

This article was downloaded by: [West Virginia University]

On: 15 December 2010

Access details: Access Details: [subscription number 918546920]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



International Journal of Production Research

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713696255>

Capacity planning through queueing analysis and simulation-based statistical methods: a case study for semiconductor wafer fabs

Jingang Liu^a; Feng Yang^a; Hong Wan^b; John W. Fowler^c

^a Industrial and Management Systems Engineering, West Virginia University, Morgantown, West Virginia 26505, USA ^b Purdue University, West Lafayette, Indiana, USA ^c Department of Industrial Engineering, Arizona State University, Tempe, AZ 85287-5906, USA

First published on: 12 October 2010

To cite this Article Liu, Jingang , Yang, Feng , Wan, Hong and Fowler, John W.(2010) 'Capacity planning through queueing analysis and simulation-based statistical methods: a case study for semiconductor wafer fabs', International Journal of Production Research,, First published on: 12 October 2010 (iFirst)

To link to this Article: DOI: 10.1080/00207543.2010.501828

URL: <http://dx.doi.org/10.1080/00207543.2010.501828>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

RESEARCH ARTICLE

Capacity planning through queueing analysis and simulation-based statistical methods: a case study for semiconductor wafer fabs

Jingang Liu^a, Feng Yang^{a*}, Hong Wan^b and John W. Fowler^c

^aIndustrial and Management Systems Engineering, West Virginia University, Morgantown, West Virginia 26505, USA; ^bPurdue University, West Lafayette, Indiana, USA; ^cDepartment of Industrial Engineering, Arizona State University, Tempe, AZ 85287-5906, USA

(Received 14 November 2009; final version received 5 June 2010)

This paper presents a comprehensive framework for the strategic capacity expansion of production equipment in semiconductor manufacturing, and the proposed approach is applied to a model of an actual wafer fabrication facility. It is the intention of this work to show that, once intelligently integrated, an analytical queueing model and a numeric computer simulation model can be used synergistically and can lead to a better alternative method than methods restricted to only one of them. The outcome of our methods is a number of good system configurations, each of which is characterised by its cycle time (CT)–throughput (TH) profile. Such profiles fully describe the system's comprehensive performance over a wide range of demand scenarios (involving varying product mix), and hence can be used to thoroughly evaluate alternative configurations in capacity expansion decisions.

Keywords: capacity planning; discrete event simulation; semiconductor manufacture; scheduling; batch scheduling

1. Introduction

With the cost of new wafer fabrication (fab) facilities now well in excess of 3.5 billion U.S. dollars, capacity planning decisions are more important than ever to the success of semiconductor firms (Chen and Chen 2010). The initial investment for building a wafer fab and installing the initial equipment is close to a few billion dollars, and, in addition, every year tool and equipment procurement could cost tens of millions of dollars per facility. This paper is concerned with capacity expansion decisions that intend to make the best use of a given budget to satisfy customer demand in the future.

Capacity planning for wafer fabs is difficult, mainly due to the volatility of customer demand and the complexity/variability inherent in fab systems (e.g., complex product flows, diverse equipment characteristics, downtime, etc.). Recent work by Geng and Jiang (2009) provides a thorough review of the current research methods in capacity planning, and here we briefly discuss these existing methods. In industry, the most widely used approach involves spreadsheet models (Ozturk *et al.* 2003), which are deterministic and cannot accommodate the stochastic behaviour of wafer fabs. In the literature and in the practice of some industries (particularly the semiconductor industry), both analytical

*Corresponding author. Email: fengyang08@gmail.com

queueing (e.g., Silva and Morabito (2009)) and computer simulation (e.g., Ayag (2007) and Kumar and Nottestad (2009)) models have been used individually to address the capacity expansion problems for stochastic manufacturing systems (Neacy *et al.* 1994). The use of queueing and simulation models in capacity planning is usually coupled with optimisation schemes: the queueing or simulation models are used to evaluate the performance of a given system configuration, while the optimisation procedure iterates in search of the best configuration (e.g., Bard *et al.* (1999) and Hopp *et al.* 2002). There are two major drawbacks with such hybrid methods. The first is associated with the performance evaluation models: queueing models are generally criticised for inaccuracy and simulation is known to have a high computational cost. The second drawback lies in the use of optimisation methods: an objective function and a number of constraints must be formulated mathematically. This hinders the ability to evaluate the system performance over a wide range of future scenarios, and the majority of the work in the literature only considers a single (e.g., Hopp *et al.* (2002)) or a set (e.g., Swaminathan (2000)) of discrete demand scenarios. Further, mathematical optimisation methods make it difficult to address trade-offs that are not easily quantified (Nelson 2004). The solution resulting from the optimisation-based hybrid method, if not impractical, may be far from sufficient to provide an adequate alternative pool for decision makers, who generally have to consider factors beyond numbers.

To overcome the drawbacks of existing capacity expansion methods, we propose in this paper a comprehensive framework that integrates queueing models and simulation-based statistical methods. Our approach is distinct in the following aspects.

- The desired number of promising configurations (scenarios) can be generated, with each one fully characterised by its cycle time (CT)–throughput (TH) performance profile. These profiles will allow us to evaluate the comprehensive performance of each expansion alternative over a wide demand (or TH) range, which enables risk analysis of investment decisions. The availability of a set of good configurations (rather than a single good configuration) will provide much more latitude in decision making.
- The strengths of queueing and simulation models are fused for the efficient generation and characterisation of the desired alternatives. Statistical methods such as ranking and selection (Henderson and Nelson 2006, Chapter 17) and metamodeling (Henderson and Nelson 2006, Chapter 18) are adopted to ensure the computational efficiency and statistical validity of simulation experiments.

It is also worth mentioning that, in the review of capacity planning by Geng and Jiang (2009), it was pointed out that new methods are yet to be developed that can (i) accommodate the uncertainty in product demand (including the uncertainty in product mix) and (ii) take the CT performance measure into consideration. This paper attempts to develop a capacity planning approach that addresses these two issues in a better way than presented in the literature.

In this case study, we apply the proposed methods to a model of a real wafer fab for capacity expansion decisions. The remainder of the paper is organised as follows. Section 2 provides an overview of the research problem and the proposed framework used to address it. Section 3 describes the simulation model of the wafer fab used to perform this case study. In Section 4, the proposed methods are applied to the wafer fab model and the results of the capacity expansion analysis are presented. The computational and statistical

tools developed in this case study are briefly described in Section 5. Section 6 gives a brief summary and discusses future work.

2. Overview of the problem and method framework

We focus our attention on the quarterly/yearly capacity expansion problem. Our interaction with industry indicates that a medium-sized wafer fab could spend tens of millions of dollars every quarter on procurement of new tools. The question is how to allocate the total budget to purchase tools/equipment so that the system's performance improvement will be maximised.

For wafer fabs, among the most important performance metrics are throughput (TH), manufacturing cycle time (CT), and work in process inventory (WIP). TH is defined as the rate at which jobs are processed by the system, CT refers to the random variable representing the time it takes a job to traverse the system, and WIP represents the average number of jobs present in the system (Hopp 2007). Since computing the mean CT and WIP is equivalent (Little's Law can be used to compute from one to the other), here we consider TH and mean CT as the two metrics of primary interest. For ease of discussion, we will use CT to refer to mean CT in the rest of the paper except in Section 6.

In semiconductor manufacturing, it is desirable to (i) match the system TH with the customer demand rate for various types of products, and, at the same time, (ii) minimise the CT. However, these are considered to be conflicting goals: for a given system, as we increase the TH to satisfy a higher demand, the CT will also increase, following a nonlinear path. Figure 1 gives an example of CT–TH profiles for single-product systems. Each curve describes the trade-off performance of a different system configuration with the upper limit of the TH being the capacity of that system (indicated by the dashed line in Figure 1). See Fowler *et al.* (2001) and Yang *et al.* (2007, 2008) for details of CT–TH curve generation. For discussions concerning the selection of the best system configuration among different scenarios based on their CT–TH curves, see Spence and Welter (1987).

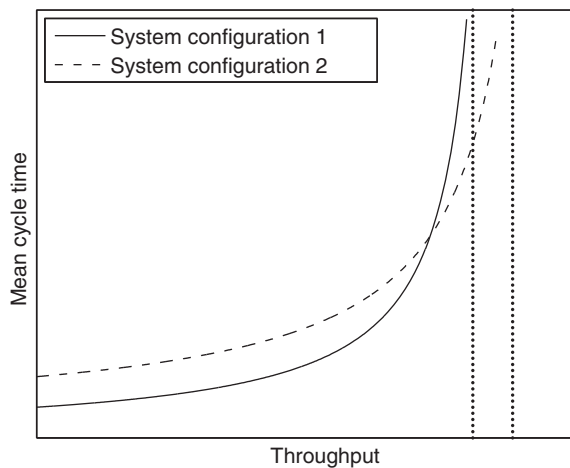


Figure 1. CT–TH profiles for two different system configurations in a single-product environment.

In this work, a multi-product manufacturing environment is investigated, and characterised by a multi-dimensional CT–TH profile, for which

- TH represents the vector of throughput for multiple products, with each element being the throughput for a certain type of product, and
- CT includes the expected cycle time of each of the multiple types of products.

For this long-term strategic planning, we consider the manufacturing system as a push system (Hopp 2007), for which production managers can control the TH (TH has to be less than the capacity for the stability of the system) by controlling the release rate of jobs into the system. Each TH corresponds to an expected cycle time for products. The trade-off relationship between CT and TH has long been recognised as providing a comprehensive performance profile for a manufacturing system (Atherton and Dayhoff 1986, Hopp and Spearman 2008). With such profiles, the system's overall performance (e.g., the total profit in the next six months), which mainly depends on the TH, CT and WIP, can be evaluated over a range of demand patterns and hence allow for the risk analysis of various investment options.

Ideally, a CT–TH profile is desirable for each feasible system alternative so that a thorough comparison can be performed across all the expansion options. However, the number of feasible alternatives may be extremely large (as shown in Section 4.1), which makes it practically impossible to generate a CT–TH profile for each alternative. In light of this, we propose a three-stage framework to address the capacity expansion problem. The analysis in the first two stages is mainly to obtain a candidate pool containing a reasonable number of good system alternatives, and this pool should be the largest possible that we can afford to explore further in Stage 3, constrained by the time available to make a decision. In Stage 3, the systems resulting from Stages 1 and 2 will be characterised by their CT–TH profiles. More specifically, the three-stage approach is outlined as follows.

Stage 1: Given the budget, generate a large number of alternatives for the expanded facility (Section 4.1). The queueing network models developed by Hopp *et al.* (2002) will be used to perform the capacity analysis and to facilitate this alternative generation. For real semiconductor fabrication systems (that involve batches, re-entrants, setups, and multi-product classes), this model is able to (i) exactly calculate the capacity of each workstation in the system, and (ii) identify the bottleneck station(s).

Stage 2: Screen the large number of alternatives down to a relatively small number of promising configurations (Section 4.2). Analytical queueing models (for capacity analysis) and/or computer simulation will be used to perform this alternative screening. The adopted queueing model of Hopp *et al.* (2002) has already been briefly introduced in Stage 1 above, and simulation-based screening will be performed following the procedure proposed by Koenig and Law (1985). Koenig and Law's procedure is one of the ranking and selection procedures (Henderson and Nelson 2006, Chapter 17), and is designed to select among a number of alternatives a subset of size ℓ that contains the m best candidates ($1 \leq m \leq \ell$). The essence of such a procedure lies in the control of the probability of correct selection and the achievement of simulation efficiency.

Stage 3: For each selected alternative from Stage 2, CT–TH profiles will be generated through simulation experiments using the statistical procedures developed by Yang (2010) (Section 4.3). To generate the CT–TH profiles, Yang (2010) developed a neural network based metamodeling approach. A metamodel, which takes the form of polynomial regressions, splines, etc., is a mathematical approximation of the quantitative relationship

implied by the simulation. Metamodelling techniques refer to the integration of computer simulation and response surface modelling (Henderson and Nelson 2006, Chapter 18). In Yang (2010), to metamodel the CT–TH surfaces, (TH, CT) data pairs were first collected by performing a selected set of simulation experiments; from the data, statistical methods were used to fit a neural network model (metamodel) which provides a functional approximation for the CT–TH relationship. Here, a metamodelling approach is adopted for CT–TH modelling because neither analytical queueing models nor pure simulation are able to provide an adequate characterisation. Queueing models, although fast and easy to use, rely on restrictive assumptions and are often not able to accurately capture the CT–TH relationship for real complex systems, particularly when multiple resources are required for processing (e.g., a machine and an operator) or when non-FCFS dispatching policies are employed. Discrete-event simulation, on the other hand, is known for its high fidelity and flexibility, but may be very time-consuming to run: models of complex manufacturing systems may take several hours for a single replication (Fowler and Rose 2004). Metamodelling aims at overcoming the major drawbacks of queueing methods and computer simulation, and is able to generate metamodels representing the CT–TH relationships for a given system. Such models are mathematical functions like those provided by a tractable queueing model while possessing the high fidelity of simulation.

In this paper, the proposed framework is applied to a model of a real wafer fab, which is described in Section 3. The detailed methods and results are given in Section 4.

3. Simulation model

We constructed the simulation model in Microsoft Visual Studio C++ and it represents the semiconductor wafer fab dataset 1 provided by the SEMATECH testbed at Arizona State University (<http://www.eas.asu.edu/~masmlab>). Those datasets were developed in the early 1990s by a group of researchers at SEMATECH to provide the public with models of real wafer fabs that can be used for testing new simulation packages, novel heuristics and factory control policies, and other newly developed approaches to wafer fab/equipment scheduling (Mason and Fowler 2000). In this work, dataset 1 was selected because of the seven SEMATECH datasets available, it is the only one with equipment cost data, which is required in our capacity planning (the costs of the equipment have been adjusted to be more consistent with their current values.)

3.1 Workstations and operators

A workstation refers to a group of functionally identical machines that perform the same operations. Dataset 1 consists of a total of 83 workstations, and each station consists of one to 18 identical machines. The times between machine failures and the times to repair follow exponential distributions. The same dispatching rule, first come first served (FCFS), is applied to all the workstations. We note that if other dispatching rules were used here, our procedures would not change.

In this simulation model, there are 31 operator groups who are responsible for loading, unloading and machine operating. Each operator group has a number of operators who work only at specified workstations. Workers within each group are assumed to have the same level of working skills. Operators prioritise their tasks based on the FCFS rule.

3.2 Product flow

Dataset 1 is composed of two different product flows that use the 83 different tool groups (workstations) mentioned above. Product 1 has 210 processing steps with a pure processing time of 293 hours, while Product 2's process routing contains 245 steps with a pure processing time of 336 hours. Both products are released into the fab in a fixed lot size of 48. As mentioned earlier, the system is treated as a push system, and thus the release rate of jobs (e.g., the weekly wafer starts) is controlled and specified in the simulation experiments.

3.3 Model verification and validation

We converted the SEMATECH dataset 1 into a simulation model coded in Microsoft Visual Studio C++. The C++ model was verified using techniques recommended by Law and Kelton (2000), such as running the model with simplified assumptions to detect logical mistakes and testing the model outputs under a variety of input settings. Also, the C++ model was validated against the SEMATECH dataset executed in Factory Explorer (<http://www.wwk.com>), which is simulation software in an Excel spreadsheet environment. The outputs from the C++ and Factory Explorer models are compared for a wide set of inputs.

The reason why we converted the dataset into C++ is so that the simulation model can be integrated as a sampling tool into the adopted statistical procedures (Sections 4.2.2 and 4.3). As will be explained in Section 5, the software we have developed for this case study is able to automatically iterate between the running of the simulation for data collection, statistical data analysis, and experimental design of future simulation experiments.

4. Decision-making methods for capacity expansion

In this section, we present in detail the proposed capacity planning method, and apply it to the wafer fab model described in Section 3. For the reader's convenience, we first define the following notation.

- b the total budget available for capacity expansion
- K number of different types of products/wafers in a semiconductor fabrication system
- \mathbf{d} $= (d_1, d_2, \dots, d_K)$. The demand rate vector with each element representing the demand per week for a product type
- $\boldsymbol{\alpha}$ $= (\alpha_1, \alpha_2, \dots, \alpha_K)$. The product mix (PM) vector with each element α_k representing the fraction of type k product in the product flow
- d $= \sum_{k=1}^K d_k$. The overall release rate (throughput) of all products
- J total number of workstations in a system
- Π the symbol used to represent a system configuration; Π_0 denotes the base system, and Π_i ($i = 1, 2, 3, \dots$) an alternative system configuration
- $\mu_j(\Pi, \boldsymbol{\alpha})$ the capacity of station j ($j = 1, 2, \dots, J$) in the system Π with a PM of $\boldsymbol{\alpha}$
- $\rho_j(\Pi, \mathbf{d})$ the utilisation of station j ($j = 1, 2, \dots, J$) in the system Π with a product flow specified as \mathbf{d}

- j_{BN} the BN station that achieves $\max\{\rho_j(\Pi, \mathbf{d}), j=1, 2, \dots, J\}$, the highest station utilisation
- $\mu(\Pi, \boldsymbol{\alpha})$ the system capacity for Π , which is equal to the capacity of the BN station $\mu_{j_{\text{BN}}}(\Pi, \boldsymbol{\alpha})$
- I the number of important stations identified by the analytical queueing analysis
- $c_k(\Pi, \mathbf{d})$ the expected cycle time of product k in system Π when the product flow is given as $\mathbf{d}=(d_1, d_2, \dots, d_K)$
- x the system utilisation, i.e. the utilisation of the bottleneck station.

Suppose that the system processes K different types of products (wafers), and $\mathbf{d}=(d_1, d_2, \dots, d_K)$ denotes the future demand rate vector with each element representing the demand per week for a product type. At the planning capacity stage, \mathbf{d} is forecasted and could potentially vary over a certain range. The impact of the demand \mathbf{d} upon the total profit of running a semiconductor facility is illustrated by Bermon and Hood (1999). Thus, it is of interest to evaluate different system configurations over a range of demand, which is equivalent to the target throughput that the system will be running at to satisfy customer needs. Hence, we will also refer to \mathbf{d} as the system TH. Alternatively, the TH (or demand vector) \mathbf{d} can be expressed as $d \cdot \boldsymbol{\alpha}$, with $d = \sum_{k=1}^K d_k$ being the overall throughput of all the products, and $\boldsymbol{\alpha}=(\alpha_1, \alpha_2, \dots, \alpha_K)$ the PM vector where each element α_k represents the fraction of type k product in the system's product flow. Obviously, the PM satisfies the following condition:

$$\sum_{k=1}^K \alpha_k = 1, \quad \alpha_k \geq 0, \text{ for } k = 1, 2, \dots, K.$$

Assuming steady state, the TH is equal to the release rate of products into the system and can be controlled in production. Let Π_0 be the current/base system consisting of J workstations, and Π_i ($i=1, 2, 3, \dots$) an alternative system that is reconfigured from Π_0 using the given budget, b .

Figure 2 illustrates the framework of our three-stage capacity expansion method, which was briefly discussed in Section 2. The output of the proposed procedure is a number of good expansion alternatives, and each of them is fully characterised by its CT–TH performance profile. The inputs of the procedure include the current configuration Π_0 of the system being investigated, the total budget, and the most likely demand forecast \mathbf{d}^* , which is assumed to be available from expert opinion (Swaminathan 2000). In our work, the analysis in the first two stages is driven by the most likely demand. If necessary, a discrete set of demand scenarios (when available) can replace the single demand \mathbf{d}^* in Stages 1 and 2, and the corresponding extension is straightforward, as will become clear later (in Sections 4.1 and 4.2). The aim of the first two stages is to obtain a candidate pool including good alternatives. The size of the pool is selected in such a way that all the systems in it can be characterised by their CT–TH profiles (Stage 3) within the time available for making the expansion decision. As explained earlier, such CT–TH profiles capture the system performance over a fairly wide range of TH (or demand). In the procedure, both queueing models and simulation-based methods are adopted. Queueing models are used to perform capacity/utilisation analysis, which is highly accurate or even exact for that purpose, whereas the CT-related performance profiles are estimated via simulation experiments, which are designed and analysed by statistical methods.

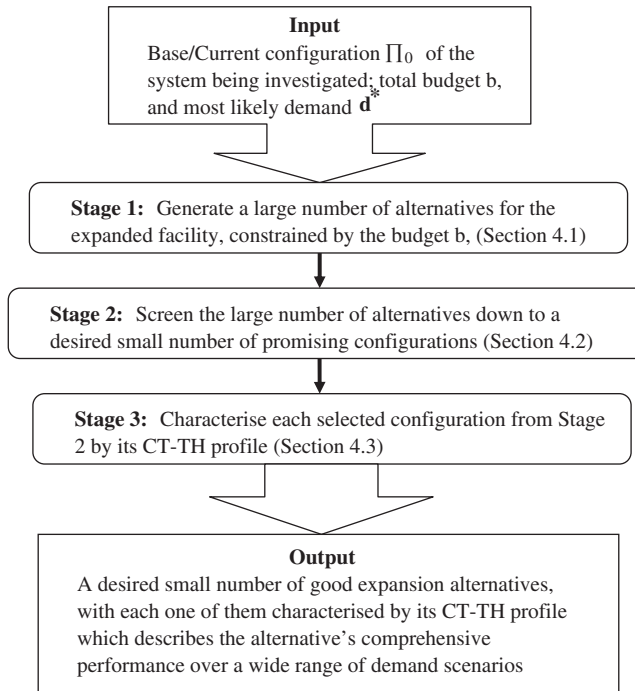


Figure 2. Flowchart for the proposed three-stage procedure.

For this case study, the inputs of the procedure are set as follows. The SEMATECH wafer fab described in Section 3 is being considered for capacity expansion. Three different budget levels, $b = 15, 30$ and 45 million dollars, have been used in our experiments, and due to space constraints, here we only present the details for the case with a budget of $b = 30$ million dollars. In the wafer fab of interest, there are $K = 2$ different types of product flows, and thus both α and \mathbf{d} are two-dimensional vectors. The most likely demand \mathbf{d}^* is set as $(49, 49)$ lots/week, or, equivalently, $\mathbf{d}^* = d^* \alpha^* \approx 98(0.5, 0.5)$.

4.1 Potential capacity expansion scenarios

In this part, a large number of feasible alternatives that satisfy the budget constraint will be generated with the assistance of analytical queueing models. A lot of work has been devoted to developing queueing approximations to model the behaviour of manufacturing systems, and Shantikumar *et al.* (2007) provide a recent review, which includes White (1983), Bitran and Tirupati (1988), Chen *et al.* (1988), Mitrani and Puhalskii (1993), Connors *et al.* (1996), Morrison and Martin (2007), etc. In our work, the open queueing network model developed by Hopp *et al.* (2002) is used to perform the capacity analysis for wafer fab systems.

Hopp's model incorporates features common to the semiconductor environment, such as batch processes, re-entrant flows, multi-product classes, and machine setups, but does not include operators, and it provides an exact calculation of the station capacity, which is denoted by $\mu_j(\Pi, \alpha), j = 1, 2, \dots, J$, a function of the system configuration and product mix

Table 1. The most heavily utilised stations in the base system Π_0 at the most likely forecasted demand \mathbf{d}^* .

| Station | Number of machines | Unit cost (\$) | Utilisation | Station | Number of machines | Unit cost (\$) | Utilisation |
|---------------|--------------------|----------------|-------------|--------------|--------------------|----------------|-------------|
| E_SINK | 3 | 3,000,000 | 1.15 | ALIGNER | 6 | 8,000,000 | 0.7 |
| MATRIX | 7 | 3,000,000 | 1.1 | DRIVE_OX | 2 | 4,000,000 | 0.7 |
| LEITZ | 8 | 1,000,000 | 1.03 | AME_8310 | 2 | 15,000,000 | 0.69 |
| CRIT_DEV | 12 | 3,750,000 | 0.85 | VWR_OVEN | 2 | 500,000 | 0.65 |
| DIFF_SINK_2 | 1 | 3,000,000 | 0.82 | INTERGATE | 3 | 4,000,000 | 0.65 |
| NONCRIT_DEV | 9 | 2,000,000 | 0.81 | REFLOW | 4 | 4,000,000 | 0.64 |
| UV_BAKE | 2 | 750,000 | 0.8 | CRIT_COAT | 12 | 3,750,000 | 0.64 |
| PEAK | 2 | 4,000,000 | 0.79 | NONCRIT_COAT | 9 | 2,000,000 | 0.62 |
| STEPPER | 11 | 15,000,000 | 0.75 | QUAESTAR | 1 | 10,000,000 | 0.61 |
| HIGH_CURR_IMP | 4 | 15,000,000 | 0.7 | | | | |

(Bermon and Hood 1999). Given Π and α , the effective process time at each workstation can be computed taking into account machine failures and repairs, setups, batches, and re-entrants, and $\mu_j(\Pi, \alpha)$ ($j = 1, 2, \dots, J$), the maximum process rate (in terms of, say, lots/week) at station j is equal to the inverse of the effective process time at that station (Hopp *et al.* 2002). For a given product flow specified by $\mathbf{d} = \alpha\mathbf{x}$, the utilisation $\rho_j(\Pi, \mathbf{d})$ can also be obtained for station j and is required to satisfy $0 < \rho_j(\Pi, \mathbf{d}) < 1$ to ensure the stability of the system. The bottleneck (BN) station j_{BN} is the station that achieves $\max\{\rho_j(\Pi, \mathbf{d}), j = 1, 2, \dots, J\}$, the highest station utilisation, and there may be multiple BN stations. The system capacity $\mu(\Pi, \alpha)$ is equal to the capacity of the BN station $\mu_{j_{\text{BN}}}(\Pi, \alpha)$.

This exact capacity analysis provided by Hopp's queueing model is used to assist the generation of the large number of potential system alternatives. As mentioned earlier, in the alternative generation (Section 4.1) and screening (Section 4.2), we assume that a most likely demand forecast $d^*(\alpha_1^*, \alpha_2^*) = 98(0.5, 0.5)$ is available. Suppose that the given budget is $b = 30$ million dollars, we proceed in the following two steps to generate system alternatives.

First, among the total of $J = 83$ workstations in the wafer fab, a number of the most heavily utilised stations were identified as important stations considered for capacity expansion. Given the most likely forecasted demand $d^*\alpha^* = 98(0.5, 0.5)$, the station utilisations $\{\rho_j(\Pi_0, \mathbf{d}^*), j = 1, 2, \dots, J\}$ were computed for the base system, and the $I = 19$ most heavily utilised stations (with a utilisation above the threshold $\rho_L = 0.6$) were selected, as shown in Table 1. Apparently, the base system is not able to satisfy the future demand, since the most heavily loaded station has a utilisation of $1.15 > 1$, which means that the system will not operate stably if its product release rate is pushed up to match the desired demand. These $I = 19$ stations will be considered important and as candidates for investment, and their unit cost is obtained by multiplying by 50 the original cost in the early 1990s given in the SEMATECH dataset. The multiplier of 50 was chosen in order to make the cost of the equipment somewhat close to the cost of today's equipment.

Second, feasible alternatives will be generated allocating the budget of $b = 30$ million dollars to the 19 important stations. Suppose that the $I = 19$ important stations are

numbered from 1 to 19. An alternative system configuration Π can be specified as $\{x(j), j=1, 2, \dots, I\}$, representing the number of machines (including the existing machines and those to be purchased) at each important station. We coded a program to search for all the feasible configurations $\{\Pi_1, \Pi_2, \Pi_3, \dots\}$, i.e. the combinations $\{x(j), j=1, 2, \dots, I\}$ that satisfy (i) the budget constraint, and (ii) $\rho_j(\Pi, \mathbf{d}^*) \geq 0.6$ with $j \in I$, $\Pi = \{x(j), j \in I\}$ and $\mathbf{d}^* = 98(0.5, 0.5)$. Constraint (ii) implies that we avoid adding more machines to a station once its utilisation drops below 0.6. In our case, the resulting number of feasible system alternatives ended up being 318,153. Due to the extremely large number, it takes about 30 hours (Intel Core 2 Duo E6850 3.0 GHz CPU) to generate these alternatives and estimate the system capacity for each one. The number of alternatives and thus the computation time required will be substantially reduced if the number of selected stations could be reduced, either by resorting to subjective judgement (e.g., considering space limitations in the fab) or slightly increasing the threshold utilisation level ρ_L . The experiment presented here is intended to show the strenuous use of our proposed approach.

In this part, we simply seek to generate the possible system configurations constrained by the total budget b . To help reduce the tremendously large number of alternatives, basic queuing knowledge (the utilisation constraints) is used to screen out the obviously inferior configurations. The utilisation of each station $\rho_j(\Pi, \mathbf{d})$ ($j=1, 2, \dots, J=83$) depends on the demand rate \mathbf{d} . Thus, as \mathbf{d} varies, the important stations selected (such as those in Table 1) and the set of generated alternatives may also vary. In this paper, a single most likely demand \mathbf{d}^* is used to drive the alternative generation. But if a set of discrete demand scenarios is available, then each one can be used to generate a set of system configurations, and the union of these configuration sets can serve as the initial candidate pool. The necessity of using multiple demand scenarios depends on the sensitivity of station utilisations to a change in demand, and has to be evaluated on a case-by-case basis.

4.2 Screening of system alternatives

The set of system configurations generated from Section 4.1 may include too many alternatives, and it may be practically impossible to characterise and evaluate each of them by its CT–TH profile. Thus, in this section, we introduce a screening mechanism that aims at obtaining, say m , good systems, on all of which the CT–TH modelling (Section 4.3) can be applied within the decision time available. We consider the number m as a value that could be several to nearly 100 in practice. In this case study, m is set as five to illustrate the proposed approach.

Depending on the number of alternatives generated in Section 4.1, one or both of the alternative screening approaches described in Sections 4.2.1 and 4.2.2 may be performed. The most likely demand $\mathbf{d}^* = d^* \boldsymbol{\alpha}^*$ is assumed given and also used in these screening processes.

4.2.1 Capacity-based pre-screening

If a large number (>100) of alternatives is generated in Section 4.1, we first perform a pre-screening based on the system capacity of each candidate configuration. We simply rank the alternatives $\{\Pi_1, \Pi_2, \dots\}$ by their system capacity $\mu(\Pi, \boldsymbol{\alpha}^*)$ ($i=1, 2, \dots$), and select the top tens of configurations that have the highest capacities. In our case, $\boldsymbol{\alpha}^*$ is assumed to be $(0.5, 0.5)$, and 60 system alternatives were selected with their capacities ranging from 121 to 125 lots/week. If multiple rather than a single demand scenario(s) are given,

each candidate system can also be pre-evaluated based on their weighted average utilisation across the different demand rates.

4.2.2 Cycle time-based screening via simulation

The resulting tens of alternatives, which may be the outputs of Section 4.1 or the survivors from Section 4.2.1, can then (i) be scrutinised by experienced personnel for the evaluation of practicality, and/or (ii) be put through the CT-based screening, which is discussed next.

As mentioned already, simulation is usually much more accurate than queueing models in terms of estimating the CT metric for realistic systems. (One can certainly use the queueing CT estimates for the following screening if they are sufficiently accurate for the manufacturing system of interest.) Since this work focuses on steady-state behaviour, the CT estimates are obtained from steady-state simulation. A warm-up period of 2 months is chosen for all the system alternatives based on the techniques suggested by Law and Kelton (2000). The total simulation length of each replication is 24 months. In a simulation replication, the cycle time of each product simulated in steady state was recorded, and the average cycle time of a certain type of product provides a CT estimate for that product type. The number of replications required will be determined by a statistical procedure in a sequential manner, as will be discussed below.

We adopt the simulation-based ‘rank and selection’ procedure of Koenig and Law (1985) to select from the 60 candidate configurations a subset of $\ell=5$ alternatives containing $m=5$ best systems. Here, the systems are evaluated based on their product mix-weighted average CT of both product types at the demand rate of \mathbf{d}^* (i.e. the throughput/release rate of \mathbf{d}^*). The best systems are those that have the lowest weighted average CT. The size of m can be selected by the user, as pointed out earlier. Koenig and Law’s procedure is able to control the probability of correct selection, and to obtain the desired subset in a most computationally efficient manner. Note that this cannot be achieved by conventional mathematical programming methods.

There are two basic input parameters to Koenig and Law’s procedure. One is the probability of correct selection, which is set as 95% in our experiments. The other is the indifference-zone (IZ) parameter δ , which is the amount of difference in expected performance that is deemed practically significant; in our context, if the expected CT of configuration Π_i is at least δ lower than that of Π_j , then alternative Π_i is considered practically superior to Π_j in terms of CT at the most likely demand \mathbf{d}^* . As in Section 4.2.1, this selection can be straightforwardly extended to multiple demand scenarios where the systems can be compared based on their weighted average CT performance over the different demand rates. In our case, δ is set as 12 hours, which is roughly 2% of the weighted cycle time. Under such settings, it takes about 20 hours (Intel Core 2 Duo E6850 3.0 GHz CPU), which is mainly simulation time, to run Koenig and Law’s procedure and select the five best systems out of 60 alternatives. The computation time is sensitive to the desired probability of correct selection and the IZ parameter, which are both user-specified parameters.

The resulting five selected system configurations $\{\Pi_1, \Pi_2, \Pi_3, \Pi_4, \Pi_5\}$ are specified in terms of the number of machines at the 19 important stations $\{x(j), j=1, 2, \dots, I\}$ ($I=19$), and are given in Table 2. Each column represents a selected alternative, and each row an important workstation considered for expansion. The cell value refers to the number of new machines purchased for the corresponding station under the corresponding alternative.

Table 2. The number of new machines purchased for the 19 important stations under the five selected alternatives.

| Station | Alternative | | | | | Station | Alternative | | | | |
|---------------|-------------|---------|---------|---------|---------|--------------|-------------|---------|---------|---------|---------|
| | Π_1 | Π_2 | Π_3 | Π_4 | Π_5 | | Π_1 | Π_2 | Π_3 | Π_4 | Π_5 |
| E_SINK | 2 | 2 | 2 | 2 | 2 | ALIGNER | 0 | 0 | 0 | 0 | 0 |
| MATRIX | 3 | 3 | 3 | 3 | 3 | DRIVE_OX | 0 | 0 | 0 | 0 | 0 |
| LEITZ | 3 | 3 | 4 | 3 | 3 | AME_8310 | 0 | 0 | 0 | 0 | 0 |
| CRIT_DEV | 1 | 1 | 1 | 1 | 1 | VWR_OVEN | 1 | 1 | 0 | 1 | 1 |
| DIFF_SINK2 | 1 | 1 | 1 | 1 | 1 | INTERGATE | 1 | 0 | 0 | 0 | 0 |
| NONCRIT_DEV | 0 | 2 | 0 | 1 | 0 | REFLOW | 0 | 0 | 0 | 0 | 0 |
| UV_BAKE | 1 | 1 | 0 | 0 | 1 | CRIT_COAT | 0 | 0 | 0 | 0 | 0 |
| PEAK | 0 | 0 | 1 | 0 | 1 | NONCRIT_COAT | 0 | 0 | 0 | 1 | 0 |
| STEPPER | 0 | 0 | 0 | 0 | 0 | QUAESTAR | 0 | 0 | 0 | 0 | 0 |
| HIGH_CURR_IMP | 0 | 0 | 0 | 0 | 0 | | | | | | |

4.3 CT–TH characterisation of system alternatives

In this part, we characterise the five selected system configurations, denoted as $\{\Pi_1, \Pi_2, \Pi_3, \Pi_4, \Pi_5\}$, by their CT–TH profiles. It is worth mentioning that there is a sequence of works in this regard, including Yang *et al.* (2007, 2008) and Yang (2010). The first two papers address the CT–TH modelling issues for a single-product manufacturing system, and the third is able to handle multiple-product environments and is hence adopted in this case study to analyse our two-product wafer fab.

Recall that $\mathbf{d} = (d_1, d_2) = d\boldsymbol{\alpha} = d(\alpha_1, \alpha_2)$ is used to represent the demand rate vector as well as the system TH, and here \mathbf{d} is a two-dimensional vector with each element representing the throughput (or, equivalently, the release rate) of a certain product type. The system utilisation $x = d/\mu(\Pi, \boldsymbol{\alpha})$, the overall release rate of products divided by the system capacity $\mu(\Pi, \boldsymbol{\alpha})$, can be calculated analytically for a given system Π and PM $\boldsymbol{\alpha}$ (Section 4.1). Thus, we have

$$c_k(\Pi, \mathbf{d}) = c_k(\Pi, \mathbf{d}, \boldsymbol{\alpha}) = c_k(\Pi, d/\mu(\Pi, \boldsymbol{\alpha}), \boldsymbol{\alpha}) = c_k(\Pi, x, \boldsymbol{\alpha}), \quad k = 1, 2. \quad (1)$$

The objective of Yang (2010) is to estimate, for a system configuration Π , the functional performance surfaces $\{c_k(\Pi, x, \boldsymbol{\alpha}); k = 1, 2\}$, from which the CT of product k can be derived for any product flow described by \mathbf{d} .

To efficiently generate the target surfaces $\{c_k(\Pi, x, \boldsymbol{\alpha}); k = 1, 2\}$, Yang (2010) integrates queueing analysis, adaptive statistical methods, and computer simulation. The queueing models play two essential parts. First, the analytical analysis of the CT–TH relationships suggests that the target surfaces are smooth and differentiable within a PM sub-region where the bottleneck (BN) station stays unchanged, and thus motivates the partition of the PM region into such constant-BN sub-regions before the CT–TH metamodelling. (Recall that the station utilisations and hence the locations of the BN station depend on PM.) Second, the queueing network models (e.g., Hopp *et al.* (2002)) are used to perform the capacity and BN analysis, and to divide the PM region into a number of constant-BN sub-regions. The neural network (NN)-based metamodelling is then applied on each PM sub-region to obtain the functional approximations of the target surfaces. The metamodelling approach of Yang (2010) is distinct from the existing NN modelling

work in three major aspects. First, instead of treating a NN as a black box, the NN geometry is investigated and utilised in the metamodelling of CT–TH surfaces. Second, a progressive model-fitting strategy is developed to achieve the parsimonious NN adequate to characterise the CT–TH relationship. Third, a design of experiment strategy, particularly suitable to NN modeling, is developed to efficiently collect simulation data for the estimation of the NN models.

The basic idea of CT–TH modelling is as follows. Simulation is performed at a set of carefully selected design points in the input space (x, α) ; from the data collected, models (i.e. NN) are fitted to obtain the functional approximation of the CT–TH surfaces within each constant-BN sub-region. The approach of Yang (2010) was demonstrated to be efficient in generating CT–TH profiles through empirical evaluation of extensive systems, including multi-product wafer fabs. Here, we apply it to the selected alternatives. For each system configuration Π , the CT–TH surfaces $\{c_k(\Pi, x, \alpha), k = 1, 2\}$ are estimated from the simulation data. The setup of the simulation experiments (e.g., warm-up period and simulation length) has already been explained in Section 4.2.2.

We take alternative Π_1 as an example, and present the experimental results for the generation of $\{c_k(\Pi, x, \alpha) = c_k(\Pi_1, x, \alpha_1), k = 1, 2\}$. Note that since $\alpha_1 + \alpha_2 = 1$, α is determined by α_1 . Hence, the system utilisation x and the fraction of product 1 α_1 are considered as independent variables. The range of interest for utilisation x is set as $[0.75, 0.85]$, which is considered as the typical range within which semiconductor manufacturers run their facility (Hopp 2007). In practical use of the method, the utilisation range can be adjusted by the user depending on the utilisation levels at which the system will be run. The PM α_1 is allowed to vary over $[0.25, 0.75]$. The two-dimensional input region spanned by x and α_1 is defined as $[0.75, 0.85] \times [0.25, 0.75]$, the Cartesian product of the two sets $[0.75, 0.85]$ and $[0.25, 0.75]$. Queueing analysis was first performed to divide the input region into constant-BN sub-regions with stations NONCRIT_DEV and DRIVE_OX (Table 1) being the BN, and the PM level with $\alpha_1 = 0.44$ was identified as the BN-shift point, and thus the resulting two constant-BN sub-regions are $[0.75, 0.85] \times [0.25, 0.44]$ and $[0.75, 0.85] \times [0.44, 0.75]$. In the entire input region, a total of 15 design points were selected following the design strategies of Yang (2010), and about 2.5 hours (Intel Core 2 Duo E6850 3.0 GHz CPU) were spent running simulations at those design points. Based on the data collected, NN models were fitted for different sub-regions. For products of type 1, the estimated CT–TH models for product 1 in the two sub-regions are given by, respectively,

$$c_1(\Pi_1, x, \alpha_1) = 573.49 + \frac{11,593}{1 + \exp(24.86 - 15.55x - 19.04\alpha_1)} + \frac{109}{1 + \exp(1.509 - 6.56x + 11.21\alpha_1)},$$

$$\alpha_1 \in [0.25, 0.44] \text{ and } x \in [0.75, 0.85], \quad (2)$$

$$c_1(\Pi_1, x, \alpha_1) = 491.39 + \frac{537}{1 + \exp(13.61 - 19.89x + 5.83\alpha_1)} + \frac{8475}{1 + \exp(349.612 - 426.79x + 39.85\alpha_1)},$$

$$\alpha_1 \in [0.44, 0.75] \text{ and } x \in [0.75, 0.85]. \quad (3)$$

For product 2, the fitted CT–TH models are given as

$$c_2(\Pi_1, x, \alpha_1) = 598.52 + \frac{712.88}{1 + \exp(14.79 - 12.99x - 8.37\alpha_1)} + \frac{834.42}{1 + \exp(139.06 - 31.04x - 254.61\alpha_1)},$$

$$\alpha_1 \in [0.25, 0.44] \text{ and } x \in [0.75, 0.85], \quad (4)$$

$$c_2(\Pi_1, x, \alpha_1) = 643.06 + \frac{64,850}{1 + \exp(15.02 - 14.60x + 4.60\alpha_1)} - \frac{20}{1 + \exp(-208.33 - 749.60x + 1218.20\alpha_1)},$$

$$\alpha_1 \in [0.44, 0.75] \text{ and } x \in [0.75, 0.85]. \quad (5)$$

With fitted models as Equations (2)–(5) for each alternative, the system configurations can be evaluated and compared based on their CT–TH profiles, which fully characterise the systems' performance over a wide range of demand scenarios represented by $\mathbf{d} = d(\alpha_1, \alpha_2)$. Figures 3 and 4 illustrate graphically how the configuration alternatives, say Π_1 and Π_2 , can be compared using their fitted CT–TH models.

For the sake of graphical clarity, two-dimensional curves at fixed PM, as opposed to three-dimensional response surfaces, are plotted in Figures 3 and 4. Note that, for any given PM $\tilde{\alpha}_1 \in [0.25, 0.75]$, characteristic curve $\{c_k(\Pi, d, \tilde{\alpha}_1), k = 1, 2\}$, which is expressed in terms of the overall release rate d , can easily be derived from the fitted models $c_k(\Pi, x, \alpha_1)$ by setting $\alpha_1 = \tilde{\alpha}_1$ and $x = d/\mu(\Pi, \tilde{\alpha}_1)$.

Figure 3 shows characteristic curves $\{c_k(\Pi, d, 0.5); \Pi = \Pi_1, \Pi_2; k = 1, 2\}$ for configurations Π_1 and Π_2 at the fixed PM level of $\alpha_1 = 0.5$. Figures 3(a) and (b) correspond to product types 1 and 2, respectively. Specifically, take Figure 3(a) as an example. The solid curve represents $c_1(\Pi_1, d, 0.5)$ and the dotted curve $c_1(\Pi_2, d, 0.5)$. The horizontal axis represents the overall release rate (or, equivalently, the system throughput) d . The range of d , $[d_L, d_U]$, shown in the graph corresponds to the utilisation range of $[0.75, 0.85]$ for the given PM $\alpha_1 = 0.5$, that is $d_L \approx 0.75\mu(\Pi_1, 0.5) \approx 0.75\mu(\Pi_2, 0.5)$ and $d_U \approx 0.85\mu(\Pi_1, 0.5) \approx 0.85\mu(\Pi_2, 0.5)$. The approximations hold here because the capacities of both alternatives, $\mu(\Pi_1, 0.5)$ and $\mu(\Pi_2, 0.5)$, are similar, which is consistent with the pre-screening performed in Section 4.2. Recall that one of the screening criteria is the magnitude of system capacity; system configurations with the highest capacity will be selected and further considered for CT–TH modelling. The vertical axis represents the expected CT of product 1 at certain values of d , which corresponds to a TH vector of $d(0.5, 0.5)$ in Figure 3(a). Figure 3(b) is the product 2 counterpart of Figure 3(a), and plots $\{c_2(\Pi, d, 0.5), \Pi = \Pi_1, \Pi_2\}$.

Figure 4 has the same interpretation as Figure 3 except that the PM α_1 is fixed at 0.25. That is, Figure 4 displays $\{c_k(\Pi, d, 0.25); \Pi = \Pi_1, \Pi_2; k = 1, 2\}$ for the two alternatives being compared. Note that the range of the horizontal axis in Figure 4 is quite different from that in Figure 3. This is because product 2 requires more resources than product 1, and thus the system capacity at the PM of $\alpha_1 = 0.25$ differs markedly from that at $\alpha_1 = 0.5$. In both Figures 3 and 4, $[d_L, d_U]$ are set to cover the utilisation range of $[0.75, 0.85]$ so that the facility will not be overloaded or underloaded regardless of the system capacity.

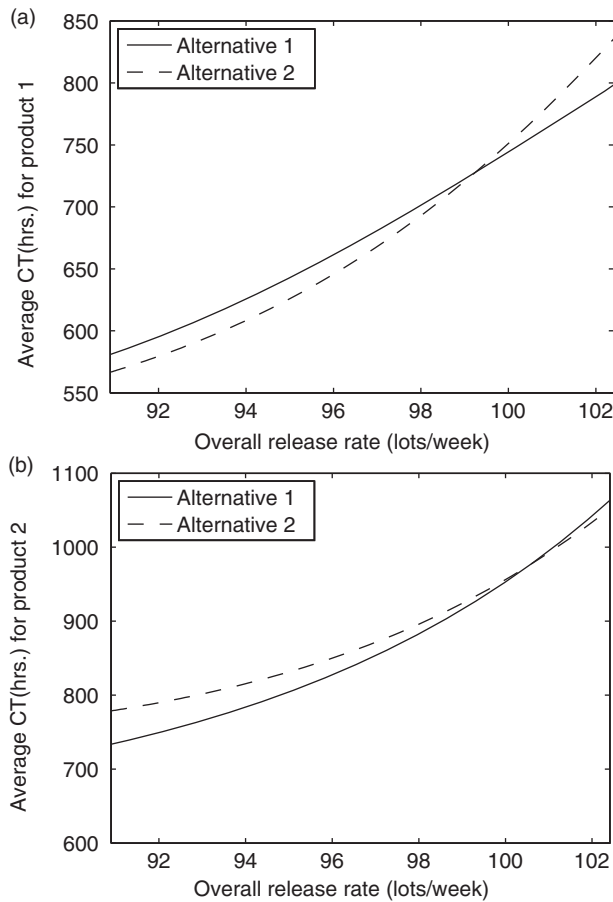


Figure 3. Performance curves at $\alpha_1=0.5$ for two different system alternatives. (a) $\{c_1(\Pi, d, 0.5); \Pi = \Pi_1, \Pi_2\}$; (b) $\{c_2(\Pi, d, 0.5); \Pi = \Pi_1, \Pi_2\}$.

Again, at any PM level, characteristic curves such as those in Figures 3 and 4 can be generated for each of the five selected system alternatives. As examples, Figures 3 and 4 provide the following insights. The performance superiority/inferiority of a system configuration heavily depends on the demand scenario $(d_1, d_2) = d(\alpha_1, \alpha_2)$. For instance, comparing the left and right sides of Figure 3(a), Π_2 excels over Π_1 across the demand range $\{d(0.5, 0.5), d \in [91, 97]\}$ with an expected CT for product 1 about 16 hours less than that of Π_1 , whereas around the demand level of 102(0.5, 0.5), Π_1 outperforms Π_2 with a CT difference of about 34 hours for product 1. For the CT performance of product 2, Figure 3(b) shows that alternative Π_1 is superior over the demand of $\{d(0.5, 0.5), d \in [91, 100]\}$ with the largest CT saving of about 50 hours. If the demand happens to fall in $\{d(0.25, 0.75), d \in [70, 80]\}$, then according to Figure 4, Π_1 and Π_2 are practically equivalent.

For each alternative Π , the CT–TH models $\{c_k(\Pi, d, \alpha_k), k=1, 2\}$ provide a comprehensive performance profile over a wide range of demand scenarios, and hence give decision makers a complete picture of the system's behaviour under future uncertainties. We emphasise that the models $\{c_k(\Pi, d, \alpha_k), k=1, 2\}$ are simple

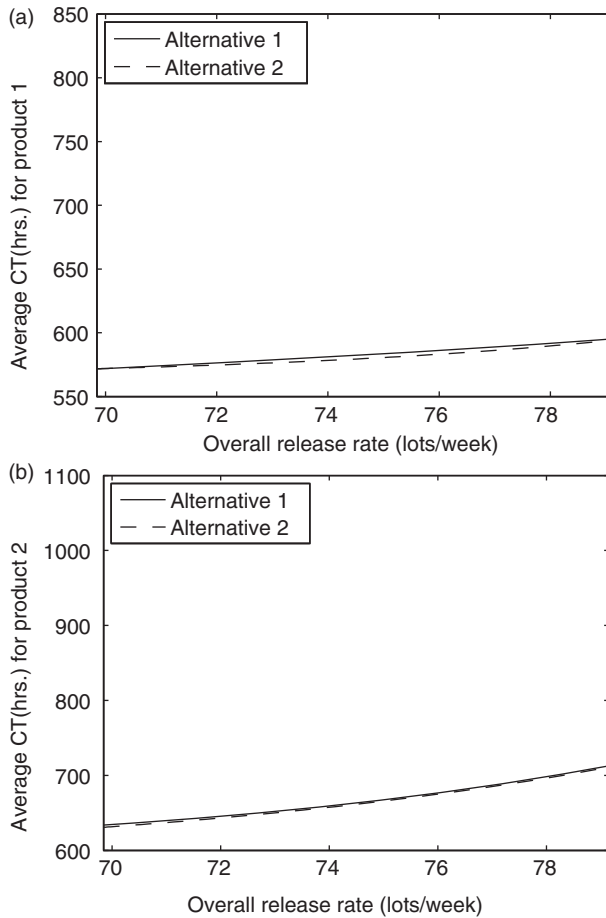


Figure 4. Performance curves at $\alpha_1 = 0.25$ for two different system alternatives. (a) $\{c_1(\Pi, d, 0.25); \Pi = \Pi_1, \Pi_2\}$; (b) $\{c_2(\Pi, d, 0.25); \Pi = \Pi_1, \Pi_2\}$.

mathematical equations, and once established, they can be used directly for performance evaluation and decision optimisation without requiring any additional simulations.

One of the important performance metrics that can be derived from the CT–TH models is the profit, which takes into account the total revenue, capacity expansion cost, production cost, cost for inventories and delivery delay, etc. All these numbers can be estimated from the CT–TH models. For instance, the WIP inventory cost is a function of WIP, which is equal to the product of the cycle time and throughput (Little’s Law); the lead time (the amount of time between the placing of an order and the receipt of the goods ordered) is dominated by the manufacturing CT, which depends on the system TH. Nazzal *et al.* (2006) provide an example of CT–TH-based economic analysis that can be directly applied to our studies to evaluate the various system alternatives once they have been characterised by their CT–TH profiles. Aside from the economic metric, the CT–TH models are also useful in evaluating the potential loss of market share. Suppose that the demand happens to be unexpectedly high, then the two basic options we may have are: (i) maintain the TH that leads to a reasonable CT, but risk losing the unsatisfied demand

to competitors; and (ii) push up the TH to match demand, but risk alienating customers with a long lead time. The CT–TH models can be used to support such trade-off decisions.

To conclude this section, it is again worth mentioning that it is the purpose of this work to provide decision makers with the original and complete quantitative profiles that fully characterise different system alternatives, i.e. the CT–TH models $\{c_k(\Pi, \mathbf{d}), k = 1, 2\}$. From these models, various performance metrics such as profit/cost can be evaluated over a wide range of demand patterns and, moreover, when we consider factors that are hard to quantify (e.g., loss of market share), the CT–TH profiles can also provide valuable support.

5. Computational and statistical tools

To perform this case study, the simulation model representing a SEMATECH wafer fab was developed in Microsoft Visual C++, and the statistical procedures involved (Sections 4.2.2 and 4.3) were implemented in Matlab. Our simulation-based statistical analysis is sequential in that (i) simulation experiments are carried out at multiple stages; (ii) interim data analyses are performed at the end of each stage; and (iii) further experiments are guided by the information already collected. The C++ simulation engine is integrated with the Matlab code, and thus the screening/modelling procedure is able to automatically iterate between sampling via simulation, analysing the collected data, and designing subsequent experiments.

6. Discussion

This work proposes a decision-making framework for the capacity expansion of a semiconductor fabrication system. The approach integrates queueing analysis, computer simulation, and adaptive statistical methods to generate a number of good reconfiguration alternatives, whose comprehensive performance is fully characterised by their CT–TH profiles. The proposed framework is demonstrated in a case study performed on a SEMATECH dataset representing a real wafer fab. Our approach is distinct from the existing capacity planning work with respect to (i) the unique fusion of queueing and simulation methods and (ii) the complete characterisation of reconfiguration alternatives by their CT–TH profiles, which allows for performance evaluation over a wide range of demand scenarios.

In this case study, attention is centred on the relationship between the first moment (mean) of CT and the TH. However, the percentiles of CT also play an important role in strategic planning for manufacturing, and are considered essential in quoting the lead time with the pre-specified customer service level (Hopp 2007). For instance, the functional relationship between the 95th percentile of CT and the TH can be used to balance the TH level in production and customer lead time while achieving 95% on-time delivery. The methods in this paper can be extended to CT percentile estimation, and the key lies in the adaptation of the Stage 3 metamodelling. As pointed out by Yang (2010), the NN metamodelling can be adapted to simultaneously estimate the first, second, third, and fourth moments of CT as functions of TH. With the first three moments, any percentile of CT can be estimated at any TH following the methods proposed by Yang *et al.* (2008). Bekki *et al.* (2010) provide a method that utilises the first four moments to estimate

percentiles of CT. Using the percentile as well as the mean CT–TH profiles to guide capacity expansion will be explored in future work.

References

- Atherton, R.W. and Dayhoff, J.E., 1986. Signature analysis: Simulation of inventory, cycle time, and throughput trade-offs in wafer fabrication. *IEEE Transactions on Components, Hybrids, Manufacturing Technology*, 9 (4), 498–507.
- Ayag, Z., 2007. A hybrid approach to machine-tool selection through AHP and simulation. *International Journal of Production Research*, 45 (9), 2029–2050.
- Bard, G.F., Srinivasan, K., and Tirupati, D., 1999. An optimization approach to capacity expansion in semiconductor manufacturing facilities. *International Journal of Production Research*, 37 (5), 3359–3382.
- Bekki, J.M., et al., 2010. Indirect cycle-time quantile estimation using the Cornish–Fisher expansion. *IIE Transactions*, 42 (1), 31–44.
- Bermon, S. and Hood, S.J., 1999. Capacity optimization planning system (CAPS). *Interfaces*, 29 (5), 31–50.
- Bitran, G.R. and Tirupati, D., 1988. Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Management Science*, 34 (1), 75–100.
- Chen, H., et al., 1988. Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Operations Research*, 36 (2), 202–215.
- Chen, J.C. and Chen, C.W., 2010. Capacity planning of serial and batch machines with capability constraints for wafer fabrication plants. *International Journal of Production Research*, 48 (11), 3207–3223.
- Connors, D., Feigin, G., and Yao, D., 1996. A queueing network model for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 9 (3), 421–427.
- Fowler, J.W. and Rose, O., 2004. Grand challenges in modeling and simulation of complex manufacturing systems. *Transactions of the Society for Computer Simulation International*, 80 (9), 469–476.
- Fowler, J.W., et al., 2001. Efficient cycle time–throughput curve generation using a fixed sample size procedure. *International Journal of Production Research*, 39 (12), 2595–2613.
- Geng, N. and Jiang, Z., 2009. A review on strategic capacity planning for the semiconductor manufacturing industry. *International Journal of Production Research*, 47 (13), 3639–3655.
- Henderson, S.G. and Nelson, B.L., 2006. *Handbooks in operations research and management science: Simulation*. Amsterdam: Elsevier.
- Hopp, W.J., 2007. *Supply chain science*. New York: McGraw-Hill/Irwin.
- Hopp, W.J. and Spearman, M.L., 2008. *Factory physics*. 3rd ed. New York: McGraw-Hill/Irwin.
- Hopp, W.J., et al., 2002. Using an optimized queueing network model to support wafer fab design. *IIE Transactions*, 34 (2), 119–130.
- Koenig, L.W. and Law, A.M., 1985. A procedure for selecting a subset of size m containing the l best of k independent normal populations. *Communications in Statistics: Simulation and Computation*, 14 (3), 719–734.
- Kumar, S. and Nottestad, D.A., 2009. Flexible capacity design for the Focus Factory—A case study. *International Journal of Production Research*, 47 (5), 1269–1286.
- Law, A.M. and Kelton, W.D., 2000. *Simulation modeling and analysis*. 3rd ed. New York: McGraw-Hill.
- Mason, S.J. and Fowler, J.W., 2000. Maximizing delivery performance in semiconductor wafer fabrication facilities. In: *Proceedings of the 2000 winter simulation conference*, Orlando, FL, 1458–1463.

- Mitrani, I. and Puhalskii, A., 1993. Limiting results for multiprocessor systems with breakdowns and repairs. *Queueing System: Theory and Applications*, 14 (3–4), 293–311.
- Morrison, J.R. and Martin, D.P., 2007. Practical extensions to cycle time approximations for the G/G/m-queue with applications. *IEEE Transactions on Automation Science and Engineering*, 4 (4), 523–532.
- Nazzala, D., Mollaghasemib, M., and Anderson, D., 2006. A simulation-based evaluation of the cost of cycle time reduction in Agere Systems wafer fabrication facility—A case study. *International Journal of Production Economics*, 100, 300–313.
- Neacy, E., Brown, S., and McKiddie, R., 1994. Measurement and improvement of manufacturing capacity (MIMAC) survey and interview results. *SEMATECH Technology Transfer*, Austin, Texas, 94052374A-XFR.
- Nelson, B.L., 2004. 50th anniversary article: Stochastic simulation research in management science. *Management Science*, 50 (7), 855–868.
- Ozturk, O., Coburn, M.B., and Kitterman, S., 2003. Conceptualization, design and implementation of a static capacity model. In: *Proceedings of the 2003 winter simulation conference*, New Orleans, LA, 1373–1376.
- Shantikumar, J.G., Ding, S., and Zhang, M.T., 2007. Queueing theory for semiconductor manufacturing systems: A survey and open problems. *IEEE Transactions on Automation Science and Engineering*, 4 (4), 513–522.
- Silva, C.R.N. and Morabito, R., 2009. Performance evaluation and capacity planning in a metallurgical job-shop system using open queueing network models. *International Journal of Production Research*, 47 (23), 6589–6609.
- Spence, A.M. and Welter, D.J., 1987. Capacity planning of a photolithography work cell in a wafer manufacturing line. In: *Proceedings of the IEEE international conference on robotics and automation*, Raleigh, NC, 702–708.
- Swaminathan, J.M., 2000. Tool capacity planning for semiconductor fabrication facilities under demand uncertainty. *European Journal of Operational Research*, 120 (2), 545–558.
- White, W., 1983. The queueing network analyzer. *The Bell System Technical Journal*, 62 (9), 2779–2815.
- Yang, F., Ankenman, B.E., and Nelson, B.L., 2007. Efficient generation of cycle time–throughput curves through simulation and metamodeling. *Naval Research Logistics*, 54 (1), 78–93.
- Yang, F., Ankenman, B.E., and Nelson, B.L., 2008. Cycle time percentile curves for manufacturing systems. *INFORMS Journal on Computing*, 20 (4), 628–643.
- Yang, F., 2010. Neural network metamodeling for cycle time–throughput profiles in manufacturing. *European Journal of Operational Research*, 205 (1), 172–185.