is also aggregate in nature, since the CF estimates the total output of the production resource in a planning period in units of time, and the ACF model allocates this aggregate output among the different products in each period.

Recently Omar et al. (2017) have proposed an innovative approach in which the expected output of the production resource is estimated for a range of possible system states, described by the expected work in process (WIP) level of each product. An optimization model then determines the releases of each product in each period to ensure a system state that will yield the minimum total cost over the planning horizon. In contrast to the CF-based approach, this can be viewed as a data driven non-parametric approach, since it does not involve estimation of parameters for a pre-specified functional form. It is also a disaggregated approach, since it is based on the expected output of each product in each system state, as opposed to the aggregated output of the resource as in the ACF approach. We shall refer to this approach as the Data-Driven (DD) approach for the remainder of the paper.

In this paper we compare the performance of the ACF formulation and the DD approach of Omar et al. (2017) on a production system representing a scaled-down semiconductor wafer fabrication facility extensively proposed by Kayton et al. (1997) and extensively studied in previous work (Irdem et al. 2010;

Kacar et al. 2012). In the following section we describe the two planning approaches. Section 3 describes the scaled-down semiconductor wafer fabrication system used as the testbed, the simulation model we use to evaluate the performance of the two planning models, and the procedures by which the CF (required by the ACF model) and the system state descriptions for the DD model are obtained. Our computational experiments and their results are discussed in Section 4. We conclude the paper with a summary of our principal findings and some directions for future research.

2 PRODUCTION PLANNING MODELS

2.1 The Allocated Clearing Function (ACF) Model

The CF-based model we use is the Allocated Clearing Function (ACF) model of Asmundsson et al. (2009), which was the first effective multiproduct formulation using CFs. To describe this model we define an operation as the processing of a product at a specific workcenter on its process routing. The index l will denote the l'th operation of a product on its routing, so each specific operation of a product g will be denoted by gl. A workcenter is a set of machines or resources capable of processing a specified set of operations. Let $L(k)$ denote the set of all operations taking place on a specific workcenter, and $k(gl)$ the workcenter where the l'th operation of product g takes place. In our testbed all operations processed on a workcenter k require the same processing time, allowing us to measure workload in number of units of product present. We define the following notation:

Indices:

t	Period index.
g	Product index.
k	Workcenter index.
l	Operation index.

Variables:

Y_{gtl}	Quantity of product g completing its l^{th} operation in period.
Y^k_{gt}	Total output of all operations of product g from workcenter k in period t; $Y^k_{gt} = \sum_{l \in L(k)} Y_{gtl}$.
X_{gtl}	Amount of product g released to the l^{th} operation on its route in period t. We assume no strategic inventory can be held in the production line, and material that completes processing at one operation is immediately transferred to the next operation in its routing, or to finished inventory if the operation is the last in the routing. Hence, $X_{gtl} = Y_{g,t,l-1}$, where operation $l-1$ is immediately precedes operation l in the routing of product g. We also assume that each operation will have exactly one predecessor and one successor in its routing, although our formulations are easily extended to relax this assumption.
X^k_{gt}	Total amount of product g released to workcenter k in period t, given by $X^k_{gt} = \sum_{l \in L(k)} X_{gtl}$.
Y_{gt}	Output of completed units of product g in period t.
X_{gt}	Quantity of product g released into the first operation on its routing in period t.
W_{gtl}	WIP of product g at its l'th operation at the end of period t.
W^k_{gt}	WIP of product g during period t at workcenter k, given by $W^k_{gt} = \sum_{l \in L(k)} W_{gtl}$.
I_{gt}	Units of product g in finished goods inventory at the end of period t.
B_{gt}	Backlog of product g at the end of period t.
Z^k_{gtl}	Fraction of workcenter k's output allocated to the l'th operation of product g in period t.

Parameters:

h_{gt}	Unit inventory holding cost of product g in period t.
b_{gt}	Unit backlogging cost of product g in period t.

ω_{gt}	Unit WIP holding cost of product g in period t.
D_{gt}	Demand for product g in period t.
$C(k)$	Index set of the line segments used to approximate the CF describing workcenter k.
μ_n^k	Intercept of the n^{th} linear segment of the CF describing workcenter k.
β_n^k	Slope of the n^{th} linear segment of the CF describing workcenter k.

The ACF formulation is given as follows:

$$\min \quad \sum_{g\in G}\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\omega_{gt}W_{gt}^{k}+\sum_{t=1}^{T}h_{gt}I_{gt}+\sum_{t=1}^{T}b_{gt}B_{gt}\right] \tag{1a}$$

s.t

$$W_{gtl}=W_{g,t-1,l}+X_{gtl}-Y_{gtl} \quad \forall g\in G,\ t=1,...,T,\ l\in L \tag{1b}$$

$$I_{gt}=I_{g,t-1}+Y_{gt}-B_{g,t-1}+B_{gt}-D_{gt} \quad \forall g\in G,\ t=1,...,T \tag{1c}$$

$$Y_{gtl}\le Z_{gtl}^{k}\mu_n^k+\beta_n^k(X_{gtl}+W_{g,t-1,l}) \quad \forall g\in G,\ t=1,...,T,\ l\in L,\ k\in K(l),\ n\in C(k) \tag{1d}$$

$$\sum_{g\in G,\ l\in L(k)} Z_{gtl}^{k}=1, \quad \forall t=1,...,T,\ k\in K \tag{1e}$$

$$W_{gtl},Y_{gtl},I_{gt},B_{gt},X_{gtl},Z_{gtl}^{k}\ge 0, \quad \forall g\in G,\ t=1,...,T,\ k\in K,\ l\in L \tag{1f}$$

The objective function (1a) minimizes the sum of WIP holding, finished inventory holding and backorder costs for all products and operations over the planning horizon. Constraints (1b) and (1c) are material balance equations for the WIP at each operation and the finished inventory of each product, respectively. Constraints (1f) restrict the output of each operation with the allocated clearing function, while constraints (1e) ensure that the total output allocated to all operations processed at a workcenter do not exceed the total output compatible with the CF. A detailed derivation and motivation of this formulation are given by Asmundsson et al. (2009), to which the interested reader is referred.

2.2 Data Driven (DD) Model

The DD model takes a rather different approach from ACF by defining a set R of system states, each of which specifies the average WIP level of each product during a planning period and the expected output of each product in the period compatible with the specified WIP levels. The WIP levels are specified for the entire system, not for individual workcenters. In principle, any one of several different methods could be used to estimate the expected output of each product g in a system state r, including queuing models, simulation or empirical observation on the factory floor. The authors elect to use the Mean Value Analysis (MVA) approach of Suri and Hildebrant (1984), details of which are given in Omar et al. (2017). The behavior of the production system is described by a suitably representative set R of system states $r\in R$. The planning model then selects a system state for each period to minimize the total cost over the planning horizon, selecting release quantities for each product in each period and enforcing material balance for both WIP and finished goods inventories. We define the following additional notation for the DD model:

Sets:

R	set of system states used to describe the behavior of the production system.

Indices:

r	state index.

Variables:

$$\Gamma_{rt} = \begin{cases} 1, & \text{if system state } r \text{ is chosen in period } t \\ 0, & \text{otherwise} \end{cases}$$

Parameters:

Q_{gr}	WIP level of of product g in system state r.
O_{gr}	Expected throughput of product g in system state r.

The DD model is given as follows:

$$\min \quad \sum_{g\in G}\sum_{t=1}^{T}[\omega_{gt}W_{gt}+h_{gt}I_{gt}+b_{gt}B_{gt}] \tag{2a}$$

s.t

$$W_{gt}=W_{g,t-1}+X_{gt}-Y_{gt} \quad \forall g\in G,\ t=1,...,T \tag{2b}$$

$$I_{gt}=I_{g,t-1}+Y_{gt}-B_{g,t-1}+B_{gt}-D_{gt} \quad \forall g\in G,\ t=1,...,T \tag{2c}$$

$$W_{gt}=\sum_{r=1}^{R}Q_{gr}\Gamma_{rt} \quad \forall g\in G,\ t=1,...,T \tag{2d}$$

$$Y_{gt}=\sum_{r=1}^{R}O_{gr}\Gamma_{rt} \quad \forall g\in G,\ t=1,...,T \tag{2e}$$

$$\sum_{r=1}^{R}\Gamma_{rt}=1, \quad \forall t=1,...,T \tag{2f}$$

$$W_{gt},Y_{gt},I_{gt},B_{gt},X_{gt}\geq 0, \quad \forall g\in G,\ t=1,...,T \tag{2g}$$

$$\Gamma_{rt}\in 0,\ 1, \quad \forall r\in R,\ t=1,...,Ti \tag{2h}$$

The objective function is the same as that of the ACF model, minimizing the sum of WIP holding, finished goods holding and backorder costs. Constraints (2b) and (2c) again represent material balance equations for WIP and finished goods inventory, respectively. In contrast to the WIP balance constraint (1b) in the ACF formulation, constraint (2b) represents WIP across the entire production system in an aggregate manner, as opposed to capturing WIP balance for each individual operation of each product. Constraints (2d) require the total WIP of product g in the system in period t to be that implied by the selected system state, and constraints (2e) define the expected throughput of each product similarly. The WIP balance constraints (2b) ensure that the releases X_{gt} of each product into the system in each period are consistent with the WIP levels implied by the system states in consecutive periods. Finally, constraints (2f) ensure that exactly one system state $r\in R$ is selected for each planning period t.

By representing only the aggregate WIP in the entire system, without considering its distribution across the workcenters, the DD model requires far fewer constraints and continuous decision variables than the ACF formulation. However, the ability of this model to represent system behavior accurately clearly depends on appropriate selection of the set R of system states, which determines the number of binary variables Γ_{rt}. By using the MVA algorithm to estimate the expected throughput O_{gr} in each state r from the WIP levels Q_{gr}, the authors are assuming that the planning periods are sufficiently long that the queues in the system will have attained steady state, and hence that the distribution of the WIP across the workcenters is the steady-state joint distribution under the assumptions of the MVA analysis, which are a closed queuing system with exponential service times. In their computational experiments Omar et al. (2017) use the same testbed as this paper, but modify it in several ways: they assume exponential service time distributions, ignore failures, and approximate batch processing machines by replacing them with parallel identical machines. However, they do not evaluate the performance of their planning approach using simulation under the original data of the Kayton et al. (1997) testbed. They compare the performance

of the DD and ACF models based on objective function value, which is difficult to interpret due to the different assumptions made by the models.

In the following sections we compare the performance of the two planning models using simulation, which allows us to observe the realized performance of the production system under the plans proposed by the different models. We first briefly describe the testbed production system and the simulation model, followed by our approach to estimating the different parameters required by the two planning models.

3 SIMULATION MODEL AND ESTIMATION

Our simulation model represents a reentrant bottleneck system built by Kayton et al. (1997) with attributes of a real-world semiconductor wafer fab and studied extensively in Kacar et al. (2012), Kacar et al. (2013), Kacar and Uzsoy (2014), and is illustrated in Figure 1. Each row represents the sequence of workcenters visited by each product; thus, for example, Product 1 first visits workcenter 1, then proceeds to workcenter 4, then to workcenter 3, and so on. The principal characteristics of wafer fabrication, including a reentrant bottleneck, unreliable machines, batch processing machines, and multiple products with different routings are included in the model. There is a distinct reentrant bottleneck workcenter representing the photolithography process. The processing times for all other workcenters are scaled to the bottleneck processing time so that no non-bottleneck workcenter has utilization close to that of the bottleneck. This is not always realistic, but simplifies the experimental design. The model has 11 workcenters (tool groups) and 3 products. The number of operations for products 1, 2, and 3 are 22, 14, and 14, respectively. Workcenter 4 is the bottleneck and has two servers, while all other workcenters have a single server. Products 1 and 2 visit the bottleneck workcenter 6 and 4 times, respectively. Product 3 does not use the bottleneck workcenter, instead visiting Workcenter 11, the only workcenter that exceeds 80% utilization without exceeding the bottleneck utilization. Workcenters 3 and 7 are unreliable machines whose failures cause starvation at the bottleneck. Workcenters 3 and 7 has an availability of 0.8 based on their failure distributions given in Table 2. All processing times in Table 1 follow a lognormal distribution and are given in minutes. The standard deviations of the processing times are less than or equal to 10% of the mean except 1 and 2. The processing times for all products on a given workcenter are the same, and we assume instantaneous material transfer between consecutive operations on a routing. Hence the processing variability is due mainly to down times at the unreliable machines. The time to failure and time to repair follow gamma distributions, with parameters given in Table 2.

The rather restrictive assumptions required by the MVA model require some modifications to the original system data to ensure a more even comparison between the two planning models. Workcenters 1 and 2 are batching (bulk service) machines in the system considered in Kacar, Irdem, and Uzsoy (2012), but since the MVA algorithm does not model batch processors , we reduce the processing times at workcenters 1 and 2 by a factor of 4 (the maximum batch size in the original system) to have a fair comparison to the DD model. For the unreliable machines, the effective number of servers in the MVA algorithm is multiplied by the mean availability of the workcenter, given by

$$A = \frac{\text{Mean time to fail}}{\text{Mean time to fail + Mean time to repair}}.$$

We now discuss our implementation of the two planning models, specifically the fitting of the CFs required by the ACF model and the selection and characterization of the set R of system states for the DD model.

3.1 Estimating Clearing Functions

In order to implement the ACF formulation we need to fit a CF to each workcenter in the production system. The Load-based CF we use in this paper aggregates releases and initial WIP into a single state variable, treating the total output Y_{kt} of workcenter k in period t as a function of the sum of releases made to that workcenter in that period and the WIP available to the workcenter at the end of the preceding period, which we shall refer to as the *workload* of workcenter k in period t given by $\Lambda_t^k = \sum_{g \in G} W_{g,t-1}^k + X_{gt}^k$ and a CF $f_k(\Lambda_t^k)$.

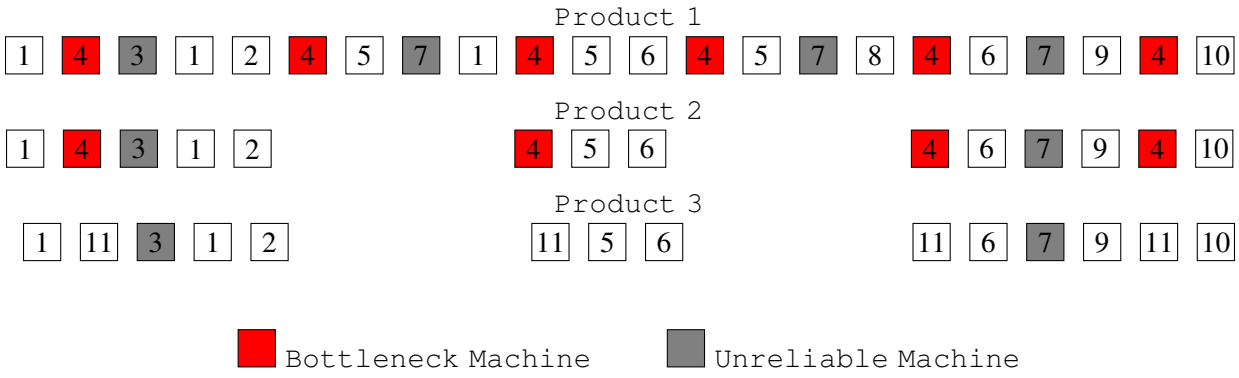

Figure 1: Reentrant bottleneck model process flows.

Table 1: Processing Time Distributions (lognormal distributed).

Workcenter #	Mean	Std.Dev	# of machines
1	20	3.5	1
2	55	8	1
3	45	4	1
4	40	4	2
5	25	2	1
6	22	2.4	1
7	20	2	1
8	100	12	1
9	50	4	1
10	50	5	1
11	70	2.5	1

Since in our simplified model all operations at a workcenter have the same processing time distribution regardless of product type or operation, W_{t-1}^k represents the total number of lots in WIP at the beginning of the period at workcenter k and X_t^k the total number of lots released to workcenter k in period t from preceding operations or as new starts into the fab. Λ_t^k thus represents the total amount of work available to the workcenter during the planning period. Hence, in regression terms, the independent variables (covariates) are the total releases X_t^k and initial WIP W_{t-1}^k for each period $t = 1, ..., T$ in the planning horizon; the dependent variables are the expected output Y_t^k. We define $\sum_{g \in G} X_{gt}^k = X_t^k$, $\sum_{g \in G} W_{g,t-1}^k = W_{t-1}^k$, and $\sum_{g \in G} Y_{gt}^k = Y_t^k$. The procedure used to generate the data is as follows:

Step 1: Seven different demand realizations specifying values of D_{gt} for all products g and periods t are randomly generated assuming normally distributed (truncated) demand in each period. The demand distributions are adjusted to obtain average bottleneck utilization levels of 49%, 60%, 70%, 77%, 87%, 94%, and 99%. A release schedule is then obtained by setting the release quantity for each product in each period to the demand for that product in that period, i.e., $X_{gt} = D_{gt}$

Step 2: For each release schedule, five independent replications of the simulation model are run for 26 periods, collecting observations of the Y_{gt}^k, X_{gt}^k and $W_{g,t-1}^k$ for each workcenter k in each period t. Note that the uncertainty in the simulation is associated with workcenter failures and processing times; demand is assumed to be deterministic and known.

Step 3: Data from all replications is combined and plotted for each workcenter to allow visual inspection.

Table 2: Failure Distribution Parameters (minutes).

	TTF				TTR			
Workcenter #	Alpha	Beta	Mean	Std.Dev	Alpha	Beta	Mean	Std.Dev
3	7200	1	7200	84.9	1200	1.5	1800	52.0
7	7200	1	7200	84.9	1200	1.5	1800	52.0

For the fitting step, we use a piecewise linear concave fitting optimization problem for each workstation k to obtain the slopes and intercepts as defined in the ACF model. The number of observations for fitting is $m = 7 \times 5 \times T$. Thus for each workcenter k we solve the following optimization problem:

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} \left| f_i - Y_i^k \right| \\ s.t \quad & f_i = \min_{n=1..p} \{\beta_n^k \Lambda_i^k + \mu_n^k\} \quad \forall i \in 1..,m \\ & f_i \in \mathbb{R}, \quad \beta^k \in \mathbb{R}^p, \quad \mu^k \in \mathbb{R}^p \end{aligned}$$

We formulate this problem as a binary mixed integer linear program following Toriello and Vielma (2012), to which we add the valid inequalities proposed by Gopalswamy, Fathi, and Uzsoy (2018). The models are solved using Gurobi 7.2 on an Intel Xeon 3.4 GHz system with 64 GB of RAM. Details of this fitting problem are given in Gopalswamy, Fathi, and Uzsoy (2018). The total number of data points collected with the above steps is 9100, which is larger than the model can handle in short CPU times. We reduce the data by taking the mean of Y_i^k for each unique level of Λ_i^k observed.

3.2 Collecting and Characterizing System States for the DD Model

The central issue in implementing the DD model is specifying an appropriate set R of system states. Omar et al. (2017) use long run simulation with releases equal to the demands to estimate the average WIP of each product in the system, which is then used as input to the MVA algorithm. In this work, we design the set R based on the WIP patterns observed during the data collection for the CF estimation described above. We average the WIP of each product in each period across the simulation replications used to estimate the CF to obtain the average WIP estimate for the MVA algorithm. These WIP levels are computed at every utilization level to provide a WIP pattern which is fed into MVA algorithm to obtain the estimated expected throughput. We have 7 utilization levels and $T = 26$ periods which gives 182 WIP patterns to which we add the trivial pattern of $(0,0,0)$. The cost values used for both the models are $\omega_{gt} = 35$, $h_{gt} = 15$ and $b_{gt} = 50$. Both the models begin with zero initial WIP and backlog while the initial finished inventory for each product set to $(1/2T)\sum_{t=1}^{T} D_{gt}$.

An interesting observation about both the models is the relationship to the cycle time. The cycle times do not appear directly in either model, but are represented indirectly. In the ACF model the planned average time of a product at a resource can be inferred using Little's Law, assuming the queues within the planning period are in steady state. The concavity of the CF implies an increasing lead time as workload increases. The average cycle time under the DD model in a planning period is directly related to the system state selected by the model in that period. Each state corresponds to a closed queuing system whose average cycle time is determined by the product mix and WIP levels. As the DD model selects different states in different periods, the average cycle time for each product will also vary.

4 COMPUTATIONAL EXPERIMENTS

Our computational experiments examine the performance of the ACF and DD models using simulation. We consider a planning horizon of $T = 26$ periods, each of length 1440 minutes. Separate, independent data sets are used to fit the CFs and to evaluate their performance. Demand in each period follows a

continuous uniform distribution $U(a,b)$ with parameters selected to achieve the utilization levels specified in the experimental design.

Table 3: Experimental Design.

Factor	**Values**	**Levels**
Bottleneck utilization	70% and 90%	2
Demand CV	0.1 and 0.3	2
Failure Distribution	None and Failure (Table 2)	2
Demand Replications		5
Simulation Replications per Release Plan		10

Our computational experiments proceed as follows:

1. Solve the ACF and DD models using the estimates of their data described to obtain a planned release schedule.
2. Simulate the operation of the fab under this release plan for 10 independent replications.
3. Compute the expected profit value from the outcomes of the simulation replications.

Since exact solution of the DD model (2a) - (2h) takes a long time, we impose a 500 second time limit on the solver based on preliminary experiments, in which we observed an optimality gap of 5% in under 500 seconds but about 3 hours was needed to reduce it to 3%.

5 RESULTS AND DISCUSSION

We begin with a comparison of the computation time for the ACF and DD models. Due to the high runtime required to run the DD model to optimality, we present the solution times for a single instance with bottleneck utilization of 70%, demand CV 0.1 % and no failures. Table 4 shows that the solution time for the DD model is several orders of magnitude higher than that required for the ACF model, all instances of which were solved in less than a minute. This seriously impacts the applicability of the DD model for large production systems and limits our ability to chose a finer grid of WIP levels to obtain refined solutions. Hence the remaining experiments were run with the 500 second CPU limit discussed above.

Table 4: Solution time comparison (in hrs).

Demand seed:	1	2	3	4	5
ACF:	0.0056	0.0094	0.0122	0.0094	0.0064
DD:	31.92	10.90	14.87	5.30	2.37

The release patterns from the ACF and DD models are simulated for each design point for 10 replications and the average total cost of operating the system is presented in Table 5 and Table 6 for the cases without and with failures, respectively. We also present the worst case results in Table 7 and Table 8.

Table 5: Average cost comparison: No failure.

Util	CV	Demand Seed 1		2		3		4		5	
		ACF	DD	ACF	DD	ACF	DD	ACF	DD	ACF	DD
0.7	0.1	43027.5	40238.5	46377	37909	46586.5	40668.5	45079	35048	45738.5	34348.5
	0.3	41335.5	36775.5	40005	34524.5	37273.5	30619	40807	30641	40102	35824
0.9	0.1	44796	51226.5	42989.5	42530.5	43459	40372.5	44561.5	42789.5	48400	40511
	0.3	114415	206595.5	48824.5	60549	50970.5	42542	61709	79395	58734.5	64743

Table 6: Average cost comparison: Failure.

Util	CV	Demand Seed 1		2		3		4		5	
		ACF	DD	ACF	DD	ACF	DD	ACF	DD	ACF	DD
0.7	0.1	44999	51160	43081.5	50417	47315.5	53547.5	47908.5	50133.5	46317	50594.5
	0.3	48292	58157	45715.5	53491	45695	42533	48459	50739.5	48044.5	53698.5
0.9	0.1	69867	79638	70873.5	71089.5	61821.5	61742	71923	67565.5	68683.5	69578
	0.3	173132	213445	79301	85464.5	64516	62265	92331	108785.5	91727.5	93597

Table 7: Worst case cost comparison: No failure.

Util	CV	Demand Seed 1		2		3		4		5	
		ACF	DD	ACF	DD	ACF	DD	ACF	DD	ACF	DD
0.7	0.1	43378	40527	46547	38346	46748	41008	45270	35331	46036	34703
	0.3	41507	37216	40362	34661	37488	30804	41147	30754	40593	35965
0.9	0.1	45170	51506	43268	42871	43714	40596	44889	43218	48687	40692
	0.3	115493	207050	49131	60957	51112	42812	61986	80005	58937	65084

Table 8: Worst case cost comparison: Failure.

Util	CV	Demand Seed 1		2		3		4		5	
		ACF	DD	ACF	DD	ACF	DD	ACF	DD	ACF	DD
0.7	0.1	47080	51924	43836	51879	48131	54656	49536	51618	48291	51265
	0.3	49624	59791	46558	55000	46682	44008	49114	52337	49058	54794
0.9	0.1	72491	82822	75985	75244	63654	65485	74881	70453	73094	74629
	0.3	175208	222203	81929	89376	65419	65583	96232	112846	94031	96225

With no failures, at low utilization, the DD model outperforms the ACF model for all demand seeds and both levels of the demand CV. However, at high utilization the performance of ACF improves considerably, yielding comparable, though slightly worse, solutions than DD when CV = 0.1, and better solutions for four out of five demand seeds for CV = 0.3. It is interesting that DD fails quite badly in demand seeds 1, 2 and 4 for CV = 0.3, but ACF can fail quite badly at low utilization levels. In the presence of failures, ACF outperforms DD in all cases except demand seed 3, where DD performed slightly better than ACF for all four cases.

We analyze the results for demand seed 2 in the no failure case to obtain further insights. At $u = 0.7$ and CV = 0.1, the average workloads of the ACF and DD models over the periods are (47.86, 15.97,15.98) and (54.91, 17.38, 17.23), respectively, whereas at $u = 0.9$ and CV = 0.3, the average workload are (61.81, 20.60, 20.59) and (77.96, 24.41, 22.85), respectively. The DD model has higher workload than the ACF model in both scenarios. While at lower utilization and under less variable demand the DD model yields lower cost, it does not perform well at higher utilization. At lower levels of utilization and variability higher workload will result in higher output, but at higher utilization overloading the system causes congestion and increases the cycle time. This is reflected in average cost of DD model which is significantly higher than the ACF model as seen in Table 5.

6 CONCLUSIONS AND FUTURE DIRECTIONS

These admittedly limited results suggest that the ability of the DD model to accurately capture the performance of multistage multi-item production systems is limited when the MVA algorithm is used to estimate the

expected output of a given WIP state. At low utilization and demand CV, DD yields better solutions than ACF within a CPU time limit of 500 s in the absence of failures. With failures implying increased variability of the effective processing times, ACF yields markedly better solutions in a fraction of the CPU time required by DD. Both approaches require extensive off-line computational effort, to fit the CFs (for the ACF model) and to estimate the throughput for the different WIP configurations for the DD model. In addition, exact solution of the DD model is computationally very demanding in comparison to the modest CPU requirements for the ACF, which is a LP.

Despite these apparent limitations of the DD approach, however, it is important to bear in mind that the ACF model has had the benefit of more than a decade of research work by several different research groups. DD is an interesting and innovative approach that deserves closer study. There is no reason the MVA approach, with its limiting assumptions, has to be used to evaluate the system states used by DD; other, more flexible queuing models, or even terminating simulation runs could be used. It is also very likely that exact solutions to the DD model are not necessary to obtain good production plans once a suitable set of system states is used. How to identify a minimal set of system states sufficient to obtaining good production plans is an important research direction.

REFERENCES

Albey, E., Ü. Bilge, and R. Uzsoy. 2014. "An Exploratory Study of Disaggregated Clearing Functions for Production Systems with Multiple Products". *International Journal of Production Research* 52(18):5301–5322.

Asmundsson, J., R. L. Rardin, C. H. Turkseven, and R. Uzsoy. 2009. "Production Planning with Resources subject to Congestion". *Naval Research Logistics* 56(2):142–157.

Atherton, L. F., and R. W. Atherton. 1995. *Wafer Fabrication: Factory Performance and Analysis*, Volume 339. New York City, NY: Springer Science & Business Media.

Buzacott, J. A., and J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*. Englewood Cliffs, NJ: Prentice Hall.

Byrne, M., and M. A. Bakir. 1999. "Production Planning using a Hybrid Simulation–Analytical Approach". *International Journal of Production Economics* 59(1):305–311.

Gopalswamy, K., Y. Fathi, and R. Uzsoy. 2018. "Valid Inequalities for Concave Piecewise Linear Regression". *NCSU ISE. Working paper*.

Graves, S. C. 1986. "A Tactical Planning Model for a Job Shop". *Operations Research* 34(4):522–533.

Hackman, S. T., and R. C. Leachman. 1989. "A General Framework for Modeling Production". *Management Science* 35(4):478–495.

Haeussler, S., and H. Missbauer. 2014. "Empirical Validation of Meta-models of Work Centers in Order Release Planning". *International Journal of Production Economics* 149:102–116.

Hung, Y.-F., and R. C. Leachman. 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.

Irdem, D. F., N. B. Kacar, and R. Uzsoy. 2010. "An Exploratory Analysis of Two Iterative Linear ProgrammingSimulation Approaches for Production Planning". *IEEE Transactions on Semiconductor Manufacturing* 23(3):442–455.

Kacar, N. B., D. F. Irdem, and R. Uzsoy. 2012. "An Experimental Comparison of Production Planning using Clearing Functions and Iterative Linear Programming-Simulation Algorithms". *IEEE Transactions on Semiconductor Manufacturing* 25(1):104–117.

Kacar, N. B., L. Monch, and R. Uzsoy. 2013. "Planning Wafer Starts using Nonlinear Clearing Functions: A Large-Scale Experiment". *IEEE Transactions on Semiconductor Manufacturing* 26(4):602–612.

Kacar, N. B., and R. Uzsoy. 2014. "A Comparison of Multiple Linear Regression Approaches for Fitting Clearing Functions to Empirical Data". *International Journal of Production Research* 52(11):3164–3184.

Kang, Y., E. Albey, S. Hwang, and R. Uzsoy. 2014. "The Impact of Lot-Sizing in Multiple Product Environments with Congestion". *Journal of Manufacturing Systems* 33(3):436–444.

Karmarkar, U. S. 1989. "Capacity Loading and Release Planning with Work-In-Progress (WIP) and Leadtimes". *Journal of Manufacturing and Operations Management* 2(105-123).

Kayton, D., T. Teyner, C. Schwartz, and R. Uzsoy. 1997. "Focusing Maintenance Improvement Efforts in a Wafer Fabrication Facility Operating under the Theory of Constraints". *Production and Inventory Management Journal* 38(4):51.

Kim, B., and S. Kim. 2001. "Extended Model for a Hybrid Production Planning Approach". *International Journal of Production Economics* 73(2):165–173.

Missbauer, H., and R. Uzsoy. 2011. "Optimization Models of Production Planning Problems". In *Planning Production and Inventories in the Extended Enterprise*, 437–507. Springer.

Omar, R. S. M., U. Venkatadri, C. Diallo, and S. Mrishih. 2017. "A Data-Driven Approach to Multi-Product Production Network Planning". *International Journal of Production Research* 55(23):7110–7134.

Pürgstaller, P., and H. Missbauer. 2012. "Rule-based vs. Optimization-based Order Release in Workload Control: A Simulation Study of an MTO Manufacturer". *International Journal of Production Economics* 140:670–680.

Spitter, J., C. A. Hurkens, A. De Kok, J. K. Lenstra, and E. G. Negenman. 2005. "Linear Programming Models with Planned Leadtimes for Supply Chain Operations Planning". *European Journal of operational research* 163(3):706–720.

Srinivasan, A., M. Carey, T. E. Morton et al. 1988. "Resource Pricing and Aggregate Scheduling in Manufacturing Systems". Technical report, Carnegie Mellon University, Tepper School of Business.

Suri, R., and R. R. Hildebrant. 1984. "Modelling Flexible Manufacturing Systems using Mean-Value Analysis". *Journal of Manufacturing Systems* 3(1):27–38.

Toriello, A., and J. P. Vielma. 2012. "Fitting Piecewise Linear Continuous Functions". *European Journal of Operational Research* 219(1):86 – 95.

Vollmann, T., W. Berry, D. Whybark, and F. Jacobs. 2005. *Manufacturing Planning and Control Systems for Supply Chain Management*. New York City, NY: McGraw-Hill.

AUTHOR BIOGRAPHIES

KARTHICK GOPALSWAMY is a Doctoral student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He holds a Masters in Industrial and Systems Engineering from North Carolina State University. His research focuses on data oriented decision analysis in production systems using stochastic simulation. His e-mail address is kgopals@ncsu.edu.

REHA UZSOY is Clifton A. Anderson Distinguished Professor in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He holds BS degrees in Industrial Engineering and Mathematics and an MS in Industrial Engineering from Bogazici University, Istanbul, Turkey. He received his Ph.D. in Industrial and Systems Engineering in 1990 from the University of Florida. His teaching and research interests are in production planning, scheduling, and supply chain management. He was named a Fellow of the Institute of Industrial Engineers in 2005, Outstanding Young Industrial Engineer in Education in 1997, and has received awards for both undergraduate and graduate teaching. His email address is ruzsoy@ncsu.edu.