Push verse pull

Liu, L.; Xu, H.; Zhu, Stuart

Published in:
European Journal of Operational Research

DOI:
10.1016/j.ejor.2020.04.033

Publication date:
2020

European Journal of Operational Research 287 (2020) 119–132

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Production, Manufacturing, Transportation and Logistics

Push verse pull: Inventory-leadtime tradeoff for managing system variability

Liming Liu[a], He Xu[b,*], Stuart X. Zhu[c]

[a] Faculty of Business, Lingnan University, Hong Kong
[b] School of Management, Huazhong University of Science and Technology, Wuhan, China
[c] Department of Operations, University of Groningen, P.O. Box 800, Groningen 9700 AV, the Netherlands

ARTICLE INFO

Article history:
Received 18 September 2018
Accepted 17 April 2020
Available online 28 April 2020

Keywords:
Push
Pull
Order response time
Base stock
Inventory-queue

ABSTRACT

We study a two-stage push–pull system in an assemble-to-order manufacturing environment. Modelling the system as an inventory-queue model, we construct a decision model to determine the optimal stock level of the semifinished base product and the optimal leadtime of the finished products that will minimize the total operational cost. We analytically characterize the structure of the optimal policy. For systems with moderate demand and upstream processing time variabilities, there exists a threshold determined purely by the tradeoff of operational costs so that when the upstream utilization is above the threshold, the push–pull strategy is optimal; otherwise the pure-pull strategy is optimal. When the interarrival time or the upstream service time follows a general probability distribution, the optimal policy depends on the demand or process variability at the upstream stage. Our results can be used to guide managers in selecting the right inventory and leadtime strategy to cope with system variability. We find that in comparison of the downstream variability, under some mild condition, the upstream variability has a more significant impact on the choice of the optimal policy, the corresponding inventory, and lead time. Further, the guaranteed/constant downstream processing time does not always benefit the supply chain performance.

1. Introduction

Push and pull systems have become cornerstones of modern manufacturing practice (Hopp & Spearman, 2008). According to Pyke and Cohen (1990), there are multiple definitions of push and pull in operations literature. The corresponding definition used in our paper is given as follows. Push systems satisfy demand from inventory and can achieve a high capacity utilization, but may cause excessive inventory. Pull systems satisfy demand by production and are flexible to cope with variabilities in demand processes, but may result in a long delivery leadtime. To provide a balance between inventory and leadtime, push–pull systems have been widely implemented in the assemble-to-order manufacturing environment where product variety and order delivery speed are two drivers of the competitiveness of a firm (see Olhager & Östlund, 1990). This hybrid strategy addresses the variety and delivery speed challenge by allowing the production of semifinished base product (SBP) or the ordering of standard items for the whole product family. A SBP represents a standard item used for the final assembly operation. An additional advantage of the push–pull strategy is for the firm to offer customers a competitive promised delivery time (PDT) when an order is received as a tool to build customers' trust (Urban, Sultan, & Qualls, 2000).

In a typical push–pull system, the production of a family of products is often organized into two basic stages. At the upstream stage, an SBP is made. At the downstream stage, order-specific components are assembled onto an SBP to generate an end product for a customer order. The upstream production control is push type. while that of the downstream stage is pull type. Phillips' computed tomography (CT) scanner supply process offers a good example (Serhadli, 2016). Phillips manufactures a basic CT model and stores it in Hamburg, Germany. Upon receiving an order from one of its dealers, a basic model unit is shipped to Amsterdam, where customer-specified components and peripheral devices are assembled to the basic unit for producing an end product, which is then tested and delivered to the customer through its distribution unit within a certain delivery deadline. Since peripheral devices can be supplied promptly by contract manufacturers, Phillips focuses on the inventory holding cost of the basic CT model which accounts for more than 70% of the total cost of a complete CT

* Corresponding author.
E-mail addresses: limingliu@ln.edu.hk (L. Liu), xuhe@hust.edu.cn (H. Xu), x.zhu@rug.nl (S.X. Zhu).

https://doi.org/10.1016/j.ejor.2020.04.033
0377-2217/

Scanner and takes a relatively long leadtime to produce. Meanwhile, the company also needs to achieve a high fulfillment standard by reducing customers' waiting cost that heavily depends on its promised order delivery leadtime. Therefore, the two key decisions faced by the manager are the stock level of the SBP and the promised delivery leadtime. The Phillips' supply system can be treated as a standard push–pull system where the basic CT model can be produced before receiving the customer order. Besides the healthcare industry, this type of assemble-to-order environment is also common for various heavy industries, such as heavy trucks and public buses.

Moreover, the supply process of an Internet retailer, like that of BestBuy.com (Maltz, Rabinovich, & Sinha, 2004) and Hewlett Packard (O'Marah, 2005), can also be seen as a two-stage push–pull process. Usually after an online order is confirmed, an e-retailer "assembles/configures" the order, from items in its warehouse/distribution center, according to the customer's specification and delivers it through a logistics service provider to the customer within a pre-specified time window. The upstream stage is the ordering process of standard products into the warehouse and, to the Internet retailer, the whole supply process clearly follows a push–pull strategy.

Although it seems that push–pull is the preferred supply strategy in a demand-driven business environment, there is another major factor that may affect the choice of the supply strategy. For instance, to cope with an increasing of specific requirements by individual customers, Phillips had to launch more product lines so that the demand rate for each product family becomes smaller, prompting the company to consider reverting the push–pull strategy back to the pure pull strategy under which there is no on-hand stock of the basic CT model. A similar trend occurs in the auto industry. General Motors tries to transform itself into a build-to-order manufacturer from one of the world's biggest build-to-stock operations (Simison, 2000). In many situations, the choice between pure pull and push–pull strategies is not obvious. Many factors are involved, especially the trade-off of inventory and delivery speed often determines the choice of the right supply strategy.

Based on the recent review by Atan, de Kok, Stegehuis, van Boxel, and Adan (2017), the academic literature has not yet proposed a clear guideline on the choice between pure pull and push–pull strategies by considering the integrated decisions of the SBP stock level and the PDT of the finished products. Furthermore, due to the complexity of the system, the literature usually considers inventory and leadtime decisions separately. However, to identify an appropriate supply strategy and manage supply operations effectively, these two decisions need be considered simultaneously. Therefore, the main purpose of this study is to provide an analytical framework for the choice between pure pull and push–pull strategies by simultaneously considering the inventory and leadtime decisions.

To capture the system dynamics, we construct a two-stage system in which an SBP is produced, stored, and then, upon receiving customer order, assembled into an end product for delivery. The supply process uncertainty is represented by the upstream and downstream processing times variability, while the demand uncertainty is represented by the order arrival process variability. Based on this system, we construct a decision model to determine the optimal PDT of the finished products and the SBP inventory level that minimize the total operational cost, including the holding costs of SBP and the finished product, the late delivery cost, and customers' waiting cost. The choice between pure pull and push–pull is implied in the decision of SBP inventory level. This model reflects the tradeoff of inventory and speed. Analyzing this model, we seek to provide insights into the following important issues. At the strategic level, how do firms choose the right supply strategy? At the operational level, what is an appropriate tradeoff of the inventory and delivery speed under a given supply strategy? Further, we also examine how the upstream and downstream capacity and process variabilities affect the supply strategy and optimal inventory and speed tradeoff?

Our main contributions to the related literature are twofold. First, we characterize the structure of the optimal policy for the control of the push–pull system. For systems with moderate demand and upstream stage supply uncertainties, we find that the optimal policy for the choice of supply strategy is a threshold type of the upstream stage capacity utilization. The threshold is completely determined by the tradeoff of different costs. When the utilization is above the threshold, the push–pull policy is optimal; otherwise, the pure pull policy should be used. Second, we explore the impact of demand and process variability on the optimal policy. When demand and or upstream stage supply process uncertainties are not moderate, the policy becomes significantly more complicated. Generally speaking, the threshold is replaced by one or two switching curves. We find that, under some mild condition, while the upstream variability has a significant impact on the tradeoff of inventory cost and speed that determines the choice of supply strategy, the downstream supply process plays a minimum role. Furthermore, we find that, contrary to intuition, eliminating only the downstream processing time variability (by providing a guaranteed constant downstream processing time) will not always benefit the supply chain performance.

The rest of the paper is organized as follows. In Section 2, we briefly review the related literature and highlight our contributions. In Section 3, we define a two-stage supply system, derive the order response time distribution, and present the decision model. Section 4 is devoted to the optimal supply strategy and control policy of the uncongested and congested downstream stage. Section 5 exam the impact of the downstream variability on the choice the supply strategy. In Section 6, we examine the impact of upstream variability on the integration strategy and optimal policy. Finally, we summarize the findings and managerial insights in detail and point out some potential extensions for further research in Section 7. All the proofs are presented in the online supplementary materials.

2. Literature review

Our study is related to two streams of research. The one is literature on the promised delivery time on an inventory-queue system. The other is the literature on setting planned leadtimes in an assembly system.

First, our work is related to research that models production systems as systems of queues or inventory-queues, and examines and compares their performance under push, pull, and push–pull production strategies (see Seimchi-Levi, Kaminsky, & Seimchi-Levi, 2008). Spearman and Zazanis (1992) compare the performance of push and pull systems and find that pull systems are less congested and easier to control. By minimizing the expected holding and backorder costs, Arreola-Risa and DeCroix (1998) consider the choice between push and pull strategies for multiple heterogeneous products in a single-stage system. They find that with a per unit backorder cost, the choice depends only on cost parameters and the arrival rate. When the backorder cost is per unit and per unit time, the production time distribution also plays a role. Glasserman and Wang (1998) study the tradeoff between finished-goods inventory and delivery leadtime in a push production system, and obtain an approximate linear relation between the stock level and the delivery leadtime in the heavy-traffic regime. By modeling the supply/configuration process of each component of a multi-component and multi-stage assembly system as a base-stock $M^X/G/\infty$ inventory-queue, Cheng, Markus, Lin, and Yao (2002) study a network of inventory-queues without

Table 1
Related literature.

Research papers	Considering factors			Decision variables	
	Performance	Optimization	Variability	BSL	Leadtime
Yano (1987)	√				
Arreola-Risa and DeCroix (1998)	√	√		√	
Glasserman and Wang (1998)	√				
Song et al. (2000)		√		√	√
Cheng et al. (2002)	√	√		√	
Liu et al. (2004)	√		√		
Gupta and Benjaafar (2004)	√	√		√	
Axsäter (2005)	√	√			
Alptekinoğlu and Corbett (2010)		√		√	√
Teo et al. (2011)	√	√			√
Cheng et al. (2012)	√	√		√	
Atan et al. (2016)		√			√
Ben-Ammar et al. (2018)		√			√
Jansen et al. (2019)		√			√
Our paper	√	√	√	√	√

congestion and quantify the tradeoff between inventory and the end-customer service level (off-the-shelf availability). Liu, Liu, and Yao (2004) consider the effective allocation of inventories to multiple stocking points in a serial production system. To incorporate the congestion effect, each stage is modeled as a $GI/G/1$ inventory-queue, and a robust job-queue decomposition approach is developed to evaluate the system performance. They investigate the tradeoff between the inventory and end-customer service level and find that effective capacity deployment is no longer the conventional "bowl shaped", but increasing along the line downstream when work-in-process inventory can be controlled.

For the optimization of inventory replenishment and lead time decisions, Alptekinoğlu and Corbett (2010) consider pricing, product variety, and promised leadtime decisions for the design of a production system which is modelled by an $M/M/1$ inventory-queue. They focus on how the tradeoff between leadtime/FGI inventory and product variety informs the design choice between push and pull supply strategies. To address the trade-off between capacity requirements and the amount of working in progress, Teo, Bhatnagar, and Graves (2011) consider how to determine the planned leadtimes and the corresponding time windows for multiple product families under an pure pull manufacturing environment. The authors formulate the model into a non-linear optimization program to find the optimal values of the planning variables. Cheng, Ettl, Lu, and Yao (2012) consider a push–pull production system with a two-stage manufacturing process. The authors develop a nonlinear optimization model to examine the tradeoff between capacity utilization and inventory cost reduction. Ahmadi, Atan, de Kok, and Adan (2019) study the choice of the lead time with a corresponding commitment cost and investigate the impact of the commitment cost on the optimal strategy under a continuous-review setting. For this stream of research,

Gupta and Benjaafar (2004) is the most related to our work. For a two-stage system where each stage is modelled by an $M/M/1$ queue, they determine the optimal base-stock level to minimize the sum of the expected holding cost and backorder cost. They find that the optimal base-stock level depends on the unit holding cost, unit backorder cost, and the utilization at Stage 1 and is independent of the distribution of order response time. By considering the leadtime as a decision variable, we use the late delivery penalty cost to characterize the impact of backorder. We show that the optimal base-stock level depends on both cost parameters and the distribution of the order response time. Further, we consider a general arrival process and a general processing time, which allows us to investigate the impact of variability on the optimal strategies and performance.

Second, our work is also related to the problem of setting planned leadtimes given order due dates and random component processing (or procurement) leadtimes and final assembly times. Yano (1987) considers a two-stage assembly systems with two components and finds numerically for some cases negative safety leadtime for at least one component and substantial safety leadtime for the final assembly. Song, Yano, and Lerssisuriya (2000) study component planned leadtime and order quantity decisions when the customer order due date is given by the quantity is random. The end-product is assembled from n different components with random procurement leadtimes but constant final assembling time. Axsäter (2005) suggests an approximation decomposition technique to determine the order release times at each stage for a multi-stage assembly system. Recently, in comparison of the existing approaches, Atan, Kok, Dellaert, Janssen, and Boxel (2016) develop a faster and more accurate heuristic to compute the optimal planned leadtimes at each stage of a multi-stage assembly system. A numerical optimization method is applied to confirm the accuracy of the heuristic. Ben-Ammar, Dolgui, and Wu (2018) focus on the determination of optimal order release dates with stochastic lead times for each component at each level, and they deveop a branch and bound algorithm to minimize the sum of inventory holding and backlogging costs. Jansen, Atan, Adan, and de Kok (2019) introduce a concept of a "blame policy" with a new holding cost accounting scheme, which can be applied to planned lead-time optimization problems for a assembly system structure.

In short, we use Table 1 to highlight our contributions to the existing literature. Different from the first research line, we explicitly investigate how integrated decisions of semi-finished product inventory and delivery leadtime inform the choice of the pure pull strategy and the push–pull strategy. Different from the above ATO literature where production/procurement leadtimes and order due date are exogenous, we consider finite production and final assembling capacities and the integrated decisions of the SBP stock level and the promised order delivery time. We focus on the choice between push–pull and pure pull strategies while the ATO literature focusing on component inventory replenishment policies or safety leadtime decisions. Further, we investigate the impact of demand, upstream and downstream variability on the optimal strategy and system performance.

3. Model formulation

We consider a two-stage system that supplies a family of products customized from a single SBP. The upstream stage produces the SBP while the downstream stage assembles auxiliary and custom components on the SBP based on specific customers'

requirements. The firm does not keep end-product inventory to maintain supply flexibility and avoid high inventory cost, but may keep SBP inventory to shorten the order response time (ORT). We assume that the inter-arrival time of customer orders follows a certain probability distribution with the mean given by $1/\lambda$. In other words, λ can be treated as the average demand rate of customers. Denoted by T, the ORT is the time period from customer order confirmation to delivery of the ordered product to the customer. By the definition of the ORT, T is a random variable that is affected by both the demand uncertainty and the random processing time. In Section 4, we will show how the probability distribution of T is derived. As stated in the Philip's example, since the availability of the auxiliary components can be guaranteed by contract manufacturers and the impact of those components on the cost efficacy is much less important than the SBPs, we do not consider the operational decisions of the components for the tractability of the model. Further, to reflect customers' differentiation at the downstream stage, the assembly time is usually a general random variable, indicating different assembly time for different orders.

To model this supply system, we note that the SBP (e.g., the basic CT unit) usually takes a very large portion of the total product cost. It is reasonable to assume that the production facility for the SBP is flexible and its setup cost is relatively insignificant. Ignoring the setup cost, we may assume a base-stock policy for the SBP at Stage 1. This makes the resulting model tractable while still capturing the tradeoff between inventory and fulfillment service. An eloquent modeling justification of the base-stock control policy for ATO systems can be found in Song and Yao (2002). The firm makes an integrated decision on PDT, denoted by l, and the base-stock level (BSL) of the SBP, denoted by B, to minimize its expected total cost rate. When the optimal B is found to be 0, the firm adopts a pure pull strategy and intentionally keeps no SBP inventory. With this model setting, the strategic-level decision on the supply strategy is implied in the integrated decision on l and B. This setup allows the supply system to be either in a pure pull mode when no SBP inventory is maintained or a push–pull mode with intermediate SBP stock.

The decision maker is conscious of the impact of waiting time on customers' buying intentions and imposes a customer waiting cost w per order per unit time in the decision objective. Other costs important to the decision include the SBP holding cost s per unit per unit time, the late delivery penalty cost p per order per unit time to compensate the customer for late delivery arising when the ORT is longer than the PDT, and the end product holding cost c per unit per unit time when the order is completed earlier than the time specified by the PDT. It is usually more costly to hold the end product than to hold SBP, and hence we assume that $c > s$. We also must have $p > w$, since otherwise the firm could just quote an unreasonably short PDT and take the risk of paying the late delivery penalty.

In summary, the assumptions are listed as follows.

- The policy for the SBP at Stage 1 is a base-stock type.
- The production capacity at Stage 2 is sufficiently large.
- The unit holding cost of a finished product is higher than that of a SBP, i.e., $c > s$.
- The unit penalty cost is higher than the unit waiting cost, i.e., $p > w$.

The notation is summarized in Table 2.

We construct the following integrated decision model

$$\min_{B\geq 0,\ \ell\geq 0} TC(B,\ell) = \lambda[cE(\ell - T)^+ + pE(T-\ell)^+ + w\ell] + sE(I), \quad (1)$$

where the first term in the brackets is the expected inconvenience cost, then the second is the expected delay penalty cost, and the third term means the waiting cost per customer. Thus, the summation of these terms represents the expected leadtime cost to serve one customer and then is multiplied by the average demand rate. $sE(I)$ is the holding cost of the expected steady-state inventory level of SBP at the downstream of Stage 1.

Table 2
Model notation.

T	The ORT
B	The base-stock level
ℓ	The PDT
w	The customer waiting cost per unit per unit time
s	The SBP holding cost per unit per unit time
p	The late delivery penalty cost per order per unit time
c	The end product holding cost per unit per unit time
λ	The average arrival rate of customers
$R(t,\ B)$	The distribution function of T
$\Phi(\cdot)$	The cumulative distribution function of process time at Stage 2
ρ	The utilization at Stage 1
μ_1	The mean production rate at Stage 1
μ_2	The mean production rate at Stage 2

Let $R(t,\ B)$ be the distribution function of the ORT T. We have the following convexity property of $TC(B,\ \ell)$ with respect to ℓ for any given B. The proof of this Lemma 1 as well as all the other proofs in this paper are given in the appendix.

Lemma 1. *For any given B, $TC(B,\ \ell)$ is strictly convex in ℓ. The optimal PDT is uniquely given by*

$$R(\ell^*, B) = \frac{p-w}{p+c}. \quad (2)$$

Remark 1. For any given B, (2) gives a newsvendor-type solution to the optimal PDT in which $c+w$ is the overage cost of increasing the PDT by one unit while $p-w$ is the underage cost of decreasing the PDT by one unit.

4. A threshold policy of upstream utilization

Consider a benchmark model with a Poisson demand process of single-unit orders at rate λ. Stage 1 has a single processing unit with exponential processing time at rate μ_1. With a base-stock policy for the SBP output, Stage 1 is then an $M/M/1$ inventory-queue, which allows endogenous/load-dependent leadtimes for the SBP. We assume that there is no congestion delay at Stage 2, i.e., the assembly operation for an order can always start immediately as long as an SBP is available. The downstream stage is then a modified $M/G/\infty$ queue as in Cheng et al. (2002). This is equivalent to assuming an exogenous processing leadtime/ample capacity at Stage 2, which is reasonable for many of today's agile/quick-response supply chains. For example, the delay to the assembly operation of customized CT units due to the number of outstanding customer orders is likely insignificant. Further, the setting of the general assembly time can represent customers' specific requirements for their own finished products.

When there is an SBP stock out at Stage 1 at the arrival of an order, a work-in-process (WIP) delay occurs to this order at Stage 2 so that Stage 2 is not a standard $M/G/\infty$ queue. With Poisson demands, the probability of a WIP delay at Stage 2 equals the stock-out probability at Stage 1 by the PASTA theorem (Wolff, 1989). Let us denote the possible WIP delay by T_B. For an arbitrary order, conditioning on the job queue length N at Stage 1, the WIP delay is given by

$$T_B = \begin{cases} 0, & if \quad N \leq B-1; \\ W_{1r}, & if \quad N = B; \\ \sum_{j=1}^{N-B} W_1^j + W_{1r}, & if \quad N \geq B+1, \end{cases} \quad (3)$$

where W_1^j is the service time of the jth job in the job queue at Stage 1 and W_{1r} is the steady state residual processing time at Stage 1. The ORT is the sum of T_B and the exogenous processing

time W_2 at Stage 2. With an ample capacity at Stage 2, the state at Stage 1 and hence the WIP delay are independent of W_2 so that the distribution function $R(t, B)$ of T is the convolution of the distribution function of W_2 and that of T_B.

Lemma 2. *The distribution function of the ORT for the above defined two-stage system is given by*

$$R(t,B) = \Phi(t) - \rho^B \int_0^t e^{-(\mu_1-\lambda)(t-u)} d\Phi(u), \tag{4}$$

where $\Phi(\cdot)$ is the cumulative distribution function of W_2 and $\rho = \lambda/\mu_1$ is the utilization at Stage 1.

In the following subsections, we first characterize the optimal policies for two scenarios: a general downstream delay and a constant downstream delay. We then compare of these two scenarios.

4.1. Random downstream delay

For the system defined above, the average SBP inventory level $E(I)$ is $B - \rho(1-\rho^B)/(1-\rho)$. Substituting (4) into (1), we have

$$TC(B,\ell) = \lambda\left[c\int_0^{\ell}(\ell - t)dR(t,B) + p\int_{\ell}^{+\infty}(t-\ell)dR(t,B) + w\ell\right] + s\left(B - \frac{\rho(1-\rho^B)}{1-\rho}\right), \tag{5}$$

The corresponding first-order conditions (FOC) are

$$\bar{\Phi}(\ell^*) + \rho^{B^*} e^{-(\mu_1-\lambda)\ell^*}\int_0^{\ell^*} e^{(\mu_1-\lambda)u} d\Phi(u) = \frac{c+w}{c+p}, \tag{6}$$

$$s + \frac{(p+s)\rho^{B^*+1}\ln(\rho)}{1-\rho} - \lambda(c+p)\rho^{B^*}\ln(\rho)\int_0^{\ell^*}\int_0^t e^{-(\mu_1-\lambda)(t-u)} d\Phi(u)dt = 0, \tag{7}$$

where $\bar{\Phi}(x) = 1 - \Phi(x)$.

By Lemma 1, for any given B, the optimal PDT can be obtained from (6). Treating ℓ as a function of B, i.e., $\ell^*(B)$, our original problem is reduced to a single-variable optimization problem $TC(B, \ell^*(B))$. We first develop a set of bounds for the optimal PDT $\ell^*(B)$.

Lemma 3. *$\ell^*(B)$ is decreasing in B and bounded, i.e., $\underline{\ell} \le \ell^*(B) \le \bar{\ell}$ where*

$$\underline{\ell} = \bar{\Phi}^{-1}\left(\frac{c+w}{c+p}\right), \tag{8}$$

and $\bar{\ell}$ is uniquely given by

$$\bar{\Phi}(\bar{\ell}) + \int_0^{\bar{\ell}} e^{-(\mu_1-\lambda)(\bar{\ell}-u)} d\Phi(u) = \frac{c+w}{c+p}. \tag{9}$$

The lower bound is obtained with $B \to \infty$, i.e., when the two stages are completely decoupled. Clearly, the minimum PDT depends only on the downstream processing rate and cost parameters. The upper bound is obtained by setting $B = 0$. As shown below, it is the optimal PDT when the supply system is pure pull.

Lemma 4. *When the cumulative distribution $\Phi(t)$ of the downstream delay is log-concave, $TC(B, \ell^*(B))$ is convex if*

$$\bar{\ell} < \Phi^{-1}\left[\frac{p+s}{p+c}\right], \tag{10}$$

and first concave and then convex otherwise.

Lemma 4 characterizes the property of TC on the assumption that the cumulative distribution of the downstream delay ($\Phi(t)$) is log-concave in t. This assumption is needed to ensure that the second derivative of $TC(B, \ell^*(B))$ with respect to B is increasing in B. This assumption is very mild. Many common distributions, such as Uniform, Normal, Erlang, Gamma, and Beta are included in this distribution family. Based on Lemma 4, we have the following results.

Theorem 1. *Under the condition that $\Phi(t)$ is log-concave, there exists a stage-1 utilization threshold $\tilde{\rho}$ uniquely given by*

$$\frac{s}{s+w} + \frac{\tilde{\rho}\ln(\tilde{\rho})}{1-\tilde{\rho}} = 0, \tag{11}$$

such that:

(i) *When $\rho \ge \tilde{\rho}$, for both the convex and concave–convex cases, the optimal strategy is push–pull with $B^* > 0$, and (B^*, ℓ^*) is uniquely determined by FOC (6) and (7);*

(ii) *When $\rho < \tilde{\rho}$ and TC is convex, the optimal strategy is pure pull with $B^* = 0$ and $\ell^* = \bar{\ell}$, where $\bar{\ell}$ is given by (9);*

(iii) *When $\rho < \tilde{\rho}$ and TC is concave–convex, the optimal strategy is pure pull with $B^* = 0$ and $\ell^* = \bar{\ell}$ if*

$$\bar{\ell} \le \Phi^{-1}\left[\frac{p-w}{p+c} - \frac{s}{p+c}\frac{(1-\rho)}{\rho\ln\rho}\right]; \tag{12}$$

otherwise, the optimal strategy could be pure pull or push–pull, i.e.,

$$(B^*, \ell^*) = \arg_{B,\ell}\min\{TC(0,\bar{\ell}), TC(\tilde{B},\tilde{\ell})\},$$

where $(\tilde{B}, \tilde{\ell})$ is determined by FOC (6) and (7).

Theorem 1 indicates that under quite general conditions, the upstream stage capacity utilization level determines the choice of supply strategy. When the demand rate relative to the upstream processing speed is sufficiently large, a push–pull strategy should be chosen. Otherwise, a pure pull strategy is usually optimal. How large the demand rate is sufficiently large depends only on the tradeoff of the SBP holding cost s and the customer waiting cost w. On the one hand, when the SBP holding cost s increases, it is more costly to hold the SBP stock, which motivates the firm to use a pure pull strategy. On the other hand, when customer waiting cost w increases, the firm intends to quote a short leadtime in order to balance the negative consequence of the waiting cost. Thus, the firm has to use a push–pull strategy for offering a short leadtime. This finding illustrates Philips' case quite closely. With the increasing of specific requirements from customers, the number of basic CT models increases. As a result, orders based on each basic CT model decrease significantly, which results in a low capacity utilization. This forces Philips reverting back to using the pure pull supply strategy.

The evolution of Amazon's distribution strategy provides another illustration of the managerial insights from Theorem 1. Amazon started with direct shipment from publishers/wholesellers after receiving online customer orders, which is essentially a pure pull strategy without its own distribution centers and stocks between suppliers and customers. This strategy worked initially, but as Amazon's business volume grew and demand rate increased, and as customers becoming more demanding on fast delivery (waiting cost increases), the pure pull strategy started hurting the business and reputation, Amazon had to switch to the push–pull strategy building its own distribution centers with stocks for rapid picking and delivery on receiving customer orders. But at what level of demand should Amazon switch from pure pull to push–pull strategy? It is obviously not a easy question to answer for Amazon as well as many other online businesses. Theorem 1 suggests a way to estimate the switching point based on $s/(s+w)$ shown in (11). We note that somewhat unexpectedly the delay penalty cost is not involved in the determination of the threshold. This may be explained by noting that the choice of supply strategy

is made for the trade off of SBP inventory and final product delivery time, so only the two corresponding costs are involved in the determination of the threshold.

The optimal switching policy can also be applied to interpret the postponement strategy implemented by manufacturing firms. When the SBP holding cost is relatively low, the required upstream capacity utilization for the push–pull strategy to be optimal will also be lower. In practice, this strategy can also be interpreted as *form postponement*. Although the product family shares a common SBP, the forms of the end products depend on customers' specific requirements. Therefore, to hedge the demand uncertainty caused by the variety of the end products, firms should be more likely to implement the form postponement strategy by holding a certain amount of SBPs and fulfilling orders by customizing the SBP to end products, as Dell Computer did for its customer orders. When the SBP holding cost is relatively high, a higher utilization is required for the push–pull or form postponement strategy to be optimal. On the other hand, the pure pull strategy is more likely suitable with the PDT being set to the upper bound $\bar{\ell}$. This situation may be called *time postponement*. The idea is to avoid starting the production of SBP in anticipation of the demand when the overall demand is relatively slow and the holding cost of SBP is relatively high. The CT scanner example discussed earlier fits in this scenario nicely. The price of each end product is around 1 million US dollars so that the SBP is also very costly while the demand rate, only from very few specific sectors, such as large hospitals and research institutions, is also rather low. Therefore, it is cost-efficient for the CT manufacturer to implement (switch to) the pure pull strategy.

4.2. Constant downstream delay

Theorem 1 shows that the policy threshold is independent of the specific distribution form of the downstream delay within the log-concave distribution family. Since the log-concave distribution family covers the whole range of coefficient of variation (CV), we conjecture that the downstream delay does not affect the policy threshold. One exception, however, is the case of constant downstream delay. Thus, we consider system with a constant downstream delay which does not satisfy the log-concave condition in Theorem 1. Let the downstream deterministic delay be μ_2^{-1}. The distribution of the ORT is reduced to

$$R(t,B)=\begin{cases}0, & t<\mu_2^{-1}\\ 1-\rho^B e^{-(\mu_1-\lambda)(t-\mu_2^{-1})}, & t\geq\mu_2^{-1}\end{cases}. \tag{13}$$

By (13), if $\ell\leq\mu_2^{-1}$ and assuming $p>w$, the optimal PDT is equal to μ_2^{-1}. In the following, we focus on the nontrivial case with $\ell\geq\mu_2^{-1}$. For a constant μ_2^{-1}, the objective function can be simplified to

$$\begin{aligned}TC(B,\ell)=&\frac{p\lambda\rho^B}{\mu_1-\lambda}+c\lambda(\ell-\mu_2^{-1})\\&+w\lambda\ell-\lambda\rho^B(c+p)\left[\frac{1-e^{-(\mu_1-\lambda)(\ell-\mu_2^{-1})}}{\mu_1-\lambda}\right]\\&+s\left[B-\frac{\rho(1-\rho^B)}{1-\rho}\right].\end{aligned} \tag{14}$$

Lemma 5. *For any given B, the optimal PDT is given by*

$$\ell^*(B)=\begin{cases}\frac{1}{\mu_1-\lambda}\ln\left(\rho^B\frac{c+p}{c+w}\right)+\mu_2^{-1}, & B\in[0,\log_\rho\frac{c+w}{c+p}]\\ \mu_2^{-1}, & B\in(\log_\rho\frac{c+w}{c+p},+\infty)\end{cases}; \tag{15}$$

and $\ell^*(B)$ is decreasing in B and bounded, i.e., $\underline{\ell}\leq\ell^*(B)\leq\bar{\ell}$, where $\underline{\ell}=\mu_2^{-1}$ and

$$\bar{\ell}=\frac{1}{\mu_1-\lambda}\ln\frac{c+p}{c+w}+\mu_2^{-1}.$$

Then, the optimal policy is given by the theorem below.

Theorem 2. *There exists a utilization threshold $\tilde{\rho}$ uniquely given by*

$$1+\frac{c+w}{s+w}\ln\frac{p+c}{w+c}=\frac{s}{w+s}\frac{(\tilde{\rho}-1)}{\tilde{\rho}\ln\tilde{\rho}}\left[1-\ln\left(\frac{s(\tilde{\rho}-1)}{\tilde{\rho}(p+s)\ln\tilde{\rho}}\right)\right], \tag{16}$$

such that

(i) *When $\rho\geq\tilde{\rho}$, the optimal strategy is push–pull with $B^*=\log_\rho\left(\frac{s(\rho-1)}{(s+p)\rho\ln\rho}\right)$ and $\ell^*=\mu_2^{-1}$;*

(ii) *When $\rho<\tilde{\rho}$, the optimal strategy is pure pull with $B^*=0$ and*

$$\ell^*=\frac{1}{\mu_1-\lambda}\ln\left(\frac{c+p}{c+w}+\mu_2^{-1}\right). \tag{17}$$

It is interesting to note that while the policy structure is very simple, the switching threshold is much more complicated. Unlike in the case of general random downstream delay when the threshold is determined by the tradeoff between the SBP holding cost and customer waiting cost, FGI holding cost and delay penalty cost are also involved in the complicated tradeoff. The reason can perhaps be found in Theorem 2 itself. When the optimal strategy is to hold the SBP stock, i.e., push–pull the firm should quotes the constant downstream service time as the optimal PDT to avoid FGI completely. However, late delivery may still occur. So it is important to consider the tradeoff between FGI holding cost and late delivery penalty cost in addition to the tradeoff between SBP holding cost and waiting cost to decide when to adopt the push–pull strategy.

4.3. Optimal policy of a congested downstream stage

This section studies the impact of downstream congestion delay on the optimal policy. When the downstage stage only has a finite capacity, it is essential to incorporate the congestion delay at Stage 2. Here, we assume that the downstream stage is modelled by a single server with the general-distributed assembly time. The upstream stage is modelled by an $M/M/1$ queue. For such a system, Lee and Zipkin (1992) suggest an approximation for evaluating the performance measure of a multistage production system by assuming that the different stages are operated independently. The authors claim that their approximation is sufficiently accurate to be used to find the optimal base-stock level. Based on their approximations, the downstream stage can be treated as an $M/G/1$ queue. Then, we can derive the distribution of ORT given by

$$R(t,B)=\Psi(t)-\rho^B\int_0^t e^{-(\mu_1-\lambda)(t-x)}d\Psi(x), \tag{18}$$

where $\Psi(t)$ is the cumulative probability distribution of the sojourn time in the $M/G/1$ queue. Note that $\Psi(t)$ includes the processing time and the waiting time. The waiting time reflects the impact of congestion on the system while $\Phi(t)$ in Eq. (4) only includes the process time.

In the following, we first present the bounds and then the optimal policy by assuming that $R(t, B)$ is approximated by (18).

Lemma 6. *$\ell^*(B)$ is decreasing in B and bounded, i.e., $\underline{\ell}\leq\ell^*(B)\leq\bar{\ell}$, where*

$$\underline{\ell}=\Psi^{-1}\left(\frac{c+p}{c+w}\right), \tag{19}$$

and $\bar{\ell}$ is uniquely given by

$$\bar{\Psi}(\bar{\ell})+\int_0^{\bar{\ell}}e^{-(\mu_1-\lambda)(\bar{\ell}-x)}d\Psi(x)=\frac{c+w}{c+p}. \tag{20}$$

By following the similar argument like Lemma 4, we can show that under the condition that $\Psi(t)$ is log concave, if

$$\bar{\ell}<\Psi^{-1}\left[\frac{p+s}{p+c}\right],$$

TC is convex; otherwise, *TC* is first concave then convex otherwise.

Analyzing the property of *TC*, we characterize the optimal policy for the congested model.

Theorem 3. *Under the condition that $\Psi(t)$ is log-concave, the optimal policy is threshold-type (the threshold $\tilde{\rho}$ is given by (11)), i.e.,*

(i) *When $\rho \geq \tilde{\rho}$, the optimal policy is push–pull with $B^* > 0$ in both the convex and concave–convex cases. (B^*, ℓ^*) is uniquely determined by*

$$s + (p+s)\frac{\rho^{B^*+1}\ln\rho}{1-\rho} - \lambda(c+p)\rho^{B^*}(\mu_1-\lambda) \times \int_0^{\ell^*}\int_0^t \bar{\Psi}(t-x)e^{-(\mu_1-\lambda)x}dxdt = 0, \tag{21}$$

$$\bar{\Psi}(\ell^*) + (\mu_1-\lambda)\rho^{B^*}\int_0^{\ell^*}\bar{\Psi}(\ell^*-x)e^{-(\mu_1-\lambda)x}dx - \frac{c+w}{c+p} = 0; \tag{22}$$

(ii) *When $\rho < \tilde{\rho}$, the optimal policy is pure pull with $B^* = 0$ and $\ell^* = \bar{\ell}$ in the convex case, where $\bar{\ell}$ is given by (20);*

(iii) *When $\rho < \tilde{\rho}$ and the cost function is concave–convex, if*

$$\bar{\ell} \leq \bar{\Psi}^{-1}\left[\frac{p-w}{p+c} - \frac{s}{p+c}\frac{(1-\rho)}{\rho\ln\rho}\right], \tag{23}$$

the optimal policy is pure pull with $B^* = 0$ and $\ell^* = \bar{\ell}$; otherwise, the optimal policy could be pure pull or push–pull, i.e.,

$$(B^*, \ell^*) = \arg_{B,\ell}\min\{TC(0,\bar{\ell}), TC(\tilde{B},\tilde{\ell})\},$$

where $(\tilde{B},\tilde{\ell})$ is determined by (21) and (22).

Theorems 1 and 3 demonstrate that the same threshold policy holds for systems with both the uncongested and the congested downstream stage. Both theorems show that the thresholds only depend on the utilization of the upstream stage. The reason may be that when ℓ is chosen as a decision variable in our model, no matter whether the downstream stage is congested or not, the quotation of ℓ has to cover the downstream service time for a single product. The choice of B is to maintain an appropriate stock level to enhance the upstream stage capacity. Thus, the optimal policy is only determined by the utilization at Stage 1.

Similar to the uncongested system discussed in Section 4.2, we also study the constant downstream delay for the congested system. Suppose that the downstream stage is modelled by an $M/D/1$ model. The approximation of the sojourn time for the $M/D/1$ model is given by

$$P(T \leq t) = 1 - \frac{1-\rho_2}{\rho_2 + r_0 - 1}e^{-r_0 t}, \tag{24}$$

where r_0 is given by $\rho_2(e^{r_0}-1) - r_0 = 0$, $r_0 > 1$, and $\rho_2 = \lambda/\mu_2$. The approximation is proposed by Roberts, Mocci, and Virtamo (1996). By numerical experiments, the authors find that this approximation is quite accurate. Since the approximation is an exponential distribution, we can prove that the optimal policy has the same structure as Theorem 3.

5. Impact of downstream-process-time variability

Here, we compare the optimal policy of these two cases to examine the impact of the downstream variability on the choice of the supply strategy.

Proposition 1. *The policy threshold is lower when the downstream delay is constant than when the downstream delay is random.*

Somewhat unexpectedly, pure pull strategy can sustain higher upstream capacity utilization when the downstream delay is random than when it is deterministic. This suggests that when the pure pull strategy is more desirable, it is counterproductive to force a rigid downstream operation/processing time, given the existence of uncertainty in the upstream operation.

We now examine how constant and random downstream delays affect performance differently under both pure pull and push–pull strategies. For tractability, exponential distribution is assumed here for the downstream delay, and we use subscripts d and u to indicate the constant delay and exponential delay, respectively. In the following two propositions, by substituting $\Phi(t) = 1 - e^{-\lambda t}$ into the results of Theorem 1, we can derive the expressions of ℓ_u^* and B_u^*. ℓ_c^* and B_d^* can be derived from Theorem 2. For the brevity, the expressions of the optimal solutions are presented in the corresponding proofs.

Proposition 2. *Under the pure pull strategy:*

(i) *When $\mu_2 < \mu_1 - \lambda$, $\ell_d^* \geq \ell_u^*$;*

(ii) *When $\mu_2 \geq \mu_1 - \lambda$, $\ell_d^* \leq \ell_u^*$ if $(c+w)/(c+p) \leq \bar{\gamma}$, where $\bar{\gamma}$ is uniquely given by*

$$\frac{\mu_2 e^{-(\mu_1-\lambda)/\mu_2} - (\mu_1-\lambda)e^{-1}\bar{\gamma}^{\mu_2/(\mu_1-\lambda)-1}}{\mu_2 - (\mu_1-\lambda)} - 1 = 0,$$

and $\ell_d^* > \ell_u^*$ otherwise.

We note that $\bar{\gamma} = e^{2-e}$ when $\mu_2 = \mu_1 - \lambda$.

We further analyze the optimal PDT under the pure pull strategy. From Lemma 1, the optimal PDT is given by $R(\ell^*, 0) = (p-w)/(c+p)$. If and only if μ_2 is smaller than the residual upstream capacity $\mu_1 - \lambda$, we can show that the ORT of the constant model is stochastically larger than that of the exponential model. By comparing (4) with (13), and the constant model has a longer optimal PDT than the exponential model, independent of cost parameters (Proposition 2(i)). When $\mu_2 \geq \mu_1 - \lambda$, the comparison of PDT depends on the cost parameters. When the unit overage cost $c+w$ is relatively small, the firm's focus is to avoid the underage cost caused by the potential delay. In this case, the exponential model may require a PDT longer than that of the constant model.

Proposition 3. *Under the push–pull strategy:*

(i) *$B_d^* \geq B_u^*$;*

(ii) *$\ell_d^* \geq \ell_u^*$ if $(c+w)/(c+p) \geq [\mu_2 e^{-(\mu_1-\lambda)/\mu_2} - (\mu_1-\lambda)e^{-1}]/[\mu_2 - (\mu_1-\lambda)]$ when $\mu_2 \neq \mu_1 - \lambda$ and $2e^{-1} \leq (c+w)/(c+p) \geq 0.736$ when $\mu_2 = \mu_1 - \lambda$;*

(iii) *$\ell_d^* \leq \ell_u^*$ if $(c+w)/(c+p) \leq 0.368$.*

Part (i) of Proposition 3 shows that the optimal BSL for the constant model is always higher than that for the exponential model. For the optimal PDT, the exponential model is more efficient except when the delay penalty cost p is significantly greater than the waiting cost w as indicated in Propositions 2(ii) and 3(iii).

To understand the effects of downstream variability better, we examine how the downstream delay variability affects the optimal PDT and BSL, and the minimal total cost numerically. We use constant, two-stage Erlang, exponential, and two-stage hyperexponential distributions in increasing coefficient of variation (CV); and set parameter values at $\lambda = 6, \mu_1 = 10, \mu_2 = 7$ and $s = 0.1, c = 1, p = 4, w = 2$. We maintain the downstream mean service time constant at $1/\mu_2$ for all the three distributions. We perform sensitivity analysis by varying μ_2 (capacity parameter) and λ (demand parameter), one at a time. The following six figures plot the effect of the variability on the PDT, BSL, and the total cost. Note that the values of parameters in numerical instance are based on the project with Philips Healthcare by Serhadli (2016).

First, we observe from Figs. 1 and 2 that PDT decreases while BSL increases in μ_2. Note that μ_2 measures the downstream production rate. When μ_2 increases, the average downstream

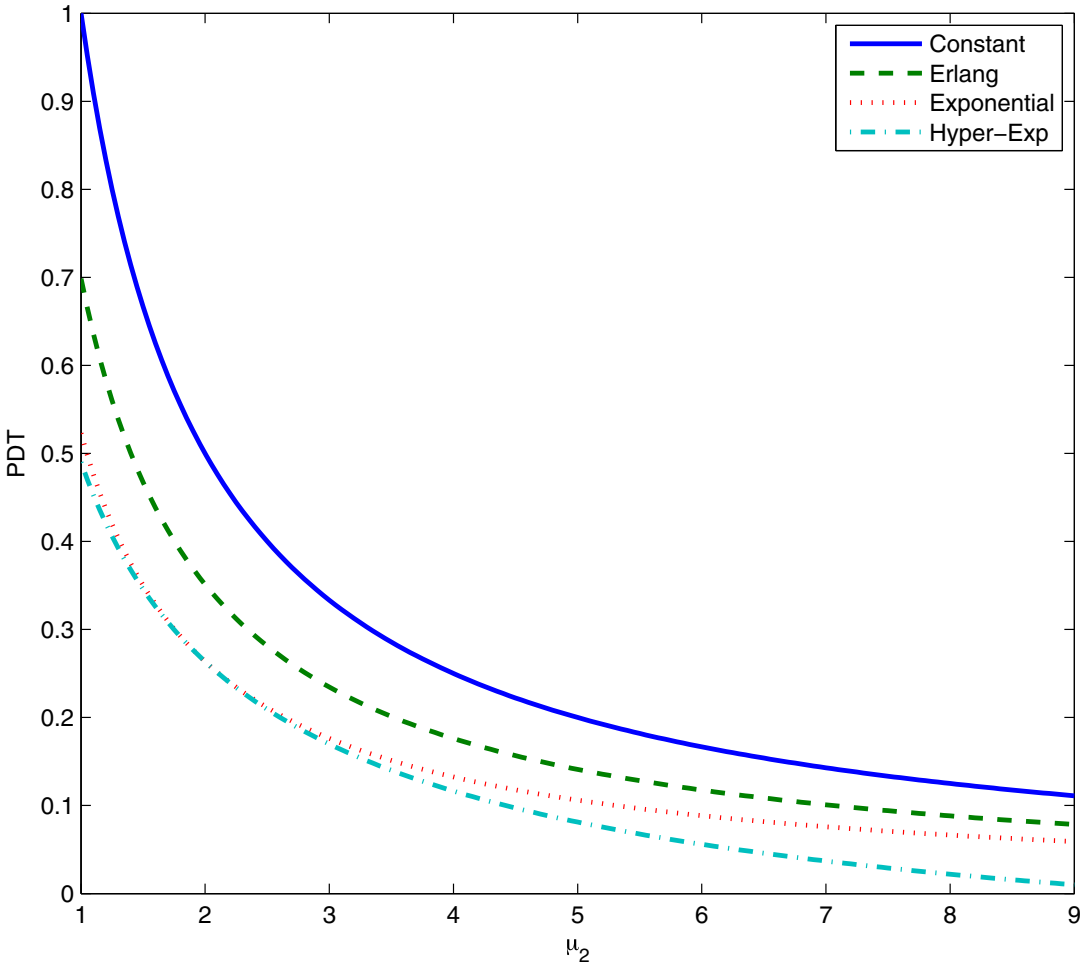

Fig. 1. Impact of the variability on PDT with respect to μ_2.

processing time becomes shorter. Thus, the firm can reduce PDT. Meanwhile, the firm has to increase BSL in order to maintain a sufficient amount of SBP stock and decrease the chance of process starvation at the downstream stage. A smaller CV usually leads to a longer PDT but a lower BSL, except reverting to the highest BSL when CV reduces to 0 (constant).

Second, with increasing λ, Fig. 3 shows that the optimal BSL increases independently of the downstream delay CV. When demand rate λ increases, the firm has to maintain a high level of SBP stock to fulfill a increasing number of orders. As shown in Fig. 4, the PDT on the other hand almost remains constant, but longer PDT is required for smaller downstream delay CV.

Third, we observe from Figs. 5 and 6 that the total cost always increases in CV except the case of constant. A deterministic downstream delay always leads to a higher total cost. From Figs. 1 to 4, it is interesting to find that the constant model requires a higher BSL and a longer PDT. Based on the cost function given by (1), we find that a higher BSL yields a higher inventory holding cost of SBP and a longer PDT yields a higher expected leadtime cost, which results in greater total cost than the three other models.

In summary of the above propositions and numerical results, one interesting managerial insight is that a deterministic downstream delay may not bring any performance advantages and cost savings compared with stochastic downstream delays. This findings is supported by the Philips case, where Philips prefers to perform assembly operations in house than outsourcing those operations to a third-party logistics provider (3PL) even if the 3PL offers a fixed processing time. There may be a simple explanation, that is, when the supply delay consists multiple interrelated components, a deterministic components among random ones is usually not positive to the overall effect of the delay.

6. The impact of demand and upstream-processing-time variability

We show in Section 4 that the downstream variability does not affect the strategy switching policy and the deterministic downstream delay is in general not preferable compared with the stochastic downstream delay. These findings are based on moderate upstream and demand variabilities (CV=1). In this section, we address the impact of demand and process variabilities on the optimal switch policy at the upstream stage. Here, we assume that demand variability is characterized by the inter-arrival time variability, and the process variability by the upstream service time variability, respectively.

6.1. Demand variability

For the impact of demand variability, we model the upstream stage by a $GI/M/1$ inventory-queue and the downstream stage by a $GI/M/\infty$ queue (with WIP delay). By the standard queueing theory (see page 395, Wolff, 1989), the distribution of the upstream order sojourn time is $\exp[(1-\alpha)\mu_1]$, where α is uniquely given by

$$\alpha = \tilde{A}[(1-\alpha)\mu_1]. \tag{25}$$

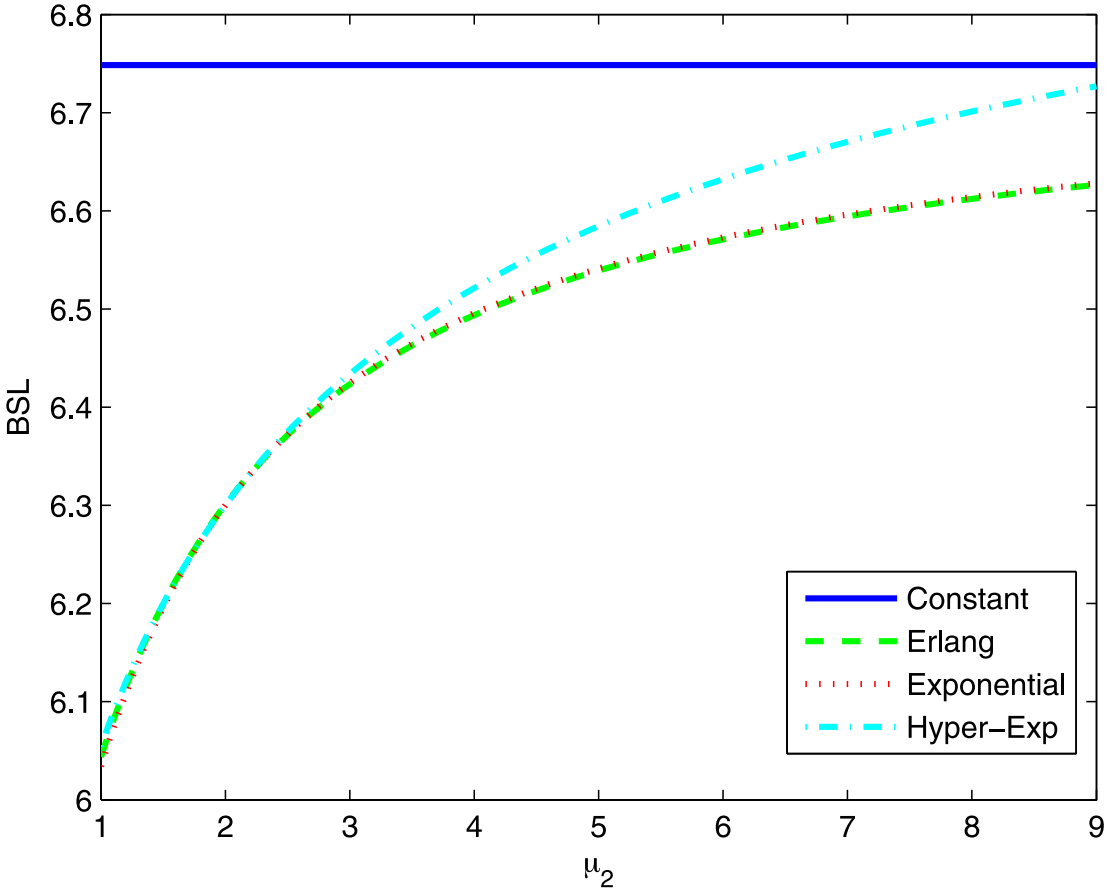

Fig. 2. Impact of the variability on BSL with respect to μ_2.

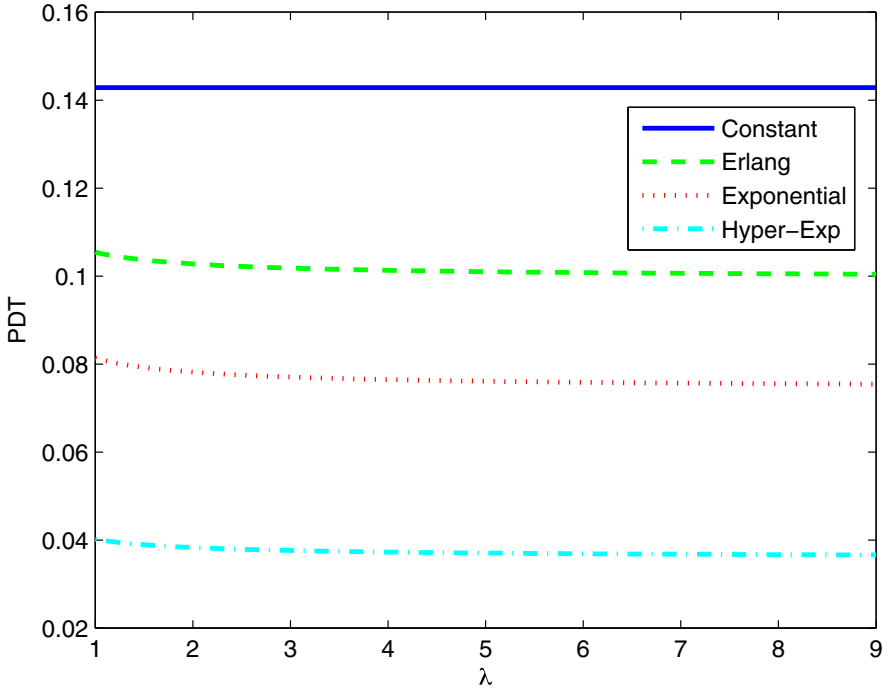

Fig. 3. Impact of the variability on PDT with respect to λ.

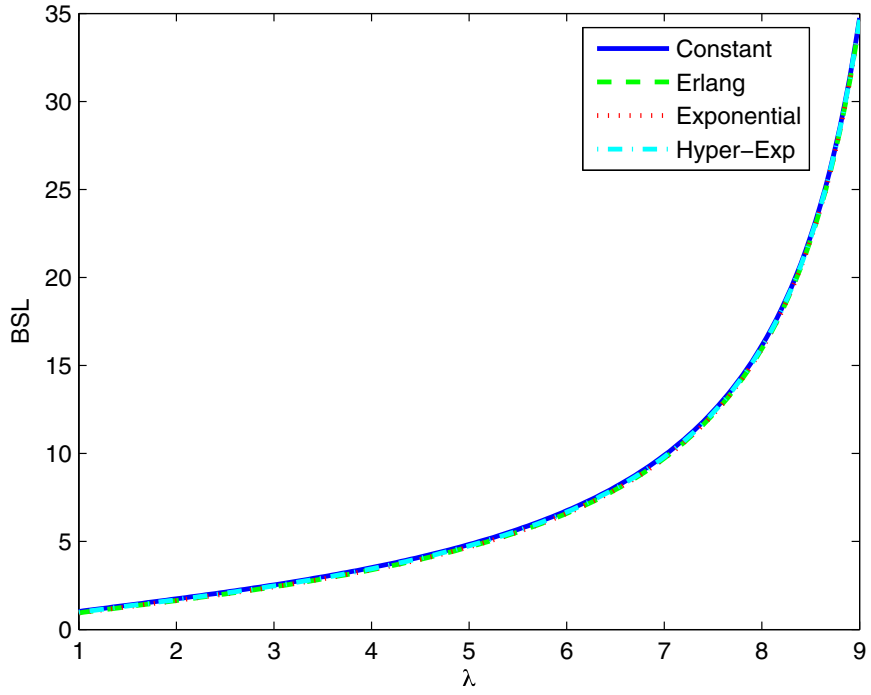

Fig. 4. Impact of the variability on BSL with respect to λ.

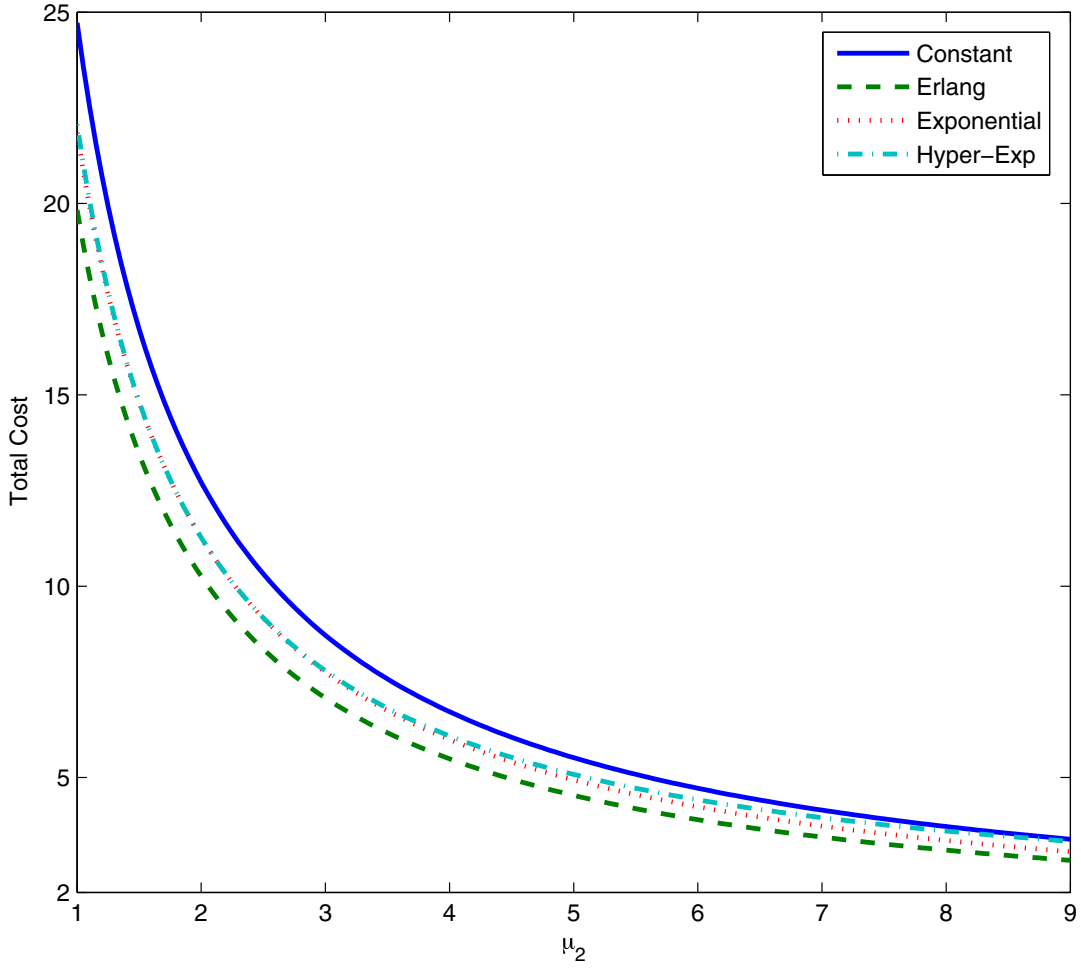

Fig. 5. Impact of the variability on the total cost with respect to μ_2.

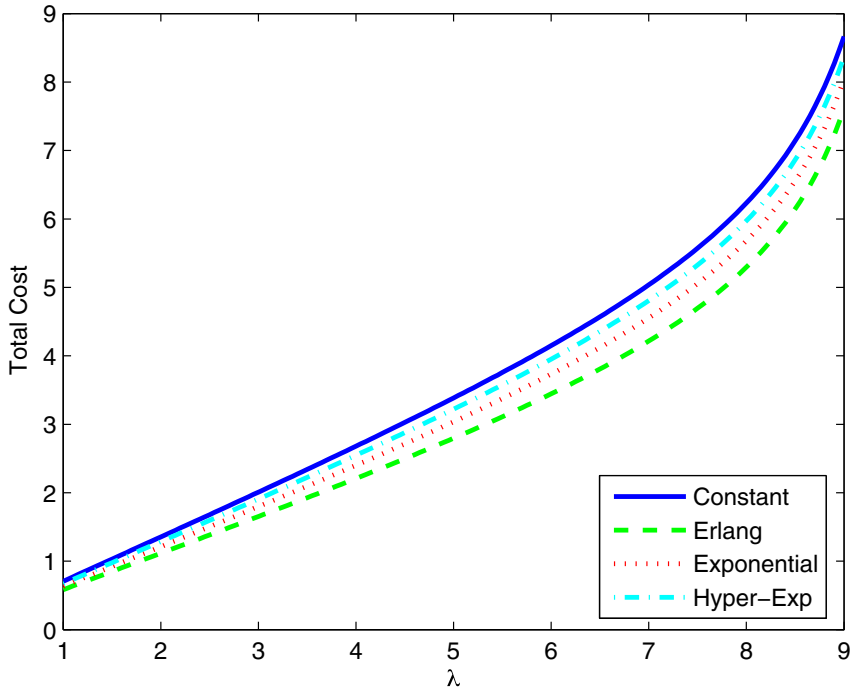

Fig. 6. Impact of the variability on the total cost with respect to λ.

Note that $0<\alpha<1$ is the probability that the system is busy at the demand arrival epoch, whereas $\tilde{A}$ is the Laplace–Stieltjes transform of the inter-arrival time distribution.

Similar to Lemma 2, we can derive the distribution of ORT, i.e.,

$$R(t,B) = \begin{cases} 1-e^{-\mu_2 t}-\frac{\mu_2\rho\alpha^{B-1}}{\mu_2-\mu_1(1-\alpha)}[e^{-\mu_1(1-\alpha)t}-e^{-\mu_2 t}], & \mu_2\neq(1-\alpha)\mu_1; \\ 1-e^{-\mu_2 t}-\mu_1(1-\alpha)t\rho\alpha^{B-1}e^{-\mu_1(1-\alpha)t}, & \mu_2=(1-\alpha)\mu_1. \end{cases} \tag{26}$$

The total cost is then given by

$$\begin{aligned} TC(B,\ell) = {} & p\lambda\left(\frac{1}{\mu_2}+\frac{\alpha^B}{\mu_1(1-\alpha)}\right) \\ & + (c+w)\lambda\ell-\lambda(c+p)\frac{1-e^{-\mu_2\ell}}{\mu_2} \\ & -\lambda(c+p)\frac{\mu_2\alpha^B}{\mu_2-\mu_1(1-\alpha)}\left(\frac{1-e^{-\mu_1(1-\alpha)\ell}}{\mu_1(1-\alpha)}-\frac{1-e^{-\mu_2\ell}}{\mu_2}\right) \\ & + s\left(B-\frac{\rho(1-\alpha^B)}{1-\alpha}\right). \end{aligned}$$

We derive the corresponding first-order conditions: for $\mu_2\neq(1-\alpha)\mu_1$,

$$e^{-\mu_2\ell}+\frac{\mu_2\alpha^B}{\mu_2-\mu_1(1-\alpha)}\left[e^{-\mu_1(1-\alpha)\ell}-e^{-\mu_2\ell}\right]-\frac{c+w}{c+p}=0, \tag{27}$$

$$\begin{aligned} & s-(c-s)\frac{\rho\alpha^B\ln\alpha}{1-\alpha} \\ & -(c+p)\frac{\rho\alpha^B\ln\alpha}{1-\alpha}\frac{\mu_1(1-\alpha)e^{-\mu_2\ell}-\mu_2 e^{-\mu_1(1-\alpha)\ell}}{\mu_2-\mu_1(1-\alpha)}=0, \end{aligned} \tag{28}$$

and for $\mu_2=(1-\alpha)\mu_1$,

$$e^{-\mu_2\ell}+\mu_1(1-\alpha)\ell\alpha^B e^{-\mu_1(1-\alpha)\ell}-\frac{c+w}{c+p}=0, \tag{29}$$

$$s-(c-s)\frac{\rho\alpha^B\ln\alpha}{1-\alpha}+(c+p)(\mu_2\ell+1)e^{-\mu_2\ell}\rho\alpha^B\frac{\ln\alpha}{1-\alpha}=0. \tag{30}$$

Lemma 7. *$\ell^*(B)$ is decreasing in B and bounded, i.e., $\underline{\ell}\le\ell^*(B)\le\bar{\ell}$, where*

$$\underline{\ell}=\frac{1}{\mu_2}\ln\frac{c+p}{c+w}, \tag{31}$$

and $\bar{\ell}$ is uniquely given by

$$\begin{cases} e^{-\mu_2\ell}+\frac{\mu_2}{\mu_2-\mu_1(1-\alpha)}[e^{-\mu_1(1-\alpha)\ell}-e^{-\mu_2\ell}]=\frac{c+w}{c+p}, & \text{if } \mu_2\neq(1-\alpha)\mu_1; \\ e^{-\mu_2\ell}+\mu_1(1-\alpha)\ell e^{-\mu_1(1-\alpha)\ell}=\frac{c+w}{c+p}, & \text{if } \mu_2=(1-\alpha)\mu_1. \end{cases} \tag{32}$$

Lemma 8. *Let $\hat{\mu}_2$ be the unique solution to*

$$\hat{\mu}_2\left(\frac{c-s}{c+p}\right)^{\mu_1(1-\alpha)/\hat{\mu}_2}-\hat{\mu}_2\frac{c+w}{c+p}+\mu_1(1-\alpha)\frac{s+w}{c+p}=0. \tag{33}$$

(i) *For $\mu_2\neq(1-\alpha)\mu_1$, TC is convex if $\mu_2\le\hat{\mu}_2$ and first concave then convex otherwise.*

(ii) *For $\mu_2=(1-\alpha)\mu_1$, TC is convex if*

$$\ln\left(\frac{c+p}{c-s}\right)\le\frac{s+w}{c-s},$$

and first concave then convex otherwise.

With the above results, we can now present the optimal policy for the supply strategy.

Theorem 4. *The optimal policy for the (GI/M/1, GI/M/∞) system is:*

(i) *When $\frac{\rho\ln\alpha}{(1-\alpha)}<-\frac{s}{s+w}$, the optimal policy is push–pull with $B^*>0$, and (B^*,ℓ^*) is uniquely determined by (27) and (28) if $\mu_2\neq(1-\alpha)\mu_1$ and by (29) and (30) if $\mu_2=(1-\alpha)\mu_1$;*

(ii) *When $\frac{\rho\ln\alpha}{(1-\alpha)}\ge-\frac{s}{s+w}$ and $\mu_2\le\hat{\mu}_2$, the optimal policy is pure pull with $B^*=0$ and $\ell^*=\bar{\ell}$;*

(iii) *When $\frac{\rho\ln\alpha}{(1-\alpha)}\ge-\frac{s}{s+w}$ and $\mu_2>\hat{\mu}_2$, the optimal policy is pure pull with $B^*=0$ and $\ell^*=\bar{\ell}$ if*

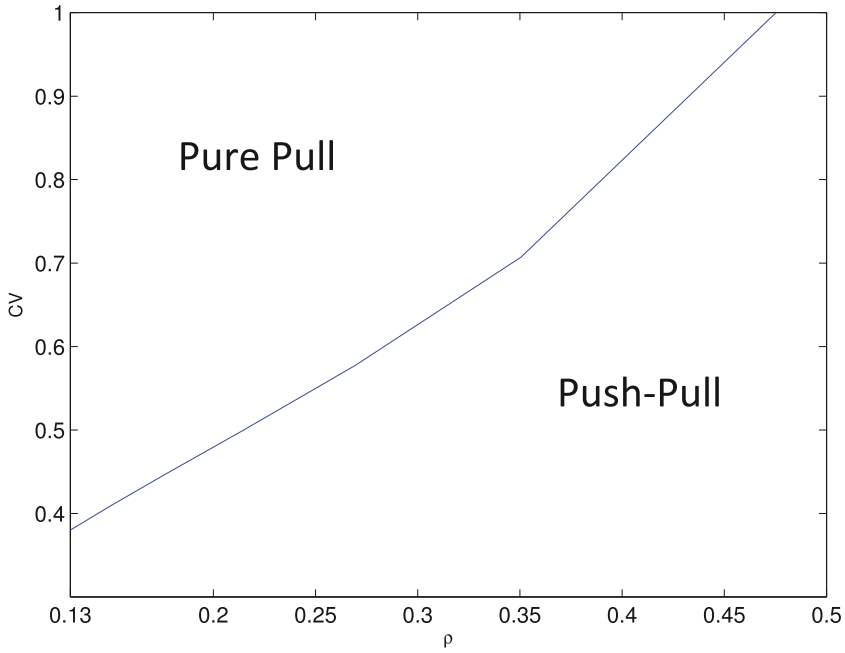

Fig. 7. Switching curve policy under demand variability.

$$\bar{\ell} \le -\frac{1}{\mu_2}\ln\left[\frac{s(1-\alpha)}{(c+p)\rho\ln\alpha}+\frac{c+w}{c+p}\right], \tag{34}$$

otherwise, the optimal policy could be pure pull or push–pull, i.e.,

$$(B^*,\ell^*) = \arg_{B,\ell}\min\{TC(0,\bar{\ell}), TC(\tilde{B},\tilde{\ell})\},$$

where $(\tilde{B},\tilde{\ell})$ *is determined by (27) and (28) for* $\mu_2 \neq (1-\alpha)\mu_1$ *and by (29) and (30) for* $\mu_2 = (1-\alpha)\mu_1$.

We note that α is (at least approximately) a function of the demand CV and the upstream stage utilization. Theorem 4 shows that the optimal policy is a switching curve type of ρ and the demand CV. We use Fig. 7 to illustrate the findings based on the switching curve policy. Here, we assume that the inter-arrival time follows a k-stage Elrang distribution. Fig. 7 shows that the policy switching curve divides the feasible policy space into two, with one for the pure pull strategy and the other for the push–pull strategy. When the demand variability increases, the demand variability dominates the choice of the optimal strategy, i.e., the pure pull strategy is preferred in order to reduce the risk of holding stock. When utilization increases, utilization dominates the choice of the optimal strategy, i.e., the push–pull strategy should be implemented to hedge the risk of stockout.

6.2. Upstream variability

We model the upstream stage by an $M/G/1$ inventory-queue and the downstream stage by $M/M/\infty$. For an $M/G/1$ queue, the explicit analytical expression for the distribution of the ORT is not available. Thus, we derive an approximation of the ORT distribution based on the approximated queue length distribution by Buzacott and Shanthikumar (1993) and Liu et al. (2004), i.e.,

$$P_n = \begin{cases} 1-\rho, & n=0,\\ \rho(1-\sigma)\sigma^{n-1}, & n\ge 1,\end{cases} \tag{35}$$

where P_n is the probability that the length of queue length is n, $\bar{C}_s$ is the coefficient of variation of the service time at the upstream stage,

$$\sigma = \frac{(1+\bar{C}_s^2)\rho}{2(1-\rho)+(1+\bar{C}_s^2)\rho}.$$

Lemma 9. *Given the queue length distribution (35), the ORT distribution is given by*

$$R(t,B) = 1 - e^{-\mu_2 t} - \mu_2\sigma^{B-1}\int_0^t \bar{G}(t-x)e^{-\mu_2 x}dx, \tag{36}$$

where $\bar{G}(t) = 1-G(t)$ *and* $G(t)$ *is the cumulative probability distribution of the stationary waiting-time distribution in the M/G/1 queue.*

From (35), the expected inventory level is given by $E(I) = B - \rho(1-\sigma^B)/(1-\sigma)$, and we can then obtain the expected total cost function

$$TC(B,\ell) = \lambda\left[p\left(\frac{1}{\mu_2}+\frac{\sigma^B}{\mu_1(1-\sigma)}\right)+(c+w)L-(c+p)\int_0^L \bar{R}(t,B)dt\right] + s\left[B-\frac{\rho(1-\sigma^B)}{1-\sigma}\right]. \tag{37}$$

By (2), we treat $\ell^*(B)$ as a function of B and obtain

$$\frac{d^2TC}{dB^2} = \sigma^{B-1}(\ln\sigma)^2 \times\left[(s+p)\frac{\rho\sigma}{1-\sigma}-(c+p)\mu_2\left(\int_0^{\ell}H(t)dt+\frac{H(\ell)}{\ln\sigma}\frac{d\ell^*(B)}{dB}\right)\right],$$

where

$$H(t) = \int_0^t \bar{G}(t-x)e^{-\mu_2 x}dx,$$

$$\frac{d\ell^*(B)}{dB} = \frac{H(\ell)\ln\sigma}{\int_0^{\ell}g(l-x)e^{-\mu_2 x}dx+\sigma^{1-B}e^{-\mu_2\ell}-\rho_1 e^{-\mu_2\ell}}.$$

The complicated d^2TC/dB^2 makes it extremely difficult (though possible) to characterize the property of TC. Here, we consider two common distributions for the upstream service time: two-stage hyper-exponential and two-stage Erlang. For these two distributions, we can prove that $TC(B, \ell^*(B))$ is unimodal in B. Then, the optimal policy depends on the signs of the two first-derivatives at

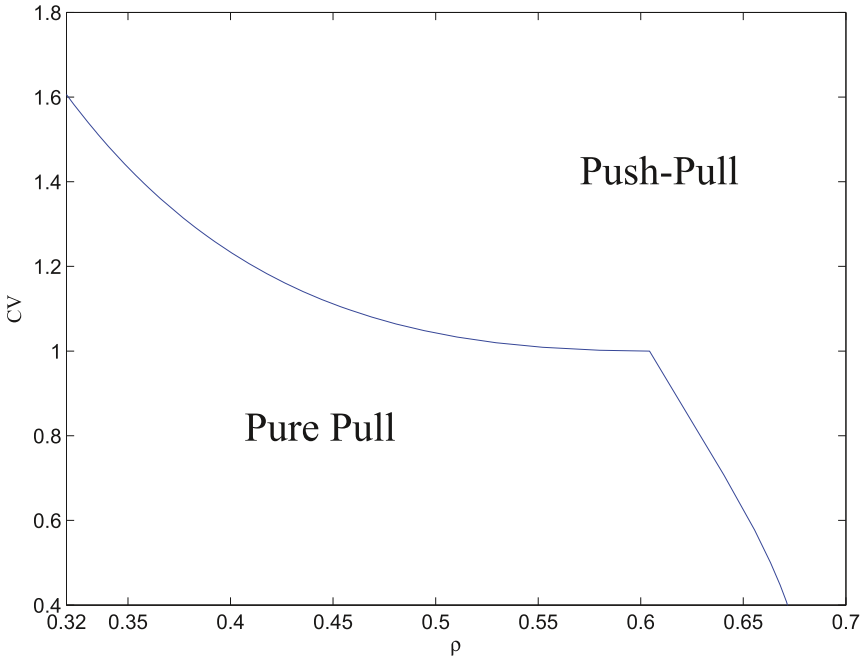

Fig. 8. Switching curve policy under upstream variability.

$B = 0$ and at $B = \hat{B}$, where $B = \hat{B}$ is obtained from $d^2TC/dB^2|_{B=\hat{B}} = 0$. Define $\Delta_1 = dTC/dB|_{B=0}$ and $\Delta_2 = dTC/dB|_{B=\hat{B}}$, where

$$\frac{dTC}{dB} = \lambda\left[p\frac{\sigma^B \ln(\sigma)}{\mu_1(1-\sigma)} - (c+p)\mu_2\sigma^{B-1}\ln(\sigma)\int_0^{\ell} H(t)dt\right] + s\left(1 + \frac{\rho\sigma^B \ln(\sigma)}{1-\sigma}\right). \quad (38)$$

Proposition 4. *Under the assumption that the service time follows a two-stage hyper-exponential or a two-stage Erlang distribution, the optimal policy is given by:*

(i) *When* $\Delta_1 < 0$, *the optimal supply strategy is push–pull with* (B^*, ℓ^*) *being uniquely given by the first order conditions based on (37);*
(ii) *When* $\Delta_1 \geq 0$ *and* $\Delta_2 \geq 0$, *the optimal supply strategy is pure pull with* $B = 0$ *and* $\ell^* = \bar{\ell}$;
(iii) *When* $\Delta_1 \geq 0$ *and* $\Delta_2 < 0$, *the optimal supply strategy could be pure pull or push–pull, i.e.,* $(B^*, \ell^*) = \arg_{B,\ell}\min\{TC(0, \bar{\ell}), TC(\tilde{B}, \tilde{\ell})\}$, *where* $\tilde{B}$ *and* $\tilde{\ell}$ *are uniquely given by the first order conditions based on (37).*

Proposition 4 indicates that when the service time is a general random variable, the optimal policy is characterized by two curves, i.e., $\Delta_1 = 0$ and $\Delta_2 = 0$. By (38), the optimal policy mainly depends on the variability of the upstream processing time and the upstream utilization. Therefore, the structure of the optimal policy is similar to a switching-curve type. Under the assumption that the processing time follows a two-stage hyper-exponential distribution, Fig. 8 shows that when the variability of processing time is relatively high and utilization is relatively low, the pure pull strategy should be applied; when the variability of processing time is relatively low and utilization is relatively high, the push–pull strategy should be used. We find that the sharp decrease is caused by the change of the function that characterizes the switching curve. When ρ is lower than 0.6, the function is like a convex function while it becomes concave for $\rho > 0.6$. Different from the findings of Fig. 7, either the system variability or utilization cannot dominate the choice of the optimal strategy.

7. Conclusion and future extensions

For a two-stage system that supplies a family of products customized from a single SBP, the key issue is to find an appropriate supply strategy so that a firm can efficiently operate this system and provide customers with a competitive PDT. We formulate a decision model for the optimal PDT and the optimal BSL for the SBP. This model integrates the strategic level integration policy and the operational level inventory and speed tradeoff decisions. We find that under some mild condition, the optimal policy is a threshold type, and the threshold is only on the utilization of the upstream stage when the variability of the upstream processing time is moderate. The thresholds are determined entirely by the system cost structure. When the utilization is above the threshold, the push–pull policy is optimal; otherwise, the pure pull policy is optimal. However, when either the service time or the inter-arrival time at the upstream stage follows a general distribution, the optimal policy depends on the demand variability or the process variability at the upstream stage.

Our second finding is about the impact of the downstream variability by the comparison between the constant and uncertain downstream service time. We would expect that the guaranteed downstream processing and delivery time is beneficiary to the supply chain performance. However, the analytical and numerical results are quite unexpected: (1) Under the push–pull strategy, the optimal BSL for SBP inventory for the constant model may be higher than that for the uncertain model, and we can find conditions under which the optimal PDT for the uncertain model is shorter than that for the constant model; (2) Under the pure pull policy, depending on system design and cost parameters, either the constant model or the uncertain model is more efficient; (3) The threshold for the constant model is always lower than that for the uncertain model; (4) the uncertain model may be more cost-efficient than the constant model.

There are a number of limitations of this work that could be potential direction for future research. First, we may consider a decentralized two-stage supply chain where BSL and PDT decisions are made independently. Under this setting, we want to see how the upstream stage chooses between pull and push modes and

how the independent decisions may be coordinated. Next, both pricing and time are key marketing decisions, affecting market shares and profitability. It is challenging to combine the pricing decision with inventory and quoted leadtime determinations and let the demand rate be affected by the price and leadtime strategies. Further, although the issue of inventory replenishment and final assembly allocations with multiple components and multiple products is out of the scope of the current study, it is worth extending my model by incorporating these issues. As a starting point, We refer readers to the excellent reviews by Song and Zipkin (2003) and Atan et al. (2017).

Acknowledgments

The authors are grateful to the referees and the editor for their constructive suggestions that significantly improved this study. The research was partly supported by Netherlands Organisation for Scientific Research under Grant 040.21.010 and National Natural Science Foundation of China under grant 71831007, 71871099, 72042017.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2020.04.033 .

References

Ahmadi, T., Atan, Z., de Kok, T., & Adan, I. (2019). Optimal control policies for an inventory system with commitment lead time. *Naval Research Logistics, 66*(3), 193–212.

Alptekinoğlu, A., & Corbett, C. (2010). Leadtime-variety tradeoff in product differentiation. *Manufacturing & Service Operations Management, 12*, 569–582.

Arreola-Risa, A., & DeCroix, G. A. (1998). Make-to-order versus make-to-stock in a production-inventory system with general production times. *IIE Transactions, 30*(8), 705–713.

Atan, Z., de Kok, A. G., Dellaert, N. P., Janssen, F. B. S. L. P., & van Boxel, R. (2016). Setting planned leadtimes in customer-order-driven assembly systems. *Manufacturing & Service Operations Management, 18*(1), 122–140.

Atan, Z., de Kok, T., Stegehuis, C., van Boxel, R., & Adan, I. (2017). Assemble-to-order systems: a review. *European Journal of Operational Research, 261*(3), 866–879.

Axsäter, S. (2005). Planning order releases for an assembly system with random operation times. *OR Spectrum, 27*(2), 459–470.

Ben-Ammar, O., Dolgui, A., & Wu, D. D. (2018). Planned lead times optimization for multi-level assembly systems under uncertainties. *Omega, 78*, 39–56.

Buzacott, J. A., & Shanthikumar, J. G. (1993). *Stochastic models of manufacturing systems*. Englewood Cliffs, NJ: Prentice Hall.

Cheng, F., Ettl, M., Lu, Y., & Yao, D. D. (2012). A production-inventory model for a push-pull manufacturing system with capacity and service level constraints. *Production and Operations Management, 21*(4), 668–681.

Cheng, F., Markus, E., Lin, G., & Yao, D. D. (2002). Inventory-service optimization in configure-to-order systems. *Manufacturing & Service Operations Management, 4*, 114–132.

Glasserman, P., & Wang, Y. (1998). Leadtime-inventory tradeoffs in assemble–to-order systems. *Operations Research, 46*, 858–871.

Gupta, D., & Benjaafar, S. (2004). Make-to-order, make-to-stock, or delay product differentiation? A common framework for modeling and analysis. *IIE Transactions, 36*, 529–546.

Hopp, W. J., & Spearman, M. L. (2008). *Factory physics: Foundations of manufacturing management* (3rd ed.). The McGraw-Hill Companies, Inc..

Jansen, S., Atan, Z., Adan, I., & de Kok, T. (2019). Setting optimal planned leadtimes in configure-to-order assembly systems. *European Journal of Operational Research, 273*(2), 585–595.

Lee, Y., & Zipkin, P. (1992). Tandem queues with planned inventories. *Operations Research, 40*, 936–947.

Liu, L., Liu, X., & Yao, D. D. (2004). Analysis and optimization of multi-stage inventory-queues. *Management Science, 50*, 365–380.

Maltz, A., Rabinovich, E., & Sinha, R. (2004). Logistics: the key to e-retail success. *Supply Chain Management Review, 8*, 56–63. April

Olhager, J., & Östlund, B. (1990). An integrated push-pull manufacturing strategy. *European Journal of Operational Research, 45*(2–3), 135–142.

O'Marah, K. (2005). The leaders' edge: driven by demand. *Supply Chain Management Review, 9*, 30–35. May/June

Pyke, D. F., & Cohen, M. A. (1990). Push and pull in manufacturing and distribution systems. *Journal of Operations Management, 9*(1), 24–43.

Roberts, J., Mocci, U., & Virtamo, J. (1996). *Broadband network traffixc: Performance evaluation and design of broadband multiservice networks*. Berlin Heidelberg: Springer.

Seimchi-Levi, D., Kaminsky, P., & Seimchi-Levi, E. (2008). *Designing and managing the supply chain: Concepts, strategies and case studies* (3rd ed.). N.Y.: McGraw-Hill.

Serhadli, E. (2016). Performance improvements in global supply chains by flexsible strategies: A case study study at philips healthcare. Master thesis. Faculty of Science and Engineering, University of Groningen.

Simison, R. L. (2000). GM aims to become build-to-order firm but custom online sales are daunting task. *The Wall Street Journal*. Feb. 22, https://www.wsj.com/articles/SB951174312103280846.

Song, J., Yano, C., & Lerssisuriya, P. (2000). Contract assembly: dealing with combined supply lead time and demand quantity uncertainty. *Manufacturing & Service Operations Management, 2*, 287–296.

Song, J., & Yao, D. D. (2002). Performance analysis and optimization of assemble–to-order systems with random leadtimes. *Operations Research, 50*, 889–903.

Song, J., & Zipkin, P. (2003). Supply chain operations: Assemble-to-order and configure-to-order systems. In T. De Kok, & S. Graves (Eds.), *Handbooks in operations research and management science, Vol. 11: Supply chain management*. North-Holland. Chapter XIII.

Spearman, M. L., & Zazanis, M. A. (1992). Push and pull production systems: Issues and comparisons. *Operations Research, 40*, 521–532.

Teo, C., Bhatnagar, R., & Graves, S. (2011). Setting planned lead times for a make–to-order production system under master schedule smoothing. *IIE Transactions, 43*(6), 399–414.

Urban, G., Sultan, F., & Qualls, W. (2000). Placing trust at the center of your internet strategy. *Sloan Management Review, 42*, 39–48. Fall

Wolff, R. (1989). *Stochastic modeling and the theory of queues*. Englewood Cliffs, N.J: Prentice-Hall.

Yano, C. (1987). Stochastic leadtimes in two-level assembly systems. *IIE Transactions, 19*, 371–378.