Computers & Operations Research 53 (2015) 353–363

Contents lists available at ScienceDirect

Computers & Operations Research

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

# Incorporating contractual arrangements in production planning

R. John Milne [a,*], Chi-Tai Wang [b], Brian T. Denton [c], Kenneth Fordyce [d]

[a] Clarkson University School of Business, 107 B.H. Snell Hall, P.O. Box 5790, Potsdam, NY 13699, USA
[b] Institute of Industrial Management, National Central University, 300 Jhongda Road, Jhongli City, Taoyuan County 32001, Taiwan
[c] Department of Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109-2117, USA
[d] Arkieva, 5460 Fairmont Drive, Wilmington, DE 19808, USA

ARTICLE INFO

Available online 18 April 2014

*Keywords:*
Linear programming
Foundry
Supply chain planning
Semiconductor manufacturing

ABSTRACT

The semiconductor supply chain is full of complexities outside of the traditional order, make/buy, and deliver process. One critical challenge occurs when part of a semiconductor fabricator's capacity is allocated to produce wafers designed by and provided to fabless companies. In this situation, linked customer requirements are expressed simultaneously at both the semiconductor level of the supply chain and the finished goods level. As a result of the complex contractual relationships between the foundry and the fabless company, a new solution model and method is needed to determine a production plan. In our approach, two linear programming (LP) models are solved sequentially where the results of a first LP are post-processed into input for a second LP. We describe the application of this approach for two different types of contracts where the goal is maintaining as much common modeling as possible while ensuring the unique features of each contract are covered. For one type of contract, the first LP model determines the minimum quantities of wafers required to be released into the fab to meet the contractual obligation; these required starts are added as a constraint for the second LP model. For the other type of contract, the first LP determines production at one level of the bills of materials and feeds these outputs into a second LP that determines production for later stages of manufacture.



## 1. Introduction

In many industries, manufacturers have traditionally both designed and produced the products they sell. However, companies have emerged that design products that are subsequently produced by contract manufacturers, referred to as *foundries* [1]. For manufacturers with high fixed costs, the economies of scale may be such that large manufacturers must conduct two modes of business to be profitable. In the first, they manufacture products of their own design. In the second, they act as a foundry, providing manufacturing services via contract to other firms. This hybrid production model has created challenges for supply chain managers who must allocate resources for these two competing purposes.

As a result of this hybrid business model, IBM and other firms with semiconductor fabrication facilities (often shortened to *fabs*) serve as foundries for fabless companies. In these relationships, the client designs the product and contracts its manufacturing to the foundry. With new fabs costing as much as five billion dollars each [2], foundry/fabless relationships are becoming increasingly common because the companies designing products do not want to incur the expense of building semiconductor fabs and foundry manufacturers must build a wide range of products to fully utilize their fabs. This mutual dependence motivates long term agreements so that semiconductor foundries can be assured of enough demand to fill their fabs while fabless companies can be assured of sufficient supply to fulfill their needs. As a result, fabless firms enter into contracts with IBM that stipulate minimum guaranteed production levels over an agreed upon time frame.

Contractual arrangements between foundries and fabless companies are influenced by the nature of the semiconductor manufacturing process. In semiconductor wafer circuit fabrication, four manufacturing steps are repeated dozens of times: deposition, photolithography, etching, and ion implantation. Through these steps, a set of three-dimensional, layered circuit structures are built on the two-dimensional surface of each wafer in lots of 4–25 wafers, where 25 is common. After these circuits have been built, they are connected through wiring (within the chip) in which the following four manufacturing steps are repeated numerous times: deposition, photolithography, etching, and metallization (wiring). Typical lead times range from 50 to 150 days to manufacture both the circuits and the wires which connect them. The time jobs spend waiting for equipment to become available comprises the largest component of semiconductor lead times.

Following wafer fabrication, finished circuits are tested, the good chips cut (diced) from the finished wafer and placed onto a

* Corresponding author. Tel.: +1 315 268 7919; fax: +1 315 268 4478.
*E-mail addresses:* jmilne@clarkson.edu (R.J. Milne), ctwang@mgt.ncu.edu.tw (C.-T. Wang), btdenton@umich.edu (B.T. Denton), kfordyce@arkieva.com (K. Fordyce).

http://dx.doi.org/10.1016/j.cor.2014.03.023
0305-0548/

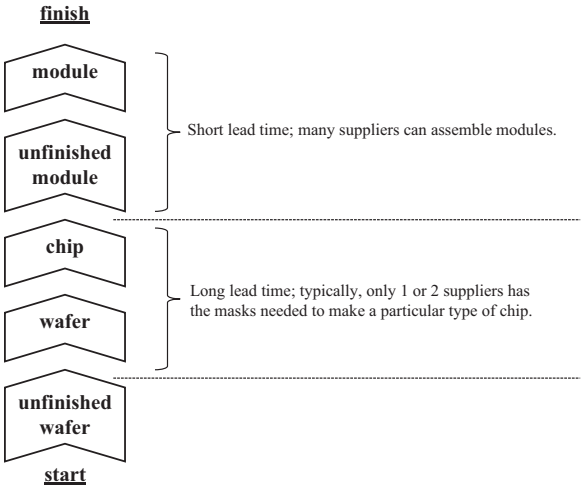


**Fig. 1.** Simplified bills of materials semiconductor manufacturing flow.

substrate and packaged to make modules. These modules are used in the assembly of a variety of products. Some IBM modules go into IBM servers such as supercomputers, mainframes, and workstations, but the chips and modules shipped to fabless companies often are assembled into consumer products such as cell phones, automobiles, global positioning systems, and game machines such as PlayStation, X-box, and Wii. Typical manufacturing lead times for module assembly and test are 10–20 days.

A set of masks (also referred to as reticles) are used in photolithography processes for etching circuit and wiring patterns onto a silicon wafer. Producing a set of masks for a single type of chip can cost several million dollars [3]. Because of the high cost of masks and other fixed costs involved in being able to manufacture a particular chip design, fabless firms will contract the manufacturing of a particular chip to a limited number of foundries—often only a single foundry. In contrast, module assembly operations are relatively few and less technologically complex resulting in short manufacturing lead times. Furthermore, as indicated in Fig. 1, the module assembly operations are more generic and can be performed economically by a large number of manufacturing contractors.

In contrast to a typical supply chain planning (SCP) process of order/make/deliver, the fabless/IBM partnerships involve the fabless company specifying their requirements simultaneously at both the start (wafer) or chip level and finished product (module) level in the supply chain (Fig. 1), and these requirements are linked. In addition to ordering finished modules, the fabless company places requirements on the production of the wafers/chips that are used to make the modules they order. This non-traditional SCP information flow (versus order/make/deliver) creates SCP challenges. While we discuss the modeling in the context of IBM's business, other semiconductor manufacturers can benefit from applying these concepts.

The remainder of this article is organized as follows. In Section 2, we review the related literature. In Section 3, we summarize the problem statement. In Section 4, we describe the core linear programming (LP) formulation used by IBM for production planning. In Section 5, we describe the usage of the core LP model and additional constraints required for minimum wafer starts contracts; these contracts require IBM to release a minimum number of wafer starts so long as there are sufficient orders placed by the fabless company to consume the output resulting from those wafer starts. In Section 6, we describe the usage of the core LP model and additional constraints required for complementary demand contracts under which customer requirements are simultaneously applied in a linked manner to both chips and modules. Section 7 provides numerical examples illustrating insights into the proposed methods and the advantages of these methods. Section 8 concludes with a summary of the insights from this article.

## 2. Literature review

Mathematical programming approaches have been applied to many contexts in which supply and demand must be matched subject to competing priorities and complex material flows such as those present in semiconductor manufacturing (e.g. product substitutions, alternative bills of materials, alternative manufacturing plant locations, and alternative capacities within a plant) within the traditional practice of order/make/deliver. For mathematical programming approaches to semiconductor SCP, see [4–22].

Kim et al. [23] propose algorithms for allocating (pegging) wafers lots to orders (demand). They peg the entire quantity of each lot to a set of orders. At IBM, it is possible for portions of lots to be pegged to portions of demands. As a result, although we use pegging in our method, the work of Kim et al. is not applicable for the IBM situation.

Hackman and Leachman [17] and Hung and Leachman [18] describe the modeling of lead times that are non-integral multiples of the LP time periods. In these papers, production starts made in one period may result in production output in multiple (typically two) periods and is often referred to as *fractional lead times*. The IBM team implemented this approach as an option in its original LP model. After extensive computational and usability exploration, IBM determined that best practices for detailed production planning models dictate that all starts made in one LP period should be modeled to arrive at stock in a single time period. This approach was preferred (but not perfect).

We illustrate the practical difficulty of the fractional lead time approach with a simple example. Suppose that production starts made in period 2 result in 40% of its production becoming available as output in period 4 and the remainder available in period 5. Further suppose there is only a single demand and it is for 100 pieces and is due in period 4. Typically, the objective function penalizes late deliveries more severely than anything else (case 1). Consequently, an LP model may recommend starting 250 pieces in period 2 so that 40% of them result in enough production output to meet the period 4 demand (assuming the yield rate is 100%); this would result in an excess inventory of 150 pieces. Conversely (case 2), if the ending inventory is penalized in the objective function to be more expensive than satisfying demand on time, then the LP may recommend starting 100 pieces in period 2 which will result in a backorder of 60% of the demand in period 4 that is not satisfied until period 5. In case 1, the planner sees 150 pieces produced and never consumed. In case 2, the planner sees the backordered demand even though capacity and components may be available. The result is unsatisfactory in both cases.

To account for demands of differing priorities, Leachman [19] and Leachman et al. [6] describe a goal programming type approach that invokes an LP run for each demand class priority in sequence of the most important demands first. During an LP run, the model is constrained to satisfy all demands more important than the current class of demands at least as well as the more important demands were satisfied during previous runs of the LP model. As a result, the on time delivery performance of satisfying the most important demands is as high as possible. For performance and “best practice” reasons, Denton et al. [4] extend the approach of Leachman et al. [6,19] so that multiple demand class priorities are accommodated within each run of an LP model. This approach involves running several demand class priorities within a single LP run and as in the Leachman et al. [6,19]

approach, constraints are added between LP runs to ensure that the on-time delivery of higher priority demands is no worse than during the previous LP run. "Best practice" refers to the fact that management and planners are focused not just on an optimal solution, but identifying repair actions to meet all demand; placing the demand from different classes in the same run allows the possibility of a trade-off between classes to facilitate this repair process.

The approach of the present paper is similar in spirit to Leachman [19] and Denton et al. [4]. The three approaches share the general idea of running an LP on the most important items first, adding some constraints, and then conducting another LP run. In the present paper, linked requirements are expressed simultaneously at different levels of the bills of materials supply chain. In contrast, in Leachman and Denton et al., unlinked requirements are expressed only at the level of end-item demand. Consequently, the present paper sets different types of constraints between LP runs than the other approaches. Furthermore, when each of the LP runs of the present method is executed in practice, the present method invokes the Denton et al. method as a subroutine for solving the LP model in a manner that favors the more important demand classes.

The present paper describes an important component of a suite of functions collectively referred to as Central Planning Engine (CPE) to support management of IBM's semiconductor demand supply network. A description of the full enterprise run and other important components can be found in related works by the authors [8–10,13,22]. Milne et al. [8] determine an optimized material requirements plan which—unlike plans created by the present paper's method—is not necessarily feasible. Fordyce et al. [9] and Degbotse et al. [10] describe a comprehensive central planning process and method that utilizes the present paper's methods as subroutines. Denton and Milne [13] describe a method for modeling demands in which a single demand may have both a commit date of one demand class priority and an earlier request date of a less important demand class priority. Milne and Wang [22] extend the approach of Denton and Milne [13] with a generalized and more efficient LP modeling approach when the importance of satisfying a demand on time increases as a function of its fulfillment date.

Knoblich et al. [24] discuss contract clauses applicable to semiconductor supply chains. Yang and Chang [25] develop a theoretical cooperative-game model where the foundry and fabless customer share the risk of yield. Kempf et al. [20] and Peng et al. [21] describe a decision support framework at Intel that supports the process of negotiating contracts with equipment suppliers and recommends which options Intel should purchase and which options to exercise to buy equipment at faster than normal lead times.

In this article, we describe models that form the basis of IBM's decision support system for SCP of its semiconductor foundry business where linked client requirements are specified at two product levels in the bills of materials supply chain.

## 3. Problem statement

This paper will consider two types of contracts that involve linked customer requirements being placed simultaneously at two levels of the supply chain: (a) minimum wafer starts and (b) complementary demand.

*Minimum wafer starts contracts*: Only a portion of the semiconductor chips fabricated on a wafer will pass the testing phase. This portion is called the *yield* and "in most [semiconductor] fabrication facilities, yield is arguably the most important metric of manufacturing performance" [26]. Yield is influenced by the product design as well as the capabilities of the manufacturing process and tooling [27]. Because the product design is provided by the fabless customer, it is partially responsible for the resulting yield.

Furthermore, chip production involves long product lead times and, as noted in [20,21], long lead times to purchase manufacturing equipment. These long lead times limit the ability of the foundry manufacturer to adjust capacities and outputs to react to unanticipated changes in yields and demands.

The long lead times and limited control over yields make the fab reluctant to agree to minimum output levels for the fabless customer. Meanwhile, the fabless customer wants assurance that the fab will allocate enough capacity for the fabless customer to have a high probability of obtaining the number of modules it needs. As noted in Fordyce et al. [28], semiconductor capacities are commonly expressed in terms of wafer starts capacities. These factors—and others—have led to contracts that focus on the minimum quantity of wafers to be released into the foundry's manufacturing line rather than on the traditional quantity of end products produced.

In minimum wafer starts contracts, IBM enters into contractual obligations to release at least a minimum number of wafer starts so long as there are sufficient orders placed by the fabless company to consume the anticipated output resulting from those wafer starts. For example, assume module M2 consumes chip C2 which consumes wafer W2 where one wafer contains enough chips to support 100 M2 modules. Suppose the daily demand for M2 is 5000 units which would require 50 wafers. If the contract calls for minimum wafer starts of 100 W2 per day, IBM would only need to start 50 per day because of the limited M2 demand. If the contract calls for a minimum of 30 W2 wafer starts per day, starting these 30 would be the highest priority for the IBM fab whereas the additional 20 wafers demanded would compete with other client demands.

*Complementary demands contracts*—refers to the customer requirements being simultaneously applied in a linked manner to both chips and modules. In this case, the foundry fab produces all of the wafers needed by the customer, but some of the post-fab processing is done by IBM and some by alternative manufacturers under contract to the fabless customer. Therefore, the relative importance of the module demand differs from that of the exploded requirements for chips.

For example, assume that module M2 requires chip C2 as a component and IBM is the only producer of C2. The customer places a requirement on IBM to produce 10,000C2 chips and requests that IBM actually assemble 2000 of those 10,000C2 chips into modules and that IBM ship the remaining 8000C2 chips to a third-party hired by the customer to assemble them into M2 modules. As the sole producer of C2 chips, it is important for IBM to produce all 10,000C2 chips required; it is less important that IBM assemble 2000 of the 10,000C2 chips into M2 modules. If IBM is unable to assemble those 2000C2 chips into M2 modules, then the customer would have a third-party conduct the required assembly operations. The 2000 M2 demands and the 10,000C2 demands are referred to as *complementary*. The 2000 M2 demands would have lower priority than the 10,000C2 demands. For the 2000 complementary units expressed at both the module and chip level, 2000 units of high priority demand are placed on C2 and complementary demand of 2000 units of lower priority is placed on M2. These two 2000 piece demands are placed simultaneously at both the M2 and C2 levels even though the total is 2000 units (rather than 4000 units). The complementary 2000 piece demand for M2 does *not* indicate that a complete set of 2000 M2 modules be built from scratch but rather only that the M2 modules be assembled from the 2000C2 chips produced in response to the complementary demand placed directly on C2. By expressing complementary demand in this manner, a higher priority can be

given to producing the C2 chips than the M2 module assembly operations that consume C2 chips. In contrast, in a traditional supply chain management relationship, the customer would order 8000C2 chips and 2000 M2 modules, essentially saying, "ship me 8000C2s and also ship me 2000 M2s." In the complementary demand contractual relationship, the customer essentially says, "build 10,000C2s and if you can assemble 2000 of them into M2s, then do so, but if you cannot, still build the 10,000C2s for me."

## 4. Core linear programming formulation

In this section, we describe the core LP formulation used by IBM. Section 5 of this paper will describe the usage of this core LP model plus additional constraints required for minimum starts contracts, and Section 6 of this paper addresses complementary demand contracts.

*Definition of subscripts*

- $j$ time period or bucket
- $m$ part number
- $n$ part number that is being substituted by another part number
- $z$ group that represents a family of part numbers
- $e$ process that can be manufacturing or purchase
- $a$ plant location within the enterprise
- $v$ receiving plant location
- $k$ customer location (note that a customer location cannot be a plant location)
- $q$ demand class which indicates relative demand priority
- $w$ capacity of a resource that can be a machine, a worker, etc.
- $u$ consuming location(s) that can be a plant within the enterprise or an external demand location

*Definition of objective function coefficients*

- $PRC_{maej}$ cost to build each piece of part number $m$ at plant $a$ using process $e$ during period $j$
- $SUBC_{amnj}$ cost to substitute each piece of part number $n$ using part number $m$ at plant $a$ during period $j$
- $TC_{mavj}$ cost to transport each piece of part number $m$ from plant $a$ to plant $v$ during period $j$
- $INVC_{maj}$ cost to hold each piece of part number $m$ in inventory at plant $a$ at the end of period $j$
- $DMAXC_{auzj}$ cost per piece of part number in group $z$ that exceeds the maximum quantity specified for the group shipment made from plant $a$ to consuming location $u$ during period $j$
- $DMINC_{auzj}$ cost per piece of part number in group $z$ that falls below the minimum quantity specified for the group shipment made from plant $a$ to consuming location $u$ during period $j$
- $BOC_{mkqj}$ cost to backorder each piece of part number $m$ at the end of period $j$ for any class $q$ demand at customer location $k$

The early work on SCP optimization attempted to integrate on time delivery (OTD) with the financial performance of the firm. A general consensus of best practice emerged to separate OTD and financial performance into two related, but different uses of the model. A simple example would be measuring the long term impact on the financial performance of not meeting a firm commitment. In this scenario, the LP model would focus on detailed optimization of OTD for demands with different relative importance that provides the planner both an optimal solution and indications of improvement opportunities. Consequently, objective function coefficients are not "real costs" such as would be recognized by a financial organization. Instead, these cost/penalty values are chosen to drive the LP to an optimal production and shipping plan that is consistent with what management wants. For example, backorder costs are chosen so that being late on the most important demands are the most expensive and as a result resources will tend to be allocated to give priority to the most important demands. Production costs are chosen to create just-in-time completion of parts when possible given material and component availability.

*Definition of parameters*

- $DEMAND_{mkqj}$ quantity requested for part number $m$ at customer location $k$ with a demand class $q$ during period $j$
- $RECEIPT_{maj}$ quantity of projected WIP and/or purchase order for part number $m$ expected to arrive at plant $a$ during period $j$
- $CAPACITY_{waj}$ capacity of resource $w$ available at plant $a$ during period $j$
- $CAPREQ_{wmaej}$ capacity of resource $w$ required to make each piece of part number $m$ using process $e$ at plant $a$ during period $j$. This description of the model assumes that capacity is consumed during the period of product release. This assumption is made for ease of exposition and because time offsets for this purpose are rarely used in practice in semiconductor manufacturing as explained in [28].
- $QTYPER_{maenj}$ quantity of component part number $m$ needed for making each piece of part number $n$ during period $j$ at plant $a$ using process $e$
- $YIELD_{maej}$ expected output for each piece of part number $m$ released (or started) at plant $a$ during period $j$ using process $e$
- $SUBQTY_{amnj}$ quantity of part number $m$ required to substitute for each piece of part number $n$ at plant $a$ during period $j$
- $MAXPCT_{auzj}$ maximum percentage of the total shipment for part number group $z$ that leaves supply plant $a$ during period $j$ to be consumed at location(s) $u$
- $MINPCT_{auzj}$ minimum percentage of the total shipment for part number group $z$ that leaves supply plant $a$ during period $j$ to be consumed at location(s) $u$
- $CT_{maej}$ cycle time (a.k.a. lead time, i.e., number of periods from the release to the completion of part number) of releasing part number $m$ using process $e$ at plant $a$ during period $j$
- $TT_{mav}$ time needed to transport part number $m$ from plant $a$ to plant $v$

*Definition of decision variables*

- $I_{maj}$ inventory of part number $m$ at plant $a$ at the end of period $j$
- $P_{maej}$ manufacturing start quantities of part number $m$ at plant $a$ using process $e$ during period $j$
- $L_{amnj}$ quantity of part number $n$ that is being substituted by part number $m$ at plant $a$ during period $j$
- $T_{mavj}$ internal logistics (i.e., shipment within the enterprise) of part number $m$ from plant $a$ to plant $v$ during period $j$
- $F_{makqj}$ customer shipment of part number $m$ that leaves plant $a$ during period $j$ to satisfy class $q$ demand at customer location $k$
- $B_{mkqj}$ backorder of part number $m$ at the end of period $j$ for class $q$ demand at customer location $k$

$H_{uzj}$ total shipment of part number group $z$ that leaves supply locations during period $j$ to be consumed at location(s) $u$

$S_{auzj}$ amount by which the total shipment of part numbers in group $z$ from plant $a$ during period $j$ to consumption location(s) $u$ exceeds the maximum amount specified in the sourcing rules

$G_{auzj}$ amount by which the total shipment of part numbers in group $z$ from plant $a$ during period $j$ to consumption location(s) $u$ falls short of the minimum amount specified in the sourcing rules

LP formulation

$$\begin{aligned}\text{Minimize } Z = &\sum_m \sum_a \sum_e \sum_j PRC_{maej} P_{maej} + \sum_a \sum_m \sum_n \sum_j SUBC_{amnj} L_{amnj} \\ &+ \sum_m \sum_a \sum_v \sum_j TC_{mavj} T_{mavj} + \sum_m \sum_a \sum_j INVC_{maj} I_{maj} \\ &+ \sum_a \sum_u \sum_z \sum_j DMAXC_{auzj} S_{auzj} + \sum_a \sum_u \sum_z \sum_j DMINC_{auzj} G_{auzj} \\ &+ \sum_m \sum_k \sum_q \sum_j BOC_{mkqj} B_{mkqj} \end{aligned} \quad (1)$$

Subject to

Material balance constraints:

$$\begin{aligned} I_{maj} = &I_{ma(j-1)} + RECEIPT_{maj} + \sum_{x \ni x + CT_{maex} = j} \sum_e YIELD_{maex} P_{maex} \\ &+ \sum_n L_{anmj} + \sum_{x \ni x + TT_{mva} = j} \sum_v T_{mvax} - \sum_n SUBQTY_{amnj} L_{amnj} \\ &- \sum_v T_{mavj} - \sum_k \sum_q F_{makqj} - \\ &\sum_{\substack{n \ni m \text{ is a} \\ \text{component of } n}} \sum_e QTYPER_{maenj} P_{naej}, \quad \forall m, a, j \end{aligned} \quad (2)$$

Backorder conservation constraints:

$$B_{mkqj} = B_{mkq(j-1)} + DEMAND_{mkqj} - \sum_a F_{makqj}, \quad \forall m, k, q, j \quad (3)$$

Capacity constraints:

$$\sum_m \sum_e CAPREQ_{wmaej} P_{maej} \le CAPACITY_{waj}, \quad \forall w, a, j \quad (4)$$

Sourcing constraints:

$$H_{uzj} = \sum_{m \in z} \sum_a \left( T_{mauj} + \sum_q F_{mauqj} \right), \quad \forall u, z, j \quad (5)$$

$$\sum_{m \in z} \left( T_{mauj} + \sum_q F_{mauqj} \right) - S_{auzj} \le MAXPCT_{auzj} H_{uzj}, \quad \forall a, u, z, j \quad (6)$$

$$\sum_{m \in z} \left( T_{mauj} + \sum_q F_{mauqj} \right) + G_{auzj} \ge MINPCT_{auzj} H_{uzj}, \quad \forall a, u, z, j \quad (7)$$

Non-negativity constraints:

$$\text{all decision variables } X_{i,j,\ldots} \ge 0 \quad (8)$$

The objective function and constraints are explained as follows:

- The LP model minimizes the overall supply chain costs (Eq. (1)), including the costs of production $P$, product/part substitution $L$, interplant logistics $T$, inventory $I$, sourcing $S$ and $G$, and backorder $B$.
- Material balance for each part number in each plant and time period (Eq. (2)) is achieved as follows. The inventory at the end of a time period is the inventory at the end of the previous time period plus the quantities coming to stock in that period resulting from activities that increase inventory and minus the quantities of those activities that reduce inventory in that period. Activities that increase inventory include projected WIP and purchase order receipts ($RECEIPT_{maj}$), production resulting from new manufacturing starts of the part number (offset by cycle time), parts substituting for it, and receipts from interplant shipments (offset by transportation lead time). Activities that decrease inventory include substitutions of the part number for other parts, interplant shipments leaving the plant, customer shipments, and manufacturing starts of part $m$'s assembly part number ($P_{naej}$).
- In Eq. (3), the demand backordered at the end of a period is the demand backordered at the end of the prior period plus any new demand due in that period minus any shipments to customers made during the time period. In other words, demand that is not satisfied in one period $t-1$ is backordered to the following period $t$.
- The manufacturing capacity consumed at a resource cannot exceed the resource's available capacity (Eq. (4)).
- Sourcing (Eqs. (5)–(7)) encourages interplant logistics $T$ and customer shipments $F$ to stay within a management-preferred range. Eq. (5) keeps track of the total shipments of a product group being received at a (consumption) location over time. Eqs. (6) and (7) keep track of the amount by which those total shipments applicable to a given supplying plant exceed ($S$) or fall short ($G$) of the desired maximum and minimum percentage of the total shipments.
- Non-negativity of the decision variables (Eq. (8)).

## 5. Modeling contracts for minimum wafer starts

In this type of contract, the minimum wafer starts contract, IBM is obligated to start a minimum amount of wafers into their manufacturing lines. However, if the customer does not place enough orders for the finished module product, then IBM can—and to avoid excess inventory should—build less than the minimum starts. This contractual situation is handled by solving two LP models sequentially as illustrated in Fig. 2.

In LP model 1, only module foundry customer orders are considered as demand. These foundry orders are demands for modules which are linked through their bills of materials to the wafers specified in the minimum wafer starts contract (see Fig. 1). These foundry orders are treated by LP model 1 of Fig. 2 as if they are due in the first time period. That is because the corresponding minimum wafer starts must be made (as an obligation of the contract) if foundry orders are available to consume them, even if the due dates of these orders are far in the future.

LP model 1 is formed by using the core LP formulation of Section 4 plus the following constraint (Eq. (9)) which prevents production from exceeding the contractually obligated minimum starts, where $P_{maej}$ is the production starts of wafer material $m$ at plant $a$ using process $e$ in time period $j$:

$$\sum_a \sum_e P_{maej} \le \text{Contracted\_minimum\_starts}_{jm}, \quad \forall j, m \quad (9)$$

The result of solving LP model 1 is a set of production starts that are the *minimum starts required to be released* based on the module foundry customer orders and the contractually specified minimum wafer starts. These *minimum starts required to be released* thus cannot exceed the starts required to meet the module foundry orders and the minimum starts specified in the contract. These starts—indicated by the additional superscript "from LP model 1" on the $P$ variable below—are capacity feasible and are aggregated across manufacturing plants ($a$) and processes ($e$) following the solution of LP model 1:

$$\begin{aligned} &Minimum\_starts\_required\_to\_be\_released_{jm}^{from\ LP\ model\ 1} \\ &\quad = \sum_a \sum_e P_{maej}^{from\ LP\ model\ 1}, \quad \forall j, m \end{aligned}$$

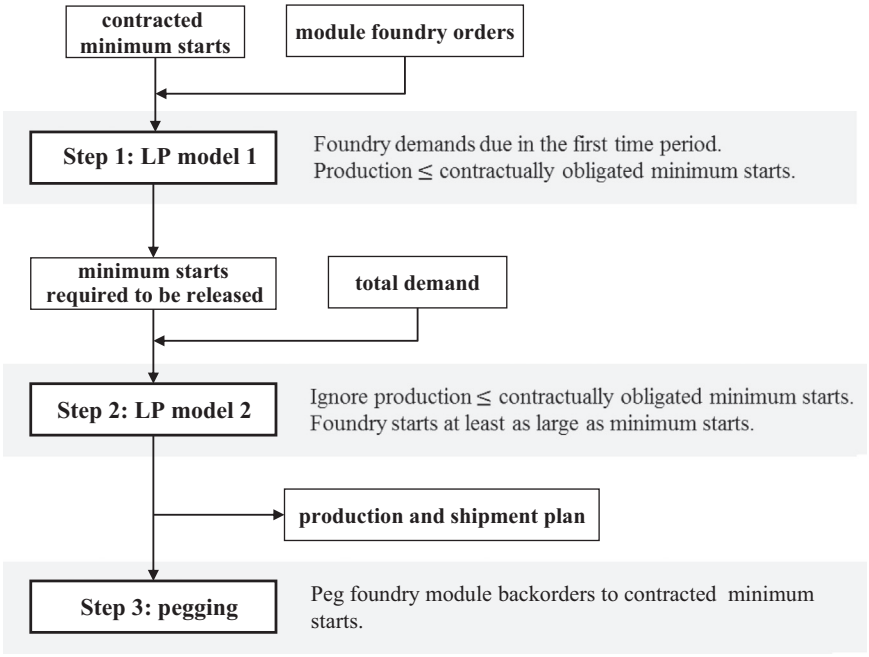


**Fig. 2.** Logic flow for modeling "Minimum Starts" contracts.

While LP model 1 considers only module foundry orders as demand, LP model 2 considers all demand (including module foundry orders) as indicated in Fig. 2. Constraint 9 above is not included in LP model 2. Instead, LP model 2 contains Eqs. (1)–(8) of the core LP formulation and the following additional constraint (Eq. (10)) to ensure that the production starts are at least as large as the minimum starts calculated from LP model 1 on a cumulative basis across time:

$$\sum_{t=1}^{j}\sum_{a}\sum_{e} P_{maet} \geq \sum_{t=1}^{j} Minimum\_starts\_required\_to\_be\_released_{tm}^{from\ LP\ model\ 1}, \quad \forall j, m \tag{10}$$

After LP model 2 has been solved, users analyze whether or not any unsatisfied module foundry orders stem from a capacity problem preventing the release of the minimum starts or from foundry orders exceeding the corresponding minimum starts. This is done through a pegging calculation that traces foundry orders from end-items (typically modules) to the wafer starts supporting them. Thus, the pegging process maps customer demand to supply upstream in the supply chain.

The pegging method is summarized as follows. The method works from the top of the bills of materials (e.g. module in Fig. 1) to the bottom of the bills of materials (e.g. unfinished wafer in Fig. 1). At each level of the bills of materials supply chain, the independent and dependent demands at that level are associated with withdrawals of inventory supporting these demands on a part number, quantity, location, and date basis. (A part's *dependent demands* are those resulting from the bills of materials explosions of planned releases of assemblies consuming the part as a component.) These consumptions of inventory are matched with the activities resulting in an increase in inventory (such as from production starts or product substitutions), assuming the sequence in which items are received into inventory is the sequence in which they are withdrawn from inventory (i.e. FIFO). Thus, by proceeding with these calculations level by level through the bills of materials supply chain, the method associates ("pegs") the end-item demands at the top of the bills of materials with the wafer starts at the bottom of the bills of materials supply chain. Often a single demand is supported by multiple wafer starts and a single wafer start supports multiple demands. See [10] for more detail on this pegging process.

When end-item foundry orders are being satisfied late, this may be caused by capacity problems or by manufacturing lead times. The pegging information and associated capacity utilization reports help the user analyst identify the cause of any late shipments. If the module foundry orders are not being satisfied due to capacity not being available to support the contractually obligated minimum starts, then the supply chain planner will negotiate with Manufacturing and/or Manufacturing Engineering to reallocate fabrication equipment toward the foundry products having the difficulty. Once the additional capacity has been allocated to the foundry products, LP models 1 and 2 are re-solved.

## 6. Modeling contracts for complementary demands

In this contractual relationship, the fabless company is primarily concerned with IBM delivering semiconductor chips that are made through a long and complex process. At the fabless company's request, some of these chips will be delivered directly to the customer or to a third party assembling the modules on behalf of the fabless company. The demands for these chips are referred to as *complementary chip demand type 1* (as illustrated in Fig. 3). Other chips will be reserved for assembly into modules by IBM and the demands for these chips are referred to as *complementary chip demand type 2*. Because of the contract and/or a desire on IBM's part to satisfy the fabless customer, it may be very important for IBM to produce all of the chips demanded by the customer—whether of type 1 or type 2. However, because many manufacturers can conduct the subsequent module assembly operations, if IBM is unable to finish these operations on time (i.e. not satisfying complementary module assembly demands entirely), that would not create a significant adverse effect on the customer (as long as IBM is able to deliver the required amount of chips.) Whereas in a traditional order/make/deliver supply chain relationship, IBM would simply provide commit dates for when it can deliver the ordered module assemblies, in this case, IBM must ensure that it can produce the chips going into these modules on time even if it cannot deliver the finished modules on time. Essentially, the

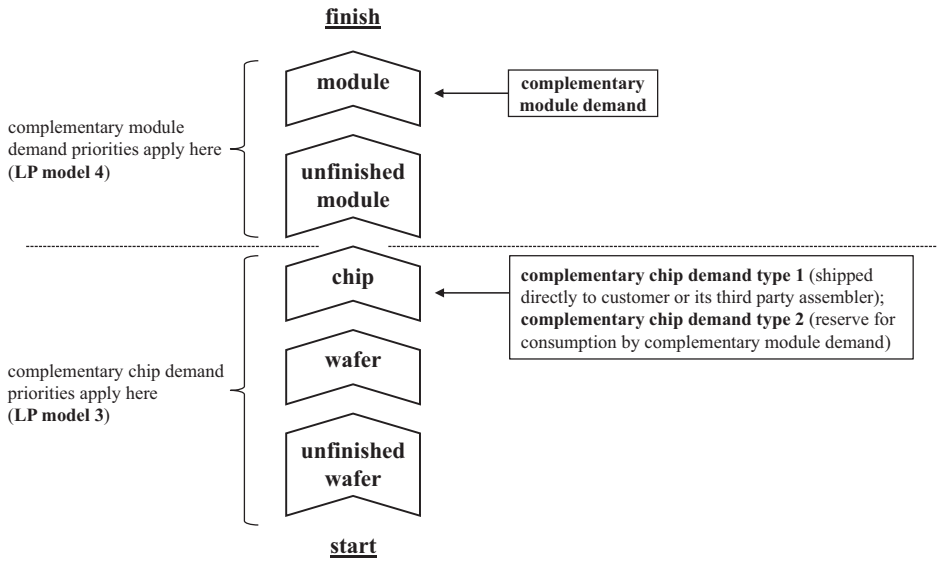


**Fig. 3.** Representative placement of complementary demands and associated LP models where they are treated as input.

fabless customer tells IBM the quantity of chips that it needs IBM to build and the quantities of modules that it requests IBM to assemble from those chips. These module requests are referred to as *complementary module demand* (Fig. 3). These demands may be satisfied only if these modules can be assembled using the supply of chips that satisfy complementary chip demand of type 2. Typically, IBM assigns high priority to the complementary chip demand and low priority to the corresponding complementary module demand. This prioritization reflects the relative importance of these demands to the fabless company and thus prioritizes IBM's capacity and material assets to maximize customer satisfaction and to fulfill contractual obligations.

Summarizing the above discussions, there are three types of complementary demand with the following characteristics:

1) Complementary module demand that may be satisfied only using chip supply that is available from satisfying complementary chip demand type 2.
2) Complementary chip demand type 1 that may be satisfied only by shipping the chips directly to the customer or its designated third party module assembly contractor.
3) Complementary chip demand type 2 that is reserved for potential use in assembling modules at IBM for which there is complementary module demand.

All demands that are not among the above three types would be considered *non-complementary demands*. Non-complementary demands consist primarily of modules being used in IBM servers and by IBM customers who place orders only for completed modules without a corresponding chip output requirement or minimum wafer starts requirement. Because the fabless company owns its products' designs and IBM owns the designs of non-complementary demanded products, each module part number and each chip part number can have complementary or non-complementary demand, but not both. Non-complementary demand must be considered when planning to satisfy complementary demand because shared capacity is used in responding to both kinds of demands.

Fig. 4 illustrates how this type of contractual relationship is solved. The first step of the method is solving LP model 3 which uses Eqs. (1)–(8) of the core LP formulation and considers complementary chip demand (types 1 and 2) and all non-complementary demand in Eq. (3). LP model 3 does not consider complementary module demand in Eq. (3). Although all non-complementary demands are considered, the primary purpose of LP model 3 is to establish a production plan for the chips with complementary demand and for all parts feeding them in the bills of materials supply chain (e.g. unfinished wafer, wafer, and chip in Fig. 3), as well as a shipment plan that satisfies complementary chip demand of type 1. (The non-complementary demand will be considered in LP model 4 and it is the production and shipment plan from LP model 4 that will determine how the non-complementary demand will be fulfilled; however, these non-complementary demands must be considered in LP model 3 because they can influence how much capacity and how many material assets should be used in fulfilling the complementary chip demand.)

LP model 4 uses Eqs. (1)–(8) of the core LP formulation plus Eqs. (11)–(18) below. The supply of chips with complementary demands resulting from the solution of LP model 3 becomes firm and therefore fixed when used in LP model 4. In particular, any activity (new production $P$, product substitutions $L$, interplant shipments $T$) that results in an increase in inventory for these chips becomes fixed (these values are denoted by a "from LP model 3" superscript in Eqs. (11)–(16)). Also, we fix the customer shipments being used to satisfy complementary chip demand type 1 (Eqs. (17) and (18)). To avoid computational difficulties stemming from small rounding errors, these decision variables are fixed by including the following constraints in LP model 4 so that the variables are within a specified tolerance $\varepsilon$ (which is on the order of 0.0001) of their values resulting from solving LP model 3:

$$P_{maej} \le P_{maej}^{from\ LP\ model\ 3} + \varepsilon \quad \forall m \in complementary\ chip\ demands, a, e, j \tag{11}$$

$$P_{maej} \ge P_{maej}^{from\ LP\ model\ 3} - \varepsilon \quad \forall m \in complementary\ chip\ demands, a, e, j \tag{12}$$

$$L_{anmj} \le L_{anmj}^{from\ LP\ model\ 3} + \varepsilon \quad \forall m \in complementary\ chip\ demand, a, n, j \tag{13}$$

$$L_{anmj} \ge L_{anmj}^{from\ LP\ model\ 3} - \varepsilon \quad \forall m \in complementary\ chip\ demands, a, n, j \tag{14}$$

$$T_{mavj} \le T_{mavj}^{from\ LP\ model\ 3} + \varepsilon \quad \forall m \in complementary\ chip\ demands, a, v, j \tag{15}$$

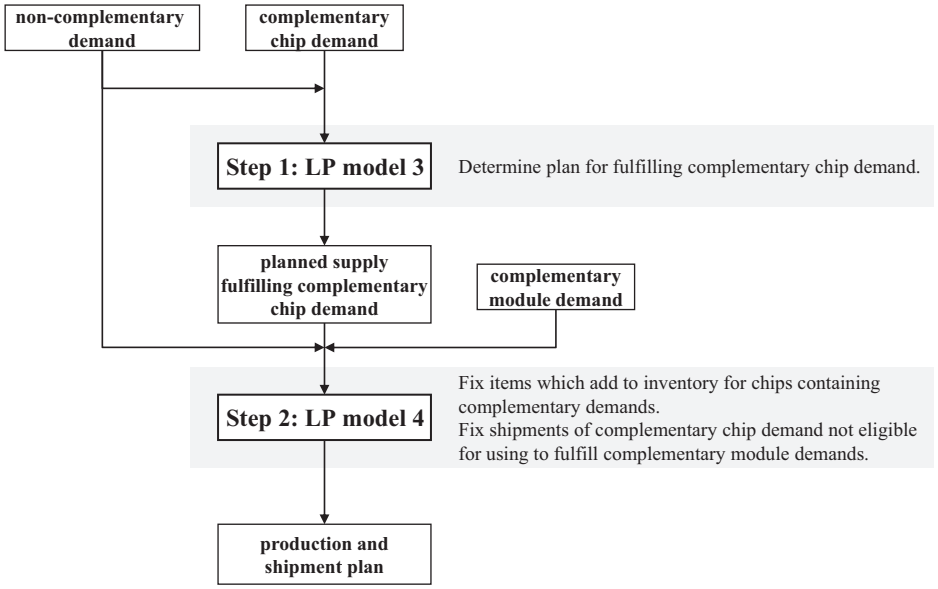


**Fig. 4.** Logic flow for modeling "Complementary Demand" contracts.

$$T_{mavj} \geq T_{mavj}^{from\ LP\ model\ 3} - \varepsilon \quad \forall m \in complementary\ chip\ demands, a, v, j \tag{16}$$

$$F_{makqj} \leq F_{makqj}^{from\ LP\ model\ 3} + \varepsilon \quad \forall m, k \in complementary\ chip\ demands \text{ of type } 1, a, q, j \tag{17}$$

$$F_{makqj} \geq F_{makqj}^{from\ LP\ model\ 3} - \varepsilon \quad \forall m, k \in complementary\ chip\ demands \text{ of type } 1, a, q, j \tag{18}$$

LP model 4 includes all non-complementary demand as well as complementary module demand. These complementary module demands typically have a less important priority than the complementary chip demands. That is because if these complementary module demands are unsatisfied, the customer can have another third party assemble the modules using the chips provided by IBM. The complementary chip demands of type 1 may be included in LP model 4, but their presence is redundant because the customer shipments satisfying these demands have been fixed. Complementary chip demands of type 2 are excluded from LP model 4.

Referring to Figs. 3 and 4, this method creates a production and shipment plan in which the production of the chip, wafer, and unfinished wafer are driven by the complementary chip demand priorities, while the module and unfinished module production and shipment plans are driven by the complementary module demand priorities. Production plans for parts at all levels of the bills of materials supply chain are driven by the need to satisfy non-complementary demand as well. As a result, this method coordinates all the respective priorities and all demands—complementary and otherwise—into a coherent production and shipment plan.

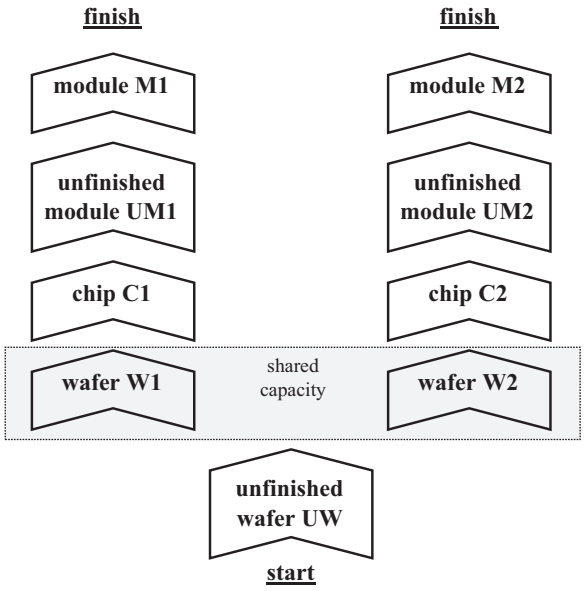


**Fig. 5.** Bills of materials for numerical examples.

## 7. Numerical examples

We illustrate our two-stage LP methods using examples that utilize the bills of materials of Fig. 5. These examples contain simplified scenarios to convey the key concepts without burdening the reader with unnecessary details and calculations.

In Fig. 5, wafers W1 and W2 are each started using unfinished wafer UW as a component. Capacities are expressed as a maximum of 100 wafers that may be started into the manufacturing line during any given time period. We assume it takes two periods of time to make chips from unfinished wafers and one period of time to make modules from finished chips. We assume 100 chips are produced from each wafer and all yields are 100%. We assume no WIP or in-stock inventories, no purchase order receipts, and no shipments in transit. The fab owns the design for module M1 (and its embedded component chip C1), and the fabless customer owns the design for module M2 (and its embedded component chip C2.)

### 7.1. Complementary demands example

Fig. 5 and Table 1 illustrate our two-stage LP method for modeling contracts which contain complementary demands. Referring to Table 1, module M1 has non-complementary demands of priority 2 of 6000 pieces in each of periods 4, 5, and 6. The complementary demands of priority 3 for module M2 in periods 4, 5, and 6 correspond to the complementary demands of type 2 for chip C2 in periods 3, 4, and 5 when accounting for the one period lead time. All complementary chip demand has the most important priority (priority 1) including the 2000 pieces of complementary demand

**Table 1**
Complementary demands example. Input data is in boldface type; output of the method is in italics font.

| Periods | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **Non-complementary demand M1 (priority 2)** | | | | **6000** | **6000** | **6000** |
| **Complementary demand M2 (priority 3)** | | | | **5000** | **5000** | **2000** |
| **Complementary demand type 1 C2 (priority 1)** | | | **2000** | **2000** | **2000** | |
| **Complementary demand type 2 C2 (priority 1)** | | | **5000** | **5000** | **2000** | |
| Total demand of C2 (all priority 1) | | | 7000 | 7000 | 4000 | |
| **Total wafer starts capacity** | **100** | **100** | **100** | | | |
| W1 unconstrained starts (exploded from C1) | 60 | 60 | 60 | | | |
| W2 unconstrained starts (exploded from C2) | 70 | 70 | 40 | | | |
| *W1 starts (2-stage LP)* | *30* | *30* | *60* | | | |
| *W2 starts (2-stage LP)* | *70* | *70* | *40* | | | |
| W1 constrained starts (via 1 LP first alternative) | 60 | 60 | 60 | | | |
| W2 constrained starts (via 1 LP first alternative) | 40 | 40 | 40 | | | |

**Table 2**
Minimum starts example. Input data is in boldface type; output of the method is in italics font.

| Periods | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **Demand M1 (priority 2)** | | | | **6000** | **6000** | **6000** |
| **Demand M2 (priority 3)** | | | | **5000** | **5000** | **2000** |
| W1 unconstrained starts | 60 | 60 | 60 | | | |
| W2 unconstrained starts | 50 | 50 | 20 | | | |
| **Total wafer starts capacity** | **100** | **100** | **100** | | | |
| **W2 minimum starts based on contract** | **50** | **50** | **50** | | | |
| *W1 starts (2-stage LP)* | *50* | *50* | *80* | | | |
| *W2 starts (2-stage LP)* | *50* | *50* | *20* | | | |
| W1 starts (1-stage LP) | 60 | 60 | 60 | | | |
| W2 starts (1-stage LP) | 40 | 40 | 40 | | | |
| W2 starts with manual modification | 50 | 50 | 50 | | | |
| Total wafer starts (1-stage LP with manual modification) | 110 | 110 | 110 | | | |

of type 1 for chip C2 in periods 3, 4, and 5. That is because IBM is the only source for chip C2.

Because of the two period lead time to make chips from unfinished wafers and 100 chips per wafer, meeting all demand on time would require starting 60 W1 wafers in each of periods 1, 2, and 3, and starting 70, 70, and 40 W2 wafers in periods 1, 2, and 3, respectively. However, there is only enough capacity to start 100 total wafers in each period. The limited capacity must be allocated among the wafers. Our two-stage LP method finds the best solution balancing demand priorities, contractual commitments, and capacity limitations. Our method allocates the capacity first to W2 in step 1 of Fig. 4 because chip C2 demands have the highest priority and the remaining capacity to W1 (which supports M1 modules of priority 2). Step 2 of Fig. 4 would determine the production and shipment plan for the modules based on the fixed chip production of step 1. The net result of our two-stage LP method would include 70, 70, and 40 starts of W2 and 30, 30, and 60 starts of W1 in periods 1, 2, and 3, respectively.

Imagine a scenario in which the two-stage LP processing did not exist and the planners had to manage this scenario with production planning software that contained only a single-stage LP. The planner would only be able to enter complementary demands at one level of the bills of materials. Consequently, the planner would either enter the complementary chip demand of type 2 or complementary demand for its corresponding module (but not both). Suppose the planner chooses to manually enter only the module complementary demand. The right side of Fig. 5 bills of materials supply chain would be driven to support the priority 3 demands on M2 and the complementary chip demand of type 1 on C2. The one-stage LP model would allocate the wafer capacity to support first the priority 1 demands on C2 (which corresponds to 20 W2 wafers per period), secondly the priority 2 demands on M1 (which corresponds to 60 W1 wafers per period), and finally the priority 3 demands on M2. The net result would be 60 starts of W1 and 40 starts of W2 in each of periods 1, 2, and 3. This would be a poor allocation of resources because it would favor the priority 2 M1 demand over the (missing from the one-stage model) priority 1 complementary demand of type 2 for C2. This misallocation illustrates the relative advantage of our two-stage LP method. Alternatively, if the planner were to keep the complementary demands of type 2 on C2, the complementary demand on M2 would need to be eliminated which would eliminate the opportunity for the business to provide the value added service of assembling the module. Consequently, no matter which complementary demand the planner chose to eliminate, the result would be poor. The reader may wonder if the planner would attempt to manually adjust the allocations outside of the core solution in which the LP method is embedded. So doing would result in suboptimization and disconnecting the foundry solution from that of the rest of the business. It is better to keep the foundry solution integrated with the rest of the business for a globally optimal solution via our two-stage LP model.

### 7.2. Minimum starts examples

We use Fig. 5 and Tables 2 and 3 to illustrate our two-stage LP method for modeling contracts that have minimum starts requirements. In Table 2, module M1 has demands of priority 2 of 6000 pieces in periods 4, 5, and 6. Module M2 has demands of priority 3 in periods 4, 5, and 6 of 5000, 5000, and 2000 pieces, respectively. Adjusting for lead times and 100 chips per wafer results in the unconstrained wafer starts of W1 and W2 as shown in Table 2 which would be enough to meet all demands on time. However, there is only enough capacity to release a total of 100 wafers per period. Assume the contractually obligated minimum starts are 50 of W2 per period for three periods. Because there are only enough M2 demands to support 50, 50, and 20 W2 wafer starts for the first three periods (and because all are subject to the

**Table 3**
Minimum starts example with starts capacity and minimum starts varying over time. Input data is in boldface type; output of the method is in italics font.

| Periods | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **Demand M1 (priority 2)** | | | | **7000** | **7000** | **5000** |
| **Demand M2 (priority 3)** | | | | **5000** | **5000** | **3000** |
| W1 unconstrained starts | 70 | 70 | 50 | | | |
| W2 unconstrained starts | 50 | 50 | 30 | | | |
| **Total wafer starts capacity** | **80** | **110** | **110** | | | |
| **W2 minimum starts based on contract** | **30** | **50** | **40** | | | |
| *W1 starts (2-stage LP)* | *50* | *60* | *70* | | | |
| *W2 starts (2-stage LP)* | *30* | *50* | *40* | | | |
| W1 starts (1-stage LP) | 70 | 70 | 50 | | | |
| W2 starts (1-stage LP) | 10 | 40 | 60 | | | |
| W2 starts with manual modification | 30 | 50 | 40 | | | |
| Total wafer starts (1-stage LP with manual modification) | 100 | 120 | 90 | | | |

contractually obligated minimum starts), those are the minimum number of starts required to be released as output by step 1 of Fig. 2. Step 2 of Fig. 2 first allocates enough capacity to W2 to meet those required W2 starts of 50, 50, and 20. Because W1 is supporting more important (priority 2) demands for M1 than the (priority 3) demands for M2, the remaining wafer starts capacity is allocated to W1 with the two-stage LP results for W1 and W2 wafer starts as shown in Table 2. The 80 piece start of W1 in period 3 supports the unconstrained 60 piece starts in period 3 plus 10 pieces backordered in both periods 1 and 2. Observe that the two-stage LP method allocated the capacity appropriately given the over-riding importance of satisfying the contractually obligated minimum starts requirement.

Suppose instead that the two-stage LP method was not available to the planner who instead used a one-stage LP model. Probably the planner would need to ignore the minimum starts requirement for the initial run of the one-stage LP and make manual adjustments afterwards outside of the formal system. Running the one-stage LP would allocate capacity first to W1 (based on W1 supporting M1 demands of priority 2 which are more important than the M2 demands of priority 3 supported by W2), resulting in 60 pieces of W1 and 40 pieces of W2 starting in each of the first three periods. After the entire solution has been created, the planner would need to manually modify the W2 starts to satisfy contractual obligations. This would involve boosting the W2 starts to 50 in periods 1, 2, and 3. This modification would be consistent with the minimum starts requirement as understood by the planner who—going outside the formal system—would not have visibility to the fact that the limited demands of M2 in period 6 would imply that only 20 pieces of W2 would need to be started in period 3 to fulfill the contractual requirement. Furthermore, the wafer releases in each of the first three periods would total 110 pieces which is over the line's capacity. The release of these wafers would result in excess WIP and in practice the manufacturing equipment that is overcapacity would be allocated to work on wafers in the line in an ad hoc manner that would not necessarily be aligned with the priorities of the demands they are supporting. This situation would be worse than that which would result from using our two-stage LP solution.

The example of Table 3 includes minimum wafer starts and capacity available that both vary over time. Module M1 has demands of priority 2 of 7000, 7000, and 5000 in periods 4, 5, and 6, respectively, and module M2 has demands of priority 3 of 5000, 5000, and 3000 in periods 4, 5, and 6, respectively. The minimum wafer starts specified in the contract for W2 are 30, 50, and 40 in periods 1, 2, and 3, respectively. Capacity is limited to a maximum total wafer starts of 80 in period 1 and 110 in periods 2 and 3. Our two-stage LP method effortlessly balances demand priorities, contractual commitments, and capacity limitations in this scenario. In contrast, a one-stage LP approach will likely result in an overcapacity situation such as illustrated in Table 3 where the total wafer starts are overcapacity by 20 and 10 wafer starts in periods 1 and 2 respectively.

## 8. Conclusion

We have presented two-stage LP based methods for planning semiconductor foundry manufacturing for two types of fabless/foundry contracts. When both types of contracts are present, the logic of Fig. 2 is embedded within the logic of Fig. 4. For example, during step 1 of Fig. 4, both steps 1 and 2 of Fig. 2 would be executed. Similarly, steps 1 and 2 of Fig. 2 would again be executed within step 2 of Fig. 4. We have not encountered or modeled a situation where a single product involves both types of contracts.

These two-stage LP examples provide insight for future practitioners and researchers when modeling supply chains in the presence of similar business relationships. A key aspect of both methods is the integration of the LP models using common equations when possible. In modeling minimum starts contracts, the use of two LP models coordinates aspects of the contract affecting multiple levels of the bills of materials. In modeling complementary demand contracts, the two LP models facilitate a coordinated approach for horizontally segregating two levels of the bills of materials supply chain while yielding in a good global solution in the presence of linked (complementary) demands.

Those seeking more detail on aspects of these methods are referred to [11,12]. See [9,10] for more information on the context of this work within a comprehensive central planning process that handles a variety of complexities in planning IBM's supply chain. As noted in [10], for daily usage, the central planning software (which includes the present paper's methods) can solve enterprise data in a few hours. The software is implemented in C++ and AIX scripts (AIX is IBM's version of the UNIX operating system) with CPLEX as the core solver of the LP equations.

While this manuscript focuses on the demand side of the contract issue, similar issues are encountered in the domain of material procurement when semiconductor manufacturers may contract with suppliers of packaging and leads. Honoring these contracts in a manner that optimizes the supply chain is an opportunity for future research.

## References

[1] Brown C, Linden G. Chips and change: how crisis reshapes the semiconductor industry. Cambridge: MIT Press; 2011.
[2] Monch L, Fowler JW, Mason SJ. Production planning and control for semiconductor wafer fabrication facilities. New York: Springer; 2013.
[3] Jhaveri T, Stobert I, Liebmann L, Karakatsanis P, Rovner V, Strojwas A, et al. OPC simplification and mask cost reduction using regular design fabrics. Proc SPIE 2009 (727417-1–727417-8).

[4] Denton B, Forrest J, Milne RJ. IBM solves a mixed-integer program to optimize its semiconductor supply chain. Interfaces 2006;36:386–99.
[5] Chou Y-C, Hong I-H. A methodology for product mix planning in semiconductor foundry manufacturing. IEEE Trans Semicond Manuf 2000;13:278–85.
[6] Leachman RC, Benson R, Liu C, Raar D. IMPReSS: an automated production-planning and delivery-quotation system at Harris Corporation—semiconductor sector. Interfaces 1996;26(1):6–37.
[7] Stray J, Fowler JW, Carlye WM, Rastogi AP. Enterprise-wide semiconductor manufacturing resource planning. IEEE Trans Semicond Manuf 2006;19:259–68.
[8] Milne RJ, Wang C-T, Yen C-K, Fordyce K. Optimized material requirements planning for semiconductor manufacturing. J Oper Res Soc 2012;63:1566–77.
[9] Fordyce K, Wang C-T, Chang C, Degbotse A, Denton B, Lyon P, et al. The ongoing challenge: creating an enterprise-wide detailed supply chain plan for semiconductor and package operations. In: Kempf, Keskinocak, Uzsoy, editors. Planning production and inventories in the extended enterprise: a state of the art handbook, vol. 2. New York: Springer; 2011. p. 313–87 ([chapter 14]).
[10] Degbotse A, Denton B, Fordyce K, Milne RJ, Orzell R, Wang C-T. IBM blends heuristics and optimization to plan its semiconductor supply chain. Interfaces 2013;43:130–41.
[11] Chang C, Denton B, Hegde S, Milne RJ, Smith S, Wang C-T, et al. Planning production for complementary demands. U.S. Patent 8,140,372; 2012.
[12] Denton B, Milne RJ, Orzell R, Vajjala S, Ward J. A method for optimizing foundry capacity. U.S. Patent 7,103,436; 2006.
[13] Denton B, Milne RJ. Method for simultaneously considering customer commit dates and customer request dates. U.S. Patent 8,234,144; 2012.
[14] Bang J-Y, Kim Y-D. Hierarchical production planning for semiconductor wafer fabrication based on linear programming and discrete-event simulation. IEEE Trans Automat Sci Eng 2010;7(2):326–36.
[15] Chen Y-Y, Chen T-Z, Liou C-D. Medium-term multi-plant capacity planning problems considering auxiliary tools for the semiconductor foundry. Int J Adv Manuf Technol 2013;64(2):1213–30.
[16] Ciriani TA, Leachman RC. Optimization in industry 2: mathematical programming and modeling techniques in practice. Chichester: John Wiley & Sons; 1994.
[17] Hackman ST, Leachman RC. A general framework for modeling production. Manag Sci 1989;35(4):478–95.
[18] Hung Y-F, Leachman RC. A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. IEEE Trans Semicond Manuf 1996;9(2):257–69.
[19] Leachman RC. Semiconductor production planning. In: Pardalos, Resende, editors. Handbook of applied optimization. New York: Oxford University Press; 2001. p. 746–62.
[20] Kempf KG, Erhun F, Herzler EF, Rosenberg TR, Peng C. Optimizing capital investment decisions at Intel Corporation. Interfaces 2013;43(1):62–78.
[21] Peng C, Erhun F, Hertzler EF, Kempf KG. Capacity planning in the semiconductor industry: dual-mode procurement with options. Manuf Serv Oper Manag 2012;14(2):170–85.
[22] Milne RJ, Wang C-T. Enhancing mathematical programming models to account for demand priorities increasing as a function of delivery date. J Ind Prod Eng 2014;31(1):51–63.
[23] Kim Y-D, Bang J-Y, Kwee-Yeon A, Lim S-K A. Due-date-based algorithm for lot-order assignment in a semiconductor wafer fabrication facility. IEEE Trans Semicond Manuf 2008;21(2):209–16.
[24] Knoblich K, Ehm H, Heavey C, Williams P.Modeling supply contracts in semiconductor supply chains. In: Proceedings of the 2011 winter simulation conference 2011; p. 2108–18.
[25] Yang Y.-N., Chang S-C. A contract of purchase commitments on shared yields as a risk-sharing mechanism among fabless–foundry partnership. In: Proceedings of the 2008 winter simulation conference; 2008. p. 2244–50.
[26] Leachman RC, Ding S. Excursion yield loss and cycle time reduction in semiconductor manufacturing. IEEE Trans Automat Sci Eng 2011;8:112–7.
[27] Chatterjee A, Gudmundsson D, Nurani RK, Seshadri S, Shanthikumar JG. Fabless–foundry partnership: models and analysis of coordination issues. IEEE Trans Semicond Manuf 1999;12:44–52.
[28] Fordyce K, Milne RJ, Fournier J, Singh H. Tutorial: illusion of capacity—challenge of incorporating the complexity of fab capacity (tool deployment & operating curve) into central planning for firms with substantial non-fab complexity. In: Proceedings of the 2012 winter simulation conference; 2012. p. 2302–17.