Optimizing a supply network with production smoothing, freight expediting and safety stocks : an analysis of tactical trade-offs

Zahraei, Seyed Mehdi; Teo, Chee-Chong

2017

Zahraei, S. M., & Teo, C.-C. (2017). Optimizing a supply network with production smoothing, freight expediting and safety stocks : an analysis of tactical trade-offs. *European Journal of Operational Research*, 262(1), 75-88. doi:10.1016/j.ejor.2017.02.045

<https://hdl.handle.net/10356/103425>

<https://doi.org/10.1016/j.ejor.2017.02.045>

© 2017 Elsevier B.V. All rights reserved. This paper was published in *European Journal of Operational Research* and is made available with permission of Elsevier B.V.

Downloaded on 26 Apr 2023 19:10:31 SGT

Optimizing a Supply Network with Production Smoothing, Freight Expediting and Safety Stocks: An Analysis of Tactical Trade-Offs

SEYED MEHDI ZAHRAEI^{a, 1}

CHEE-CHONG TEO^{a, c}

^a*School of Civil & Environmental Engineering,
Nanyang Technological University, Nanyang Avenue, Singapore 639798.
E-mail: {teocc, zahraei}@ntu.edu.sg*

¹*Present Address: Embry-Riddle Aeronautical University, Asia
75 Bukit Timah Road, Singapore 229833
E-mail: zahraeis@erau.edu*

^c*Corresponding author.*

Abstract

Production smoothing, safety stocks and freight expediting are common tactics employed to mitigate demand variability in supply chains. However, in complex supply chains, it is not obvious how best to deploy them in a coordinated approach to minimize cost and to provide a high service level. In this paper, we consider production smoothing and its effect on inter-stage demand flow in a supply network that is capable of expediting shipments between its stages. We model a general network of production and inventory stages, which allows production smoothing to be optimized in realistic network configurations. We incorporate parameters that can significantly affect smoothing, expediting and safety stock decisions but are not considered in most multi-echelon models, including reorder frequencies and inter-stage demand correlation. We assume that services from inventories and production are guaranteed by performing expediting (at additional cost) in times of shortages. Under such a setting, the model quantifies the trade-offs between production smoothness, expediting and safety stocks, and determines their optimal deployment in a network. Through numerical experiments, we demonstrate the model's application and capability; in addition, we provide insights on the choice of freight expediting *versus* safety stock and their interplay with smoothing.

Keywords: supply chain management, multi-echelon inventory, production smoothing, expedited freight, guaranteed service.

1. Introduction

Economic volatility in recent years has encouraged businesses to reduce supply chain assets in order to be more resilient to economic downturns. Linebaugh (2011) reported that companies, such as Honeywell and United Technologies, are still looking for new ways to downsize in this uncertain economic environment, even though most had already done so substantially since the 2008 financial crisis.

Although companies with fewer assets are generally more robust to economic adversity, their supply chains tend to be more vulnerable to uncertainties. Besides the common use of safety stocks, outsourcing and overtime production are the typical practices to hedge against unexpected surges in demand. In addition, production smoothing is widely employed to ease capacity loading in the face of demand variability. Increasingly, firms are also employing expedited freight to enable their supply chains to be more responsive to uncertainties. Terry (2012) reports that expedited freight services are currently being used for many more products; she provides several examples on how logistics service providers now offer a wide range of such services. The costs incurred for expedited freight, e.g. by air freight, can often be justified by reduced inventories and capacities. A classic example is Beyer and Ward (2002), which employs a combination of sea and air modes efficiently to reduce inventories in HP's network server supply chain. Despite its growing importance, companies face difficulties in incorporating expedited freight in their planning due to the complex interdependencies in a typical supply chain. As a result, expedited freight is often viewed as unplanned action needed to resolve emergencies such as stock-outs (see Aimi 2009 for related discussion).

Our intent is to provide a framework for modeling a supply chain with flexibilities to smooth production and to expedite freight between its stages. In a typical supply chain that comprises production and inventory stages, the amount of freight expediting depends not only on the safety stock levels but also the extent of production smoothing. For a make-to-stock manufacturer, a greater extent of smoothing entails that production is less responsive to the level of processed inventory; this would necessitate more safety stock or otherwise more frequent stock-outs and a greater need for expediting.

The planning parameters are inter-related between stages via demand propagation, and it is thus important to capture the spread of demand variability in the network. If more smoothing is carried out at a production stage, requirements at its component inventories will be less variable and the resulting lower demand variability may spread to other upstream stages. Furthermore, the optimal extent of smoothing at each stage depends on factors that can only be meaningfully captured by a general network model, where every stage is allowed to serve and/or be supplied by multiple stages. These factors (to be explored in this paper) include the network size and configuration, demand correlation between stages, as well as inter-stage differences in ordering frequencies, inventory holding costs and expediting costs.

Putting the aforesaid trade-offs together, it is evident that one would need to consider the effect of smoothing on how much safety stock and expediting to deploy at each stage. In this paper, we address these trade-offs at the tactical planning level. We assume the strategic decisions (e.g., number of facilities, facility locations, capacity sizing and supply chain topology) have been made and we focus herein on the subsequent set of tactical decisions. We assume the supply chain faces stochastic external demand and consists of stages performing either production or distribution functions. Each production stage has a finite nominal capacity and can choose to smooth production when replenishing its inventory. Other inventories operate with a standard base-stock policy. Ordering at every stage is carried out according to a stationary nested policy. We assume that services from production stages and inventories are guaranteed; in times of shortages in production capacities or inventories, expediting actions are performed at additional cost that are guaranteed to meet the excess requirements. The model minimizes the expected total cost of the supply chain, consisting of freight and production expediting costs plus inventory holding cost

We summarize the paper's contribution as follows. The main contribution is the development of a new integrated approach of optimizing the three commonly employed tactics for mitigating demand variability, namely production smoothing, safety stocks and freight expediting. To our knowledge, this is the first work that looks into such integrated planning in a multi-echelon setting. In contrast to most multi-echelon models which consider safety stock as the only buffer against demand uncertainty, we incorporate the tactic of expediting. The model determines the optimal deployment of expediting *versus* safety stock at each inventory site by considering inter-stage interdependencies under the effect of smoothing. In addition,

it is widely known that production smoothing can substantially affect upstream demand variability, and hence, optimal smoothing decisions depend heavily on network configuration. Despite its significance, we believe that this is the first model that optimizes production smoothing for a general supply network. From a broader context of technical contribution, we establish a linear-system approach for general network modeling of supply chains, mainly based on the modeling framework of Graves (1986) and Teo et al. (2011). It provides modeling flexibility that allows the consideration of parameters such as capacitated production, demand correlation and reorder intervals, which we will show are important parameters that can considerably affect the optimal solution.

Regarding insights, we perform experiments using the model to show how the optimal smoothing and the resulting cost benefits depend on inter-stage cost differences besides network size and configuration. We highlight the role of the downstream production stage in reducing upstream demand volatility to achieve global optimality, and we quantify its resulting local sub-optimality from the smoothing. We also find that under the effect of smoothing and at high service requirements, downstream freight expediting tend to be more critical to the overall supply chain cost, as compared to safety stocks and upstream freight expediting. These insights provide guidance on the deployment of expediting resources when implementing smoothing.

Our paper also contributes to the application of the guaranteed-service assumption, which is a key feature in the stream of models defined as the *guaranteed-service models* by Graves and Willems (2003). Although our model is of a different intent, it offers an application of the assumption to a new context and introduces a richer interpretation of it. We incorporate the demand bound in our model, which is interpreted as a design parameter in the guaranteed-service models. In addition, we include freight expediting (besides safety stocks in typical guaranteed-service models) as protection against variability and characterize the bound as a function of the decision variables. The bound enables us to incorporate capacitated expediting and to characterize the effect of capacity tightness on smoothing decisions.

The rest of the paper is organized as follows. We review related literature in the next section. Next, for ease of exposition, we first describe the model development for a simpler single-product serial supply chain in §3. In §4, we discuss how we model a general network and embed the model in optimization procedures. In §5, we discuss how the model can be extended to incorporate finite capacities for freight expediting. In §6, we report the numerical experiments that we carry out to show the use of the model and to draw related insights. We then present conclusions and directions for future research in §7.

2. Related Work

From a broad viewpoint, this work is related to the vast literature on multi-echelon inventory systems. Refer to survey articles such as Swaminathan and Tayur (2003), Diks et al. (1996), van Houtum et al. (1996), and Graves and Willems (2003). Among this extensive literature, our paper is most closely related the research stream on production smoothing. We first discuss the work related to this stream while not restricting our discussion to multi-echelon models. The literature on production smoothing involves a vast literature in both operations management and economics fields. See Cachon et al. (2007) for a review in both fields or refer to Boute and Van Mieghem (2014) and Bray and Mendelson (2015) for more recent reviews. We believe our paper is the first that considers simultaneous planning of smoothing, expediting and safety stock. Most of the studies assume single-stage or two-echelon systems. In contrast, we model a general network as our intent is to support smoothing decisions by considering the spread of variability in more realistic configurations. To achieve this, we employ the linear-system modeling approach. We discuss herein the papers related to this approach. In their early seminal work, Holt et al. (1955) propose a linear decision rule that minimizes a quadratic cost function and it prescribes how demand variability should be buffered by inventory, overtime production and adjustment of workforce size. Graves et al. (1986) employ a simpler linear smoothing rule that is a function of only inventory level. The rule leads to a linear-system model, referred to as the Tactical Planning Model (TPM), which is of adequate tractability to model a job shop. Subsequent TPM-based models include Graves (1986), Graves (1988), Graves and Fine (1989), Graves et al. (1998) and Teo et al. (2011). Balakrishnan et al. (2004) and, more recently, Boute and Van Mieghem (2014) use the same linear rule for characterizing order smoothing; notably, Boute and Van Mieghem show that the rule “closely tracks” the optimal but intractable dual base-stock policy.

We develop our model based on the same linear-system framework of the TPM, which enables us to address the modeling challenges offered by the research problem. A key challenge in developing the model is how to capture the reorder intervals that are often different among the supply chain stages. The reorder interval is an important parameter as it affects the regularity of stock-outs and thus influences the frequency of expediting. It also influences the demand and production variability, as will be shown. The problem is further complicated by the consideration of a general network in which each stage can receive orders from stages with different reorder intervals. In our model, we employ the same linear matrix-vector system in the TPM, which allows us to address the main features of the problem. However, due to contextual differences, our model focuses on demand flows and ordering actions in a supply chain rather than job flows in a production shop. Consequently, the model presented herein departs from the TPM-based models in many aspects, which are explained as follows.

First and foremost, we characterize inventory variables as a function of reorder intervals and consider replenishment lead-times when characterizing safety stocks; this differs from the prior TPM-based models that do not consider the replenishment lead-times (simply because replenishments are unrelated to job flows). Second, we re-define the stages as production, input and distribution stages, and we characterize the production or inventory variables for each type of stages; this is opposed to the existing TPM models that consider a homogeneous network of workstations. Third, we adopt a new interpretation of the linear production function in Teo et al. (2011) where we interpret the *time interval between job movements* as the reorder interval at a supply chain stage; when combined with the nested ordering assumption, we are able to establish the ordering patterns among the stages.

Our paper is also related to the research stream on expediting. We refer to Minner (2003) for a review of this stream. These models typically assume expedited shipments are characterized by shorter lead-times but higher costs. They attempt to capture different problem settings related to expediting, e.g., batching effect of expedited shipments (Jain et al. 2010), lead-time flexibility in transport mode choices (Bhatnagar et al. 2011), order crossing in expediting (Kim et al. 2009) and expediting for capacitated production (Veeraraghavan and Scheller-Wolf 2008). Some models consider dynamic expediting decisions based on real-time information. Groenevelt and Rudi (2003) consider real-time supply information; Foreman et al. (2010) consider the dynamic choice of transport modes for Dell's inbound shipments based on real-time inventory status. Similar to our paper, some works consider expediting in a supply chain. Lawson and Porteus (2000) derive the optimal base-stock policies that decide the amount of inventory to expedite, to hold or to ship with regular lead-time at every stock location. Muharremoglu and Tsitsiklis (2003) extend Lawson and Porteus (2000) by permitting supermodularity ordering costs. Berling and Martinez (2011) consider a continuous-stage serial system, which determines at real-time the speed of each shipment.

Our paper differs in that our concern is not on operational or real-time decisions but rather to support tactical planning. In particular, our model supports deployment of expediting (in relation to safety stocks) in mitigating variability, and considers how it interacts with smoothing behaviour. Another main difference with the literature is in our treatment of expedited lead-time. Existing literature derive policies based on specific assumptions on the location and/or lead-time flexibility of expediting. We cite a few examples of these specificities. Veeraraghavan and Scheller-Wolf (2008) and Groenevelt and Rudi (2003) assume regular orders cannot be expedited once shipments are in transit. Lawson and Porteus (2000) and Muharremoglu and Tsitsiklis (2003) allow flexibility in expediting in-transit inventory but restrict regular and expediting lead-time to one or zero period. Kim et al. (2009) also restrict lead-times to be one or zero period, and further assume that orders can only be expedited directly to customers (bypassing intermediary

facilities). In Foreman et al. (2010), Dell’s in-transit stocks can only be expedited by faster inland modes after being transported by fixed lead-time ocean mode.

Our paper differs in that we do not specify the lead-time or location flexibility but assume requirements can always be met by expediting (if necessary) at a quantity dependent cost. This is similar to the guaranteed-service assumption, which forms the foundation for guaranteed-service models. The guaranteed-service models typically assume that when demand exceeds a bound for which safety stock is designed to meet, other responses (not addressed by the model) are always available to achieve the guaranteed service. Comparatively, we are more explicit in our assumption in that demand in excess of safety stock is met by expedited shipments and we characterize the expediting cost as a function of the planning parameters. Furthermore, we show that the model can be extended to incorporate capacitated expediting by defining a demand bound at each stage for which both expediting and safety stock are intended to meet. But, in the same spirit as the guaranteed-service models, we do not attempt to specify exactly how the capacitated expediting should respond to stock-outs to achieve the guaranteed service.

3. Serial Supply Chain

In this section, we consider a single-product, serial supply chain. We first present the assumptions in §3.1 and then discuss the components of the expected total cost in §3.2. Subsequently in §3.3, we develop the model to characterize the key variables needed to compute the expected total cost.

3.1. Assumptions

NETWORK

There are q stages in the serial network, indexed by $i = 1, 2, \dots, q$. External demand is received only at the utmost downstream stage (stage 1). The network consists of three types of stages, namely *production stages*, *input stages* and *distribution stages*. We let the respective sets of stages be denoted by *PROD*, *INP* and *DIST*. The production stages perform processing activities to convert raw material or assemble components into semi-finished or end product, which is then stored as *processed inventory*. Each input stage represents inventory of input material held by a production stage. The distribution stages carry out handling and storage of intermediate or end product (e.g., a warehouse or a distribution center).

Every input stage is immediate upstream of a production stage. Effectively, we assume every production stage holds both processed and input inventories. (We will discuss how we can allow absence of decoupling processed and/or input inventories in §4.) The input and distribution stages operate with a standard base-stock policy (for which justification will be discussed). At the production stages, the processed inventories are not regulated by the base-stock policy but are replenished by production that

can be smoothed. We assume instantaneous replenishment at each production stage from its corresponding input stage, i.e., replenishment lead-time is zero. In effect, this implies that the input inventory is located near or at the same location as the production facility. (It may seem more intuitive to define each production-input pair of stages as one entity but we categorize them as such as they will be characterized separately in the model.)

With the above setting, only the input and distribution stages face positive replenishment lead-times. Specifically, we assume a fixed replenishment lead-time L_i at input or distribution stage i for shipments from its immediate upstream stage. We illustrate in Figure 1 an *example* of a serial supply chain.

Figure 1: An example of a serial supply chain

GUARANTEED-SERVICE ASSUMPTION

We assume that services from inventories and production stages are guaranteed by performing expediting in times of stock-outs or capacity shortfalls. If a stock-out occurs at an inventory location, we assume shipment *equivalent to the shortage quantity* is expedited at extra costs from upstream such that it is always able to fulfill demand. Effectively we assume every stock location is always able to recover from a stock-out such that its immediate customer stage does not experience any of its effect. Expediting action for capacity shortfalls at the production stages may include overtime production.

The guaranteed-service assumption applies to all stages, including all internal customers. Graves and Willems (2000) provide some justifications for this assumption for both external and internal customers. For the input and distribution stages, the guaranteed-service assumption implies that every stage observes a single reliable replenishment lead-time and is subject to a stock-out penalty in the form of expediting cost (instead of the usual backorder or lost sales cost). This setting is equivalent to the basic inventory problem characterized by a deterministic replenishment lead-time and a trade-off between stock-out penalty and holding cost, which calls for the standard base-stock policy.

DEMAND PROCESS AND REORDER INTERVALS

We assume external demand at utmost downstream stage (stage 1) is i.i.d. (independent and identically distributed) over time. We index the time series by t and let D_{it} denotes demand at stage i in period t .

We utilize two time scales to capture the dissimilar reorder intervals among the stages. Specifically, we assume all stages operate with a common planning time period in which demand quantities and production capacities are defined; within each period, sub-periods are employed to model the shorter intervals for

reordering and production re-planning. Without loss of generality, we assume each period t has a length of one time unit. Each period of stage i is sub-divided into p_i sub-intervals indexed by s ($s = 1, 2, \dots, p_i$), where $p_i = 1/\Delta_i$. We interpret Δ_i as the fixed reorder interval at stage i . For production stages, we define Δ_i as the time interval between successive production re-planning.

We assume the supply chain adopts a stationary nested policy, i.e., whenever one stage orders, the downstream stage also orders. The nested ordering requires that:

$$\Delta_i = r_i \Delta_{i-1}, \quad \text{for } i = 2, 3, \dots, q. \quad (1)$$

where stage $i-1$ places orders to stage i and r_i is a positive integer. In addition, we assume external demand arrives uniformly *within* each period at stage 1, i.e., demand in each reorder interval is D_{1t}/p_1 . Combining the assumptions of nested ordering policy and uniform intra-period external demand, the ordering pattern is established across successive stages, where demand within each reorder interval at *every* stage is also uniform and equals D_{it}/p_i . We defer a detailed explanation on how the ordering pattern leads to uniform intra-period demand at every stage when we introduce the relevant parameters in §3.3.

The dual time scales and the intra-period uniform demand assumption merit some discussion and justification. The use of dual time scales coincides with most *Advanced Planning Systems* (APS), where time buckets for demand and capacity planning are larger than the sub-intervals for reordering and production re-planning. Our assumption of intra-period uniform demand is also consistent with most APS, where demand is often disaggregated uniformly when converted into smaller time units e.g., from monthly to weekly demand (see e.g., Meyr 2012, p. 94). In addition, our research focuses on tactical planning, in which the granularity of the time buckets should be sufficient to capture the essential variability of demand and capacity requirements. For the same reason, we argue that the uniform demand assumption is reasonable as we are more concerned with inter-period demand volatility, rather than the detailed intra-period variations.

3.2. Expected Total Cost

The cost components incurred by each type of stages are explained as follows.

Input stages and distribution stages. We assume every input and distribution stage holds a base-stock whose level is denoted by B_i . Every unit of inventory shortage that is expedited incurs a unit cost of c_{fi} . We consider the *demand over risk period* at these stages. As commonly defined, *risk period* is the sum of replenishment lead-time L_i plus reorder interval Δ_i , and it signifies the total time span of demand realization that the base-stock should be designed to meet. We let rp_i denotes the length of risk period at stage i and let $D_{it}^p(s)$ denotes the demand over the risk period that begins in sub-interval s of period t . By letting x^+

$= \max(x, 0)$, $E[D_{it}^p(s) - B_i]^+$ is the expected inventory shortage *over the risk period* for which expediting is needed. We obtain the expected expedited quantity *per period* over all s (i.e., by averaging over $s = 1, 2, \dots, p_i$) as the following:

$$\frac{1}{r p_i} \frac{1}{p_i} \sum_{s=1}^{p_i} [D_{it}^p(s) - B_i]^+$$

Production stages. Each production stage has a nominal capacity m_i per period. If its production requirements R_{it} , which is the total scheduled production quantities in period t , exceed m_i , expediting actions are carried out for the excess requirements. A unit cost of c_{pi} is incurred for every expedited unit. Further, we let B_i denotes the initial level of the processed inventory (i.e., at the start of $t = 0$); $X_{it}(s)$ denotes the shortfall of processed inventory at production stage i , which we define as the difference between B_i and the actual net inventory level in sub-period s of period t . Therefore, $E[X_{it}(s) - B_i]^+$ is the expected stock-out quantity at the end of sub-interval s for which expediting is required.

All stages. EI_i denotes the expected stock level at stage i that incurs a unit holding cost of h_i per period.

We express the expected total cost per period as follows:

$$\sum_{i \in \text{PROD}} \left(c_{pi} E[R_{it} - m_i]^+ + c_{li} \sum_{s=1}^{p_i} E[X_{it}(s) - B_i]^+ \right) + \sum_{i \in (\text{INP} \cup \text{DIST})} c_{li} \frac{1}{r p_i} \frac{1}{p_i} \sum_{s=1}^{p_i} E[D_{it}^p(s) - B_i]^+ + \sum_{\text{all } i} h_i EI_i. \quad (2)$$

The expected total cost (2) forms the objective function of the optimization to be discussed in §4. The decision variables are safety stock levels and production smoothness, which determine the optimal expected expedited quantities. To find the expected total cost, we need to characterize $X_{it}(s)$, $D_{it}^p(s)$, R_{it} and EI_i . As will be shown, due to the model's linearity, if the external demand is normally distributed, $X_{it}(s)$, $D_{it}^p(s)$ and R_{it} will also be normally distributed due to the property whereby a linear combination of independent normal random variables is also normally distributed. The core of the model is to characterize the first two moments of $X_{it}(s)$, $D_{it}^p(s)$ and R_{it} , which allows the expected costs in (2) to be computed using the normal loss function. Note that although we assume a normal distribution herein, there is no distributional assumption needed in characterizing the first two moments.

3.3. Model Development

In this sub-section, we characterize R_{it} , $X_{it}(s)$, $D_{it}^p(s)$ and EI_i . First, we develop the linear functions that characterize the production and replenishment quantities of the different stage types. Next, we discuss the

dynamics of demand flow and then combine the linear functions for the individual stages into a single matrix-vector network model. Based on the model, we then derive R_{it} , $X_{it}(s)$, $D_{it}^p(s)$ and EL_i .

LINEAR FUNCTIONS OF PRODUCTION AND REPLENISHMENT VARIABLES

We first derive the linear function for production requirements at production stages and then the corresponding function for the replenishment quantities at the input and distribution stages.

Production stages. The derivation for the production requirements R_{it} is the same as that in Teo et al. (2011), except differences in context and in interpretation of the parameters. At each production stage i , orders arrive from its immediate downstream stage $i - 1$ and served by its processed inventory. The stage reviews its inventory shortfall at the end of every sub-period and then decides on its *production starts*, which we define as the production quantities scheduled to start at the end of the sub-period and be completed in the subsequent sub-period. The production starts is set according to a linear production function given by:

$$Y_{it}(s) = (\Delta_i/n_i) X_{it}(s) \quad \text{for } s = 1, 2, \dots, p_i, \quad (3)$$

where $Y_{it}(s)$ is the production starts at stage i at the end of sub-period s of period t , $X_{it}(s)$ is the inventory shortfall at stage i at the end of sub-period s of period t after all demand arrivals in the sub-period, and n_i ($n_i \geq \Delta_i$) is a control parameter for stage i .

Equation (3) is the linear rule in the TPM. Like most TPM-based models, we interpret n_i as the planned (average) lead-time at production stage i . In (3), production stage i processes (Δ_i/n_i) of the inventory shortfall, so that each order does not wait more than n_i periods on the average. This aligns with our assumption of expediting, which is undertaken to preserve the integrity of the planned lead-time when necessary. The parameter n_i can also be viewed as a smoothing parameter, a larger n_i results in smoother production output. We can show this by substituting (3) into the balance equation $X_{it}(s) = X_{it}(s-1) - Y_{it}(s-1) + D_{it}/p_i$ to obtain $Y_{it}(s) = (1 - \Delta_i/n_i) Y_{it}(s-1) + (\Delta_i/n_i) D_{it}/p_i$, which is the first-order smoothing equation. Equation (3) is consistent with how production smoothing is carried out in some ERP (Enterprise Resource Planning) systems. For example, in SAP's *Supply Chain Management* module, production smoothing is performed by equal distribution of planned orders over a planning period (see SAP Library). Equation (3) is comparable to such smoothing behavior by spreading $X_{it}(s)$ "equally" over n_i periods at every interval Δ_i . Furthermore, Fine and Graves (1988) and Teo et al. (2012) provide partial validations of (3) in their respective applications.

Besides being a descriptive function, (3) possesses some desirable properties. It has been well established that if production costs are convex, it is optimal to only partially adjust production output in response to a change in inventory position, and inventory should instead be held to buffer against the excess variability. We use (3) to approximate such output adjustments. Balakrishnan et al. (2004) prove

that (3) maximizes order smoothing among all *Moving Weighted Average* policies. Despite these positive properties, it is noteworthy that (3) is sub-optimal in minimizing the total expected cost of the production stage (i.e., for holding of processed inventory and expediting of production). However, Boute and Van Mieghem (2014) find that the scaled total cost of (3), which they employ for order smoothing, “closely tracks” that of the optimal dual base-stock policy. The average error over the range of test values is less than 3.8% (based on our readings of their graphical results). They also point out that the optimal dual base-stock policy is not analytically tractable and has to be solved numerically, even for a single production stage that serves i.i.d. demand from its inventory. In contrast, (3) allows significant tractability and depicts first-order smoothing behavior for tactical planning.

We consider the R_{it} as the sum of all production starts in period t ; that is:

$$R_{it} = \sum_{s=1}^{p_i} Y_{it}(s). \quad (4)$$

By (3) and the prior assumption of intra-period uniform demand (i.e., demand in every sub-period equals D_{it}/p_i), and by utilizing the inventory balance equation, Teo et al. (2011) derive R_{it} , which is given by:

$$R_{it} = \beta_i S_{it} + \gamma_i D_{it}, \quad (5)$$

where $\beta_i = 1 - \left(1 - (\Delta_i/n_i)\right)^{p_i}$ and $\gamma_i = 1 - \beta_i \left(1 - (\Delta_i/n_i)\right)/(1/n_i)$. The variable S_{it} in (5) denotes inventory shortfall in the first sub-interval of t prior to any demand arrival, i.e., $S_{it} = X_{it}(s=1) - D_{it}/p_i$; it is related to the production requirements and demand by the balance equation given by:

$$S_{it} = S_{i,t-1} - R_{i,t-1} + D_{i,t-1}. \quad (6)$$

For completeness, we present the derivation of (5) based on the context of this paper in *Online Appendix A*. Note that R_{it} in (5) depends on inventory shortfall at the start of the period and demand in the period; β_i and γ_i depend on the size of the reorder intervals and the planned lead-times.

Input and distribution stages. By the standard base-stock policy, the input and distribution stages order at the end of every sub-period a quantity equal to the orders received in the sub-period. Relating to (3) and by interpreting $Y_{it}(s)$ as the order placed at the end of sub-period s (while $X_{it}(s)$ still denotes inventory shortfall), it follows that $Y_{it}(s) = X_{it}(s)$. As a result, we have a linear function of the same form as (5), that is $R_{it} = S_{it} + D_{it}$, wherein R_{it} is the replenishment quantities at these stages and both β_i and γ_i equal 1.

A CLOSER LOOK AT ORDERING PATTERNS AND DEMAND FLOW

Recall in §3.1 that the assumptions of uniform intra-period external demand and nested policy results in uniform intra-period demand D_{it}/p_i at every stage i . We discuss herein the underlying demand dynamics.

We let ζ_i denotes the external demand (at stage 1) with mean μ and standard deviation σ (and is i.i.d. over time as defined earlier). The internal demand at the other stages ($i = 2, 3, \dots, q$) is given by:

$$D_{it} = \varphi_i R_{i-1,t} \quad (7)$$

The parameter φ_i is a positive scalar that indicates the number of units of demand at stage i for each unit of production or replenishment at stage $i - 1$. We expect $\varphi_i = 1$ for demand arrivals at distribution stages and production stages; however, at input stages, φ_i denotes the number of units of input material required for every unit processed at production stage $i - 1$.

Equation (7) characterizes the demand flow in each time period where demand is subsequently disaggregated uniformly into sub-period demands at each stage. As stated in §3.1, external demand is assumed to be uniformly distributed within each period. By the assumptions of nested policy (1), and that reordering and production starts occur at the end of every sub-period, the ordering pattern is established for successive stages (starting from stage 1) such that the intra-period demand at every stage is also uniformly distributed and equals D_{it}/p_i . Figure 2 depicts an example which shows the ordering pattern between two successive stages and the resulting uniform demand within each period.

Figure 2: Uniform intra-period demand

We note that the assumption of uniform intra-period demand contradicts the smoothing function (3). That is, if $\Delta_i < 1$ at a production stage and production is smoothed over the sub-periods, production quantities will be non-uniform within each period. Consequently, this assumption will be less valid if input material is withdrawn from the input stage based on the sub-period production, as this may result in non-uniform intra-period demand at the input stage and subsequently its upstream stages. We carry out experiments to show the model's robustness to this assumption in *Online Appendix B*.

CHARACTERIZING R_{it} AND X_{it} BY MATRIX-VECTOR FORMULATION

We use the derived linear functions (5) to develop the network model. Due to smoothing, demand at stages upstream of production stages are generally correlated over time. The network model developed in matrix-vector enables the correlation to be captured. Note that the characterization of the first two moments of R_{it} (equations 8 to 15) follows the TPM approach.

We express (5) and (6) in matrix notation by:

$$\mathbf{R}_t = \mathbf{F}\mathbf{S}_t + \mathbf{G}\mathbf{D}_t \quad (8)$$

and

$$\mathbf{S}_t = \mathbf{S}_{t-1} - \mathbf{R}_{t-1} + \mathbf{D}_{t-1}, \quad (9)$$

where $\mathbf{R}_t = \{R_{1t}, \dots, R_{qt}\}'$ and $\mathbf{S}_t = \{S_{1t}, \dots, S_{qt}\}'$ are column vectors. In addition, $\mathbf{F}$ and $\mathbf{G}$ are diagonal matrices with elements β_i and γ_i ($i = 1, 2, \dots, q$), respectively.

We combine ζ_t at stage 1 together with (7) for the other stages into matrix notation to obtain:

$$\mathbf{D}_t = \mathbf{\Phi} \mathbf{R}_t + \boldsymbol{\zeta}_t, \quad (10)$$

where $\mathbf{D}_t = \{D_{1t}, \dots, D_{qt}\}'$ and $\mathbf{\Phi}$ is a square matrix with φ_i as element $(i, i-1)$. By plugging (10) into (8), we can find $\mathbf{S}_t$ as:

$$\mathbf{S}_t = \mathbf{F}^{-1}(\mathbf{I} - \mathbf{G}\mathbf{\Phi})\mathbf{R}_t - \mathbf{F}^{-1}\mathbf{G}\boldsymbol{\zeta}_t, \quad (11)$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{F}^{-1}$ exists since every of its diagonal element is positive. By substituting (9) and (10) into (11), we obtain:

$$\mathbf{R}_t = \mathbf{M}\mathbf{R}_{t-1} + \mathbf{H}\mathbf{G}\boldsymbol{\zeta}_t + \mathbf{H}(\mathbf{F} - \mathbf{G})\boldsymbol{\zeta}_{t-1}, \quad (12)$$

where $\mathbf{H} = (\mathbf{I} - \mathbf{G}\mathbf{\Phi})^{-1}$ and $\mathbf{M} = \mathbf{I} - \mathbf{H}\mathbf{F}(\mathbf{I} - \mathbf{\Phi})$. We note that $(\mathbf{I} - \mathbf{G}\mathbf{\Phi})$ is invertible as long as the spectral radius of $\mathbf{\Phi} < 1$, which we will show is also a necessary condition for the system to reach steady state. By repeatedly iterating $\mathbf{R}_t$ and assuming an infinite history of the system, we obtain the expression:

$$\mathbf{R}_t = \sum_{s=0}^{\infty} \mathbf{M}^s \mathbf{K} \boldsymbol{\zeta}_{t-1-s} + \mathbf{H}\mathbf{G}\boldsymbol{\zeta}_t, \quad (13)$$

where $\mathbf{K} = \mathbf{H}\mathbf{F} - (\mathbf{I} - \mathbf{M})\mathbf{H}\mathbf{G}$. The vector $\boldsymbol{\zeta}_t$ has a mean $\boldsymbol{\mu}$ and a covariance $\boldsymbol{\Sigma}$; $\boldsymbol{\mu}$ is a column vector with μ as the only non-zero element while $\boldsymbol{\Sigma}$'s only non-zero element is the variance of external demand.

We can obtain the expectation and variance of $\mathbf{R}_t$ from (13) by:

$$\mathbf{E}[\mathbf{R}_t] = (\mathbf{I} - \mathbf{\Phi})^{-1} \boldsymbol{\mu}, \quad (14)$$

and

$$\text{Var}(\mathbf{R}_t) = \sum_{s=0}^{\infty} \mathbf{M}^s \mathbf{K} \boldsymbol{\Sigma} \mathbf{K}' \mathbf{M}'^s + \mathbf{H}\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}'\mathbf{H}'. \quad (15)$$

We observe that $\mathbf{E}[\mathbf{R}_t]$ does not rely on the planning parameters but depends only on $\mathbf{\Phi}$ and $\boldsymbol{\mu}$. Equation (15) provides the variances of production and replenishments at each stage as well as the covariance for each pair of stages. We can compute the power series in (15) analytically or numerically using the methods discussed in Graves (1986), which also shows that the series equivalent to (13) and (15) will converge if and only if the spectral radius of $\mathbf{\Phi} < 1$. This condition assures that every unit ordered or produced by any stage cannot eventually result in an average of more than one unit ordered or produced at the same stage; otherwise, the order or production quantity grows indefinitely. Both (13) and (15) converge for serial systems since orders do not flow back to the same stage (and we expect such perpetually growing back-flowing orders is unlikely to happen for general networks)

Next we derive the first two moments of $X_{it}(s)$ at each production stage. Here we are interested to find the expectation and the variance of $X_{it}(s)$ that is unconditional on s . To find its expectation, we substitute (3) into the balance equation and by repeated substitutions, we can obtain:

$$\begin{aligned} X_{it}(s) &= \alpha_i^{s-1} S_{it} + \left(1 + \alpha_i + \dots + \alpha_i^{s-1}\right) \frac{D_{it}}{p_i} \\ &= \alpha_i^{s-1} S_{it} + n_i \left(1 - \alpha_i^s\right) D_{it}, \end{aligned} \quad (16)$$

where $\alpha_i = (1 - \Delta_i/n_i)$. By plugging (5) into (6) to replace R_{it} , and by repeated substitution, we can obtain

$S_{it} = (1 - \gamma_i) \sum_{k=1}^{\infty} (1 - \beta_i)^{k-1} D_{i,t-k}$. We can then substitute S_{it} into (16) to find the expected value of $X_{it}(s)$:

$$\mathbb{E}[X_{it}(s)] = \mathbb{E}\left[\alpha_i^{s-1} (1 - \gamma_i) \sum_{k=1}^{\infty} (1 - \beta_i)^{k-1} D_{i,t-k} + n_i (1 - \alpha_i^s) D_{it}\right] = n_i \mathbb{E}[D_{it}]. \quad (17)$$

We observe from (17) that the expected value of $X_{it}(s)$ is independent of the sub-periods. We can find $\mathbb{E}[D_{it}]$ in (17) by taking the expectation of (7) in matrix notation, i.e., $\mathbb{E}[\mathbf{D}_t] = \mathbf{\Phi} \mathbb{E}[\mathbf{R}_t] + \mathbf{\mu}$.

For brevity, we drop the notation s for the variance of $X_{it}(s)$ that are unconditional on s . We can write the variance of X_{it} as:

$$\text{Var}(X_{it}) = \mathbb{E}[\text{Var}(X_{it}(s) | s)] + \text{Var}(\mathbb{E}[X_{it}(s) | s]).$$

Recognizing that $\text{Var}(\mathbb{E}[X_{it}(s) | s])$ is zero and since each sub-period occurs $1/p_i$ times over an infinite horizon, we can then obtain $\text{Var}(X_{it})$ by taking the average over all s ; that is:

$$\text{Var}(X_{it}) = \frac{1}{p_i} \sum_{s=1}^{p_i} \text{Var}(X_{it}(s)). \quad (18)$$

To find $\text{Var}(X_{it})$ in (18), we can express the variance of $X_{it}(s)$ in (16) as:

$$\text{Var}(X_{it}(s)) = \text{Var}(\alpha_i^{s-1} S_{it} + n_i (1 - \alpha_i^s) D_{it}). \quad (19)$$

We can express (19) in matrix notation as:

$$\text{Var}(\mathbf{X}_t(s)) = \text{Var}(\mathbf{v}_{1s} \mathbf{S}_t + \mathbf{v}_{2s} \mathbf{D}_t), \quad (20)$$

where matrices $\mathbf{v}_{1s}$ and $\mathbf{v}_{2s}$ are diagonal matrices with elements $v_{1i} = \alpha_i^{s-1}$ and $v_{2i} = n_i (1 - \alpha_i^s)$ respectively, as their diagonal elements. For each production stage, the above coefficients are defined for $s = 1, 2, \dots, p_i$, and zero else. By substituting (9) and (10) into (20), we can find that:

$$\text{Var}(\mathbf{X}_t(s)) = \mathbf{W}_{1s} \text{Var}(\mathbf{R}_t) \mathbf{W}_{1s}' + \mathbf{W}_{2s} \Sigma \mathbf{W}_{2s}', \quad (21)$$

where $\mathbf{W}_{1s} = \mathbf{v}_{1s} \mathbf{F}^{-1} (\mathbf{I} - \mathbf{G} \mathbf{\Phi}) + \mathbf{v}_{2s} \mathbf{\Phi}$ and $\mathbf{W}_{2s} = \mathbf{v}_{2s} - \mathbf{v}_{1s} \mathbf{F}^{-1} \mathbf{G}$. For computational convenience, we can obtain $\text{Var}(X_{it})$ at all production stages by expressing (18) in matrix notations by:

$$\text{Var}(\mathbf{X}_t) = \mathbf{\Omega} \sum_{s=1}^{\tilde{p}} \text{Var}(\mathbf{X}_t(s)), \quad (22)$$

where we sum the series of variances up to $\tilde{p}$, which is the maximum p_i among stages $i \in \text{PROD}$. Matrix $\mathbf{\Omega}$, is a diagonal matrix with $1/p_i$ as its diagonal elements.

The $\text{Var}(R_{it})$ and $\text{Var}(X_{it})$ are measures of production and inventory variability respectively. In Figure 3, we consider a single production stage and graph the standard deviation of production requirement, $\sigma(R_{it})$, and the standard deviation of inventory shortfalls, $\sigma(X_{it})$, as a function of n_i (with $\mu_i = 1$, $\sigma_i = 0.4$). Figure 3 highlights the fundamental tradeoffs between lead-time, capacity and safety stock. As one smoothes the production by increasing n_i , less capacity or expediting actions are required as indicated by the decreasing $\sigma(R_{it})$; however, the inventory shortfall becomes more variable as shown by the increasing $\sigma(X_{it})$, and thus a higher safety stock level is required for a given service level. From a network perspective, a smoother production leads to lower upstream variability as ordering becomes less variable. We can also observe that for a given n_i , increase in Δ_i results in significant increase in $\sigma(R_{it})$ and $\sigma(X_{it})$. For example, at $n_i = 0.6$, $\sigma(R_{it})$ and $\sigma(X_{it})$ for $\Delta_i = 0.5$ are 1.23 and 1.32 times respectively larger than that for $\Delta_i = 0.125$. This demonstrates the importance of incorporating the reorder interval in the model.

Figure 3: Changes in $\sigma(R_{it})$ and $\sigma(X_{it})$, as a function of n_i (with $\mu_i = 1$, $\sigma_i = 0.4$)

With the obtained expectation and variance given by (17) and (22) respectively, one may set B_i (i.e., the stock level at the start of $t = 0$) for the processed inventory by:

$$B_i = E[X_{it}] + z_i \sqrt{\text{Var}(X_{it})} \quad i \in \text{PROD}, \quad (23)$$

where z_i is the safety factor at stage i . We define EI_i ($i \in \text{PROD}$) as the expected processed inventory *before* any replenishment in the sub-period. In effect, we do not account for cycle inventory since it only depends on Δ_i , which is assumed constant. We can obtain EI_i by:

$$EI_i = z_i \sqrt{\text{Var}(X_{it})} \quad i \in \text{PROD}. \quad (24)$$

CHARACTERIZING $D_{it}^p(s)$ BY MATRIX-VECTOR FORMULATION

We characterize $D_{it}^p(s)$ in consistence with the assumption of uniform intra-period demand and nested ordering policy. Since demand arrives from immediate downstream stage $i-1$ at the end of sub-interval Δ_{i-1} , we measure the risk period at stage i in time resolution of Δ_{i-1} . As such, we round down the risk period to the nearest Δ_{i-1} . We express the risk period rp_i , at stage $i \in INP \cup DIST$ as:

$$rp_i = a_i + b_i \Delta_{i-1},$$

where a_i and b_i are nonnegative integers. We need to consider two features of the demand over the risk period. First, production output are correlated over time due to smoothing, which leads to correlated demand at upstream input stages and distribution stages. Second, the demand over the risk period depends on the timing of ordering within each time period, which we illustrate in the following example.

Example: Suppose $rp_i = 2 + 5\Delta_{i-1}$ at stage i , $\Delta_{i-1} = 1/8$ and $\Delta_i = 1/4$. Figure 4 illustrates the demand over the risk period for each of the four ordering epochs at $s = 1$ to 4 for stage i .

Figure 4: Demand over risk period at different ordering epochs

In Figure 4, for $s = 1$ in period t , the demand over risk period $D_{it}^p(s = 1)$ covers all demands received in t and $t + 1$, and $5/8$ of demand received in $t + 2$; thus, the coefficient for each of D_{it} and $D_{i,t+1}$ is 1, and $5/8$ for $D_{i,t+2}$. As illustrated, each ordering epoch leads to different coefficients of demand. Thus the coefficient of demand represents the proportion of a time bucket covered by the risk period for a particular ordering epoch. We express the demand over risk period that starts from sub-period s of period t as a sum of demands across t to $t + l + 1$, given by:

$$D_{it}^p(s) = \sum_{k=0}^{l+1} \pi_{ik}(s) D_{i,t+k} \quad \text{for } s=1, 2, \dots, p_i.$$

We derive the coefficients $\pi_{ik}(s)$ by analyzing the demand over risk period at the different ordering epochs in the same way shown in Figure 4. For each stage i , the above coefficients are defined for $s = 1, 2, \dots, p_i$ and zero else. We show the values of the coefficients in Table 1, where we let $d_i(s) = b_i + r_{ij}(s-1-p_i)$ and $\mathcal{A}(l)$ denotes the set of stages with $a_i = l$ ($i \in INP \cup DIST$). We present the derivations of $\pi_{ik}(s)$ in *Online Appendix C*.

Table 1: Coefficients $\pi_{ik}(s)$ for demand over risk period $\forall i \in \mathcal{A}(l)$

	$k = 0$	$1 \leq k \leq l-1$	$k = l$	$k = l+1$
$l = 0$	$\frac{b_i - [d_i(s)]^+}{r_i p_i}$	-	-	$\frac{[d_i(s)]^+}{r_i p_i}$
$l = 1$	$1 - \frac{s-1}{p_i}$	-	$\frac{s-1}{p_i} + \frac{b_i - [d_i(s)]^+}{r_i p_i}$	$\frac{[d_i(s)]^+}{r_i p_i}$
$l \geq 2$	$1 - \frac{s-1}{p_i}$	1	$\frac{s-1}{p_i} + \frac{b_i - [d_i(s)]^+}{r_i p_i}$	$\frac{[d_i(s)]^+}{r_i p_i}$

The expected demand over risk period that is unconditional on s is given by:

$$\begin{aligned} \mathbb{E}[D_{it}^p] &= \mathbb{E}[\mathbb{E}[D_{it}^p(s)|s]] = \mathbb{E}\left[\frac{1}{p_i} \sum_{s=1}^{p_i} D_{it}^p(s)\right] \\ &= \frac{1}{p_i} \sum_{s=1}^{p_i} \mathbb{E}[D_{it}^p(s)]. \end{aligned} \quad (25)$$

By substituting the coefficients $\pi_{ik}(s)$ in Table 1, we can solve for (25) to find the following:

$$\mathbb{E}[D_{it}^p] = r p_i \times \mathbb{E}[D_{it}]. \quad (26)$$

We can write the variance of $D_{it}^p(s)$ that is unconditional of s as:

$$\begin{aligned} \text{Var}(D_{it}^p) &= \mathbb{E}[\text{Var}(D_{it}^p(s)|s)] + \text{Var}(\mathbb{E}[D_{it}^p(s)|s]) \\ &= \frac{1}{p_i} \sum_{s=1}^{p_i} \text{Var}(D_{it}^p(s)). \end{aligned} \quad (27)$$

We have to express (27) in matrix-vector form in order to find the variance that captures the demand correlation over time; we find that:

$$\mathbf{D}_t^p(s) = \sum_{k=0}^{\tilde{l}+1} \boldsymbol{\pi}_{ks} \mathbf{D}_{t+k}, \quad (28)$$

where $\tilde{l}$ is the maximum value of a_i among all stages ($i \in \text{INP} \cup \text{DIST}$), $\boldsymbol{\pi}_{ks}$ are diagonal matrices with elements $\pi_{ik}(s)$. Now by substituting (10) and (12) and repeatedly iterating, we can find $\mathbf{D}_t^p(s)$ as follows:

$$\mathbf{D}_t^p(s) = \mathbf{N}_s \mathbf{R}_t + \sum_{k=0}^{\tilde{l}+1} \mathbf{Y}_{ks} \boldsymbol{\zeta}_{t+k}. \quad (29)$$

The matrices $\mathbf{N}_s$ and $\mathbf{Y}_{ks}$ and their derivations can be found in *Online Appendix D*. Now by substituting (13) into (29), we can obtain:

$$\mathbf{D}_t^p(s) = \mathbf{N}_s \sum_{s=0}^{\infty} \mathbf{M}^s \mathbf{K} \boldsymbol{\zeta}_{t-1-s} + (\mathbf{N}_s \mathbf{H} \mathbf{G} + \mathbf{Y}_{0s}) \boldsymbol{\zeta}_t + \sum_{k=1}^{\tilde{l}+1} \mathbf{Y}_{ks} \boldsymbol{\zeta}_{t+k},$$

where we can then express the variance of demand over risk period as:

$$\text{Var}(\mathbf{D}_t^p(s)) = \mathbf{N}_s \left(\sum_{s=0}^{\infty} \mathbf{M}^s \mathbf{K} \mathbf{\Sigma} \mathbf{K}' \mathbf{M}'^s \right) \mathbf{N}_s' + (\mathbf{N}_s \mathbf{H} \mathbf{G} + \mathbf{Y}_{0s}) \mathbf{\Sigma} (\mathbf{N}_s \mathbf{H} \mathbf{G} + \mathbf{Y}_{0s})' + \sum_{k=1}^{\tilde{l}+1} \mathbf{Y}_{ks} \mathbf{\Sigma} \mathbf{Y}_{ks}'. \quad (30)$$

To find $\text{Var}(D_{it}^p)$ that is unconditional on s and for computation convenience, we can solve (25) by a single matrix, which is given by:

$$\text{Var}(\mathbf{D}_t^p) = \mathbf{\Omega} \sum_{s=1}^{\tilde{p}} \text{Var}(\mathbf{D}_t^p(s)), \quad (31)$$

where $\mathbf{\Omega}$ is a diagonal matrix with $1/p_i$ for all stages as defined earlier, and $\tilde{p}$ is the maximum value of p_i among stages $i \in \text{INP} \cup \text{DIST}$.

With the mean and variance of $D_{it}^p(s)$ given by (26) and (31) respectively, we can set the required base-stock level and average inventory for each input and distribution stage by:

$$B_i = E[D_{it}^p] + z_i \sqrt{\text{Var}(D_{it}^p)} \quad i \in \text{INP} \cup \text{DIST}, \quad (32)$$

and

$$EI_i = z_i \sqrt{\text{Var}(D_{it}^p)} \quad i \in \text{INP} \cup \text{DIST}. \quad (33)$$

We do not account for cycle inventory in (33) for the same reason as for the processed inventory in (24).

4. General Network and Optimization

We discuss how we model a general network where each stage is allowed to have multiple immediate upstream and/or downstream stages. Furthermore, external demand can occur at any stages. Thus each stage can receive external demand, internal demand or both. We express the demand arrival for stage i with multiple immediate downstream stages j as:

$$D_{it} = \sum_j \varphi_{ij} R_{jt} + \zeta_{it}. \quad (34)$$

Notation ζ_{it} denotes external demand received by stage i . External demands are allowed to be correlated among the demand stages, where covariances of external demand is modeled by $\mathbf{\Sigma}$. Notation φ_{ij} denotes the number of units of demand at stage i for each unit of production or replenishment at stage j . If a production stage needs more than one type of input material, each type of input material is modeled as a separate input stage. In the matrix-vector formulation, (34) is modeled by (10), i.e., the same as that for the serial system. This implies that the model's linearity is preserved and the formulations for characterizing the first two moments of R_{it} , $X_{it}(s)$ and $D_{it}^p(s)$ follow those of the serial system model. However, additional considerations are needed for the nested policy, reorder intervals and risk periods, which we discuss in the following.

The assumption of the nested policy still applies. For the typical nested policy in a general network, (1) can be re-stated as $\Delta_i = r_{ij} \Delta_j$, where stage j places orders to stage i and r_{ij} is a positive integer. In characterizing the first two moments of the demand over risk period, if an input or a distribution stage supplies multiple immediate downstream stages with different reorder intervals, we have to duplicate the stage such that each serves a downstream stage. We let the input or distribution stage indexed by i' and each represented by a group of duplicated stages $G(i')$. The expectation and variance of $D_{i',t}^p(s)$ is given by:

$$\mathbb{E}[D_{i',t}^p] = \sum_{i \in G(i')} \mathbb{E}[D_{it}^p], \quad (35)$$

and

$$\text{Var}(D_{i',t}^p) = \sum_{i \in G(i')} \text{Var}(D_{it}^p) + \sum_{(i,j) \in G(i') \cap (i \neq j)} \text{cov}(D_{it}^p, D_{jt}^p), \quad (36)$$

where $\mathbb{E}[D_{it}^p]$, $\text{Var}(D_{it}^p)$ and $\text{cov}(D_{it}^p, D_{jt}^p)$ can be obtained from (26) and (31), respectively. With (35) and (36), we can set the aggregate base-stock level and average inventory, similar to (32) and (33), where risk-pooling and the demand correlations at downstream stages are considered.

For a stage that replenishes from multiple suppliers, the risk periods associated with each supplier may be different, for which we need to characterize in the model. We let $D_{ij,t}^p$ denotes the demand over risk period at stage j associated with supplier stage i . We assume a fixed proportion φ_{ij} of quantity ordered to each supplier i for every unit of demand at stage j . Hence φ_{ij} in (34) represents the fraction of order quantity placed by stage j to each supplier i . For example, a stage j that sources 30% of a part from supplier $i = 1$ and the rest from supplier $i = 2$ lead to $\varphi_{1j} = 0.30$ and $\varphi_{2j} = 0.70$. To characterize $D_{ij,t}^p$, we first obtain the expectation and variance of $D_{ij,t}^p$ in the same fashion described for the serial supply chain. The expectation and variance of D_{jt}^p can then be obtained by:

$$\mathbb{E}[D_{jt}^p] = \sum_i \varphi_{ij} \mathbb{E}[D_{ij,t}^p],$$

and

$$\text{Var}(D_{jt}^p) = \left(\sum_i \varphi_{ij} \sqrt{\text{Var}(D_{ij,t}^p)} \right)^2.$$

Thus far, we have assumed all production stages hold both decoupling input and processed inventories. One could conceive a supply chain in which some production stages do not hold processed inventory (e.g., assemble-to-order production) and/or input inventory. We offer ways to extend the model to include such systems in *Online Appendix E*.

We formulate a nonlinear program that minimizes the expected total cost (2) for the general network. The decision variables are the planned lead-times n_i and the safety factors z_i . The only constraint is $n_i \geq \Delta_i$ as defined in (3). We have not analytically established the convexity of the objective function. But by searching from multiple start points, we encountered multiple local minima in some of the test problems, thereby indicating that the problem is nonconvex. Despite the nonconvexity, the continuous decision variables and differentiable functions enable efficient gradient-based methods to be employed to find the local minimum for each start point. We implement the optimization in the Global Search solver in MATLAB that searches from multiple start points and utilizes the gradient-based algorithm for each selected start point.

5. Demand Bounds and Capacitated Expediting

We discuss how we can incorporate demand bounds and capacitated freight expediting in the model. For brevity, we consider only the input and distribution stages in the discussion; by similar analysis, one can derive the equivalent functions for the production stages by replacing D_{it}^p by X_{it} in the formulations. We let m_i^E denotes the freight expediting capacity at stage i . We regard m_i^E as the maximum quantity that can be expedited in time for stage i to achieve guaranteed service when it faces stock-out. For instance, it may imply capacity of privately owned transport fleet or cargo space that can be routinely sourced from transport service providers. We consider the following constraint for the optimization:

$$\Pr(B_i + m_i^E \geq D_{it}^p) \geq \delta_i \quad i \in INP \cup DIST. \quad (37)$$

The constraint assures base-stock B_i plus capacity m_i^E are able to meet D_{it}^p at a probability greater than a predetermined δ_i . Presumably, δ_i is high (say, 0.99) as one would expect a high desired service level for using both inventory and expediting. We can regard the expediting considered in the paper as *regular* expediting; like the guaranteed-service models, if the bound is exceeded, we assume *extraordinary* expediting (which is not addressed by the model) are carried out to meet the infrequent demand windfall. One can interpret (37) as a demand bound similar to that in guaranteed-service models, which ensures guaranteed service can be accomplished by finite base-stocks and additionally expedited shipments in this paper. By the same interpretation as Graves and Willems (2000), the demand bound is a design parameter which indicates the manager's preference of how demand variation should be countered: what proportion of it by safety stock and regular expediting, and what proportion by extraordinary expediting. One can re-express (37) as $B_i + m_i^E \geq D_i^B$, where D_i^B is the demand bound given by:

$$D_i^B = E[D_{it}^p] + z_i^T \sqrt{\text{Var}(D_{it}^p)} \quad i \in INP \cup DIST, \quad (38)$$

where $\bar{z}_i^T$ is the given safety factor for the use of both safety stock and expediting. As D_{it}^p is assumed to be normally distributed, $\bar{z}_i^T$ is the inverse of the normal cumulative function of δ_i . Unlike typical guaranteed-service models, the demand bound in (38) is not exogenous but a function of the decision variables. By substituting (32) and (38) into $B_i + m_i^E \geq D_i^B$, we can re-express (37) as:

$$(\bar{z}_i^T - \bar{z}_i) \sqrt{\text{Var}(D_{it}^p)} \leq m_i^E \quad i \in INP \cup DIST. \quad (39)$$

Given (39), if the expediting capacity is tight (i.e., m_i^E is small), the optimal solution would recommend a larger $\bar{z}_i$ and/or more upstream smoothing to reduce $\text{Var}(D_{it}^p)$. Note that $m_i^E = 0$ implies unavailable option to expedite and hence demand uncertainty is wholly buffered by safety stocks (i.e., $\bar{z}_i = \bar{z}_i^T$).

In addition, we define $(\bar{z}_i^T - \bar{z}_i) \sqrt{\text{Var}(D_{it}^p)}$ in (39) as *expediting bound*, which is the *maximum* amount of freight expediting for a specific demand bound. We note that the expediting bound may be smaller than m_i^E , i.e., if (39) is not binding. With the demand bounds, the expected cost of freight expediting in the objective function (2) must exclude those incurred for extraordinary expediting, i.e., the expedited quantity must not be larger than the expediting bound. We incorporate this condition by subtracting the objective function by the cost $c_{it} E[D_{it}^p - D_i^B]^+$, which is the expected expediting cost for the quantity in excess of the demand bound.

6. Numerical Experiments

In this section, we report on the experiments that are carried out to obtain managerial insights as well as to demonstrate the model's application in analyzing supply networks.

6.1. Insights on Production Smoothing in a Serial Network

We first focus on production smoothing in which we consider this tactic in the context of system-wide optimization and inter-stage cost trade-offs. (We will consider the interactions between smoothing, expediting and safety stocks in subsequent sub-sections.) We employ a network that comprises three production-input pairs of stages in series. This simple setup facilitates us to draw insights that can be readily extended to more complex networks. We consider three different sets of cost inputs (Table 2), which are designed to show how inter-stage cost differences affect optimal smoothing. We assume $\mu = 500$ and $\sigma = 200$. We let $m_i = 680$ at all production stages, and $\Delta_i = 1.00$, $L_i = 0$ and $\varphi_{ij} = 1$ at all stages.

To contrast the system-wide optimization performed by the model, we compare the model's results with the case in which each production-input pair of stages is locally optimized. The local optimization can be solved by an extension based on the analytical solution in Boute and Van Mieghem (2014) (primarily to include the input stage) or by our model. In the local optimization, the utmost downstream production-input pair is first optimized. The results provide the demand input needed to optimize the next upstream production-input pair. Eventually, all decision variables z_i and n_i acquired are used to compute the expected total cost.

Table 2: Cost sets for experiments

Stage No.	Set A			Set B			Set C		
	c_{Pi}	c_{Li}	h_i	c_{Pi}	c_{Li}	h_i	c_{Pi}	c_{Li}	h_i
1	1.00	20.00	0.60	1.00	15.00	0.50	1.00	8.00	0.40
2	-	14.00	0.30	-	12.00	0.45	-	8.00	0.40
3	1.00	10.00	0.25	1.00	10.00	0.40	1.00	8.00	0.40
4	-	8.00	0.20	-	8.00	0.30	-	8.00	0.40
5	1.00	5.00	0.15	1.00	5.00	0.15	1.00	8.00	0.40
6	-	3.00	0.10	-	3.00	0.10	-	8.00	0.40

Table 3 shows the comparison between the model and the local-optimized case. In the local-optimized case, the trade-off in optimizing n_i at each production-input pair involves the holding and expediting costs at the processed inventory *versus* the production expediting cost plus holding and expediting costs at the input stage; a larger n_i (i.e., more smoothing) leads to higher cost at the processed inventory but lower costs in production expediting as well as holding and expediting cost at the input stage.

Table 3: Model versus local-optimized case (local-optimized results are in parentheses)

Stage no	n_i			$\sigma(R_{ii})$			Expected Total Stage Cost		
	Set A	Set B	Set C	Set A	Set B	Set C	Set A	Set B	Set C
1	2.82(1.53)	3.82(1.99)	6.02(2.13)	93(140)	78(116)	60(111)	357(297)	330(260)	299(197)
2	-	-	-	93(140)	78(116)	60(111)	67(100)	76(114)	50(92)
3	1.03(1.80)	1.00(1.78)	1.00(2.01)	92(101)	78(90)	60(84)	56(108)	67(139)	50(139)
4	-	-	-	92(101)	78(90)	60(84)	43(47)	51(59)	50(69)
5	1.00(1.66)	1.00(1.66)	1.00(2.00)	92(85)	78(80)	60(73)	32(49)	27(45)	50(120)
6	-	-	-	92(85)	78(80)	60(73)	21(19)	17(18)	50(60)
Expected Total Cost (\$ per period)							576(620)	568(635)	549(677)

However, in the system-wide optimization, the trade-off involves the additional role of reducing variability at the upstream stages. The results highlight the abovementioned trade-offs. The values of n_i in the local-optimized case are closer among the stages. In contrast, for each of the three cost sets, the model's (system-wide) optimal n_i at the utmost downstream production stage (Stage 1) is much larger than the rest, suggesting Stage 1 should assume responsibility in smoothing demand for its upstream stages. The ensuing

lower variability causes $\sigma(R_{it})$ at the upstream stages to be lower than those in the local-optimized case (and thus leads to lower capacity requirements), despite less smoothing (smaller n_i) at these stages than in the local-optimized case. The lower variability also leads to smaller optimal safety stocks at the upstream stages. Furthermore, in the model's results, Stage 1 suffers local sub-optimality in its expected cost in order to achieve global optimality.

The cost trade-offs are further illustrated by the results computed using the different cost sets. The extent of smoothing at Stage 1 for Set A is the lowest among the cost sets due to the higher unit holding and freight expediting costs at the stage than those upstream. In contrast, the amount of smoothing is larger for Set B at Stage 1 due to its lower unit holding and freight expediting costs. Set C, which comprises the same holding and expediting unit costs at all stages, lead to the largest extent of smoothing at Stage 1. Clearly, based on the trade-offs, the network configuration also affects the optimal extent of smoothing. For example, by using a smaller (serial) network of two production-input pairs (Stage 1 to 4), the optimal n_i at Stage 1 is considerably smaller at 2.20, 3.32 and 4.50 respectively for the cost sets. This demonstrates that network configuration can significantly affect the extent of optimal smoothing.

Despite the insights, it is evident that selecting the extent of smoothing in a typical general-network supply chain is not straightforward given the intricacies of inter-stage cost differences and network configurations. As such, the results highlight the model's application in supporting system-wide optimization that incorporates such complexities.

6.2. Analysis of Tactics in a Supply Network

We study the tactics of smoothing, expediting and safety stocks in a supply network. We report on one of the supply chains that are experimented in our collaboration with DHL Supply Chain. As a solution provider, the company hopes to gain managerial insights on the aforementioned tactics to better serve their clients which are from diverse industrial sectors. The results presented are broadly representative of all other experiments carried out. The high-level, scaled down representation of the supply chain in terms of the production, input and distribution stages is illustrated in Figure 5.

Figure 5: Network configuration

The external demands occur at the distribution stages 1, 2 and 3. The expected demand (standard deviation) at these stages are 6,500 (3,100), 5,300 (2,300) and 11,400 (5,500) respectively. Each of the external demand process is normally distributed, independent between stages and i.i.d. over time. Each underlying period represents a four-week time bucket. Table 4 shows the relevant (disguised) data.

Table 4: Input data for base case

Stage no. & type	L_i	Δ_i	m_i	h_i	c_{pi}	c_{li}	Stage no. & type	L_i	Δ_i	m_i	h_i	c_{pi}	c_{li}
1. DIST	0.21	0.25	-	0.50	-	8.46	8. INP	0.64	0.5	-	0.29	-	16.8
2. DIST	0.14	0.25	-	0.49	-	5.30	9. INP	0.36	0.5	-	0.14	-	9.70
3. DIST	0.18	0.25	-	0.51	-	7.22	10. PROD	-	0.5	15,300	0.13	2.10	2.60
4. PROD	-	0.25	15,300	0.47	3.30	4.00	11. INP	0.21	0.5	-	0.11	-	6.13
5. PROD	-	0.25	16,400	0.48	4.35	4.20	12. PROD	-	0.5	29,300	0.23	2.55	6.80
6. INP	0.18	0.5	-	0.15	-	3.54	13. INP	0.29	1.0	-	0.12	-	5.88
7. INP	0.54	0.5	-	0.28	-	12.9	14. INP	0.39	1.0	-	0.09	-	4.84

Effect of Higher Unit Holding Costs Relative to Unit Expediting Costs

We consider the trade-off between expediting and safety stocks, and their interaction with smoothing. In particular, we carry out this study by varying the values of unit holding costs relative to that of unit expediting costs. The holding-to-expediting cost ratio relates to the characteristics of products distributed by a supply chain. For example, as compared to mass-market furniture, high-tech consumer electronics usually involve higher unit holding costs (owing to higher value and obsolescence rate) relative to unit freight expediting cost (due to smaller physical sizes). The ratio also corresponds to the different phases of a product's life-cycle in which its value (and thus its unit holding cost) changes across time. Hence this set-up enables us to study how different products influence the planning parameters.

We experiment by fixing both the unit freight and production expediting costs while increasing the unit holding cost by the following factors: 1.25, 1.50, 1.75 and 2.00. Table 4 shows the relevant results.

Table 4: Results for varying unit holding costs

		Multiplicative Factors in h_i				
		1.00	1.25	1.50	1.75	2.00
Average Optimal z_i		2.13	2.03	1.96	1.89	1.83
Average Optimal B_i		16,782	16,516	16,302	16,119	15,965
Expected Total Freight Expediting Cost (\$ per period)		2,980	3,817	4,676	5,568	6,474
Expected Total Production Expediting Cost (\$ per period)		611	670	717	755	784
Expected Total Inventory Holding Cost (\$ per period)		16,714	19,876	22,866	25,678	28,354
Optimal n_i	Stage 4	1.435	1.369	1.326	1.289	1.264
	Stage 5	1.375	1.315	1.271	1.239	1.216
	Stage 10	0.500	0.500	0.500	0.500	0.500
	Stage 12	0.500	0.500	0.500	0.500	0.500

Expectedly, at higher unit holding costs, it is optimal for the stages to curb their inventory holding costs with smaller B_i and opt for more frequent expediting, as indicated by both the higher expected total production expediting and freight expediting costs. However, without performing the optimization, it is not obvious how the optimal planned lead-times at production stages 4 and 5 should be set at different h_i . One might think that, at higher holding costs, the production stages 4 and 5 should adopt smaller n_i in order to achieve lower levels of processed inventories. But by doing so, the resulting higher production variability will lead to higher demand variability at upstream stages, which will be more costly to cope given the higher unit holding costs. The solution shows that it is optimal to reduce n_i at both stages. Given that the demand variability increases in the upstream stages as a consequence, it is optimal for these upstream stages to reduce their base stocks (with lower $\bar{x}_i$ values) and use more expediting to deal with the greater variability. Note that this result applies to this set-up but in general, n_i at production stages 4 and 5 also depend on the holding and expediting costs at the upstream stages (which relates to the inter-stage trade-offs discussed in §6.1). Through additional experiments carried out, we observe that if these upstream costs are adequately high, it is instead optimal to increase n_i at production stages 4 and 5 as h_i increases so as to reduce upstream demand variability.

Even though the experiments have demonstrated how smoothing should be carried out at different levels of holding costs (relative to expediting costs), it is not apparent how to set the extent of smoothing according to the different cost structures (or whether to raise or reduce smoothing as the costs change, e.g., over a product's life cycle). The model incorporates the underlying complexities when providing such decision support. In addition, to demonstrate how the cost structure can affect the model's performance, we compare the model with two base settings that are proxies of the scenarios in which expediting and safety stocks are not simultaneously planned. The settings are defined as follows:

Setting 1: 95% of availability at all inventory stages. This setting represents the tactic to hold reasonably healthy levels of safety stocks throughout the supply chain.

Setting 2: 99% of availability at distribution stages 1, 2 and 3; 97% at production stages 4 and 5; 95% at all other stages. This setting depicts the tactic in which higher safety stock levels are carried at downstream stages to achieve a high fill rate to the external customers.

In both settings, we assume smoothing is not carried out. We compare the expected total cost of each setting with that of the optimal solution at different unit holding costs. To set apart the cost reduction caused by production smoothing in the optimal solution (since there is no smoothing in both the settings), we perform an additional optimization without smoothing. Same as in the previous experiment, we increase the unit holding costs (with unit expediting costs fixed) but by a different set of multiplicative

factors $\{2, 3, \dots, 8\}$. We graph in Figure 6 the percentage savings in the expected total cost achieved by our model (with and without smoothing) as compared to both settings.

Figure 6: Percentage savings resulting from the optimal solution

For optimal solutions with smoothing, the highest (lowest) percentage savings are 37% and 36% (16.4% and 19.8%) for Setting 1 and 2 respectively, at multiplicative factor of 1 (4). For those without smoothing, the highest (lowest) savings are 16.5% and 15.1% (2.3% and 5.3%) for Setting 1 and 2 respectively, at multiplicative factor of 1 (3). The results demonstrate that the savings achieved by the model depends significantly on the cost structure, or relatedly, the type or life-cycle phase of the product.

Demand Bounds and Capacitated Freight Expediting

In the following experiments, we use the model extension in §5 to model the demand bounds and capacitated freight expediting. We show the use of the model extension in “what-if” analyses from which we also draw related insights.

We consider the probabilities for the demand bounds $\delta_i = \{0.98, 0.99, 0.999\}$. For simplicity, in every test scenario, we assume the same δ_i value applies to all stages. To vary the extent of expediting deployed so as to evaluate its benefit, we vary the freight expediting capacities m_i^E from 0 to 1,000 in increments of 250 units. We assume the same m_i^E for every inventory in all test scenarios. We graph the expected total cost at different m_i^E in Figure 7, where we show the results separately for the cases with smoothing (where both n_i and z_i are decision variables) and without smoothing (where $n_i = \Delta_i$ at all production stages and z_i are the only decision variables). In addition, to facilitate subsequent discussion, we define the cases in Table 5. We indicate the expected total costs of the cases in Figure 7.

Table 5: Cases for smoothing and freight expediting

Case	Smoothing	m_i^E
Smoothing Only	Yes	0
Expediting Only	No	1,000
Smoothing & Expediting	Yes	1,000
No Smoothing & No Expediting	No	0

Figure 7: Expected total cost with and without smoothing

Analyses such as those shown in Figure 7 allow one to study the sensitivity of the expected total cost to m_i^E as well as the cost advantage of smoothing and freight expediting. In the following, we experiment with the optimal deployment of freight expediting and safety stocks at different demand bounds under the effect of smoothing.

Distinct role of expediting freight at higher demand bounds

We compare the cases of “Smoothing & Expediting” to “Smoothing Only”. The differences in the expected total cost increases with δ_i , which are given by \$148, \$813 and \$1,805 for $\delta_i = 0.98, 0.99$ and 0.999 respectively. The widening cost gap when δ_i increases illustrates the greater importance of freight expediting at higher demand bounds. The results also highlight the distinct roles of safety stock and expedited freight in countering demand variability. Without expedited freight, considerable amount of safety stock have to be constantly held to cover for the event of very high demand that has low probability of realization. Although expedited freight incurs a higher unit cost than stock holding, the cost is incurred on “charge-on-use” basis; hence it plays a crucial role in covering the low-probability demand windfalls.

Freight expediting is more critical downstream at higher demand bounds

We look into how smoothing affects the relative importance of freight expediting (compared to safety stocks) at different stock locations. We first compare “Smoothing & Expediting” with “Expediting Only” at $\delta_i = 0.98$. The results (Table 6) indicate that smoothing causes the variability buffer of freight expediting

and safety stocks to shift to the downstream production stages. To be more specific, by using the expediting bound as a measure of requirements for freight expediting, smoothing at the utmost downstream production stages (Stage 4 and 5) in “Smoothing & Expediting” causes the expediting bounds and optimal safety stocks of the processed inventories at these stages to be much larger than those in “Expediting Only”. However, at the upstream stages, the expediting bounds and optimal safety stocks are significantly lower in “Smoothing & Expediting” than in “Expediting Only” due to the reduced variability from the smoothing.

When we raise δ_i to 0.99, the expediting bounds increase more significantly than the safety stocks. Specifically, the expediting bounds at production stages 4 and 5 increase from 489 and 628 to 999 and 995 respectively (104% and 58% respectively). In contrast, the safety stocks at these stages increase from 5,704 and 7,889 to 6,046 and 8,939 (6% and 13% respectively), constituting smaller percentage increases. The analysis demonstrates that when smoothing is performed at high demand bounds, even though requirements for both safety stocks and expediting at the downstream stages increase, freight expediting tends to be more vital than safety stocks in minimizing the expected total cost.

Table 6: Comparison between “Smoothing & Expediting” and “Expediting Only” at $\delta_i = 0.98$

Smoothing & Expediting			Expediting Only		Smoothing & Expediting			Expediting Only	
Stage	Exp. bound	Safety stock	Exp. bound	Safety stock	Stage	Exp. bound	Safety stock	Exp. bound	Safety stock
1	102	1,489	104	1,488	8	0	5,935	0	9,782
2	148	1,033	149	1,032	9	0	4,474	0	7,471
3	247	2,577	247	2,577	10	99	2,025	196	3,768
4	489	5,704	155	1,827	11	0	2,123	0	3,964
5	628	7,889	209	2,615	12	0	3,759	0	6,900
6	168	1,956	311	3,653	13	387	7,132	719	13,081
7	103	3,969	175	6,690	14	392	7,127	719	13,081

7. Concluding Remarks

This paper presents a model that supports the integrated planning of production smoothing, freight expediting and safety stocks. We first describe how to model a serial supply chain as a linear matrix-vector system. We characterize the production and inventory random variables needed to compute the expected total cost. We then discuss how the model can be extended to model a general network, and to incorporate demand bounds and capacitated freight expediting. We report on numerical experiments to draw managerial insights on smoothing and deployment of expedited freight.

We summarize the insights as follows. We show the importance of downstream smoothing in reducing upstream variability. From the system-wide viewpoint, the optimal smoothing at a production

stage depends on the trade-off between its holding and expediting costs at the processed inventory versus its production expediting cost plus all upstream costs. The trade-off depends significantly on the inter-stage cost differences and network configuration. The optimal amount of smoothing should be adjusted according to the changes in the network cost structure such as over the life of a product. Despite the insights, it may not be obvious how to set the extent of smoothing according to the changing costs for typical supply networks where network cost structure and configuration can be complex. This demonstrates the need for decision support, which provides the motivation for the model's application. In addition, we show that the optimal solution tends to suggest deployment of more safety stock and freight expediting at the downstream smoothing stages. Furthermore, if the service required from both safety stock and freight expediting is high (as defined by the demand bounds), more freight expediting tends to be deployed relative to safety stocks.

There remain limitations and questions regarding the model. First and foremost, although the model includes the consideration of the reorder intervals, we did not provide any guidance on how best to set their values. We have discussed how the model can be extended to include the absence of decoupling inventories but have effectively assumed that these decisions are predetermined and thus did not address how the positioning of such inventories interacts with the decision variables. We have resorted to the approach of employing multiple start points for the optimization, which has served well for the problems we encountered. However, we have not explored in depth the relationship between the problem structure and the computational complexity. We have assumed a stationary external demand but demand tends to be nonstationary for short-life cycles or seasonal products. We have not addressed how to dynamically re-plan the tactical parameters in response to a nonstationary demand process. All of these issues deserve further research.

References

- Aimi, G. 2009. Logistics News: Getting a Handle on Expedited Freight. Comments made by AMR Research in Supply Chain Digest, http://www.scdigest.com/assets/On_Target/09-02-03-1.php?cid=2231 (accessed 30th December 2016).
- Balakrishnan, A., J. Geunes, M.S. Pangburn. 2004. Coordinating supply chains by controlling upstream variability propagation. *Manuf. Serv. Oper. Mgmt.* **6** 163–183.
- Berling, P., V. Martinez. 2011. Optimal expediting decisions in a continuous-stage serial supply chain. Working Paper, IESE Business School, University of Navarra, Barcelona.
- Beyer, D., J. Ward. 2001. Network server supply chain at HP: A case study. J. S. Song & D. Yao, eds. *Supply Chain Structures: Coordination, Information and Optimization*. Kluwer Academic Publishers, Boston, MA. Chapter 8.

- Bhatnagar R., P. Mehta, C.C. Teo. 2011. Coordination of planning and scheduling decisions in global supply chains with dual supply modes. *Internat. J. Production Econom.* **131** 473–482.
- Boute R., J. A. Van Mieghem. 2014. Global dual sourcing and order smoothing: The impact of capacity and leadtimes. *Mgmt. Sci.* Articles in advance. doi: 10.1287/mnsc.2014.1992
- Bray R.L., H. Mendelson. 2015 Production smoothing and the Bullwhip effect. *Manuf. Serv. Oper. Mgmt.* Articles in advance. doi: 10.1287/msom.2014.0513
- Cachon, G.P., T. Randall, G.M. Schmidt. 2007. In search of the bullwhip effect. *Manuf. Service Oper. Mgmt.* **9** 457- 479.
- Diks, E. B., A. G. de Kok, A. G. Lagodimos. 1996. Multi-echelon systems: A service measure perspective. *Eur. J. Oper. Res.* **95** 241–263.
- Fine, C. H., S. C. Graves. 1989. A tactical planning model for manufacturing subcomponents in mainframe computers. *J. Manuf. & Oper. Mgmt.* **2** 4–34.
- Foreman, J., J. Gallien, J. Alspaugh, F. Lopez, R. Bhatnagar, C.C. Teo, C. Dubois. 2010. Implementing supply-routing optimization in a make-to-order manufacturing network. *Manuf. Serv. Oper. Mgmt.* **12** 547-568.
- Graves, S. C. 1986. A tactical planning model for a job shop. *Oper. Res.* **34** 522–533.
- Graves, S. C., H.C. Meal, S. Dasu, Y. Qiu. 1986. Two-stage production planning in a dynamic environment. S. Axsater, C. Schneeweiss, E. Silver, eds. *Multi-Stage Production Planning and Inventory Control*. Springer-Verlag, Berlin. 9–43.
- Graves, S. C. 1988. Safety stocks in manufacturing systems. *J. Manuf. and Oper. Mgmt.* **1** 67– 101.
- Graves, S. C., D.B. Kletter, W.B. Hetzel. 1998. A dynamic model for requirements planning with application to supply chain optimization. *Oper. Res.* **46** 35–49.
- Graves, S. C., S. P. Willems. 2000. Optimizing strategic safety stock placement in supply chains. *Manuf. Service Oper. Mgmt.* **1** 68–83.
- Graves, S. C., S. P. Willems. 2003. Supply chain design: safety stock placement and supply chain configuration. A.G. de Kok, S.C. Graves, eds. *Handbooks in Oper. Res. & Mgmt. Sci. Vol. 11., Supply Chain Management: Design, Coordination & Operation*. Elsevier, Amsterdam, The Netherlands. Chapter 3.
- Groenevelt, H., N. Rudi. 2003. A base stock inventory model with possibility of rushing part of order. Working paper, Simon School of Business, University of Rochester, Rochester, NY.
- Holt, C. C., F. Modigliani, H. A. Simon. 1955. A linear decision rule for production and employment scheduling. *Mgmt. Sci.* **2** 1-30.
- Jain, A., H. Groenevelt, N. Rudi. 2010. Continuous review inventory model with dynamic choice of two freight modes with fixed costs. *Manuf. Service Oper. Mgmt.* **12** 120–139.
- Kim C., D. Klabjan, D. Simchi-Levi. 2009. Optimal policy for a periodic review inventory system with expediting. Working paper, MIT.
- Lawson, D. G., E. Porteus. 2000. Multi-stage inventory management with expediting. *Oper. Res.*, **48** 878-893.
- Linebaugh, K. (2011, Oct 24). Lean companies ready to cut. The Wall Street Journal. Retrieved from <http://online.wsj.com/article/SB10001424052970203911804576649093203840446.html>

- Meyr, H. 2012. Demand planning. H. Stadtler, B. Fleischmann, M. Grunow, H. Meyr, C. Sürie, eds. *Advanced Planning in Supply Chains: Illustrating the Concepts Using an SAP® APO Case Study*. Springer-Verlag, Berlin. 67–108.
- Minner, S. 2003. Multiple-supplier inventory models in supply chain management: A review. *Internat. J. Production Econom.* **81-82** 265–279.
- Muharremoglu, A., J. N. Tsitsiklis. 2003. Dynamic lead-time management in supply chain. Working paper, Columbia University, New York.
- SAP Library. Heijunka: Setting for the planning profile, http://help.sap.com/erp2005_ehp_02/helpdata/en/45/6becf0dded72ace10000000a1553f7/content.htm (accessed 10 May 2012).
- Swaminathan, J., S. Tayur. 2003. Tactical planning models for supply chain management. A. G. de Kok, S. C. Graves, eds. *Handbooks in Oper. Res. & Mgmt. Sci. Vol. 11., Supply Chain Management: Design, Coordination & Operation*. Elsevier, Amsterdam, The Netherlands. Chapter 8.
- Teo, C. C., R. Bhatnagar, S. C. Graves. 2011. Setting planned lead times for a make-to-order production system with master schedule smoothing. *IIE Transactions* **43** 399–414.
- Teo, C. C., R. Bhatnagar, S. C. Graves. 2012. Application of master schedule smoothing and planned lead times control. *Production & Oper. Mgmt.* **21** 211–223.
- Terry, L. 2012. Expedited Transport: It's All in the Timing, <http://www.inboundlogistics.com/cms/article/expedited-transport-its-all-in-the-timing/> (accessed 30th December 2016).
- van Houtum, G. J., K. Inderfurth, W.H.M. Zijm. 1996. Materials coordination in stochastic multi-echelon systems. *Eur. J. Oper. Res.* **95** 1–23.
- Veeraraghavan, S., A. Scheller-Wolf. 2008. Now or Later: A simple policy for effective dual sourcing in systems. *Oper. Res.* **56** 850–864.