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.

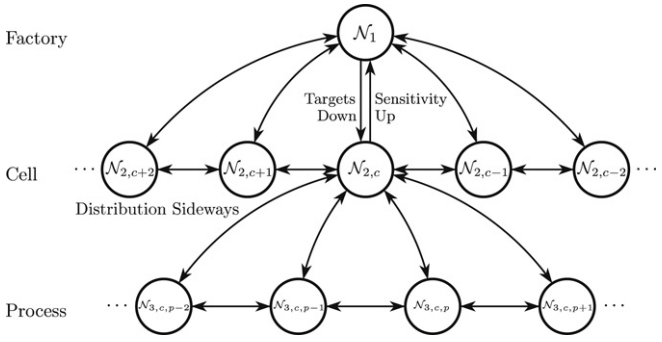

Fig. 1. Example of information architecture.

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

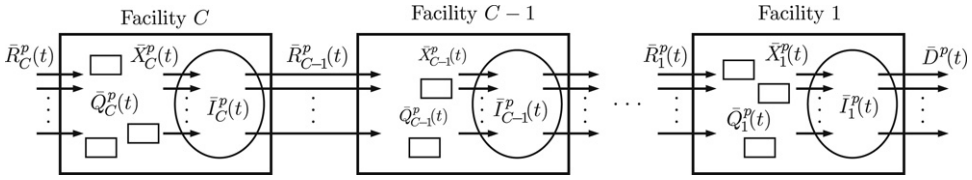

Fig. 2. A multi-class supply chain with limited production capacity at each facility.

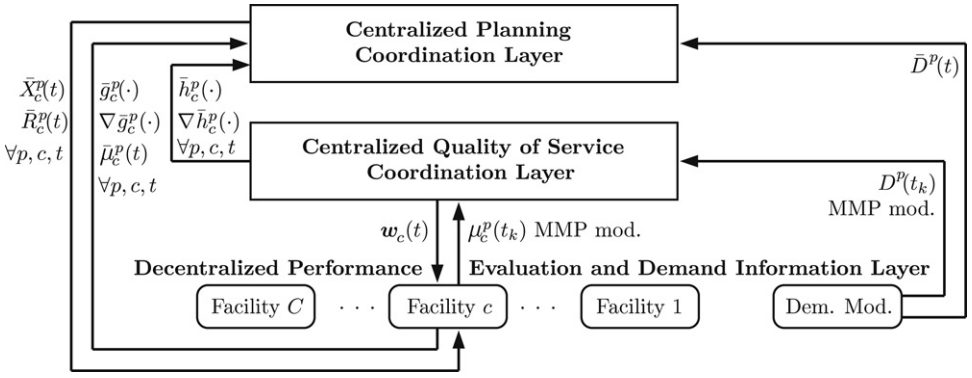

Fig. 3. Information exchange among the coordination layers and the decentralized layer.

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].

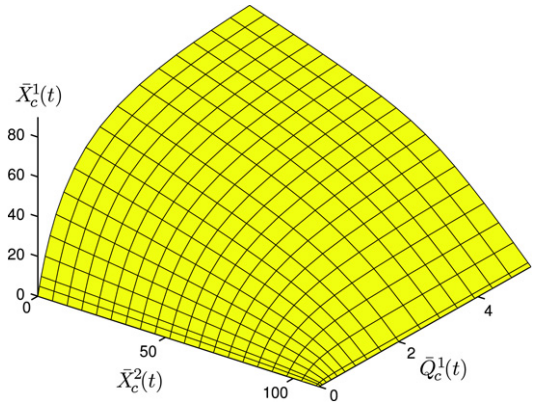

Fig. 4. Feasible production $\bar{X}_c^p(t)$ as a function of WIP $\bar{Q}_c^p(t)$.

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.

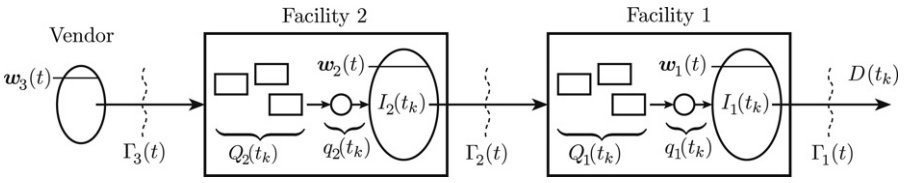

Fig. 6. Monte Carlo simulation of two-facility, one-part type SC.

(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 $\mathrm{scv}_c\left(\Gamma_{c+1}(t), \bar{\mathbf{X}}_c(t), \mu_c(t), \mathbf{P}_c(t), \pi_c(t)\right)$ with the argument list showing the significant dependencies. For simplicity is written $\mathrm{scv}_c(t) = \mathrm{scv}_c(\Gamma_{c+1}(t), \rho_c(t))$, 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 $\mathrm{scv}_c(\Gamma_{c+1}(t), \rho_c(t))$, whose values can be interpolated from the table entries.

Simulation results verify the a priori expectation that $\mathrm{scv}_c(\Gamma_{c+1}(t), \rho_c(t))$ 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 $\mathrm{scv}_c(t)$, which is then used to model the nonlinear QoS constraints in the planning coordination layer described in Section 3.4. Fig. 7 graphs the $\mathrm{scv}_c(t)$ function whose coefficients are estimated to fit the simulation table entries. Its algebraic representation is:

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

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 $\mathrm{scv}_c(t)$ 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.