Computers & Operations Research 53 (2015) 288–300

Contents lists available at ScienceDirect

Computers & Operations Research

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

The throughput rate of serial production lines with deterministic process times and random setups: Markovian models and applications to semiconductor manufacturing

Woo-sung Kim, James R. Morrison *

KAIST, Department of Industrial and Systems Engineering, Republic of Korea

ARTICLE INFO

Available online 4 April 2014

Keywords:
Serial production line
Flow line
Tandem queues
Clustered photolithography tools
Semiconductor manufacturing

ABSTRACT

We conduct exact analysis of serial production lines with deterministic service durations and various classes of random state-dependent setups between customers. Our focus is on assessing the production rate, or just-in-time (JIT) throughput. We demonstrate that such systems can be modeled as a Markov chain and consider various types of setups inspired by clustered photolithography tools in semiconductor manufacturing. We deduce when exact closed form expressions for the production rate are possible and when a numeric solution to the Markov chain balance equations are required. As these systems have shown promise for modeling process bound clustered photolithography tools, we study their accuracy versus detailed simulation for predicting the tool throughput. Various practical features such as the capacity of a pre-scan buffer and batch customers (to model wafer lots) are investigated.

1. Introduction

Serial production lines, also known as flow lines, with random behavior can serve as prototypical models for a host of manufacturing systems including production and assembly lines; cf., [1–4] or, more recently [5–11]. As such, they have a rich history of research focus and have been applied in many contexts. However, excepting limited classes of systems, exact results do not exist and approximations and simulation methods are typically employed. We endeavor to obtain exact results for a class of such systems with practical application potential in semiconductor manufacturing.

Our performance metric of choice is the maximum production rate or just-in-time (JIT) throughput. The JIT throughput is the long term average rate at which customers exit the system when the input is never starved. That is, it is the throughput when customers arrive just-in-time to the flow line. This is equivalent to the maximum rate at which the system can produce customers. In [12], it was shown that the JIT throughput for two stage serial production lines can be expressed as $1/E[\max\{S_1, S_2\}]$, where E is the expected value and S_i denotes the random service duration for stage i. Based on this result, in [12–14], explicit expressions for the JIT throughput under exponential, Erlang and uniform service distributions were obtained. For the three stage case, it was demonstrated in [15] that the JIT throughput is $1/E[\max\{S_1, S_2, S_3\}]$, where S_i is the random service time for stage i. A variety of papers pursued exact performance results based on this fact under different service time distributions; cf., [15–20]. Beyond these cases, exact analysis does not seem possible for stages with overlapping random service durations; cf., [1] or [4]. Refer also to [21].

Much recent work has focused on approximations based on aggregation [22–26] or decomposition methods [27–31]. These approximations can possess astonishing accuracy and have been used in many practical contexts; cf., [7]. However, like simulation (which can be computationally intensive relative to other methods), they do not as readily provide the same qualitative insights which may be possible via exact results. In addition, these approximations have not explicitly incorporated issues such as setups.

In clustered photolithography tools in semiconductor wafer manufacturing, various types of setups may be required. These setups have a large influence on the JIT throughput of system. While serial production lines have been used to study practical semiconductor manufacturing [32–35], no closed form expressions considering setups have been obtained. Our efforts are motivated by this lack of results. We focus on obtaining exact results and closed form expressions where possible. To this end, we will restrict attention to serial production lines with deterministic service times. While the JIT throughput of such a deterministic system is trivially the inverse of the longest process time, allowing random setups dramatically complicates analysis. Yet, setups are of practical importance and essential for models of certain manufacturing environments.

* Corresponding author.
E-mail address: james.morrison@kaist.edu (J.R. Morrison).
URL: http://xS3D.kaist.edu. (J.R. Morrison).

http://dx.doi.org/10.1016/j.cor.2014.03.022
0305-0548/

For certain classes of randomly occurring setups inspired by clustered photolithography tools (CPTs), we will demonstrate that the maximum production rate can be exactly calculated. This is possible by exploiting the fact that, for JIT customer arrivals, the delay customers experience in the system can be modeled as a Markov chain. In certain cases, explicit expressions for the system throughput can be obtained. In others, the throughput can be calculated by solving for the equilibrium probabilities of the Markov chain. With these results in hand, we endeavor to apply them to models of CPTs. Such tools have a structure that is particularly amenable to our approach. CPTs are essential components of semiconductor wafer fabrication and can be modeled as a serial production line when in the process bound region; cf., [32–35]. We incorporate practical features such as a wafer buffer before the scanner of the CPT and batch arrivals. Generally, the scanner has the longest process time in a CPT and there is a buffer to ensure its throughput. We compare the throughput obtained via our models with detailed simulations; they are acceptable. Despite limitations associated with the model assumptions, we hope that the results may prove useful in practical contexts where analytic throughput expressions are of value.

Though we describe the systems under consideration in detail in the next section, we require a few ideas now. A serial production line consists of M stages, labeled $m_1, \ldots, m_M$. Each stage consists of a single server or machine, which provides service. The process time for a customer at stage m_i is deterministic with duration τ_i. There are no buffers between stages, excepting the arrival buffer before the first stage (it has infinite capacity). This restriction is easy to relax and we will do so in the next section. There is a bottleneck stage; it is the first stage m_B whose τ_B is greater than or equal to all other process times.

Aside from any additional time that may be required for a setup, the service time for every customer in stage m_i is τ_i. Recent theoretical results [36] suggest that it may be possible to relax this assumption and allow service times to depend on the customer. However, significant work will be required to allow this. We hope that the methods developed here will then be applicable in that context.

The classes of setups we consider are as follows. These classes are those for which exact analysis is possible and/or are relevant to modeling setups in CPTs.

- *Type I setup*: *Bottleneck only*. A customer may require the bottleneck stage alone to conduct a setup once it arrives to m_B. The process time at the bottleneck for this customer will appear greater than τ_B.
- *Type II setup*: *A sequence of stages* $m_1, \ldots, m_{S_{II}}$. Before entering the system, a customer may require that all stages prior to and including some distinguished stage $m_{S_{II}}$ be vacant before a setup of those stages commences. Once the setup of those stages is complete, the customer may enter m_1 and start its service. We will assume $S_{II} \le B$.
- *Type III setup*: *Full-flush*. In a full-flush setup, all stages should be vacant before the setup begins. After the setup is complete, the customer may enter the system.
- *Type IV setup*: *A sequence of stages* $m_{S_{IV}}, \ldots, m_M$. Before entering some distinguished stage $m_{S_{IV}}$, a customer may require that all stages $m_{S_{IV}}, \ldots, m_M$ conduct a setup. They must be vacant of customers prior to the commencement of the setup. Once complete, the customer may enter $m_{S_{IV}}$ and proceed with service.

Naturally, a customer may require no setup from any stage. A Type I, II or III setup will be required for a customer with fixed probability independent of all other customers. In the sequel, we will describe how these various setups occur in CPTs. There may be other classes of setups that are amenable to analysis that we did not identify.

Multiple classes of setups may be required for a single customer. For example, some customer may demand a Type II setup from the stages $m_1, \ldots, m_{S_{II}}$, $S_{II} < B$, followed by a Type I setup at the bottleneck. Note that a Type III setup includes both Type II and Type IV setups. As such, we need not consider them together. We consider

- *Type V setup*: *Both Type I and Type II*.
- *Type VI setup*: *Both Type I and Type III*.

We study Type I, II and III setups separately. A Type IV setup that is not subsumed by another class of setup cannot be modeled by our methods. Detailed simulation will be required. We address Type V and VI setups separately. As mentioned, our focus is on the JIT throughput, the maximum production rate, of such a system; call it α. The contributions of the work are as follows.

- For Type I and III setups and Type II setups (with $S_{II} = B$ or $B-1$), we determine α explicitly (Propositions 1–4).
- For Type II setups, with certain conditions on S_{II} to be detailed later, we show that the system can be modeled as a Markov chain (Proposition 6), identify its recurrent states (Lemma 4) and determine when the equilibrium probabilities, and thus α, can be obtained explicitly (Propositions 7 and 8). Otherwise, we give the balance equations that will allow the calculation of α (Proposition 9).
- For Type V and Type VI (with $S_{II} = B$ or $B-1$) setups, we obtain α explicitly (Propositions 10–12).
- Incorporate practical features relevant to CPTs (Section 6).
- Compare the results against those from detailed simulation of a process bound CPT with wafer handling robots (Section 7).

The results are summarized for ease of reference in Table 1.

The remainder of the paper is organized as follows. In Section 2, we describe our systems of interest. We also introduce relevant known results. In Section 3, we derive the JIT throughput under Type I and Type II setups. In Section 4, we focus on Type II setups with a certain condition on S_{II}. We identify cases where the throughput can be explicitly obtained. In Section 5, we consider combination setup Types V and VI. We apply the results to serial production line models of CPTs in Section 6. In Section 7, a detailed simulation of a process bound CPT is compared with our models. The simulation includes wafer transport robots and is based on industrial CPT data from the literature. Concluding remarks are provided in Section 8.

Hereafter, we use the term flow line. It is synonymous with serial production line. An early abbreviated version of a subset of the work reported here appeared in conference form in [37].

2. Preliminaries

We first describe deterministic flow lines and relevant known results. Type I, II and III setups are also detailed.

2.1. Deterministic flow lines

A flow line is composed of a series of stages from which customers receive service in sequence. In each stage, there is a single server or machine which provides service. There is an infinite capacity buffer before the first stage. A finite capacity buffer may be provided for each stage after the first. As detailed in [38], the finite capacity intermediate buffers can be modeled as a