European Journal of Operational Research 264 (2018) 106–118

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Production, Manufacturing and Logistics

Measurement and optimization of responsiveness in supply chain networks with queueing structures☆

Sin-Hoon Hum[a], Mahmut Parlar[b], Yun Zhou[c,*]

[a] NUS Business School, National University of Singapore, 1 Business Link, 117592, Singapore
[b] DeGroote School of Business, McMaster University, Hamilton, Ontario L8S 4M4, Canada
[c] Rotman School of Management, University of Toronto, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Received 20 October 2015
Accepted 8 May 2017
Available online 12 May 2017

Keywords:
Supply chain management
Responsiveness
Queueing

ABSTRACT

In this paper we consider supply chains with multiple stages of serial or network structure. The supply chains are endogenous in the sense that they involve queues because each order's lead-time is dependent on the orders already in the system. We define supply chain responsiveness as the probability of fulfilling customer orders within a promised lead-time and study the problems of measuring and optimizing supply chain responsiveness using queueing network models. We first consider a single-server multi-stage serial supply chain and find a closed form expression for the fulfilment time distribution. For the multi-server multi-stage problem, the closed form evaluation of the fulfilment time distribution becomes intractable due to the dependency of the lead-times in different stages. We circumvent this difficulty by proposing a novel FCFS discipline which enables a closed-form analysis. For the multi-server multi-stage Jackson-type supply chain network, to enable analysis, we convert the system into an equivalent single server single stage system with state-dependent rates. For each case, we present detailed numerical examples for both measurement and the optimization of supply chain responsiveness.

1. Introduction

In today's overwhelmingly intricate world, supply chains play essential roles in coordinating various business entities and connecting supply with demand. During the past few decades, the increasing complexity of supply chains due to globalization has imposed many challenges on the practice of supply chain management, among which is to build responsive supply chains to satisfy customer demands.

With a faster pace of life, customers now value time more than they ever did. Consequently, it is common that manufacturing firms quote a promised delivery lead-time to customers. One example is the available-to-promise systems used by many manufacturers such as Dell and Maxtor, which quote delivery due dates for requested orders to customers (Ball, Chen, & Zhao, 2004). Online retailers have also followed suit by offering faster delivery times to their customers. For example, `ebay.com`, `amazon.com` and `walmart.com` launched their same-day delivery services in 2012. Such services, however, have been controversial because maintaining these services has proved to be costly and may not be sustainable; see, Bensinger (2012). Indeed, Blackburn (2012) points out that, in contrast to the theme that "faster is better," time-based competitions are constrained by the marginal value of reducing customer waiting time. As he puts it, "Faster operations are desired only up to the tipping point at which the marginal benefits from additional time reduction equals the marginal cost of the additional speed."

Therefore, it is of particular importance to evaluate and improve the possibility of fulfilling customer orders within a promised lead-time under a realistic budget limitation. To this end, the problems of measurement and optimization of supply chain responsiveness are introduced and studied by Hum and Parlar (2014). Their work defines supply chain responsiveness as the probability of fulfilling a customer order within a quoted lead-time. Compared with other measures of responsiveness such as expected lead-time (see e.g., You & Grossmann (2011)), this definition reflects the likelihood of a customer receiving her order at a preferred due date. Hum and Parlar (2014) primarily focus on exogenous supply chains for which congestion effects are negligible. In such systems, each order arriving at a given stage of the supply chain is relatively small and thus has a minimal effect on overall loading at that stage. As a result,

☆ The authors thank three referees for their constructive comments on earlier versions of the paper that helped improve the exposition. The second author gratefully acknowledges the research support received from the Natural Sciences and Engineering Research Council of Canada.

* Corresponding author.
E-mail addresses: bizhumsh@nus.edu.sg (S.-H. Hum), parlar@mcmaster.ca (M. Parlar), yzhou.zj@gmail.com (Y. Zhou).

http://dx.doi.org/10.1016/j.ejor.2017.05.009
0377-2217/

the lead-time at each stage is assumed to have a fixed and exogenous distribution.

As opposed to exogenous supply chains, Zipkin (2000) defines endogenous supply chain as one where queueing effects are present, i.e., "each order's lead-time depends on the congestion it finds in the supply system." Endogenous supply chains describe the reality more accurately by taking congestion effect into account, one example being the assembly job shop (see e.g., Azaron & Kianfar (2006)).

Exogenous supply chains serve as a good approximation of reality when the capacity of processing orders at each stage of the chain is sufficiently large compared with demands, or both the demand and supply processes are relatively stable so that the waiting plus processing time at each stage of the supply chain is constantly distributed and can be estimated empirically. Endogenous supply chains, on the other hand, enable us to examine the impact of each individual order on the entire system and in turn how the supply chain responsiveness reacts to variations of demand and supply patterns in a capacitated system.

In this paper, we generalize Hum and Parlar (2014) by explicitly modeling the responsiveness of endogenous supply chains involving queueing effects. While this paper considers networks with similar topologies (i.e., serial and Jackson network) as in Hum and Parlar (2014), it is significantly different from Hum and Parlar (2014) by including queueing effects. When queueing of customer orders exists in the system, the fulfilment time distributions becomes much more difficult to derive, and a queueing model is necessary to enable the analysis. Except for a serial system with single servers at each stage (in which case the model is essentially the same as in Hum and Parlar (2014) because of the independent and exponentially distributed duration at different stages), the analyses in the current paper is more involved than Hum and Parlar (2014).

More specifically, we contribute to the literature by studying measurement and optimization of multi-stage supply chains with either serial or network structures using queueing network models. For a single server multi-stage serial supply chain, we obtain the closed-form expression of the c.d.f. of the order fulfilment time. For this type of supply chain, under a budget limitation, we explore structural properties of the optimal service rates at the stages. We find that if the marginal costs for increasing service rates at the stages are ordered, then the optimal rates are in reverse order.

For a multi-server multi-stage serial system, closed-form evaluation of order fulfilment time distribution is intractable. This is so because the presence of multiple servers may result in a later order to overtake an earlier order which makes the lead-times at different stages dependent on each other. This dependency of lead-times has been the main difficulty preventing researchers from exact analysis of sojourn time distribution in queuing networks as we will review in Section 2 below. We circumvent this difficulty by imposing a fair first-come-first-served (FCFS) discipline, under which the system yields the same fulfilment time distribution as one where each stage is a single server system with state-dependent service rates. With this observation, we obtain exact expressions for the fulfilment time distribution in terms of its Laplace transform (LT), which we then invert to evaluate the c.d.f. and to find the optimal rates in each stage.

For a multi-server multi-stage Jackson-type supply chain network, we convert the system into a single stage queueing system with state-dependent service rates that equal the throughput rates of the queueing network. We define the capacity of a supply chain network as the maximum number of orders allowed to be processed simultaneously in the system. Once the number of orders being processed reaches the capacity, new orders will not be accepted (and thus lost) until the next fulfilment of an order. We show that the fulfilment time of a supply chain with finite capacity follows a phase-type distribution. Moreover, as the capacity increases, this fulfilment distribution converges to that of the system (which has the same parameters other than the capacity) with infinite capacity. For this problem we determine the optimal rates numerically. In Section 2 we will review papers which have presented numerical methods for evaluating sojourn time distributions of queueing systems. We believe our research is novel in, (i) providing closed-form expressions for sojourn time (i.e., responsiveness) of serial and Jackson-type supply chain networks with fixed capacity, and (ii) optimization of such networks.

We also study the effect of ignoring queueing in endogenous systems through numerical experiments. We show that ignoring this effect could significantly overestimate the fulfilment probabilities and thus be undesirable. These findings further distinguish our paper from Hum and Parlar (2014).

The rest of the paper is organized as follows. In Section 2, we briefly review the relevant literature. In Section 3, we define the problems of measurement and optimization of supply chain responsiveness where queueing effects are present. In Section 4, we analyze a simple serial supply chain with a single server at each stage and explore the structural properties of the system. Section 5 generalizes the serial system in Section 4 by allowing multiple servers at each stage. In Section 6, we analyze the measurement and optimization of a considerably more general supply chain represented by a Jackson-type queueing network. In Section 7, we examine the impact of ignoring queueing effects in endogenous supply chain systems. Section 8 concludes the paper with a brief summary and possible extensions. Appendix A can be consulted for a list of notation used in the paper. The remaining Appendices B to H present the proofs for lemmas and propositions. The appendices are provided as an online supplement.

2. Literature review

We review the literature on sojourn time distributions in queueing network systems, which are closely related to the methodology used in the current paper. In queueing networks with a general structure, possibility of overtaking has been the main issue preventing researchers from obtaining exact results for the sojourn time distribution. Simon and Foley (1979) show that the sojourn times of a job along a path in a three-node queueing network can be correlated, which makes evaluation of overall sojourn time in the system difficult. Burke (1969) shows that in a serial queueing system with multiple servers, the sojourn time in successive stages can be correlated. Sufficient conditions are provided by Reich (1963) and Burke (1968) for serial queueing systems to have independent sojourn times along the path. Lemoine (1987) develops a set of recursive equations with regard to the Laplace transforms of the sojourn time distributions at the stages in a single-server queueing network. Melamed and Yadin (1984) uses randomization procedures to compute sojourn time distributions approximately, which can result in a very large state space. Kiessler, Melamed, Yadin, and Foley (1988) develop an approximation method by assuming independent sojourn times at different stages in a three-node network and illustrate the efficiency of such an approximation. Grottke, Apte, Trivedi, and Woolet (2011), decompose the queueing network into response time blocks for calculating sojourn time distributions at each individual stage, and obtain a phase-type response time for a given path based on the approximation of independent sojourn times at different stages. Boxma and Daduna (1990) provide a comprehensive review on sojourn times studies of queueing networks as published prior to 1990.

Our work is also related to the literature on workload control in make-to-order systems. Studies on workload control date back to the 1980s (see e.g., Kingsman, Tatsiopoulos, and Hendry, 1989 and Tatsiopoulos & Kingsman, 1983) and focus on controlling order