European Journal of Operational Research 255 (2016) 345–356

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Discrete Optimization

Prioritized customer order scheduling to maximize throughput

Yaping Zhao, Xiaoyun Xu*, Haidong Li, Yanni Liu

Department of Industrial Engineering and Management, College of Engineering, Peking University, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 27 July 2015
Accepted 31 May 2016
Available online 10 June 2016

Keywords:
Scheduling
Customer order
Priority
Throughput

ABSTRACT

This study is concerned with a throughput maximization problem of prioritized customer orders. Customer orders with different priorities arrive at a server station dynamically. Each order consists of multiple product types with random workloads. These workloads will be assigned to and processed by a set of unrelated servers. Two commonly applied assignment schemes, named Workload Assignment Scheme (WAS) and Server Assignment Scheme (SAS) are considered. The objective is to determine the optimal assignments under the two assignment schemes to maximize the long-run throughput. Mathematical programming models with relaxed stability constraints are developed for the two assignment schemes, and the adequacy of the programes is guaranteed through fluid limit model analysis. It is shown that these two mathematical programes share the same optimal value, and that there exists a one-to-one correspondence between the optimal assignments. Numerical experiment verifies that the two proposed mathematical programes yield the same optimal throughput, and demonstrates that the corresponding optimal assignments under the two assignment schemes can be transformed into each other.

1. Introduction

Throughput is a widely used performance measure in both manufacturing and service industries. The control of work order throughput times is one of the main management issues in order to achieve a high customer service (Bertrand & Van Ooijen 2002). However, with the ever-increasing diversity and uncertainty in customer demand, businesses today are facing the increasing challenge in improving throughput while keeping service stable in the long run.

To meet the above challenge, this study considers the stochastic customer order scheduling problem to maximize long-run stable throughput. Customer orders arrive at a server station dynamically with rate λ. Every customer order has a randomly assigned priority which is known at the time of its arrival. Each customer order consists of F product types with independent and random workloads. The server station has a total of M servers whose processing speeds are predetermined and unrelated across all product type and server combinations. The workload of any product type can be split arbitrarily and processed independently by each server. The processing of workloads is performed on the preemptive-resume basis according to their corresponding order priorities. A customer order will not leave the system until its entire workloads have been finished.

Two commonly applied assignment schemes, named Workload Assignment Scheme (WAS) and Server Assignment Scheme (SAS), are studied in this paper. Under WAS, all servers are in fixed positions and incoming workloads of product types are assigned to and processed by these servers. A typical example of WAS is the dicing process of semiconductor production (Kim, Kim, Jang, & Chen, 2002). In the dicing process, different types of wafers move as a lot and are diced on a set of parallel equipments. The workloads of different wafer types are distributed among these equipments and processed in parallel. In contrast to WAS, Server Assignment Scheme (SAS) does not require servers to be fixed. Under SAS, workloads of different types constitute their individual queues and servers rotate among these queues to process. The most well-known scheduling algorithm under SAS is called "round-robin" policy. Ward visit in hospitals is a typical example of SAS. In a ward, inpatients usually wait in bed to be attended to, while nurses cruise among the wards to provide health care service. In each patrolling cycle, nurses move from one bed to another until all inpatients have been taken care of. The above two examples are illustrated in Fig. 1. In this study, both WAS and SAS are examined in the context of stochastic customer order scheduling with the presence of order priority.

The objective of this study is to maximize the long-run stable throughput under both WAS and SAS. The decision variable under WAS is the amount of workload of each product type assigned to

* Corresponding author. Tel.:+86 10 82529028.
E-mail addresses: Yaping.zhao@pku.edu.cn (Y. Zhao), xiaoyun.xu@pku.edu.cn (X. Xu), haidong.li@pku.edu.cn (H. Li), liuyanni@pku.edu.cn (Y. Liu).

http://dx.doi.org/10.1016/j.ejor.2016.05.066
0377-2217/

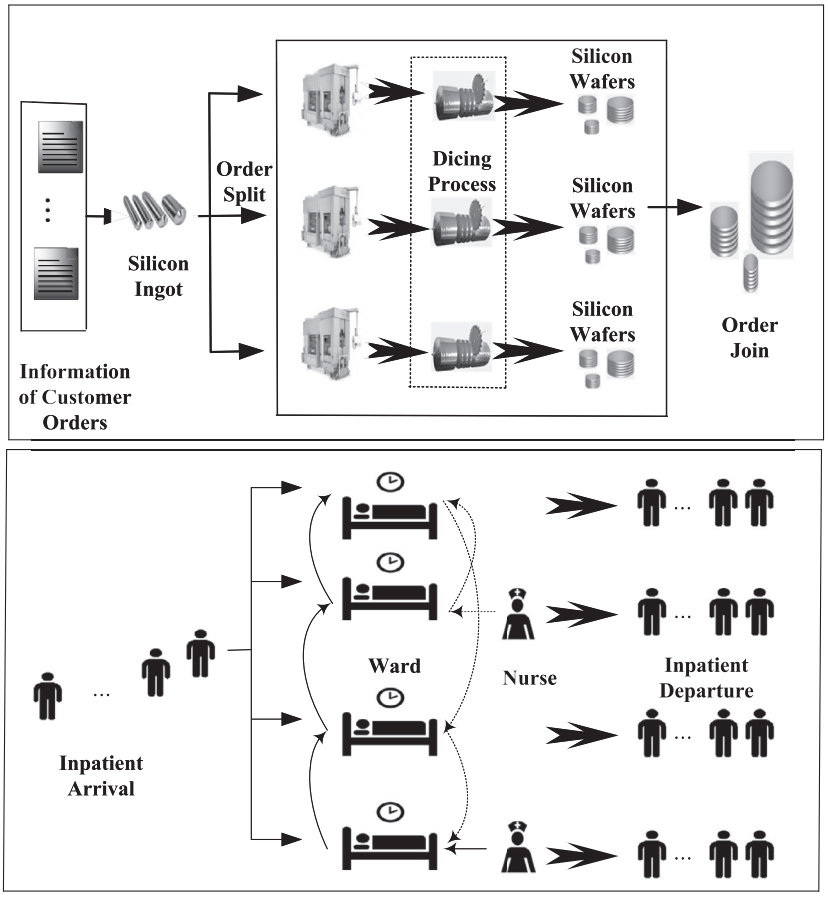

Fig. 1. Compound semiconductor wafer production (WAS) vs. nurses in wards (SAS).

each server, while under SAS, the decision variable is the portion of time each server spends on the workload of each product type. For both assignment schemes, the stability of corresponding queuing systems must be maintained in pursuit of the maximum throughput.

Throughput analysis of scheduling problems has a significant amount of literature (Andradóttir, Ayhan, & Down, 2001, 2003; Askin & Chen, 2006; Dai & Weiss, 1996; Delasay, Kolfal, & Ingolfsson, 2012; Down & Karakostas, 2008; Li, 2004; Li, Blumenfeld, Huang, & Alden, 2009; Liu, Yang, Wu, & Hu, 2012; Patchong & Willaeys, 2001). However, much less exists for priority disciplines, and the majority of the literature focuses on individual jobs instead of customer orders. Morris (1981) presents exact results of mean throughput for closed two node preemptive priority networks with negative exponential service time distributions. These results are extended to the case of state dependent service rates in Rumsewicz and Henderson (1989). As for the extremum of throughput, the maximum throughput of a two-priority queueing system is determined in Chen and Guerin (1991) under assumptions of balanced traffic and unrelated arrivals. The effects of the two priorities on the total maximum throughput are also discussed. The similar problem is further studied in Li, Hu, and Liu (1994) where jobs have fixed service time. In order to determine the stability region where throughput is involved, programmatic procedures are applied in Kumar and Meyn (1995) for queueing systems under certain buffer priority policies.

When customer orders are considered instead of individual jobs, the problem becomes much more challenging. Despite the great number of studies on deterministic customer order scheduling problems (a recent review on this subject can be found in Leung, Li, & Pinedo (2005)), no existing literature on stochastic customer order throughput optimization problem has been found to our knowledge. The difficulty in solving stochastic customer order throughput optimization problem arises from three sources. First, since product types within each customer order have simultaneous arrival and departure, and workload of the same type can be assigned to and processed by more than one server, there exist strong correlations among the servers. This prevents the application of many prevailing queuing analysis approaches as they typically require independence between queues. Second, customer orders may have different priorities, and the one with low priorities may be preempted during the processing by those with higher priorities. The frequent processing interruptions make it more challenging to establish the stability condition for the entire system. Finally, the server station consists of multiple unrelated servers. This suggests that the processing time of the same product type on different servers may vary, and vice versa. To achieve stable throughput, the workloads on all servers have to be simultaneously balanced. The corresponding optimal assignment is particularly difficult to determine when the variation in service speed is large.

To overcome these difficulties, this study utilizes fluid limit model analysis as the primary methodology to address system stability. In a fluid limit model, queueing network is modeled by a piecewise-linear fluid system where the workload arrives and is depleted in a deterministic and continuous manner in each queue (Lan & Olsen, 2006). Fluid limit model analysis is a conventional tool to demonstrate stability of queueing systems operating under a given assignment scheme (see, e.g., Bouchentouf & Sakhi, 2014; Chen & Zhang, 2000; Dai & Li, 2003; Down & Lewis, 2006; Dumas, 1997). Throughout this study, the fluid limit model is used in