Journal of Manufacturing Systems 26 (2007) 116–134

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

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

Technical paper

Tractable supply chain production planning, modeling nonlinear lead time and quality of service constraints

Osman Murat Anli [a,*], Michael C. Caramanis [b,c], Ioannis Ch. Paschalidis [b,c]

[a] Industrial Engineering Department, Isik University, Sile, Istanbul 34980, Turkey
[b] Center for Information and Systems Engineering (CISE), Boston University, Boston, MA, USA
[c] Department of Manufacturing Engineering, Boston University, Boston, MA, USA

ARTICLE INFO

Article history:
Received 15 November 2006
Received in revised form
29 January 2008
Accepted 2 May 2008

ABSTRACT

This paper addresses the task of coordinated planning of a supply chain (SC). Work in process (WIP) in each facility participating in the SC, finished goods inventory, and backlogged demand costs are minimized over the planning horizon. In addition to the usual modeling of linear material flow balance equations, variable lead time (LT) requirements, resulting from the increasing incremental WIP as a facility's utilization increases, are also modeled. In recognition of the emerging significance of quality of service (QoS), that is, control of stockout probability to meet demand on time, maximum stockout probability constraints are also modeled explicitly. Lead time and QoS modeling require incorporation of nonlinear constraints in the production planning optimization process. The quantification of these nonlinear constraints must capture statistics of the stochastic behavior of production facilities revealed during a time scale far shorter than the customary weekly time scale of the planning process. The apparent computational complexity of planning production against variable LT and QoS constraints has long resulted in MRP-based scheduling practices that ignore the LT and QoS impact to the plan's detriment. The computational complexity challenge was overcome by proposing and adopting a time-scale decomposition approach to production planning, where short-time-scale stochastic dynamics are modeled in multiple facility-specific subproblems that receive tentative targets from a deterministic master problem and return statistics to it. A converging and scalable iterative methodology is implemented, providing evidence that significantly lower cost production plans are achievable in a computationally tractable manner.

1. Introduction

1.1. Motivation and objectives

Modern manufacturing enterprises are becoming more global than ever. They encompass owned or contract manufacturing and transportation facilities, suppliers, distributors, and customer service centers scattered over the globe. Manufacturers are no longer the sole drivers of the supply chain (SC). A shift from a "push" to a "pull" environment is well on its way. Customer needs and preferences influence the SC's inner workings: product functionality, quality, speed of production, timeliness of deliveries, flexibility in adjusting to demand changes. In today's highly competitive marketplace, companies are challenged with achieving shorter order-to-delivery times while allowing customers to customize their orders. Manufacturers recognize the significance of short lead times (LT) and high quality of service (QoS) provisioning for control of stockout probability. Furthermore, time-based competition has had a significant impact on the design of production facilities (product cells) and their operation (just in time, zero in-process inventory, lean manufacturing, and so on). Finally, supplier–consumer information sharing has been looked on as a means to reduce inventories needed to provide a desired service level. Although these efforts, together with the wider use of enterprise-wide transactions databases, have achieved remarkable productivity gains, further improvements in global SC lead times and QoS are critically required. The revolution in computational intelligence and communication capabilities, assisted more recently by the emergence of sensor networks with dynamically reconfigurable topology, has brought these improvements within reach.

The lead time at each link of a SC contains information that is critical for effective coordination. Lead times change across weeks in the planning horizon. In fact, they vary nonlinearly with load, production mix, lot sizes, detailed scheduling, and other operational practices adopted during each week of the planning horizon. Nevertheless, widely used material requirements planning

* Corresponding author.
E-mail addresses: omanli@isikun.edu.tr (O.M. Anli), mcaraman@bu.edu (M.C. Caramanis), yannisp@bu.edu (I.Ch. Paschalidis).

0278-6125/$ – see front matter © 2008 The Society of Manufacturing Engineers. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jmsy.2008.05.001

(MRP) systems assume that lead times are constant across the whole planning horizon to avoid the task of estimating and communicating variable lead time information. The use of limited information in the current state-of-the-art industrial practice is responsible for inefficient planning and often chaotic and unstable operations hampered by chronic backlogs and widely oscillating inventories. Two major barriers preventing more extensive use of information are (i) the cost of collecting, processing, and communicating the requisite information and (ii) computational and algorithmic challenges in using this information to plan and manage SCs optimally. A time-scale decomposition and information communication architecture framework is proposed that is capable of exploiting sensor networks, and overcome the communication barrier. An iterative decentralized coordination algorithm is also proposed that provides proof of the concept that the computational barrier can be overcome as well.

1.2. Current industry practice

Whereas capacity is ignored and dynamics are modeled by constant lead times in the "vanilla" version of materials requirement planning (MRP) models [1], advanced planning system (APS) approaches include adequate representation of material flow dynamics and detailed representation of effective (or expected) capacity. APS models rely on mathematical programming techniques and hierarchical decomposition [2,3] to overcome combinatorial complexity explosion barriers while capturing the details of capacity restrictions. This task is particularly onerous in the face of discrete part integrality and complex production rules and constraints, which together with uncertainty, render stochastic integer programming formulations computationally intractable.

Past approaches employed to bypass these hurdles include the theory of constraints, scheduling algorithms, and fluid model approximations. The theory of constraints [4–6] approximates the model of the production system by a small number of bottleneck components that are modeled in great detail; production is scheduled around those components through constraint propagation over time. Two main shortcomings of the theory of constraints approach are: first, the difficulty in identifying and modeling bottleneck components, and second, the fact that delays or lead time dynamics along part routes are nonlinear and difficult to model. The systematic modeling of individual facilities could be possibly used to alleviate the first shortcoming. However, it is very difficult to overcome the second shortcoming. A variety of scheduling algorithms ranging from mathematical programming and Lagrangian relaxation to genetic algorithms have been used, often effectively [7–15]. Fluid model approximations have also been used extensively and with considerable success [16–21] but have not adequately addressed dynamic lead time modeling. System dynamics simulation models have been proposed [22] that capture nonlinearly increasing lead times as functions of the production facility utilization. It has been shown that deterministic fluid model approximations of stochastic discrete production networks can be employed to predict the qualitative nature of optimal scheduling rules [8,9,16,18] and to determine the stability and robustness of the approximated stochastic discrete networks [23–26]. The proposed algorithm exploits this line of research with particular emphasis on extending fluid network approximations to improve the dynamic lead time modeling capabilities.

Past efforts to model lead time in production planning are noteworthy [27,28] but are limited to static lead times estimated for average or typical production conditions. Incorporating dynamic lead times into production planning poses modeling analysis and computational difficulties leading to deliberate choices on simplifying approximations and relaxations. A mixed-integer production planning model has been proposed [29] that employs piecewise-linear functions to capture the effect of alternative routings and subcontractors on load-dependent lead times. The quantitative relationships between work in process and production are estimated via Monte Carlo simulation and used as constraints in a nonlinear production planning model [30] solved through linearization. The effect of inventory on the quality of service has been also studied in an uncapacitated single-product multi-class-QoS supply chain through queuing approximations [31] and in a multi-product single-facility fixed lead time setup [32]. A recent literature survey [33] provides an extensive overview of dynamic lead time modeling in production planning and points out the use of the aforementioned nonlinear relationship in supply chain production planning. This paper explores further in that direction.

1.3. Overview of the proposed approach

The time-scale-driven decentralized information estimation and communication architecture that are proposed in Section 2 enable coordination, planning, and operational decisions of manufacturing cells, transportation activities, inventory, and distribution facilities in a SC. It is shown that this can be achieved through optimal and consistent production targets and safety-stock levels scheduled for each part type produced by each SC facility. Proposed is a framework of iterative information exchange between three decision-making/performance-evaluation layers that is indeed capable of achieving this coordination. The framework consists of a centralized planning coordination layer, a centralized QoS coordination layer, and finally a decentralized performance evaluation and demand information layer. The planning layer determines facility-specific production targets using performance and sensitivity information it receives from the decentralized performance evaluation and information layer. The QoS layer combines interacting facility production capabilities and requirements (that is, targets) to determine hedging inventory requirements that achieve exogenously specified QoS levels. The decentralized performance evaluation and demand information layer analyze short-term (hourly) stochastic dynamics of each facility to derive expected (weekly) work-in-process and safety-stock inventory for each facility and their sensitivity w.r.t. planning level targets.

The major objective of the proposed framework is to capture second-order effects of the steady-state cell dynamics in order to model dynamic lead time effectively at the coarse (varying weekly) production planning dynamics layer. Weekly time averages are a statistic with relatively low variance due to the law of large numbers effects, and they can be effectively modeled as deterministic quantities within the planning layer. Furthermore, detailed information on machine-specific queue and setup states is not globally available, hence, it is practical to share state information that is (i) time averaged to the coarse time scale and (ii) grouped by facility. To this end, capacity, work-in-process, and production requirements are facility-specific aggregates. Production planning dynamics are thus constrained to satisfy minimum weekly average lead time requirements. Note that although facility lead times and interfacility hedging inventory requirements are averages over the fine (hourly) time-scale dynamics modeled at the decentralized performance evaluation layer, they are dynamic relative to the coarse (weekly) time scale of the planning layer. Lead times and hedging inventory requirements are modeled as functions of production planning decisions (loading and mix). This constitutes the second-order information that has been shown can be used [34] to significantly decrease inventory and backlog costs. The planning coordination layer employs an iterative interaction of a single production planning master problem on the one hand with the hedging

policy QoS layer and the performance evaluation layer's multiple decentralized facility-specific subproblems on the other.

The effectiveness of our planning layer model depends crucially on the quality with which the operational dynamics of the production facilities are modeled in the performance evaluation and information layer. To this end, the framework relies on the following two building blocks:

1. *Dynamic lead time modeling:* Performance analysis results for stochastic queuing networks are used to accurately estimate average weekly lead times as functions of capacity utilization, production mix, production policies, and distributions of stochastic disturbances such as failure and repair times. This provides delivery requirements to upstream facilities and available supply to downstream facilities, which are necessary for efficient planning of production over a multi-week horizon. These nonlinear lead time functions, denoted by $\bar{g}(\cdot)$, are incorporated as weekly constraints on decision variables in production scheduling.

2. *Provisioning of quality of service (QoS) guarantees:* Constraints are introduced that bound the probability of backlog at a SC facility. It is believed that probabilistic constraints reflect customer satisfaction considerations and follow closely the industry practice of providing QoS guarantees. These guarantees are modeled as nonlinear constraints in the production scheduling framework, denoted by $\bar{h}(\cdot)$.

The main purpose of this article is to demonstrate that dynamic lead times and hedging inventory requirements can be modeled and included in the tractable determination of faster SC production plans while maintaining the desired quality of service guarantees. The aim is to provide a proof of concept regarding the feasibility of modeling dynamic lead times and quality of service guarantees as part of the production planning process. Through a variety of numerical examples, the potential cost savings achievable by the proposed approach are studied relative to traditional constant lead time based production planning approaches that represent the bulk of today's industry practice. The comparison supports the viability of implementing the approach in real life provided that the cost of estimating and processing the required lead time information is affordable. It is not claimed that the proposed model is a perfect model of reality, particularly as far as the decentralized queuing network subproblem model is concerned. More general and accurate models have been developed in the queuing network and simulation literature, and undoubtedly further extensions are forthcoming from the formidable and research-active community studying these topics. Moreover, the adoption of radio frequency identification tag (RFID) and sensor network technologies will also contribute to the affordability of dynamic lead time information. This contribution is for showing how this information can be used in a tractable, computationally efficient, and robust production planning algorithm.

The demonstration of significant reduction in inventory costs when the nonlinear relationship of facility lead times is modeled in the SC production planning process is not the major contribution of this paper. Most practitioners will argue from experience that this is hardly surprising given the widely observed inadequacy of MRP-based production schedules that rely on the constant lead time assumption. The major contribution of this paper is in its proposal and implementation of a practical, efficient, tractable, and robust algorithm capable of actually achieving these cost savings. The aim is to prove the concept that *SC production planning on constant lead times is not a necessary evil imposed by the incorrect presumption of insurmountable computational complexity*. In fact, it is claimed that SC planning no longer has to live with the undesirable consequences of the constant lead time assumption impeding today's industry practice. This contribution supports the notion that detailed production facility models and/or adoption of RFID technologies can provide additional value added through their ability to extract reliable, albeit affordable, dynamic lead time information and make it available to the production planning optimization process.

The next section introduces the proposed time-scale-driven data communication architecture. The SC problem and the performance evaluation, QoS, and planning layers are then described, following by computational experience that shows the value of dynamic lead time and probabilistic QoS constraint information in the determination of a SC's coordinated production schedule. A three-facility SC producing five different part types is used to develop various representative examples of SCs. Comparison to production schedules that are characteristic of current industry practice indicates that substantial improvements are possible.

2. Time-scale-driven decentralized data communication and decision support architecture

The multitude of strategic, planning, and operational decisions made routinely by SC participants are far too complex and the requisite information is far too large to handle in a centralized manner. Decentralized decision making has therefore been the norm. However, since the consequences of various decisions are interdependent, it follows that appropriate coordination can foster desirable efficiencies. Consider a decentralized decision-making agent as "a decision node" in a network of communicating decision nodes. A key determinant of successful coordination is the systematic conversion of data available at a certain decision-making node i to a compact representation of information "relevant" to the decision-making process at node j. Relevant is construed here to mean *incorporating all information about the state, dynamics, and decision policies in node i that may contribute to efficient decision making in node j*. Compact representations of relevant information may take, for example, the form of a *statistic*: the time-averaged lead time in a production system, the probability distribution and autocorrelation of a demand process, or a *performance target*, such as the desired weekly output of a manufacturing process. These compact representations provide key enabling efficiencies in both the estimation of the relevant information (which can be done in a decentralized distributed manner) as well as in its communication (the transmission of a statistic requires less bandwidth and energy than the time series it describes). Although several issues are still to be resolved, intelligent communicating mobile sensor networks have the potential to both estimate and communicate relevant information in ways that are superior to conventional alternatives in terms of cost, flexibility, and reliability.

Proposed is a time-scale-driven assignment of SC decisions to nodes that is suggestive of the "relevant" information exchange architecture. The idea of time-scale-driven decomposition is not new. In fact, it has been widely used to great advantage in control theory [35]. The main idea here is the fact that decisions are characterized by a characteristic frequency and its corresponding time scale. For example, while machine operating decisions are made every few minutes, major resource acquisition decisions are made every few months or years. It is further noticed that supply chain decisions characterized by functionality (for example, resource allocation, planning, sequencing) and scope (for example, enterprise, plant, cell, process) are associated with a decreasing time scale as the scope narrows and the functionality changes from resource allocation to sequencing. Table 1 provides such a classification example where time scales decrease as the decision of interest moves to the southeast.

The SC planning algorithm proposed here employs a decentralized decision-making and information exchange architecture

Table 1
Example of time-scale-driven classification

Functionality \ Scope	Enterprise	Factory	Cell	Process
Resource Allocation	Hardware Investment	Group Technology	Layout; Overtime	Tools; Time
Contingency Planning	Risk Diversification	Outsourcing Safety Stock	Operational Policies; Performance Evaluation	Statistical Quality Control; Maintenance
Sequencing	Plant Production Schedule	Cell Production Schedule	Machine Schedule	Real-Time Process Control

that is an instantiation of the time-scale-driven approach of Table 1. Fig. 1 presents the information exchange architecture that supports factory-scope production planning decisions, cell-level performance evaluation, and process-level operation control. Note that:

- The factory production planning node passes down weekly production targets to each cell.
- Each cell evaluates its performance during each week in the planning horizon, determines variability distributions, and aggregates its hourly dynamics to weekly time averages of relevant performance measures such as work in process (WIP), lead time (LT), and their sensitivity with respect to weekly production targets passed down from the factory planning node.
- Contiguous cells coordinate horizontally to determine safety inventory of semi-finished and finished goods that assures desired quality of supply levels at each cell and quality of service to customers.
- Each cell communicates weekly averages and sensitivities up to the factory planning node and variability distributions horizontally to upstream and downstream cells.
- Using WIP, LT, and sensitivity information, the factory planning node adjusts production targets so as to achieve material flows across cells that meet required WIP and safety-stock levels while minimizing SC WIP and LT.

While the remainder of the paper proposes algorithms that can make practical use of the information flow described above and reach a stable and optimal production plan, it must be emphasized that the proposed information architecture, in addition to distributing computational effort (performance evaluation and handling of short-time-scale stochastic dynamics modeling are done in a decentralized manner at each cell) also reduces communication requirements to the relevant information. For example, the factory planning node does not need to know the cell production details: labor and other resources available, machine capacities, and manufacturing process specifics. It needs to know, however – and it does know – the weekly lead times at each cell and the hedging inventory between cells that are consistent with the production targets that the planning node sends to each cell.

The general philosophy of the time-scale-driven communication and decision support architecture described in this section provides useful guidelines but not a mindless recipe. In the rest of this paper, these guidelines are used to propose a SC production planning algorithm that is computationally tractable and outperforms two proxies of the state of the art in industry practice that it is compared to.

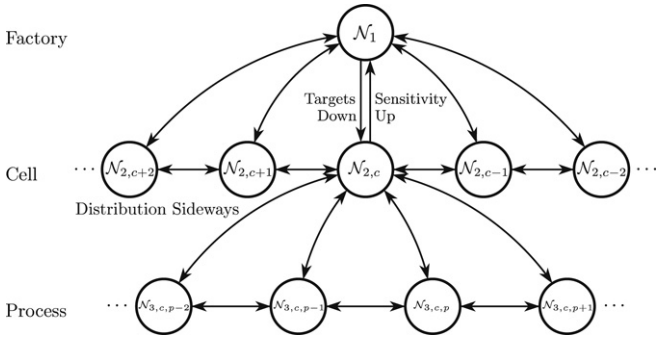

Fig. 1. Example of information architecture.

3. Supply chain management problem

This section develops a supply chain (SC) planning algorithm that utilizes the principles of the time-scale-driven architecture discussed above. Following an overview, the three layers employed are described in detail, as well as their interaction in providing the optimal SC production plan.

3.1. Supply chain problem overview

To describe our SC model and establish the notation, the system depicted in Fig. 2 is considered, and the associated information exchange and decision layers are shown in Fig. 3. Although a tree network of SC links or facilities can be modeled, C production facilities connected in series are considered here, for ease of exposition but without loss of generality. Production planning decisions and the resulting WIP and QoS hedging inventory requirements vary in the medium term (say, across weeks), and the characteristic scale of their dynamics is called a period and denoted by $t \in \{1, 2, \ldots, T\}$. On the other hand, performance evaluation and demand dynamics vary many times within a period (say, across hours) and their characteristic scale is called a time slot denoted by the subscript k of $k \in \{1, 2, \ldots, K\}$.

The QoS layer determines a hedging point or safety-stock inventory policy at each facility c, which guarantees that the probability of stockout or starvation of facility $c - 1$ does not exceed $1 - \Gamma_c(t)$, where $\Gamma_c(t)$ is the quality of service that facility c offers to facility $c - 1$. In other words, the probability that the material release requirements of facility $c - 1$ are met on time equals or exceeds $\Gamma_c(t)$. The QoS layer models random behavior of short-term facility production capacity and final demand, while the planning layer models expected values or time averages during

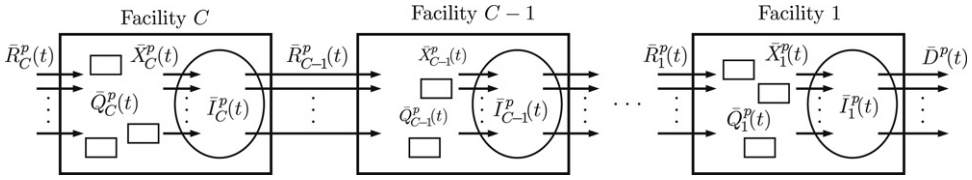

Fig. 2. A multi-class supply chain with limited production capacity at each facility.

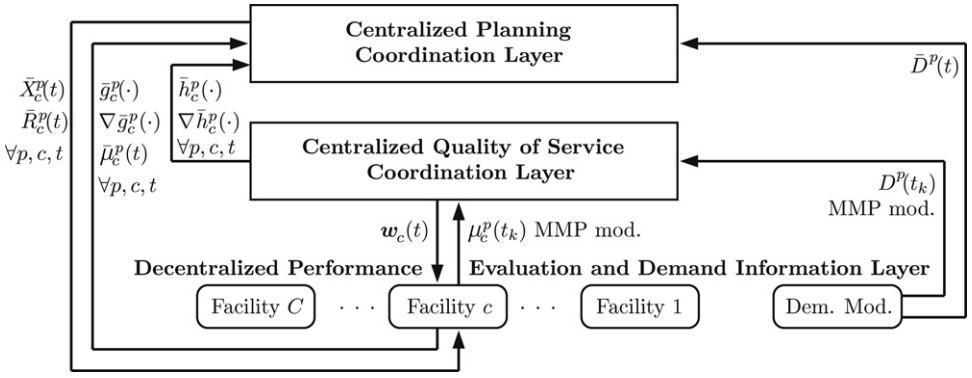

Fig. 3. Information exchange among the coordination layers and the decentralized layer.

a period (for example, a week) under the underlying assumption that the period is long enough for the time-slot stochastic process dynamics to reach steady state.

External demand is met from the available finished goods inventory in facility 1, and it is backordered if finished goods inventory (FGI) is not available. Every facility $c \in \{1, 2, \ldots, C\}$ produces a set of products and has a limited production capacity. Whereas facility C can draw from an infinite pool of inventory, production of facilities $(C-1, \ldots, 2, 1)$ is constrained by the production capacities of workstations and in addition by the work in process (WIP) in each facility. WIP is in turn constrained by (i) the FGI available upstream to release input into a facility that replenishes WIP and (ii) the production that depletes WIP. Again, for ease of exposition and without loss of generality, it is assumed that all facilities process the same set of part types, {1, 2, …, P}, and one part of part type p is required from an upstream facility $c+1$ to produce one unit of the same part type at facility c. The serial SC problem presented in Fig. 2 retains the most salient features of the more general problem, particularly in terms of the general demand and service distributions allowed. As suggested by past experience in the literature [36–38], results for the simpler system can be routinely generalized to accommodate assembly/disassembly features.

$D^p(t_k)$ denotes the amount of external orders for product of class p arriving during time slot k of period t. $\mu^p_{c,m}(t_k)$ denotes the part type p production capacity in isolation of facility c workstation m during time slot k of period t. $X^p_c(t_k)$ denotes the number of type p parts facility c produces during time slot k of period t. $R^p_c(t_k)$ is the amount of WIP released into facility c from facility $c+1$ FGI. Finally, $Q^p_c(t_k)$ denotes the type p WIP at facility c, and $I^p_c(t_k)$ denotes the type p FGI at facility c available at time slot k of period t. Only at facility 1, the FGI is allowed to take negative values to denote backordering. Following standard conventions, $\left(I^p_1(t_k)\right)^+$ and $\left(I^p_1(t_k)\right)^-$ denote, respectively, $\max\left\{0, I^p_1(t_k)\right\}$ and $\max\left\{0, -I^p_1(t_k)\right\}$.

Because the period containing K time slots is the relevant time scale in the planning layer's dynamics, and because it is assumed that it is long enough for the stochastic processes active at the time-slot scale to reach steady state, the following time-averaged variables are defined: $\bar{R}^p_c(t) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\left[R^p_c(t_k)\right]$, and similarly $\bar{X}^p_c(t)$, $\bar{Q}^p_c(t)$, $\bar{I}^p_c(t)$, $\bar{\mu}^p_{c,m}(t)$, and $\bar{D}^p(t)$. We use vector notation $\bar{\mathbf{X}}_c(t) = \left(\bar{X}^1_c(t), \ldots, \bar{X}^P_c(t)\right)$.

The SC management problem is implemented in three layers, exchanging information as shown in Fig. 3 and described below.

3.2. Performance evaluation and information layer

The performance evaluation and information layer shown in Fig. 3 models the short-term stochastic dynamics of production facilities at the operational level and develops the steady-state or time-averaged performance measure estimates of interest at the longer time scale of the planning layer. More specifically, performance evaluation means:

1. The transformation of production targets in each period to estimates of minimum average WIP required during that period in each facility to meet the production targets set by the planning layer. This estimate will generally depend on production targets, $\bar{\mathbf{X}}_c(t)$, the probability distribution of all relevant random variables $\mathbf{P}_c(t)$, and other operational policies, $\pi_c(t)$, during that period. The mapping of these inputs to the average WIP in facility c, $\bar{\mathbf{Q}}_c(t)$, is implicitly represented by function $\bar{g}^p_c\left(\bar{\mathbf{X}}_c(t), \mathbf{P}_c(t), \pi_c(t)\right)$.

2. The estimation of sensitivities (or derivatives) of $\bar{g}^p_c(\cdot)$ with respect to production targets. This is needed for tractable representation of the highly nonlinear relationship embodied in the $\bar{g}^p_c(\cdot)$ function.

3. The transformation of production targets, hedging inventory levels, $\mathbf{w}_c(t)$, and operational policies to the minimum average FGI required to meet the QoS constraint. The minimum average FGI requirements are represented by function $\bar{h}^p_c\left(\bar{\mathbf{X}}_c(t), \bar{\mathbf{X}}_{c-1}(t), \mathbf{P}_c(t), \mathbf{P}_{c-1}(t), \mathbf{w}_c(t), \pi_c(t)\right)$. For purposes of demonstrating the concept of dynamic lead times associated with dynamic QoS guarantee provisioning, a limited pairwise coupling of upstream and downstream facilities presented in Section 3.3 is considered here.

4. The estimation of sensitivities (or derivatives) of the function $\bar{h}^p_c(\cdot)$ with respect to production targets. Again, to serve the proof of concept objective, relatively simple analytic approaches are used for the determination of $\bar{h}^p_c(\cdot)$ and its sensitivity requirement (see Section 3.3).

5. A representation of an aggregate probabilistic model of facility c short-term capacity availability for use by the QoS layer.

Whereas this can be in general a Markov-modulated process (MMP) model, a simple, weighted bottleneck machine capacity exponential model (see Section 3.3) is used here, again to capture the correct factory physics and demonstrate the proof of concept in the planning layer algorithm. In practice, MMP models for production cells as well as for final demand can be estimated by analyzing possibly autocorrelated production and shipment transaction databases [39].

The estimates produced by the performance evaluation algorithm are merely intended to demonstrate qualitatively appropriate behavior, because the objective is to concentrate on an iterative planning layer. In real applications, the performance evaluation and information layer can be implemented using more accurate approaches in a distributed/decentralized manner where efficiency and robustness are important but not crucial.

Production target decisions are realizable at the desired QoS level only if the requisite WIP and FGI are available at various facilities in a manner consistent with material conservation dynamics. Functions $\bar{g}_c^p(\cdot)$ and $\bar{h}_c^p(\cdot)$ establish the minimum WIP and FGI constraints employed at the planning layer discussed in Section 3.4.

The evaluation of $\bar{g}_c^p(\cdot)$ and $\bar{h}_c^p(\cdot)$ functions is a formidable task. Analytic models have been used (mean value analysis [40] in this paper and the queuing network analyzer elsewhere [34]) to quantify them and introduce their nonlinear characteristics explicitly in an iterative decomposition planning algorithm for the purpose of investigating convergence properties of the multi-layer interactions depicted in Fig. 3. However, it is recognized that Monte Carlo simulation or complex analytical models involving Markovian or even more general stochastic decision processes may be relevant and more accurate in practice and may be indeed used in place of the more convenient models that were selected for the purpose of proof of concept, and without loss of generality since the qualitative behavior of the models is similar to that of the more accurate models that may be selected in practice. The Schweitzer–Bard Approximation [41,42] is used with our mean value analysis algorithm, which enables calculation of $\bar{g}(\cdot)$ values for real valued $\bar{\mathbf{X}}_c(t)$ and $\bar{\mathbf{Q}}_c(t)$ vectors. Similar fluid approximation enhancements as those used in the deterministic algorithms of the planning layer are also relevant in the context of stochastic models used at the decentralized layer [43,44]. These extensions are not trivial. For example, key events that are responsible for the efficiency of event-driven simulation algorithms (e.g., a buffer fills or a buffer empties) proliferate (a buffer fills or empties partially with multiple partial full/empty states) requiring more sophisticated models [45]. The important advantage of fluid production stochastic models (whether simulation based or analytic) is their ability to provide sensitivity estimates more tractably than finite differencing of stochastic discrete production models.

Finally, the convexity of the feasible regions defined by the $\bar{g}(\cdot)$ and $\bar{h}(\cdot)$ functions is crucial to the convergence of the planning layer. Fig. 4 shows a realistic example of the feasible region boundaries for a two-part type stochastic production network. More specifically, the maximum value of $\bar{X}_c^1(t)$ subject to $\bar{Q}_c^1(t) \geq \bar{g}_c^1\left(\bar{\mathbf{X}}_c(t), \mathbf{P}_c(t), \pi_c(t)\right)$ is plotted versus $\bar{Q}_c^1(t)$ and $\bar{X}_c^2(t)$.

Although the above constraints exhibit generally convex feasible regions, nonconvex feasible regions have been observed that arise when either operational policies are flagrantly suboptimal or facility designs are far from homogeneous (e.g., product classes impose diverse production time requirements on facility workstations) [46]. Consider Fig. 5 depicting mildly nonconvex and severely nonconvex feasible regions in contrast to the convex example in Fig. 4. Robust iterative master problem subproblem algorithms have been constructed that converge even under rather severe nonconvexity conditions [47].

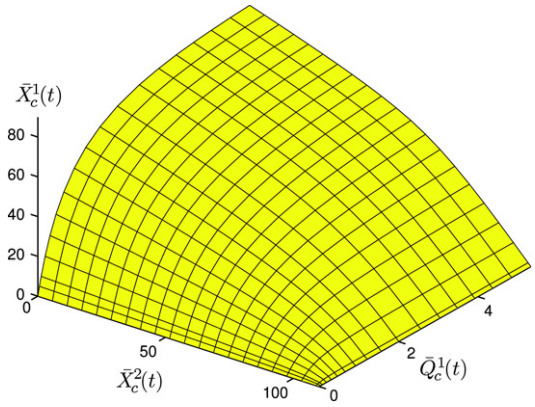

Fig. 4. Feasible production $\bar{X}_c^p(t)$ as a function of WIP $\bar{Q}_c^p(t)$.

3.3. Quality of service coordination layer

The QoS Layer interacts with other layers, as shown in Fig. 3. Its objective is to estimate a production policy that achieves the desired probabilistic QoS guarantees. Accurate modeling of QoS coordination policies is an important research problem in itself being monitored [48–50]. A number of methodologies have been used, including multi-class queuing network analysis (QNA), Monte Carlo simulation, stochastic system approximations, and large deviations asymptotics.

Given the stated objective to demonstrate the ability to construct a robust and efficient planning layer, a simple but certainly near-optimal [9] hedging policy is elected, which works as follows: Facility c produces at full capacity as long as the amount of work in its output buffers, $\tau_c^p(t)I_c^p(t)$, where $1/\tau_c^p(t)$ is the bottleneck capacity of facility c for part type p, is below the hedging inventory level expressed in units of work, $\mathbf{w}_c(t)$, set for week t by the QoS layer. The hedging inventory level is selected by the QoS layer so that the probability of a stockout of the downstream facility $c-1$ does not exceed the desired level $1-\Gamma_c(t)$. The idea is implemented using the following model:

1. The desired stockout probabilities, $1-\Gamma_c(t)$, at intermediate FGI positions and the final demand ($c = 0, 1, 2, \ldots, C$ and $t = 1, 2, \ldots, T$) are determined exogenously by the SC planner.
2. The hedging inventory level is estimated so as to achieve a maximally allowed stockout probability specified for facility $c-1$ as $1-\Gamma_c(t)$ under item 1 above by using the large deviations approach described in [49] and summarized in 5. This approach provides an efficient and accurate method for determining the parameters of a hedging point policy and the associated average inventory of semi-finished goods inventory in the output buffer of facility c as a function of the QoS required at facility $c-1$, and the first two moments of (i) the effective service time of facility c and (ii) the effective demand for material release into facility $c-1$. The associated average inventory in $\bar{I}_c^p(t)$ is then estimated by a $G/G/1$ approximation of the interaction of facilities c and $c-1$ where each multi-machine facility is approximated by a fictitious single machine with a general service time distribution. 5 describes the modeling of the hedging inventory requirements, including the special case of $c = 1$.
3. The simulation quantified the functional relationship between QoS and the coefficient of variation. The first two moments of the effective interrelease times of parts processed by each facility c and released into FGI c are estimated through simulation for a set of representative production targets and hedging points. To this end, multiple Monte Carlo simulation runs are employed as follows:

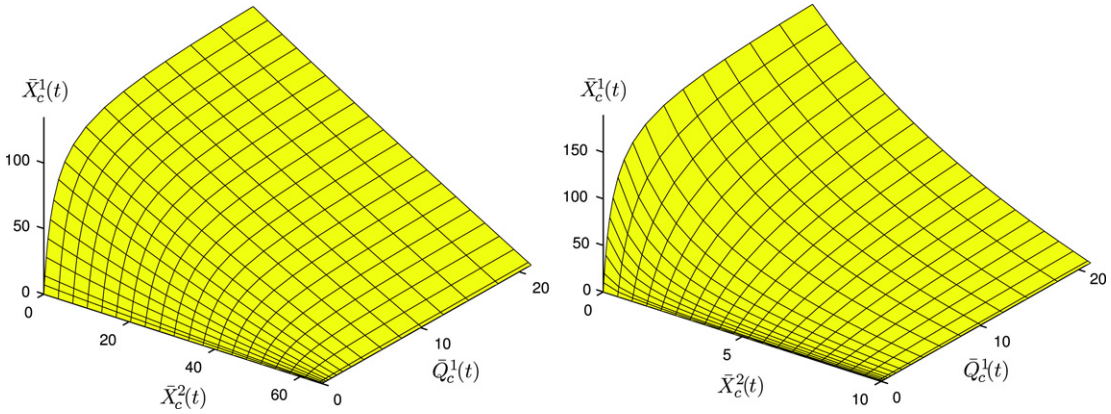

Fig. 5. Examples of mildly nonconvex and severely nonconvex feasible regions.

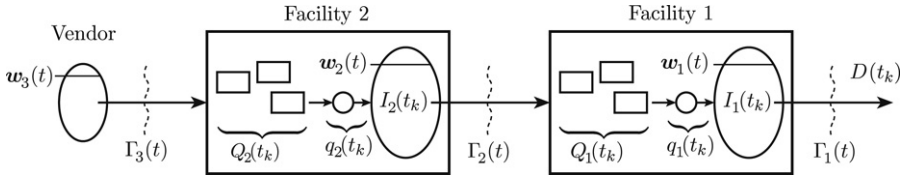

Fig. 6. Monte Carlo simulation of two-facility, one-part type SC.

(i) Selected production target values for each part type produced in facility c, $\bar{X}_c^p(t)$, are fixed as inputs. In each simulation run, the production targets across facilities are set equal to each other. The approximate mean value analysis (MVA) algorithm that was used to calculate $\bar{g}(\cdot)$ in Section 3.2 is employed here again to determine the required constant work-in-process (ConWIP) vector $\mathbf{K}_c(t)$ that guarantees facility c can produce in isolation at an average rate $\bar{\mathbf{X}}_c(t) = \left[\bar{X}_c^1(t), \bar{X}_c^2(t), \ldots, \bar{X}_c^P(t)\right]$.

(ii) The SC is simulated for a range of hedging point $\mathbf{w}_c(t)$ values that correspond roughly to QoS levels in the range of 80%–99%. Each facility is modeled as a fixed routing proportion queuing network with the material release protocol described below using as a key parameter the MVA-calculated ConWIP vector $\mathbf{K}_c(t)$.

(iii) The system of Fig. 6, where $q_c(t_k)$ is defined as a bin holding fully processed parts remaining inside the facility, is then simulated with the following material release policy described for simplicity for the special case of a one-part-type SC:

- if $Q_c(t_k) + q_c(t_k) < K_c(t) + \lceil \mathbf{w}_c(t)/\tau_c(t) \rceil - I_c(t_k)$, where $\lceil x \rceil$ is the smallest integer greater than x, then facility c absorbs material from $I_{c+1}(t_k)$ until there is equality in the expression above or until $I_{c+1}(t_k)$ empties. (Note: under this rule it is possible to temporarily accumulate parts inside facility c that exceed $\mathbf{K}_c(t)$)
- when a part's processing is completed in facility c, the facility
  - proceeds to increment $I_c(t_k)$ if $I_c(t_k) < \lceil \mathbf{w}_c(t)/\tau_c(t) \rceil$
  - or it proceeds to increment $q_c(t_k)$ if $I_c(t_k) = \lceil \mathbf{w}_c(t)/\tau_c(t) \rceil$ with contents $q_c(t_k)$ counted as part of WIP. Note that $q_c(t_k) = 0$ when $I_c(t_k) < \lceil \mathbf{w}_c(t)/\tau_c(t) \rceil$.

(iv) Interarrival times into $I_c(t_k)$ are sampled conditional on $I_c(t_k) < \lceil \mathbf{w}_c(t)/\tau_c(t) \rceil$ and used to estimate the conditional variance of effective service times of facility c denoted by $\sigma_c(t)$. Conditional variance values are stored in a table entry together with the value of exogenous input quantities used in that simulation.

Several simulations are performed, each corresponding to different values of exogenous inputs of facility production rate targets and hedging points. Each entry of the resulting table stores a value of the squared coefficient of variation of interrelease times into $I_c(t_k)$. Note that this is a function $\text{scv}_c\left(\Gamma_{c+1}(t), \bar{\mathbf{X}}_c(t), \mu_c(t), \mathbf{P}_c(t), \pi_c(t)\right)$ with the argument list showing the significant dependencies. For simplicity is written $\text{scv}_c(t) = \text{scv}_c\left(\Gamma_{c+1}(t), \rho_c(t)\right)$, where $\rho_c(t)$ represents the utilization level of facility c, defined more precisely later. The table is in effect a representation of the function $\text{scv}_c\left(\Gamma_{c+1}(t), \rho_c(t)\right)$, whose values can be interpolated from the table entries.

Simulation results verify the a priori expectation that $\text{scv}_c\left(\Gamma_{c+1}(t), \rho_c(t)\right)$ is increasing in $\bar{\mathbf{X}}_c(t)$ and decreasing in $\Gamma_{c+1}(t)$. Associating the largest index facility with the raw material vendor, a range of vendor hedging point values is simulated for the same hedging point values at the remaining facilities. Because each hedging point value at the vendor results in a different QoS level for the production facility that the vendor supplies, the impact of QoS on that facility can be calibrated. In general, the tabulated results were able to fit a smooth nonlinear function that represents the behavior of $\text{scv}_c(t)$, which is then used to model the nonlinear QoS constraints in the planning coordination layer described in Section 3.4. Fig. 7 graphs the $\text{scv}_c(t)$ function whose coefficients are estimated to fit the simulation table entries. Its algebraic representation is:

$$\begin{aligned}\text{scv}_c(t) = {} & -12.339\left(\rho_c(t)\right)^3 - 25.522\left(\Gamma_{c+1}(t)\right)^3 \\ & + 26.205\left(\rho_c(t)\right)^2\Gamma_{c+1}(t) - 19.602\rho_c(t)\left(\Gamma_{c+1}(t)\right)^2 \\ & + 85.276\left(\Gamma_{c+1}(t)\right)^2 + 1.722\rho_c(t)\Gamma_{c+1}(t) \\ & - 82.355\Gamma_{c+1}(t) + 27.425.\end{aligned}$$

For simplicity, and because the purpose of the paper is to show that optimal production planning optimization can account for nonlinear QoS constraints, the reasonable approximation is employed that the $\text{scv}_c(t)$ depends significantly only on $\Gamma_{c+1}(t)$ and $\rho_c(t)$, while dependence on the utilization or QoS of other facilities is negligible. This assumption is indeed supported by simulation experience.

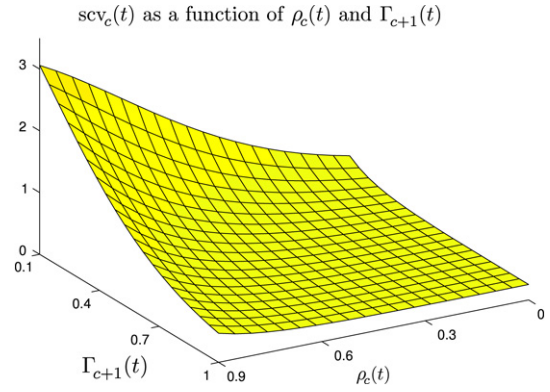

Fig. 7. Smooth interpolation of simulation results of hedging inventory replenishment process coefficient of variation.

4. The SC production planning algorithm described in Section 3.4 employs interpolation to determine variance values. The large deviations results of [49] are then used to determine $\mathbf{w}_c(t)$ values as a function of $\mathrm{scv}_c\,(\Gamma_{c+1}(t), \rho_c(t))$, $\mathrm{scv}_{c-1}\,(\Gamma_c(t), \rho_{c-1}(t))$, and $\rho_c(t)$. The $G/G/1$ approximation is finally employed to determine the required average inventory levels $\bar{I}_c^p(t)$.
5. Recall that the function $\bar{h}_c(\bar{\mathbf{X}}_c(t), \bar{\mathbf{X}}_{c-1}(t), \mathbf{P}_c(t), \mathbf{P}_{c-1}(t), \mathbf{w}_c(t), \pi_c(t))$ represents the minimum average FGI requirements that the planning layer must observe in order to model dynamic lead time and QoS requirements. This average must hold both during the fast time-slot scale that characterizes production capacity variations (for example, failures and repairs) as well as during the coarse time scale of the production planning function. The approach in [48] is adopted and the fast stochastic dynamics are modeled with a discrete Markov-modulated process model whose characteristic time scale is a day. The results in [48] are then used to obtain the hedging point required to obtain the desired QoS level as a multiple of the average daily production capacity. For a QoS level that is reasonably close to one, a good approximation for $\bar{h}_c(\cdot)$ is a quantity proportional to $\mathbf{w}_c(\cdot) - \bar{L}_c(t)$, where $\bar{L}_c(t)$ denotes the average queue length of a $G/G/1$ queuing system corresponding to a model of the reverse material flow relative to the actual system. In the reverse flow $G/G/1$ model, starvation is equivalent to blocking that occurs in the actual system when the inventory level reaches the hedging point level. Thus, the zero queue event corresponds to the event where the inventory equals the hedging point in the actual system. Furthermore, a strictly positive queue event in the $G/G/1$ model corresponds to an inventory level strictly below the hedging point in the actual system. The Krämer and Langenbach-Belz two-moment approximation is used to calculate $\bar{L}_c(t)$ as in [48]. This gives:

$$\bar{L}_c(t) = \frac{(\rho_c(t))^2\,(\mathrm{scv}_c(t) + \mathrm{scv}_{c-1}(t))}{2\,(1-\rho_c(t))} \times \exp\left\{\frac{-2\,(1-\rho_c(t))\,(1-\mathrm{scv}_{c-1}(t))}{3\rho_c(t)\,(\mathrm{scv}_c(t) + \mathrm{scv}_{c-1}(t))}\right\} + \rho_c(t)$$

if $\mathrm{scv}_{c-1}(t) \le 1$, otherwise

$$\bar{L}_c(t) = \frac{(\rho_c(t))^2\,(\mathrm{scv}_c(t) + \mathrm{scv}_{c-1}(t))}{2\,(1-\rho_c(t))} \times \exp\left\{\frac{-\,(1-\rho_c(t))\,(\mathrm{scv}_{c-1}(t) - 1)}{4\mathrm{scv}_c(t) + 4\mathrm{scv}_{c-1}\,(t)}\right\} + \rho_c(t)$$

where $\rho_c(t)$ is the weighted sum of utilization levels of the workstations in facility c caused by production targets $\bar{\mathbf{X}}_c(t)$. The weight used for each workstation equals the workstation's utilization normalized so that the weights sum to unity. The squared coefficient of variation of the interarrival times into $I_c(t_k)$ is denoted by $\mathrm{scv}_c(t)$.

Using the large deviations results of [48], the hedging point is calculated as:

$$\mathbf{w}_c(t) = -\frac{1}{\theta_c^*\,(t)} \log\left(\frac{1-\Gamma_c(t)}{\theta_c^*(t)\bar{L}_c(t)}\right)$$

where $\theta_c^*(t)$ is the solution of Eq. (11) in [48], assuming that the demand and service processes in the $G/G/1$ model are sampled at the beginning of each time slot from a normal distribution where the mean and variance values of the effective processing times are $\left(\rho_c(t);\, \mathrm{scv}_c(t)\,(\rho_c(t))^2\right)$ and $(1;\, \mathrm{scv}_{c-1}(t))$, respectively. In general, autocorrelations may be modeled and a numerical solution can be easily obtained as described in [48]. This paper uses the simple case of a single-state Markov-modulated model where the solution can be written explicitly as:

$$\theta_c^*(t) = \frac{2\,(1-\rho_c\,(t))}{\mathrm{scv}_c(t) + \mathrm{scv}_{c-1}\,(t)\,(\rho_c(t))^2}.$$

Note that the variance values reflect the stochastic dynamics at facilities c and $c-1$ that are compatible with the planned production targets and the specified QoS levels. Average utilization rates correspond to production targets in the optimized production schedule and are expressed in terms of work assigned to facility c and the production capacity of facility c.

In case of multiple part types, the minimum average FGI requirement calculated above is divided among part types proportional to the workload each imposes on facility c.

The hedging inventory policy described above achieves the specified QoS levels between facilities and decouples the implementation of operational decisions across facilities. A single fictitious part type aggregation is employed, and virtual workstations that behave similarly in each SC facility are utilized. The aggregation of part types to a single fictitious part is done so as to preserve the probabilistic behavior of work needed to produce all individual part types. Therefore, QoS, and hence stockout probability, is defined in terms of the probability that the work incorporated in the intermediate FGI of facility c falls short of the work incorporated in the part-type quantities required during the relevant time period by downstream facility $c-1$. In that sense, the hedging point quantity is a scalar that is not part-type specific and is defined in terms of work corresponding to the quantity of the fictitious part that is sufficient to achieve the desired stockout probability threshold for the fictitious part. It should be noted that the proposed supply chain production planning algorithm does not provide part-type-specific hedging FGI quantities needed for production floor implementation. Part-type-specific quantities are approximated for holding-cost accounting purposes by disaggregating the fictitious part work-hedging quantity to part-type-specific work in proportion to the work rate incorporated in each part's production target rate relative to the total, and then converting the work parts. However, it is not proposed that these part-type-specific quantities be used on the production floor. Among others, the reason is that the material flow protocol is not fully defined because a dynamic policy allocating production capacity among specific part types in real time would be required under these circumstances. It is clear that not any such policy, as for example a priority policy, will achieve the desired QoS for each part type. Such a policy is not specified, as it is considered to be out of the scope of this paper. Instead, what is used is the reasonable assumption that an appropriate policy that achieves part-type-specific QoS guarantees does exist. Furthermore, it is conjectured that, given the aggregate fictitious part-hedging quantity that provides the requisite excess capacity, an adequate real-time policy can be easily devised on the production floor.

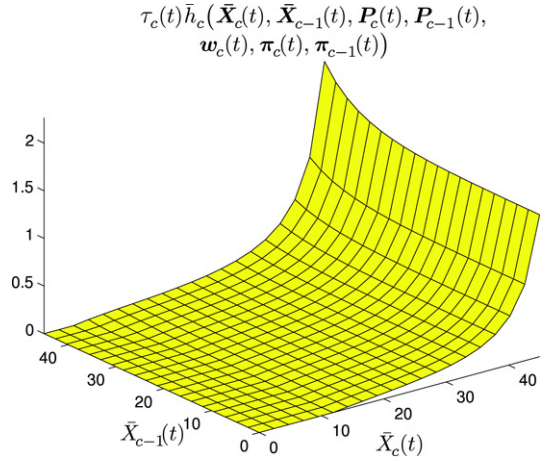

Fig. 8. Horizontal coordination constraint function scaled by $\tau_c(t)$, $\tau_c(t)\bar{h}_c(\cdot)$, for a single-part-type system.

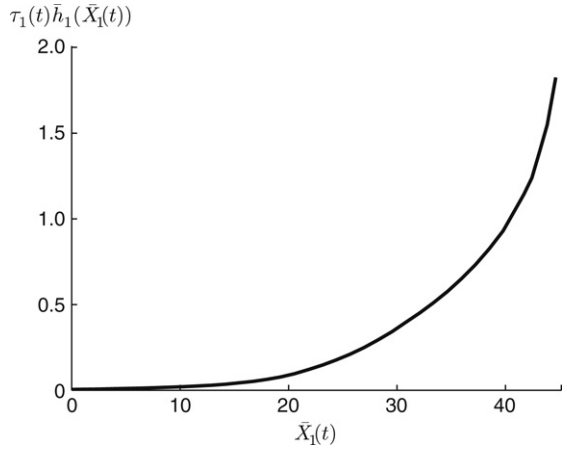

Fig. 9. FGI hedging constraint function scaled by $\tau_c^p(t)$, $\tau_1(t)\bar{h}_1(\cdot)$, for a single-part-type system.

The large deviation estimates of the required hedging point and average queue given above characterize the FGI required to meet the desired quality of service, assuming that the hedging point material flow policy is implemented in real time and is based on the single-part variability of production and demand. However, in practice, the hedging point material flow policy is not implemented in real time but over a longer period that lies between the hourly time scale of production and demand events and the weekly planning period time scale. This reduces the coefficient of variation of total production and demand over the longer material flow implementation period. Hence, the minimum FGI required to meet the desired quality of service over the longer material flow period is smaller. To model this issue, the real-time material flow policy FGI is divided by a factor of d that must be larger than or equal to one. For the purposes of the numerical examples, and to avoid the tedious work of estimating explicitly FGI under a reduced-variability assumption, $d = 6$ is used, which is approximately the number of working days per week. Thus, $\bar{h}_c(\cdot) = \frac{1}{d}\left(\mathbf{w}_c(t) - \bar{L}_c(t)\right)$. Finally, distributing this aggregate quantity among part types using weights $\tau_c^p(t)\bar{X}_c^p(t)/\sum_{r=1}^{P}\tau_c^r(t)\bar{X}_c^r(t)$ that sum up to one and converting from work units back to part units by multiplying with $1/\tau_c^p(t)$, the minimum average FGI requirements function for part type p is obtained:

$$\bar{h}_c^p\left(\bar{\mathbf{X}}_c(t), \bar{\mathbf{X}}_{c-1}(t), \mathbf{P}_c(t), \mathbf{P}_{c-1}(t), \mathbf{w}_c(t), \pi_c(t)\right) = \frac{1}{\tau_c^p(t)} \frac{\tau_c^p(t)\bar{X}_c^p(t)}{\sum_{r=1}^{P}\tau_c^r(t)\bar{X}_c^r(t)} \frac{1}{d}\left(\mathbf{w}_c(t) - \bar{L}_c(t)\right).$$

As mentioned, the coefficients $\tau_c^p(t)$ convert each unit of FGI, $\bar{I}_c^p(t)$, to units of work defined as the processing time needed at the upstream facility to produce the required amount of inventory. $\tau_c^p(t)$ is used to equal the reciprocal of the minimum production capacity for part type p of all machines in facility c, $\tau_c^p(t) = \left[\min_{\forall m \in M_c}\bar{\mu}_{c,m}^p(t)\right]^{-1}$. Fig. 8 shows an instance of the FGI constraint surface corresponding to upstream and downstream facilities, each consisting of three identical machines with production capacity of 50 parts per time period and $\Gamma_{c+1}(t) = \Gamma_c(t) = 96\%$.

The hedging inventory requirements of facility $c = 1$ is calculated by treating $\mathrm{scv}_0(t)$, the squared coefficient of variation of interarrival times of the demand for final products at the first facility, as an exogenous parameter. The QoS provided to the customers, $\Gamma_1(t)$, is also determined by the user. The required hedging inventory at $c = 1$, $\bar{h}_1^p\left(\bar{\mathbf{X}}_1(t), \bar{D}^p(t), \mathbf{P}_1(t), \mathbf{P}_0(t), \mathbf{w}_1(t), \pi_1(t)\right)$, is therefore a function of only one endogenous variable, the production target vector of the first facility. An instance of the FGI constraint surface corresponding to facility $c = 1$ comprising three identical machines with production capacity of 50 parts per time period, $\Gamma_2(t) = 96\%$, $\Gamma_1(t) = 95\%$, and $\mathrm{scv}_0(t) = 0.6$, is plotted in Fig. 9.

3.4. Planning coordination layer

The planning layer shown in Fig. 3, and its role in the collaborative framework where the master and subproblem layers interact in an iterative algorithm that produces the optimal production plan, are described next.

3.4.1. Master problem optimization algorithm

Given the longer (weekly) time scale of the planning layer, a linear programming (LP) based fluid model approximation of the discrete part production planning problem is used. Moreover, the fluid model is extended to represent WIP and FGI-driven dynamic lead times through the nonlinear constraint surfaces defined in Sections 3.2 and 3.3. These constraints are key components of the planning layer model (see inequalities (6)–(8) in Exhibit 1).

Exhibit 1 (*Planning Layer Optimization Problem*).

$$\min_{\bar{X}_c^p(t), \bar{R}_c^p(t)} \sum_{p,t}\left[\sum_{c=1}^{C}\bar{\alpha}_c^p\bar{Q}_c^p(t) + \sum_{c=2}^{C}\bar{\beta}_c^p\bar{I}_c^p(t) + \beta_1^p\left(I_1^p(t)\right)^+ + \gamma_1^p\left(I_1^p(t)\right)^- + \delta_1^p\left(J_1^p(t)\right)^-\right] \quad (1)$$

subject to:

$$Q_c^p(t+1) = Q_c^p(t) + \bar{R}_c^p(t) - \bar{X}_c^p(t) \quad \forall p, c, t \quad (2)$$

$$I_c^p(t+1) = I_c^p(t) + \bar{X}_c^p(t) - \bar{R}_{c-1}^p(t) \quad \forall p, c \geq 2, t \quad (3)$$

$$I_1^p(t+1) = I_1^p(t) + \bar{X}_1^p(t) - \bar{D}^p(t) \quad \forall p, c = 1, t \quad (4)$$

$$\sum_{p=1}^{P}\frac{\bar{X}_c^p(t)}{\bar{\mu}_{c,m}^p(t)} \leq \eta_c(t) \quad \forall m, c, t \quad (5)$$

$$\bar{Q}_c^p(t) \geq \bar{g}_c^p\left(\bar{\mathbf{X}}_c(t), \mathbf{P}_c(t), \pi_c(t)\right) \quad \forall p, c, t \quad (6)$$

$$\bar{I}_c^p(t) \geq \bar{h}_c^p\left(\bar{\mathbf{X}}_c(t), \bar{\mathbf{X}}_{c-1}(t), \mathbf{P}_c(t), \mathbf{P}_{c-1}(t), \mathbf{w}_c(t), \pi_c(t), \pi_{c-1}(t)\right) \quad \forall p, c \geq 2, t \quad (7)$$

$$I_1^p(t) - J_1^p(t) \geq \bar{h}_1^p\left(\bar{\mathbf{X}}_1(t), \bar{D}^p(t), \mathbf{P}_1(t), \mathbf{P}_0(t), \mathbf{w}_1(t), \pi_1(t)\right) \quad \forall p, c = 1, t. \quad (8)$$

The planning layer's objective is to determine targets of weekly production and release that minimize weighted WIP, FGI, and backlog costs over the horizon T, subject not only to the capacity and material conservation constraints, but also the nonlinear constraints on weekly average WIP and hedging inventory that capture the essence of faster time-scale stochastic dynamics driven by QoS provisioning and other operational policies that may be in place at some facilities.

The planning layer solves the optimization problem given in Exhibit 1 where minimization takes place over weekly production target, $\bar{X}_c^p(t)$, and release, $\bar{R}_c^p(t)$, decision variables. For all facilities, $\bar{\alpha}_c^p$ and $\bar{\beta}_c^p$ denote, respectively, WIP and FGI holding cost rates, while β_1^p, γ_1^p, and δ_1^p denote respectively, holding, shortage, and hedging constraint penalty cost rates[1] at $c = 1$ applied to the end-of-period values of FGI. Average WIP and FGI values at facilities $c > 1$ are related to the end-of-period values via the approximations $\bar{Q}_c^p(t) = 0.5Q_c^p(t-1) + 0.5Q_c^p(t)$ and $\bar{I}_c^p(t) = 0.5Q_c^p(t-1) + 0.5Q_c^p(t)$. $\eta_c(t)$ is the maximum allowed utilization level for workstations in facility c during time period t. The optimization problem of Exhibit 1 is subject to the usual positivity (with the exception of $I_1^p(t)$ and $J_1^p(t)$ being unrestricted $\forall p, t$), initial condition, capacity (5), material conservation (2)–(4), and nonlinear constraints (6)–(8) that capture lead time and QoS dynamics. $\mathbf{P}_c(t)$ and $\pi_c(t)$ are as defined in Section 3.2. Vector $[\mathbf{w}_c(t)\ c = 1, 2, \ldots, C]$ is determined at the QoS layer. It denotes the hedging inventory level implemented by facility c to control its production schedule so as to prevent the downstream stockout frequency from violating the desired probabilistic QoS constraint. Notice that the hedging inventory level is strictly enforced on $\bar{I}_c^p(t)$ in constraint (7), whereas constraint (8) is a soft constraint. Violation of the hedging inventory requirement on $I_1^p(t)$ is defined as $J_1^p(t) = I_1^p(t) - \bar{h}_1^p(\cdot)$, and its negative part, $\left(J_1^p(t)\right)^-$, is penalized in the objective function at a rate of δ_1^p. This allows partitioning the FGI cost into three parts: (i) $I_1^p(t) \geq \bar{h}_1^p(\cdot)$, with a cost rate of β_1^p, (ii) $0 \leq I_1^p(t) \leq \bar{h}_1^p(\cdot)$, with a cost rate of δ_1^p for each unit of negative deviation from $\bar{h}_1^p(\cdot)$, and (iii) $I_1^p(t) \leq 0$, with a cost rate of $\gamma_1^p + \delta_1^p$ (because both $\left(I_1^p(t)\right)^-$ and $\left(J_1^p(t)\right)^-$ are positive). This constraint is defined on the FGI end-of-period values, $I_1^p(t)$, rather than the average time period values, $\bar{I}_1^p(t)$. The justification of this modeling choice is that it constitutes a reasonable simplification (indeed, a conservative cost estimate) given the nonlinear nature of the inventory/backlog cost trajectory).

3.4.2. Iterative algorithm for layer collaboration and coordination

An iterative single master problem (centralized planning coordination layer) multiple subproblem (decentralized performance and information layer) algorithm has been developed to model the nonlinear constraints and derive optimal production plans that explicitly account for variable lead time and QoS constraints (see Fig. 3). The efficient representation of constraints (6) and (7) in Exhibit 1 requires point estimates and sensitivity estimates so as to approximate the nonlinear constraint boundaries. In fact, the iterative fine-tuning of a finite number of appropriately selected local approximations leads to convergence under mild convexity or quasi-convexity conditions [46]. This iterative algorithm is summarized below and also shown in Fig. 3:

1. The planning layer, at iteration n, calculates tentative production and material release schedules ${}^n\bar{X}_c^p(t)$, ${}^n\bar{R}_c^p(t) \forall p, c, t$, and conveys them to facility subproblems along with the WIP and FGI levels ${}^n\bar{Q}_c^p(t)$, ${}^n\bar{I}_c^p(t)$, that result from the production schedule through the material flow constraints (2)–(4) in Exhibit 1.

2. For the tentative production target assigned to it by the planning layer and the QoS it receives from the upstream portion of the SC, each facility conveys to the quality of service layer a Markov-modulated model that represents the stochastic behavior of its production capacity. For the same tentative production targets, the quality of service layer proceeds to calculate tentative hedging points $\mathbf{w}_c(t)$ that each facility needs to employ to provide the required QoS to its downstream facility.

3. Each facility evaluates its performance and calculates its average WIP as well as the WIP's sensitivity with respect to the facility's tentative production target. For each pair of upstream and downstream facilities, the pair's performance is also evaluated to calculate the average FGI needed to achieve the QoS requirements as well as the FGI's sensitivity with respect to the tentative production targets of both facilities. The planning layer receives the following feedback: (i) the required WIP and FGI $\left(\bar{g}_c^p(\cdot), \bar{h}_c^p(\cdot)\right)$, and (ii) their sensitivities, namely $\nabla\bar{g}_c^p(\cdot)$ and $\nabla\bar{h}_c^p(\cdot)$ w.r.t. the production targets.

4. The planning layer's linear constraint set is augmented using the hyperplanes tangent to functions $\bar{g}_c^p(\cdot)$ and $\bar{h}_c^p(\cdot)$ at the most recent iteration's production target values ${}^n\bar{\mathbf{X}}_c(t)$. More specifically, the additional constraints are:

$$\bar{Q}_c^p(t) \geq \bar{g}_c^r\left({}^n\bar{\mathbf{X}}_c(t)\right) + \nabla\bar{g}_c^r\left({}^n\bar{\mathbf{X}}_c(t)\right)'\left[\bar{\mathbf{X}}_c(t) - {}^n\bar{\mathbf{X}}_c(t)\right] \quad \forall p, c, t \tag{9}$$

$$\bar{I}_c^p(t) \geq \bar{h}_c^r\left({}^n\bar{\mathbf{X}}_c(t), {}^n\bar{\mathbf{X}}_{c-1}(t)\right) + \nabla\bar{h}_c^r\left({}^n\bar{\mathbf{X}}_c(t), {}^n\bar{\mathbf{X}}_{c-1}(t)\right)' \times \begin{bmatrix} \bar{\mathbf{X}}_c(t) - {}^n\bar{\mathbf{X}}_c(t) \\ \bar{\mathbf{X}}_{c-1}(t) - {}^n\bar{\mathbf{X}}_{c-1}(t) \end{bmatrix} \quad \forall p, c \geq 2, t \tag{10}$$

$$I_1^r(t) - J_1^r(t) \geq \bar{h}_1^r\left({}^n\bar{\mathbf{X}}_1(t)\right) + \nabla\bar{h}_1^r\left({}^n\bar{\mathbf{X}}_1(t)\right)'\left[\bar{\mathbf{X}}_1(t) - {}^n\bar{\mathbf{X}}_1(t)\right] \quad \forall p, c = 1, t. \tag{11}$$

Note that this set of hyperplanes (linear constraints) (9)–(11), are generated at iteration n to approximate the nonlinear WIP constraints (6) and FGI constraints (7) and (8), respectively, around the point ${}^n\bar{\mathbf{X}}_c(t)$. Note also that the probability distributions $\mathbf{P}_c(t)$, operational $\pi_c(t)$, and the hedging points $\mathbf{w}_c(t)$ are omitted for notational simplicity.

5. Using the tangent hyperplanes (9)–(11) accumulated over past iterations 1 through n and the master problem constraints (2)–(5), the planning layer solves again the master problem as a linear program to produce a new set of tentative targets, ${}^{n+1}\bar{X}_c^p(t)$, ${}^{n+1}\bar{R}_c^p(t) \forall p, c, t$. Iterations continue to convergence defined by a tolerance with respect to which the nonlinear constraints are satisfied.

Fig. 10 depicts this iterative process for iterations n and $n+1$ on a single-part-type WIP constraint. Tentative planning layer solution $\left[\left({}^n\bar{Q}_c^p(t), {}^n\bar{X}_c^p(t)\right)\right]$ is associated with the nominal point $\left[\bar{g}_c^p\left({}^n\bar{\mathbf{X}}_c(t)\right), {}^n\bar{X}_c^p(t)\right]$ on the surface of the nonlinear constraint where the tangent hyperplane constraint is generated. The next iteration's tentative solution, $\left[{}^{n+1}\bar{Q}_c^p(t), {}^{n+1}\bar{X}_c^p(t)\right]$, satisfies all constraints added so far. A new tangent hyperplane constraint is generated after iteration $n+1$ at the point $\left[\bar{g}_c^p\left({}^{n+1}\bar{\mathbf{X}}_c(t)\right), {}^{n+1}\bar{X}_c^p(t)\right]$.

Note that the approximation accuracy of the nonlinear constraint surfaces increases monotonically with the addition of tangent hyperplanes. As the algorithm approaches convergence, tentative solutions are close to each other, resulting in the addition of an increasing number of tangent hyperplane constraints in the vicinity of the optimal solution, rendering the linear approximation error of the nonlinear constraints arbitrarily small. This process of outer linearization of the nonlinear constraints is a *smart* linearization because it is not a uniform linearization that is, in general,

[1] Note that penalizing for back orders at the planning layer does not contradict the probabilistic QoS guarantees provided by the hedging point policy implemented through production policy thresholds $\mathbf{w}_c(t)$. The planning layer's constraints (7) and (8) assure that weekly production targets are consistent with probabilistic QoS requirements.

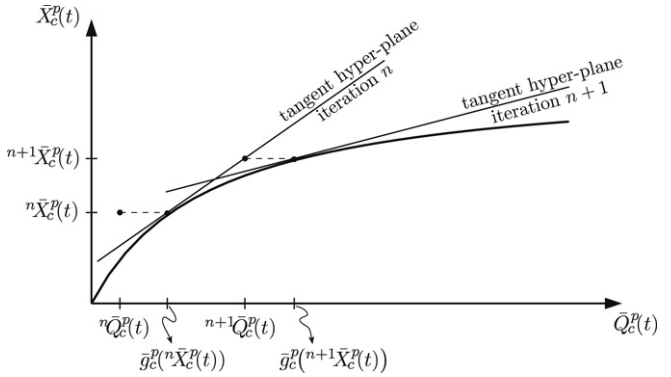

Fig. 10. Tentative target $\left({}^{n}\bar{Q}_c^p(t)\,,\,{}^{n}\bar{X}_c^p(t)\right)$ and the associated tangent hyperplane depicted on a single-part-type WIP constraint.

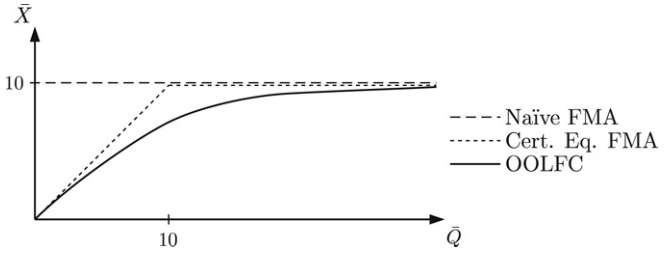

Fig. 11. Feasible production as a function of WIP with various deterministic fluid and stochastic model approximations.

computationally intractable, but is progressively dense and hence accurate where it matters, that is, in the vicinity of the optimal solution.

An example to introduce the basic idea of lead time dynamics at the coarse time scale of the planning layer and its significance in yielding a superior production plan is employed next. Consider a simple manufacturing facility consisting of a 10-workstation transfer line. Each workstation has stochastic processing time and average production capacity of 10 parts per week. This is also the capacity of the facility.

Fig. 11 shows the average weekly facility production rate, $\bar{X}$, that is achievable as a function of the average available WIP, denoted by $\bar{Q}$, under three successively more accurate fluid model approximations (FMAs). The *naive* FMA does not impose any average WIP constraints, and the production rate is simply constrained by the average production capacity. The *certainty equivalent* FMA considers deterministic workstation capacities equal to the average capacity availability. As a result, it disregards queuing delays. It underestimates the average WIP needed for sustaining a certain average production rate, setting it equal to the sum of the 10 machine utilizations. This is a linear constraint involving production and WIP. Finally, the *optimal open loop feedback controller* (OOLFC) employs a stochastic model that captures the fact that effective workstation processing times are random variables. This results in the correct estimate of average WIP, and hence of average lead time, which reflects the average queue levels in front of workstations in addition to the average number of parts being processed.

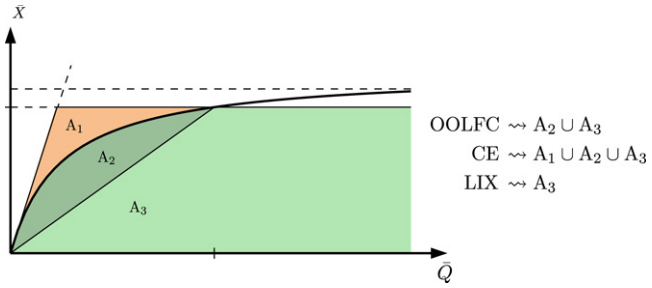

Fig. 13. OOLFC, CE, and LIX.

In the context of the SC planning coordination layer, the OOLFC model approximation is implemented via the inclusion of nonlinear constraints on production $\bar{X}_c^p(t)$ represented by the WIP and FGI functions, $\bar{g}_c^p(\cdot)$ and $\bar{h}_c^p(\cdot)$, introduced in Sections 3.2 and 3.3. Several approaches have been successfully implemented to handle these constraints, so far based on iterative tangent or piecewise-linear approximations [34] that allow LP optimization. When used in a manner of an open-loop control, errors introduced via shortcomings of modeling choices in estimating these functions, $\bar{g}_c^p(\cdot)$ and $\bar{h}_c^p(\cdot)$, will propagate over the planning horizon. The magnitude of errors depend on: (i) accuracy of the model used in estimation, and (ii) the time-scale separation required for steady-state convergence [35]. The propagation and impact can be evaluated through simulation. Use of the rolling horizon approach requires implementation of only the first-period results, thus mitigating the impact of error propagation. Although the preliminary evaluations show that the impact is relatively small, additional work in this area would be useful [33].

The nonlinear constraints render the production plans generated not only feasible but also superior to plans obtained by alternative deterministic approaches based on certainty equivalent formulations [20,51,46]. The introduction of the nonlinear constraints is equivalent to formulating and solving the open-loop optimal controller known to dominate the certainty equivalent controller [52]. The certainty equivalent controller, possibly adjusted to model a larger, worst-case, constant lead time delay in each activity, represents today's industry practice, and, as such, it is compared to the proposed open-loop optimal controller in Section 4.

4. Algorithmic implementation and numerical evaluation of benefits

This section discusses the computational experience of the proposed dynamic lead time and QoS hedging inventory modeling algorithm, referred to as the optimal open loop feedback controller (OOLFC). OOLFC performance is compared to two algorithms that represent industry practice to explore the value that one may attribute to the dynamic lead time and QoS hedging inventory information employed by the OOLFC approach. Whereas production planning algorithms have been previously reported that model explicitly dynamic lead times [34,53,51,46,54,55], results reflecting probabilistic QoS policies are presented here for the first time.

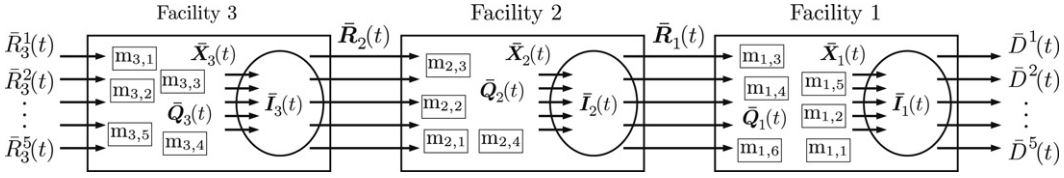

Fig. 12. An example supply chain with three facilities producing five part types.

Table 2
Average workstation production capacities

	Facility 3					Facility 2				Facility 1					
	$m_{3,1}$	$m_{3,2}$	$m_{3,3}$	$m_{3,4}$	$m_{3,5}$	$m_{2,1}$	$m_{2,2}$	$m_{2,3}$	$m_{2,4}$	$m_{1,1}$	$m_{1,2}$	$m_{1,3}$	$m_{1,4}$	$m_{1,5}$	$m_{1,6}$
p.t. 1	70	70	55	55	55	55	55	55	55	40	35	40	40	35	35
p.t. 2	75	65	60	70	55	55	55	65	55	40	50	40	35	50	50
p.t. 3	75	65	75	75	80	55	60	55	65	60	50	55	50	60	55
p.t. 4	85	70	90	80	65	65	55	70	70	55	65	65	55	55	65
p.t. 5	70	85	80	75	90	70	70	65	65	65	65	70	70	65	70

Table 3
WIP, FGI holding, and FGI backlog cost coefficients

	Facility 3		Facility 2		Facility 1			
	WIP $\bar{\alpha}_3^p$	FGI$^+$ $\bar{\beta}_3^p$	WIP $\bar{\alpha}_2^p$	FGI$^+$ $\bar{\beta}_2^p$	WIP $\bar{\alpha}_1^p$	FGI$^+$ $\bar{\beta}_1^p$	FGI$^-$ γ_1^p	FGH$^-$ δ_1^p
p.t. 1	20	27	30	55	60	150	1200	1800
p.t. 2	15	20	25	40	50	110	880	1320
p.t. 3	10	12	15	40	50	120	960	1440
p.t. 4	10	23	30	42	45	100	800	1200
p.t. 5	5	8	10	26	30	70	560	840

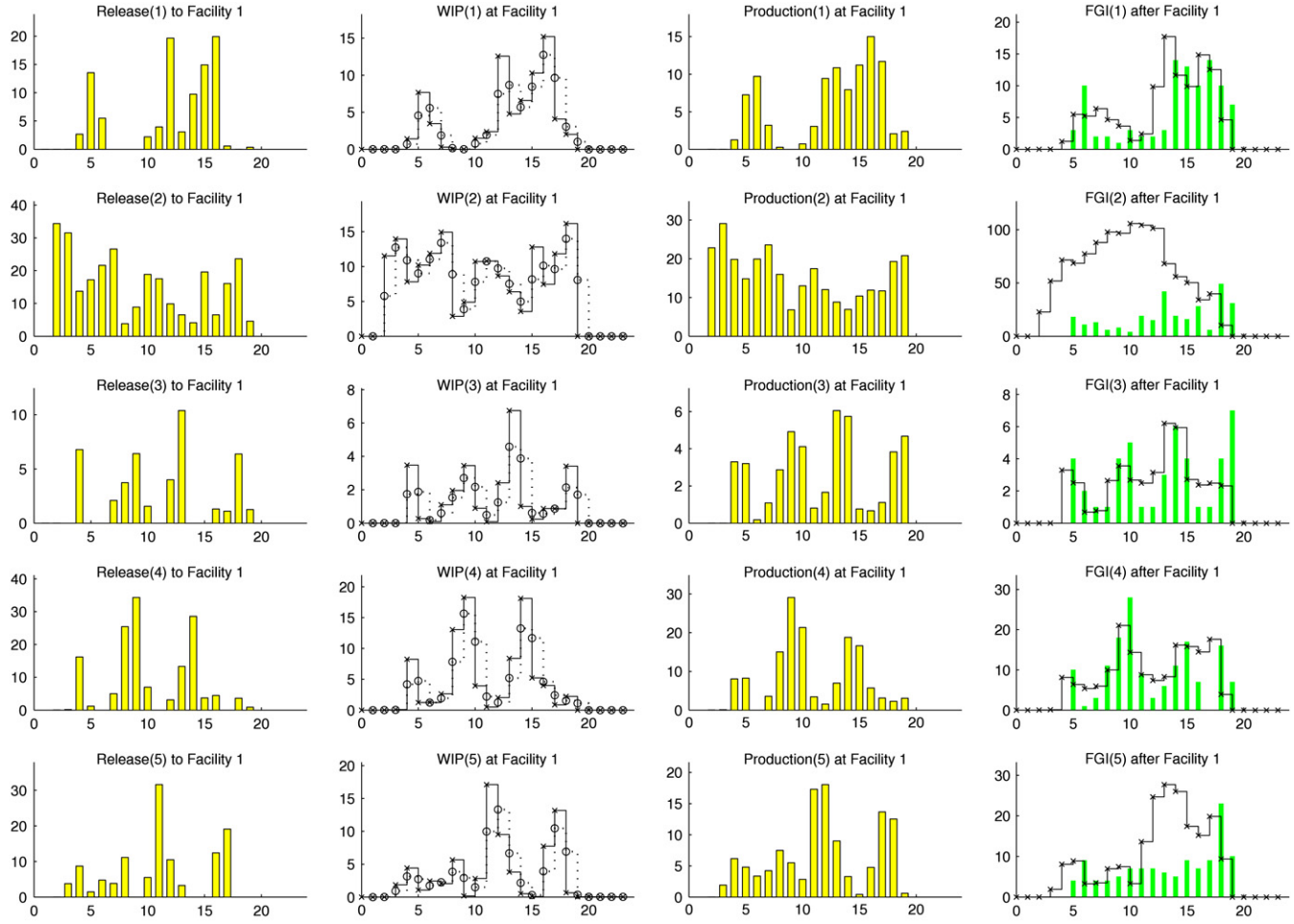

Fig. 14. Facility 1 optimal solution details of OOLFC under demand scenario 4.

Presented below is numerical experience from six demand scenarios applied to a five-part type 3 facility SC depicted in Fig. 12. Facilities 3, 2, and 1 having five, four, and six workstations, respectively, with average production capacities, are shown in Table 2, with part types corresponding to columns and workstations to rows. Facilities are allowed a maximum workstation utilization of $\eta_c(t) = 0.9 \ \forall c, t$. QoS levels are set at $\Gamma_\cdot(t) = \begin{bmatrix} 95\% & 97\% & 97\% & 97\% \end{bmatrix}$ for facility indices 1 through 4. $\Gamma_4(t)$ is the QoS provided to facility 3 by the raw material vendor. The demand-squared coefficient of variation is set to 0.6 throughout the planning horizon. Initial levels of WIP and FGI are set to zero in all facilities. Cost coefficients are shown in Table 3.

All demand scenarios consist of a 23-week planning horizon. The first four weeks in scenarios 1 and 2, the first two weeks in scenarios 3, 4, 5, and 6, and the last four weeks in all scenarios were on purpose assigned zero demand to allow the algorithm to fill and empty the system optimally without disadvantaging the suboptimal industry practice alternative planning algorithms

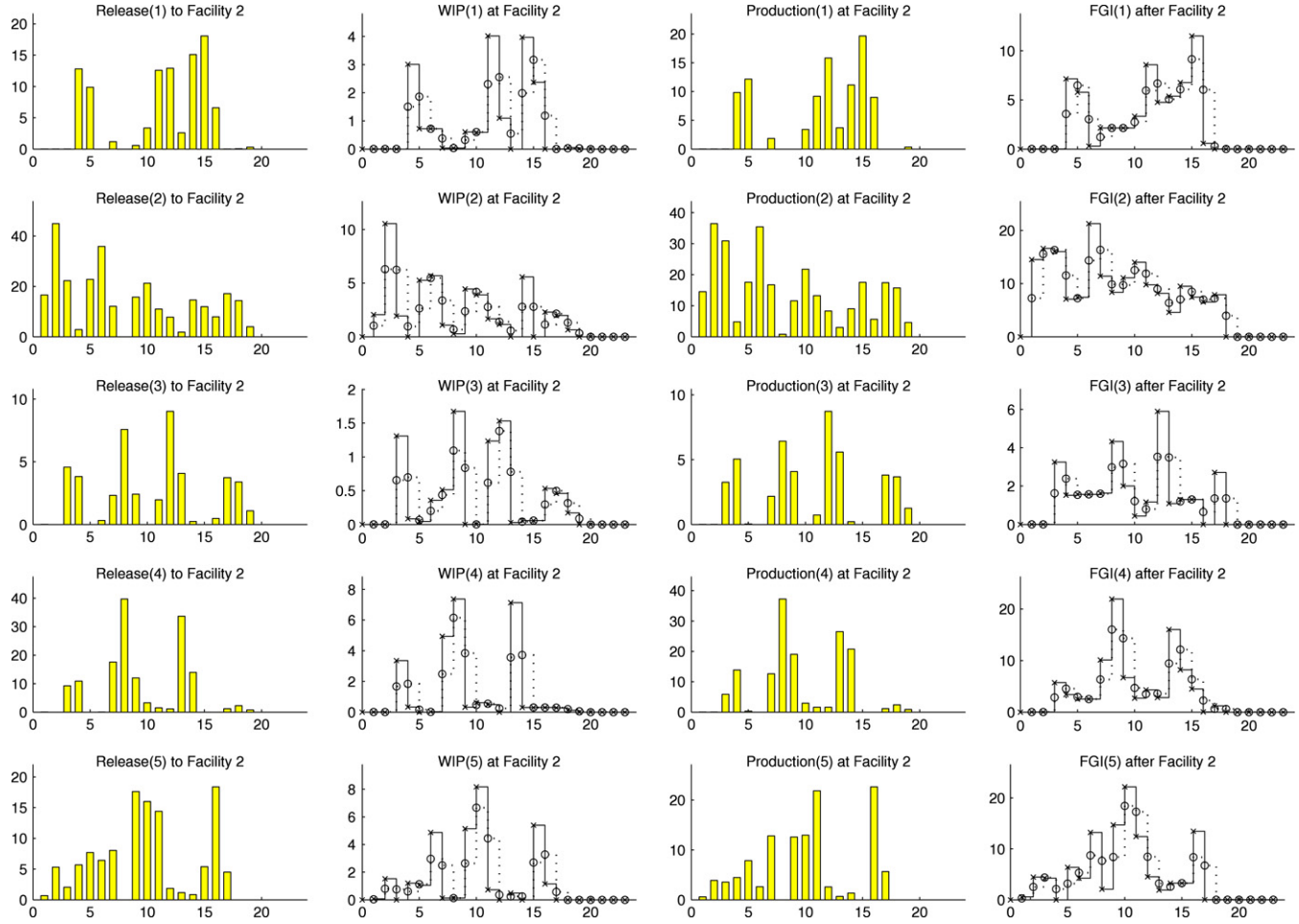

Fig. 15. Facility 2 optimal solution details of OOLFC under demand scenario 4.

Table 4
Weekly demand for scenarios 3, 4, 5, and 6; weeks 3 through 19

		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Demand scenario 3	p.t. 1	3	9	4	1	1	1	1	2	2	2	2	7	11	10	8	10	10
	p.t. 2	21	15	10	8	5	6	4	8	14	12	29	23	17	13	21	11	19
	p.t. 3	4	3	1	1	1	3	3	3	1	1	2	4	5	3	1	1	3
	p.t. 4	11	3	1	4	9	14	15	15	6	2	4	7	16	14	7	2	7
	p.t. 5	5	9	5	4	4	4	4	6	6	6	5	4	5	7	6	6	11
Demand scenario 4	p.t. 1	0	0	9	30	6	6	3	9	6	6	9	42	39	30	42	30	21
	p.t. 2	0	0	18	11	13	6	8	4	19	15	42	19	16	28	6	49	31
	p.t. 3	0	0	4	2	1	1	4	5	1	1	3	6	4	1	1	4	7
	p.t. 4	0	0	10	1	3	11	18	28	9	3	6	11	17	7	0	16	7
	p.t. 5	0	0	4	9	4	4	5	7	7	7	6	5	9	7	9	23	10
Demand scenario 5	p.t. 1	0	4	8	4	2	1	1	2	2	2	3	10	14	10	10	11	9
	p.t. 2	0	23	11	11	8	6	5	7	16	13	34	22	16	19	17	14	39
	p.t. 3	0	4	2	1	1	2	3	4	1	1	2	5	5	3	1	1	4
	p.t. 4	0	12	1	2	7	13	18	21	7	3	5	9	16	11	3	3	13
	p.t. 5	0	5	8	5	4	4	4	6	6	6	5	5	8	8	8	10	18
Demand scenario 6	p.t. 1	0	0	31	21	7	9	35	29	6	4	14	31	13	4	4	13	21
	p.t. 2	0	0	9	5	3	15	28	17	7	0	5	14	10	5	0	2	1
	p.t. 3	0	0	8	7	6	10	4	2	11	6	5	6	4	8	6	4	1
	p.t. 4	0	0	7	3	18	18	4	1	4	1	5	2	2	7	8	11	5
	p.t. 5	0	0	4	1	33	13	1	4	3	18	7	7	1	4	3	4	9

considered. The first two scenarios have a constant demand for each of the five part types during weeks 5–19 equal to vectors of (7, 14, 8, 6, 10) and (11, 9, 8, 6, 3). Demand scenarios 3, 4, 5, and 6 during weeks 3–19 are given in Table 4. Part types correspond to rows.

Scenarios 3, 4, 5, and 6 represent increasingly more costly situations for the SC where, for a range of intermediate time periods in the planning horizon, namely periods (13–16), (12–17), (14–19), and (7–19), respectively, the cumulative demand load exceeds the cumulative production capacity.

To compare the proposed SC planning algorithm to the state-of-the-art practice in industry, three planning approaches are evaluated, the one proposed in this paper and two additional ones that represent advanced versions of planning algorithms employing the standard feature of the fixed lead time approximation. Given that the fixed lead time approximation is a basic foundation of industry

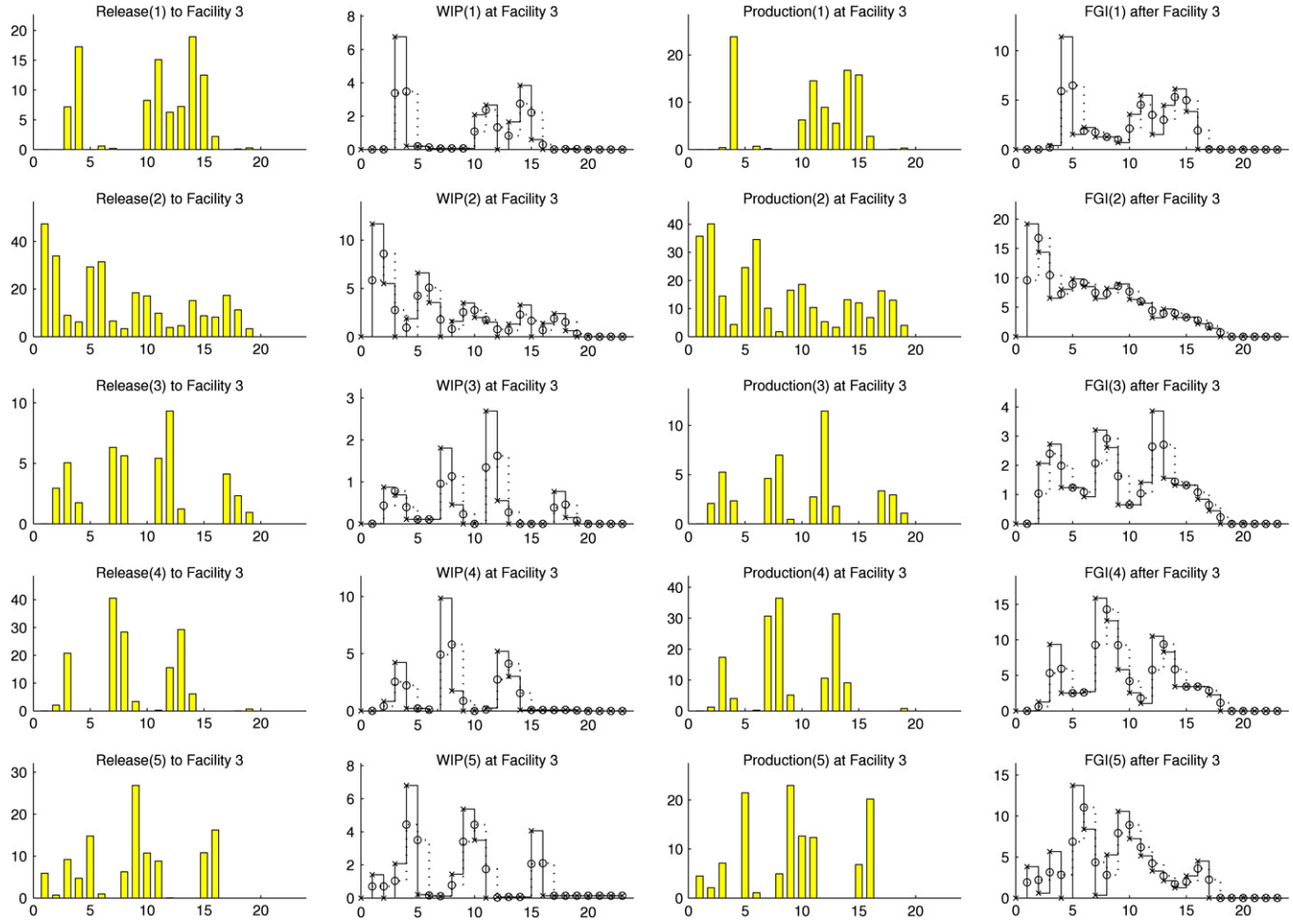

Fig. 16. Facility 3 optimal solution details of OOLFC under demand scenario 4.

Table 5
OOLFC, LIX, and CE relative cost comparison

Approach	OOLFC	LIX	CE	OOLFC	LIX	CE	OOLFC	LIX	CE
	Scenario 1			Scenario 2			Scenario 3		
T.C.	124K	130	2698	77K	158	4102	120K	204	2397
WIP	18	26	15	25	45	24	18	28	15
FGI	14	15	12	17	18	17	15	15	13
FGI^{+}	62	83	9	57	92	0	62	83	8
FGI^{-}	0	0	1614	0	0	2472	0	3	1457
FGH^{-}	6	7	1048	1	3	1588	5	75	903
	Scenario 4			Scenario 5			Scenario 6		
T.C.	242K	136	1604	169K	126	1780	1141K	152	518
WIP	10	15	8	14	19	12	2	4	2
FGI	10	9	8	12	11	11	2	2	1
FGI^{+}	74	66	21	69	84	18	4	3	1
FGI^{-}	0	10	989	0	0	1072	33	54	326
FGH^{-}	6	37	579	4	11	667	59	89	188

practice algorithms, these two algorithms are treated as proxies of industry practice. To improve the performance of the industry proxy algorithms so as to provide a fair comparison to the proposed algorithm, after the material release schedule to the upstream most facility is determined, the fixed lead time approximation is relaxed in modeling the actual production performance of the supply chain. Thus, it is asserted that these two industry practice proxies are actually representative of the state of the art in industry practice. By contrast, the "vanilla" industry practice in widespread MRP scheduling uses a fixed lead time to model delays at all production facilities, giving rise to further work-in-process inventory mismatches along the supply chain. The three methodologies used for comparison purposes in the numerical experience section are described next in greater detail:

(i) **Optimal open loop feedback controller approach.** Recall that the optimal open loop feedback controller approach (OOLFC) is the short name used for the algorithm proposed in this article and described in Section 3.4. It models the dynamics of required lead times and QoS hedging inventory by employing a smart, progressively increasing in accuracy outer linearization approximation of the nonlinear lead time and hedging inventory constraints. For a one-dimensional case, this can be visualized

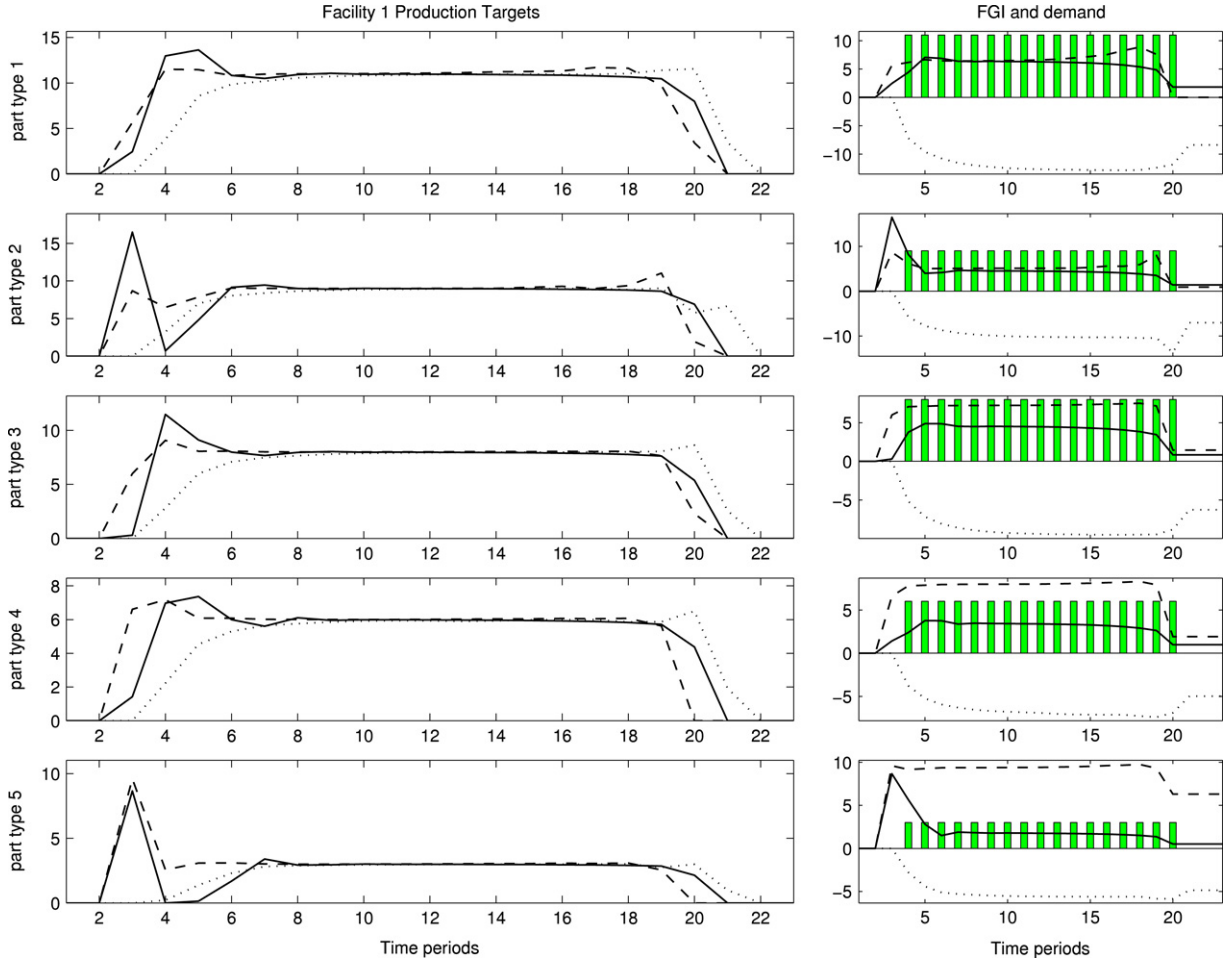

Fig. 17. Facility 3 production targets and FGI levels comparison among OOLFC, LIX, and CE approaches under demand scenario 1.

as using a tentative polyhedral feasible decision variable region area corresponding to an approximation of the union of areas A_2 and A_3 in Fig. 13. In Fig. 13, the top horizontal line represents production capacity and the one below represents the maximum allowed utilization, namely the linear constraint (5) in Exhibit 1. Little's law is then employed to convert the WIP, denoted by $\bar{Q}$, versus production, denoted by $\bar{X}$, space to the equivalent lead time versus production space. The OOLFC approach is compared to the following two approaches.

(ii) **Certainty equivalent approach.** The certainty equivalent (CE) approach provides a conservative upper cost bound to the state of the art in industry practice. It generates a raw material release schedule based on an optimistic, that is, relaxed, feasible region that corresponds to the ideal situation in which lead times equal the fixed sum of required processing time with no accounting for queuing delays. The resulting raw material release schedule generates shortfalls and almost surely provides a higher planning cost (especially material shortage costs) than that of the second proxy presented below. The relaxed feasible decision space used to generate the raw material release schedule can be visualized as the union of areas A_1, A_2, and A_3 in Fig. 13. Under the CE approach, the production planning problem is first solved under this optimistic fixed lead time constraint, ignoring QoS constraints. The resulting production schedule violates the actual lead time and QoS constraints. Under the CE approach, the raw material release schedule associated with the infeasible production plan is then fixed, and facility-specific production is reoptimized subject to the fixed raw material release schedule but with a full implementation of the nonlinear WIP and FGI constraints. The feasible production plan thus obtained is reported as the production planning cost of the CE approach. A second, more favorable proxy of the state of the art in industry practice that is also based on the fixed lead time approximation is considered below.

(iii) **Limited information exchange approach.** The limited information exchange (LIX) approach is finally used to generate production schedules that are the best representatives of the state of the art in industry practice. It models limited information exchange among planners and operators by generating a production and raw material release schedule based on constant lead times that are representative of expected facility loading conditions. More specifically, the level of the constant lead times is set to the worst-case level, namely the one required to sustain the maximum allowable utilization. The resulting feasible decision space can be visualized as area A_3 in Fig. 13. The LIX approach is a reasonable abstraction of the limited information about actual lead times that production facility managers may provide on a yearly or quarterly basis to legacy MRP-based planning systems in place. Again, to render the LIX approach a state of the art of industry practice representative, further action is taken to ensure compatibility of WIP requirements and actual production levels and to enforce interfacility QoS requirements. To this end, the raw material release rates are again fixed to the schedule determined by the linear fixed lead time constraints, and the production schedule is reoptimized subject to the union of the linear fixed lead time and the nonlinear lead time and QoS constraints. Note that the LIX approach provides feasible and virtually optimal production schedules when the production facilities in the SC are level loaded, whereas it is still disadvantaged when demand and hence loading varies over time. It may therefore be concluded

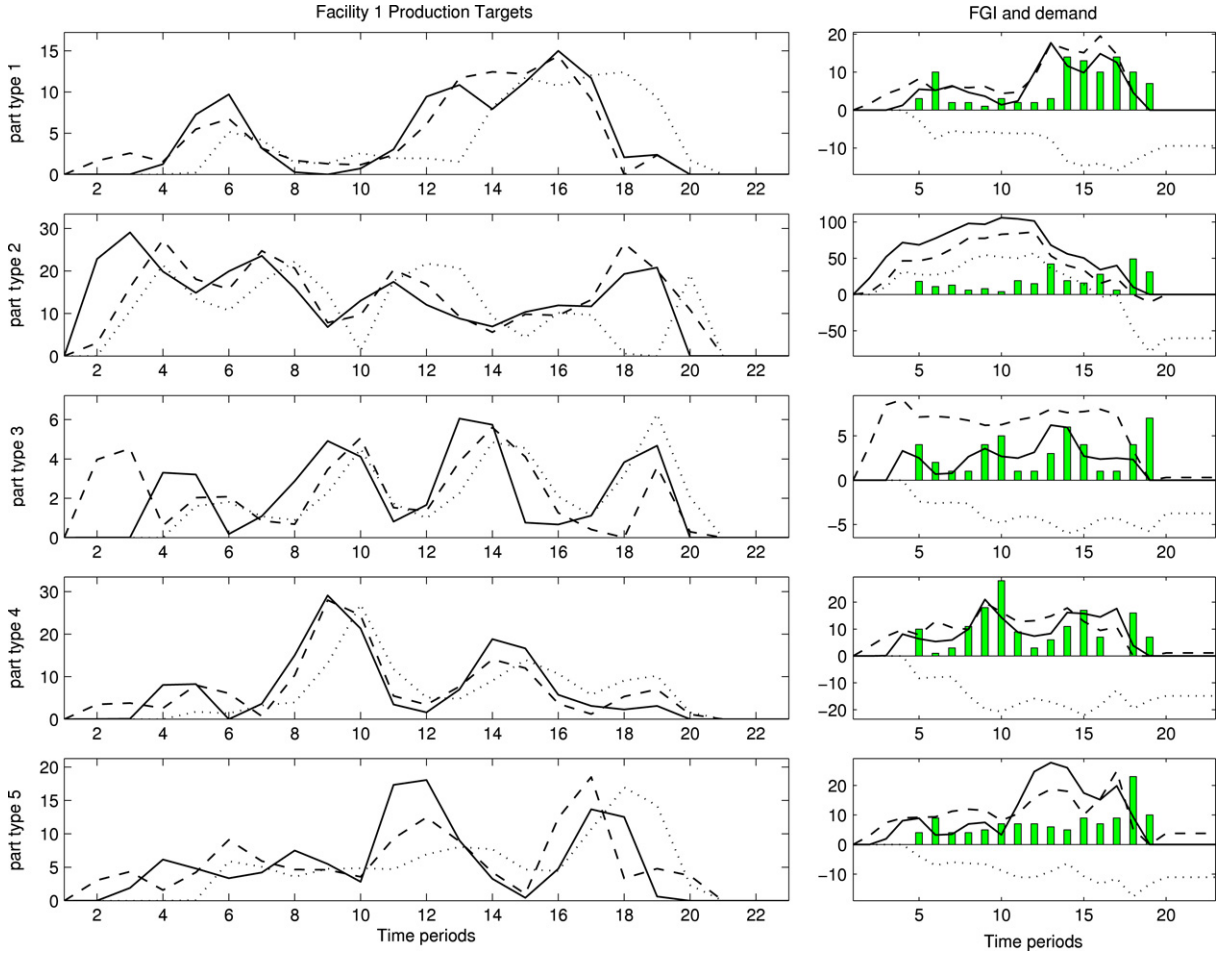

Fig. 18. Facility 3 production targets and FGI levels comparison among OOLFC, LIX, and CE approaches under demand scenario 4.

that the performance of state of the art in industry practice today is well represented by the LIX and CE approaches. Cost comparisons reported below may therefore be interpreted accordingly.

Figs. 14–16 present the optimal production schedules for facilities 1, 2, and 3 of the proposed OOLFC approach under demand scenario 4. Part-type-specific schedules are shown from top to bottom. The release rate of parts into each facility, average WIP in each facility, facility-specific production rates, and FGI (facility 1) or hedging inventory (facilities 2 and 3) schedules are also shown. Demand trajectories are superimposed for easy reference on FGI trajectories as a bar graph on the same scale. It is observed that for part types 1, 2, and 5 production plans at facility 1 are characterized by significant preproduction during the first half of the planning horizon for part type 2 and in the second half for the other two. FGI trajectories of part types 3 and 4 seem to balance the requirements of the FGI hedging constraint and the implied cost of sharp changes in FGI level.

In Figs. 15 and 16, the last column depicts the hedging inventory located downstream to the associated facility. Past computational experience indicates that the trajectory of the hedging inventory resembles closely the upstream production schedule, which is the major determinant of the prescribed hedging level. For the reasonable cost coefficients and demand loads employed, only under extreme cases does the optimization algorithm choose to stock material in FGI for future use.

In Figs. 17–19, the production targets of facility 1 obtained by each of the three approaches are compared. Solid, dashed, and dotted lines are used to distinguish OOLFC, LIX, and CE approach weekly production schedules for three different demand scenarios. Part type 1 through 5 production is shown from top to bottom, respectively.

Demand scenario 1 schedule shown in Fig. 17 corresponds to a constant workload of about 79% on facility 1 and a constant part type demand mix. Because this load is close to the 90% for which the constant lead time and hedging inventory constraint used in the LIX approach has been calculated, one would expect a relatively small cost difference in the performance of the LIX approach relative to the OOLFC approach. The differences in performance between LIX and OOLFC observed here (30% higher cost than OOLFC as reported in Table 5) are primarily caused by (i) higher FGI holding costs at the first facility due to LIX approach's need to build up inventory earlier than necessary and then not being able to reduce the inventory until the end of the planning horizon because demand load remains high till the end of the horizon, and (ii) higher WIP holding costs due to the lower production target levels that are needed during time periods at the beginning and the end of the planning horizon when the WIP and hedging inventory levels are "ramped up" from zero levels or "ramped down" to zero levels. In a production system processing a single-part-type, or equivalently a family of perfectly similar part types, level loading that is close to the maximum allowed utilization, $\eta_c(t)$, yields minimal performance differences [56]. On the other hand, scenario 2 has a constant workload of about 66%, which differs significantly from the 90% utilization to which LIX fixed lead times are calibrated. As a result, LIX performance is worse—58% higher as opposed to 30%. The level-loading, constant-production mix observed in Fig. 17 verifies the proximity of the LIX and OOLFC-generated production plans during the middle portion of

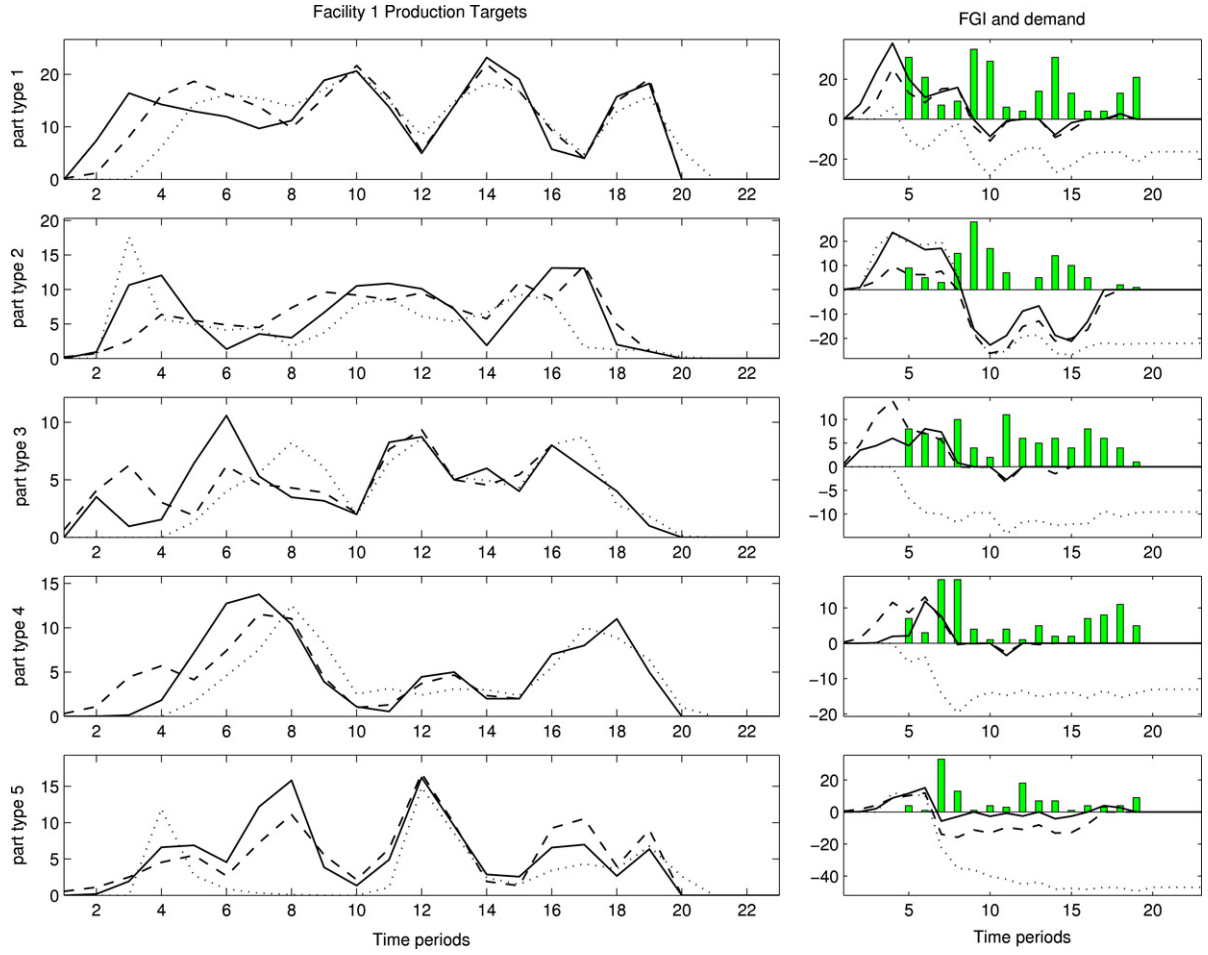

Fig. 19. Facility 3 production targets and FGI levels comparison among OOLFC, LIX, and CE approaches under demand scenario 6.

the planning horizon. The remaining differences are due to the nonlinear lead time and hedging inventory requirements during the transition period at the beginning/end of the planning horizon from/to zero demand activity.

Fig. 18 shows the production target and FGI differences among the three approaches under scenario 4. Although the overall horizon workload is feasible, demand scenario 4 contains periods of very high demand. Production schedules for each of the three approaches attempt to balance preproduction and shortages in different ways. There are multiple "ramp-up" and "ramp-down" portions of the production schedule during which the LIX approach results follow the OOLFC approach results with delay. This behavior of production schedules is even more conspicuous in Fig. 19, which presents the results of each of the three approaches on scenario 6. Due to the fluctuating demand pattern of scenario 6, none of the approaches can avoid temporary shortages of FGI at facility 1. Furthermore, the delay of the LIX approach solution in responding to scenario 6 demand has an amplification effect on backlog and results in very high backlog costs that are in absolute terms much higher than WIP costs. This is the main reason why the LIX approach reports 42% higher total cost than OOLFC in scenario 6, in contrast to 26% in scenario 4.

As mentioned already, while OOLFC corresponds to the proposed modeling of dynamic lead times through the incorporation of nonlinear constraints (6)–(8) of Exhibit 1, the LIX approach is used as a representative of the current practice of planning with a fixed lead time and safety stock assumptions, usually corresponding to level-loading conditions at the maximum allowed efficiency (worst-case analysis). Finally, the CE approach is representative of an over-optimistic, that is, too low, lead time allowance that assumes no queuing delays. Obviously, CE schedules are hampered by shortfalls and backlogs, rendering the CE approach one order of magnitude worse than the more advanced state of the art in industry practice represented by LIX.

Table 5 summarizes the potential benefits of the proposed OOLFC framework. Note that the Total Cost (T.C.) column shows the value of the optimal objective function for OOLFC and the relative percentage of the cost produced by the CE and LIX approaches. The remaining columns represent a decomposition of the total cost into its WIP, positive FGI, and backlog components. The superiority of the OOLFC dynamic lead time approach is clear.

In terms of computational effort, all three approaches under all six scenarios converged in no more than 16 master problem subproblem iterations. In 15 runs out of the total of 18, convergence was achieved in at most 12 iterations. The software implementation involved compiled C++ objects, Matlab's interpreted scripts, and Ilog's CPLEX optimizer. On a Pentium 4 PC running a Linux operating system, the calculations for each iteration takes about five seconds (there is ample room for improvement if the whole implementation is done in compiled objects), and the LP solution takes less than half a second. Extensive empirical experience with varying sizes of problem instances, up to eight part types, seven facilities, and 40 time periods indicates that: (i) the number of iterations to convergence remains flat over all cases run, and (ii) the total computational effort grows linearly in the number of time periods and number of facilities but polynomially in the number of part types.

5. Conclusion

Computationally tractable and robust algorithms for SC coordination and planning are presented that are capable of incorporating nonlinear lead time performance and probabilistic quality of service requirements to reduce SC inventory and increase its speed. Nevertheless, demonstration that significant reduction in inventory costs is possible when the nonlinear relationship of facility lead times is modeled in the SC production planning process is not the major contribution of this paper. The major contribution of this paper is the proposal and implementation of a practical, efficient, tractable, and robust algorithm that is capable of achieving these cost savings. In doing so, the concept is proved that planning on constant lead times is not a necessary evil imposed by the presumption of insurmountable computational complexity. In fact, this presumption is shown to be incorrect and that industry does not have to live any longer with the undesirable consequences of the constant lead time assumption impeding today's production planning practice.

Furthermore, the results leverage the value of research in stochastic systems and, in particular, queuing network models of production systems, including analytical probabilistic approaches as well as Monte Carlo simulation approaches that have attracted and continue to attract a substantial portion of the research community. Ongoing research is focusing on improving QoS layer algorithms and enhancing fluid Monte Carlo simulation models applicable to multi-echelon production systems.

Finally, it is emphasized that the inclusion of production system stochastic dynamics in SC coordination and production planning is particularly desirable in view of the emergence of sensor network and related technologies such as RFIDs. The value of these technologies' ability to supply reliable, affordable, and timely information about production conditions, transportation, receiving, warehousing, and retail activities will be undoubtedly enhanced if this information is translated to significant efficiency gains in SC management.

Acknowledgment

NSF Grant DMI-0300359 is acknowledged for partial support of the research reported here.

References

[1] Meal HC, Wachter MH, Whybark DC. Material requirements planning in hierarchical production planning systems. International Journal of Production Research 1987;25(7):947–56.
[2] Hax AC, Meal HC. Hierarchical integration of production planning and scheduling. In: Geisler MA, editor. Logistics, TIMS studies in the management sciences. North Holland; 1975. p. 53–69.
[3] Bitran GR, Tirupati D. Hierarchical production planning. In: Graves SC, Rinnooy Kan AHG, Zipkin PH, editors. Logistics of production and inventory, vol. 4. Amsterdam: N. Holland; 1993 [Chapter 10].
[4] Chen H, Mandelbaum A. Discrete flow networks: Bottleneck analysis and fluid approximations. Operations Research 1991;16(2):408–47.
[5] Lambrecht M, Decaluwe L. JIT and constraint theory: The issue of bottleneck management. Production and Inventory Management Journal 1988;29(3): 61–6.
[6] Goldratt E, Cox J. The goal: Excellence in manufacturing. North River Press; 1984. 262 pp.
[7] Sharifnia A, Caramanis M, Gershwin SB. Dynamic set-up scheduling and flow control in manufacturing systems. Discrete Event Dynamic Systems: Theory and Applications 1991;1(2):149–75.
[8] Caramanis M, Liberopoulos G. Perturbation analysis for the design of flexible manufacturing system flow controllers. Operations Research 1992;40(6): 1107–25.
[9] Liberopoulos G, Caramanis M. Dynamics and design of a class of parameterized manufacturing flow controllers. IEEE Transactions on Automatic Control 1995; 40(6):1018–28.
[10] Kaskavelis C, Caramanis M. Efficient Lagrangian relaxation algorithms for real-life-size job-shop scheduling problems. IIE Transactions on Scheduling and Logistics 1998;30(11):1085–97.
[11] Khmelnitsky E, Caramanis M. One-machine n-part-type optimal set-up scheduling: Analytical characterization of switching surfaces. IEEE Transactions on Automatic Control 1998;43(11):1584–8.
[12] Deleersnyder J, Hodgson T, King R, O'Grady P, Savva A. Integrating Kanban type pull systems and MRP type push systems: Insights from a Markovian model. IIE Transactions 1992;24(3):43–56.
[13] Jain S, Johnson M, Safai F. Implementing setup optimization on the shop floor. Operations Research 1996;43(6):843–51.
[14] Goncalves JF, Leachman RC, Gascon A, Xiong Z. A heuristic scheduling policy for multi-item, multi-machine production systems with time-varying, stochastic demands. Management Science 1994;40(11):1455–68.
[15] Brandimarte P, Alfieri A, Levi R. LP-based heuristics for the capacitated lot sizing problem. CIRP Annals - Manufacturing Technology 1998;47(1):423–6.
[16] Veatch M, Caramanis M. Optimal average cost manufacturing flow controllers: Convexity and differentiability. IEEE Transactions on Automatic Control 1999; 44(4):779–83.
[17] Chen H, Yao DD. Dynamic scheduling of a multiclass fluid network. Operations Research 1993;41(6):1104–15.
[18] Bertsimas D, Paschalidis ICh, Tsitsiklis JN. Optimization of multiclass queuing networks: Polyhedral and nonlinear characterizations of achievable performance. The Annals of Applied Probability 1994;4(1):43–75.
[19] Chen H, Yao DD. A fluid model for systems with random disruptions. Operations Research 1992;40(2):239–47.
[20] Connors D, Feigin G, Yao DD. Scheduling semiconductor lines using a fluid network model. IEEE Transactions on Robotics and Automation 1994;10(2): 88–98.
[21] Feng Y, Leachman R. A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. IEEE Transactions on Semiconductor Manufacturing 1996;9(3): 257–69.
[22] Orcun S, Uzsoy R, Kempf K. Using system dynamics simulations to compare capacity models for production planning. In: Proceedings of the 38th winter simulation conference. 2006. p. 1855–62.
[23] Sharifnia A. Stability and performance of distributed production control methods based on continuous-flow models. IEEE Transactions on Automatic Control 1994;39(4):725–37.
[24] Kumar PR, Meyn SP. Stability of queuing networks and scheduling policies. IEEE Transactions on Automatic Control 1995;40(2):251–60.
[25] Kumar PR, Seidman TI. Dynamic instabilities and stabilization methods in distributed realtime scheduling of manufacturing systems. IEEE Transactions on Automatic Control 1990;35(3):289–98.
[26] Dai JG. On the positive Harris recurrence for multiclass queuing networks: A unified approach via fluid models. The Annals of Applied Probability 1995;5: 49–77.
[27] Graves SC. A tactical planning model for a job shop. Operations Research 1986; 34(4):522–33.
[28] Lambrecht MR, Ivens PL, Vandaele NJ. ACLIPS: A capacity and lead time integrated procedure for scheduling. Management Science 1998;44(11): 1548–61.
[29] Woodruff DL, Voss S. A model for multistage production planning with load dependent lead times. In: Proceedings of the 37th Hawaii international conference on system sciences. 2004. p. 1–9.
[30] Asmundsson J, Rardin RL, Uzsoy R. Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Transactions on Semiconductor Manufacturing 2006;19(1):95–111.
[31] Ettl M, Feigin GE, Lin GY, Yao DD. A supply network model with base-stock control and service requirements. Operations Research 2000;48(2):216–32.
[32] Graves SC. Safety stocks in manufacturing systems. Journal of Manufacturing and Operations Management 1988;1(1):67–101.
[33] Pahl J, Voss S, Woodruff D. Production planning with load dependent lead times: An update of research. Annals of Operations Research 2007;153(1): 297–345.
[34] Caramanis M, Pan H, Anli OM. A closed-loop approach to efficient and stable supply-chain coordination in complex stochastic manufacturing systems. In: Proceedings of the American control conference. 2001. p. 1381–8.
[35] Saksena VR, O'Reilly J, Kokotovic PV. Singular perturbations and time-scale methods in control theory: Survey 1976–1983. Automatica 1984;20(3): 273–93.
[36] Clark AJ, Scarf H. Optimal policies for a multi-echelon inventory problem. Management Science 1960;6:475–90.
[37] Lambrecht MR, Muckstadt JA, Luyten R. Protective stocks in multi-stage production systems. International Journal of Production Research 1984;22(6): 1001–25.
[38] Glasserman P, Tayur SR. Sensitivity analysis for base-stock levels in multiechelon production inventory systems. Management Science 1995; 45(2):263–81.
[39] Paschalidis ICh, Vassilaras S. On the estimation of buffer overflow probabilities from measurements. IEEE Transactions on Information Theory 2001;47(1): 178–91.
[40] Suri R, Hildenbrant RR. Modeling flexible manufacturing systems using mean value analysis. Journal of Manufacturing Systems 1984;3(1):27–38.
[41] Schweitzer P. Approximate analysis of multiclass closed network of queues. In: International conference on stochastic control and optimization. 1979.
[42] Bard Y. Some extensions to multiclass queuing network analysis. In: Arato AM, Butrimenko, Gelenbe E, editors. Performance of computer systems. North Holland; 1979.
[43] Caramanis M. Production system design: A discrete event dynamic system and generalized Benders' decomposition approach. International Journal of Production Research 1987;8(25):1223–34.
[44] Kouikoglou VS, Phillis YA. An exact discrete-event model and control policies for production lines with buffers. IEEE Transactions on Automatic Control 1991;36(5):515–27.

[45] Caramanis M, Wang J, Paschalidis ICh. Enhanced fluid approximation models discrete part dynamics while enabling Monte Carlo simulation-based I.P.A. Boston University Center for Information and Systems Engineering working paper; March 2003.
[46] Caramanis M, Anli OM. Modeling work in process versus production constraints for efficient supply chain planning: Convexity issues. In: Proceedings of the IEEE CDC. 1999. p. 900–6.
[47] Anli OM. Iterative approximation selection algorithms for supply chain production planning. Boston University Center for Information and Systems Engineering working paper; 2002.
[48] Bertsimas D, Paschalidis ICh. Probabilistic service level guarantees in make-to-stock manufacturing systems. Operations Research 2001;49(1):119–33.
[49] Paschalidis ICh, Liu Y. Large deviations-based asymptotics for inventory control in supply chains. Operations Research 2003;51(3):437–60.
[50] Paschalidis ICh, Liu Y, Cassandras CG, Panayiotou C. Inventory control for supply chains with service level constraints: A synergy between large deviations and perturbation analysis. The Annals of Operations Research (Special Volume on Stochastic Models of Production-Inventory Systems) 2004; 126:231–58.
[51] Caramanis M, Anli OM. Dynamic lead time modeling for JIT production planning. In: Proceedings of the IEEE robotics and automation conference, vol. 2. 1999. p. 1450–5.
[52] Bertsekas DP. Dynamic programming and optimal control, vol. I and II. Belmont (MA): Athena Scientific; 1995.
[53] Caramanis M, Pan H, Anli OM. Is there a trade off between lean and agile manufacturing? A supply chain investigation. In: Proceedings of the third aegean international conference on design of manufacturing systems. 2001.
[54] Caramanis M, Paschalidis ICh, Anli OM. A framework for decentralized control of manufacturing enterprises. In: Proceedings of the DARPA-JFACC symposium on advances in enterprise control. 1999. p. 99–109.
[55] Caramanis M, Anli OM. Manufacturing supply chain coordination through synergistic decentralized decision making. In: Proceedings of Rensselaer's international CAICIM. 1998.
[56] Anli OM. Supply chain production planning: Modeling dynamic lead times and efficient intercell inventory policies. Boston University Center for Information and Systems Engineering Technical Report and Ph.D. dissertation; September 2003.

Osman Murat Anli holds a B.Sc. degree (1994) in industrial engineering from Istanbul Technical University, a M.Sc. degree (1997) in operations research and statistics from Rensselaer Polytechnic Institute, and a Ph.D. degree (2003) in manufacturing engineering from Boston University. He worked for the Mitsubishi Electric Research Laboratories (MERL) (Cambridge, MA) in 2003 as a research intern and as a research associate for the Manufacturing Engineering Dept. at Boston University in 2004. He joined the Industrial Engineering Dept. at Isik University in 2005 as an instructor. Dr. Anli's current research interests include stochastic models of supply chain manufacturing and online applications of e-learning in engineering education focusing on virtual enterprise models.

Michael C. Caramanis (BS, Ch. Eng. Stanford University; MS and Ph.D., Engineering and Decision and Control, Harvard University) is a professor in the College of Engineering at Boston University, affiliated with the Dept. of Mechanical Engineering, the Division of Systems Engineering, and the Center for Information and Systems Engineering. He has served as editor of *IIE Transactions on Design and Manufacturing* and is associate editor of several journals. Dr. Caramanis has been the principal investigator on numerous National Science Foundation funded projects and has collaborated widely with industry. His current research interests are in the area of production planning and control of complex manufacturing systems operating under uncertainty, with particular emphasis on the decomposition of these systems to tractable subsystems and their coordination for enterprise integration and supply chain management. Dr. Caramanis' first career was in the area of spot pricing of electricity and the liberalization of electricity markets, and he is coauthor of the influential field initiating text *Spot Pricing of Electricity*, Kluwer 1988. From 2004 until recently, he served as the chair of the Greek Regulatory Authority for Energy. Building on his energy systems research experience, he is currently focusing on the creation of cyber infrastructure based real-time markets capable of exploiting demand-side management's potential to provide the requisite reserve services for significantly higher adoption of clean energy generation and the electrification of the transportation sector.

Ioannis Ch. Paschalidis received the Diploma in ECE from the National Technical University of Athens in 1991 and the SM and Ph.D. degrees in EECS from the Massachusetts Institute of Technology in 1993 and 1996, respectively. He joined Boston University in 1996, where he is an associate professor of electrical and computer engineering and the codirector of the Center for Information and Systems Engineering (CISE). He is also the academic director of the Sensor Network Consortium (SNC)—an industry consortium with companies active in sensor networking. He has held visiting appointments with MIT and the Columbia University Business School. Dr. Paschalidis' current research interests lie in the fields of systems and control, optimization, networking, operations research, and computational biology. The main application areas he is targeting include communication and sensor networks, supply chains, and protein docking. He has received an NSF CAREER award (2000), the second prize in the 1997 George E. Nicholson paper competition by INFORMS, and was an invited participant at the 2002 Frontiers of Engineering Symposium, organized by the National Academy of Engineering. He is an associate editor of *IEEE Transactions on Automatic Control* and of *Operations Research Letters*.