Matching Assets with Demand in Supply-Chain Management at IBM Microelectronics

PETER LYON — *IBM Microelectronics Division, 1000 River Road, Essex Junction, Vermont 05452*

R. JOHN MILNE — *IBM Microelectronics Division*

ROBERT ORZELL — *IBM Microelectronics Division*

ROBERT RICE — *IBM Microelectronics Division*

In the early 1990s, the IBM Corporation decided that its microelectronics division should expand from producing parts exclusively for other IBM locations to producing a range of products for diverse customers. To overhaul its supply-chain-management applications to handle the new business, it developed intelligent models to match assets with demand to determine which demands it could meet when and to provide manufacturing guidelines. In 1994, the PROFIT team began applying OR techniques to build these tools, interweaving linear programming with a traditional material resource planning algorithm and a heuristic matching process based on clues established in the explosion algorithm. The team has deployed three core applications: a weekly division run that determines customer commitments and manufacturing requirements, daily manufacturing runs that identify the best use of manufacturing resources to meet division requirements, and a division available-to-promise application that facilitates fast response to customers placing orders (not described). This work has improved manufacturing utilization and customer-order response time.

0092-2102/01/3101/0108/$05.00
1526–551X electronic ISSN

INDUSTRIES—COMPUTER-ELECTRONIC
INVENTORY-PRODUCTION—APPLICATIONS

INTERFACES **31:** 1 January–February 2001 (pp. 108–124)

The IBM Microelectronics Division is a leading-edge producer of semiconductor and packaged solutions for the networked world supplying a wide range of customers in such market segments as application-specific components, embedded controllers, wireless, microprocessors, memory, and storage. Customers are divided into two major groups: internal and external. Internal refers to other IBM manufacturing facilities that produce such products as mainframes, workstations, storage devices, supercomputers, and controllers. External refers to all other customers. These customers make such products as workstations, video games, GPS trackers, medical equipment, and cellular phones.

The core manufacturing flow for microelectronic parts is wafers to devices to modules to cards. The wafer is a round thin piece of silicon that looks similar to a CD. The wafers go through an elaborate process that has cycle times (time to complete the task) of 30 to 90 days in which thousands of circuits are carefully etched onto it. When the wafer is completed, it is cut into small individual rectangular shaped parts called devices. Typically cycle times are a few days. The devices are then placed on a substrate and packaged to create a module, which takes between five and 20 days. Modules are then combined together on a card. Depending on the customer, IBM Microelectronics may ship wafers, devices, modules, or cards. Within each major manufacturing activity are many individual operations and extensive testing. Operations are grouped into sectors, and sectors are grouped into levels.

The actual matching models deployed by the PROFIT team work with dynamically established levels as the core manufacturing activity or decision point. The number of levels ranges from four to 40, depending on the specific model. We will use four levels (wafer, device, module, and card) in all examples. Sullivan [1989] provides more details about the manufacturing process and how it relates to scheduling.

IBM Microelectronics and its critical suppliers have between 10 and 15 manufacturing facilities in North America, Europe, and the Far East. Typically, a manufacturing facility will specialize in building wafers, wafers and devices, devices and modules, modules, or cards. The allocation of products across manufacturing facilities is quite variable and changes regularly. A typical example of a product flow is the following: part of the wafer is made at location A, the wafer is completed and the device is created at location B, the device is tested at location C, the module is created at location D, the module is tested at location C, and then it is shipped to the customer.

In 1999, it improved its use of assets by $80 million.

The many characteristics of semiconductor manufacturing within IBM that make planning and scheduling a challenge can be divided into two categories: scope and scale. Scope includes complex manufacturing flows, long cycle times, variable cycle times, instability in demands, short product life cycles, and yield (percentage of good parts after manufacturing is com-

pleted). Scale includes number of parts, number of customers, and number of manufacturing locations, and very expensive equipment or tools. As a result of this complexity, a critical component of effective utilization of manufacturing resources and of customer response is matching assets with demand intelligently.

Matching (matching assets with demand) refers to aligning assets with demand in an intelligent manner for a variety of purposes within the supply-chain-management (SCM) process. The alignment or match occurs across multiple manufacturing facilities within the boundaries established by the manufacturing specifications and process flows and business policies. Assets include but are not limited to starts, work in progress (WIP), inventory, purchases, and capacity (manufacturing equipment and manpower). Demands include but are not limited to firm orders, forecasted orders, and inventory buffer. The matching must take into account manufacturing or production specifications and business guidelines. Manufacturing specifications and process flows include but are not limited to build options, bills of material (BOM), yields, cycle times, receipt dates (the anticipated dates units of WIP will complete certain stages of manufacturing), capacity consumed, substitution (substitutability of one part for another), binning or sorting (determination of actual part types after testing, and shipping times. Business guidelines include but are not limited to frozen zones (no change can be made on supplies requested), demand priorities, priority trade-offs, preferred suppliers, and inventory policy. Many of the manufacturing specification and business guideline values will change often during the planning horizon (called date effectivity). Matching is a critical component of four SCM activities

The best-can-do (BCD) SCM activity involves determining how to best meet prioritized demand without violating temporal, asset, or capacity constraints. This application minimizes prioritized demand tardiness in establishing commitments and synchronized targets for each manufacturing location. It creates a projection of what can be produced to meet demand that is a key element of the available-to-promise (ATP) type of matching.

The optimal-manufacturing-resource-planning (OMRP) SCM activity is based on the assumption that a manufacturing facility must meet all demands on time. In theory, the BCD activity has to provide each facility with targets that are achievable. The OMRP activity provides detailed instructions about what manufacturing activities must be accomplished and when they must be completed. The instructions concern work in progress (WIP), work to be started, purchases, purchase orders, planned substitutions, and shipments. The objective of the optimization portion of this activity is to select production assignments that minimize new starts and starts in negative time. A sister process reviews the optimized assignments and alerts managers to possible latenesses.

It has increased on-time deliveries to 97 percent in 1999.

The projected-supply-planning (PSP) ac-

tivity creates an estimated supply of finished parts by forecasting the completion date of WIP and starts using standard cycle time and capacity restrictions.

The available-to-promise (ATP) activity enables an organization to dynamically reallocate projected supply in response to incremental changes in the demand statement (new orders arriving, orders being filled, and order changes or cancellations) according to business-policy guidelines, to identify projected shortfalls with respect to committed orders, and to provide real-time order commitments and status.

Business Challenge and the Birth of the PROFIT Team

Between 1992 and 1999, the microelectronics industry went through a dual transformation in core technology and use (or market). On the technology side, chip size, speed, and versatility took quantum leaps. For example, IBM pioneered copper circuits, RISC-based CPU processors, silicon-on-insulator and silicon-germanium technologies, and innovative insulation techniques for copper circuits. The market for microelectronic devices expanded from an initial base in computers to a wide range of products, such as cell phones, car security systems, advanced GPS-based trackers, greeting cards, and aids for the handicapped. Microelectronic devices pervade the world. This dual expansion transformed manufacturing from making large quantities of just a few parts to making varying quantities of numerous parts. To quantify this change, in the mid-1980s, IBM Microelectronics had about 100 active part numbers with demand; in 2000, the number is 6,000.

In addition, the microelectronics industry, like other industries, is under pressure to be more responsive to its customers.

The IBM Microelectronics Division of today—a tightly coupled set of manufacturing facilities that supply diverse products and employ a centralized SCM process that weekly determines which orders the division can meet and responds rapidly to customer requests regarding orders—was only a dream in 1994.

Historically, the manufacturing facilities that produce wafers, devices, modules, and cards, and that are located all over the world, and that are now part of the IBM Microelectronics Division were semi-independent of one another. They supplied the parts for downstream IBM manufacturing facilities for mainframes, workstations, printers, networks, and storage equipment. Their geographic links were strong. Typically, a facility in Europe would provide component parts for an IBM box plant in Europe. Supply chains consisted of individual manufacturing facilities linked directly to the box plants they supported with no centralized control. Also, each manufacturing facility produced far fewer core products and therefore managed far fewer products than it does today.

During the early 1990s, IBM Microelectronics needed to successfully compete in the business of supplying components to original equipment manufacturers (OEMs), to continue to provide the IBM box facilities with state-of-the-art (and often custom) components, and to reduce costs and improve customer satisfaction. To accomplish these goals, IBM Microelectronics had to transform the confederation of manufacturing facilities that supplied

components to the IBM box plants in their areas into a unified division that operated manufacturing facilities worldwide to produce a variety of products for a variety of customers. To transform the division and to improve its performance, IBM Microelectronics launched a major reengineering effort in 1994 to completely restructure its supply-chain-management (SCM) applications (sometimes referred to as customer-order-fulfillment (COF) applications).

In the summer of 1994, the division formed teams to improve specific aspects of the SCM processes and supporting applications and to work together improve the entire system. Separate teams worked on data, business processes, billing and shipping, and demand forecasting. The division asked the PROFIT team to design, build, and deploy a suite of applications to improve manufacturing utilization and responsiveness to customer requests by developing matching tools that could handle the complexity created by the scope and scale of the new unified IBM Microelectronics Division. The authors are the original founding members of the PROFIT team, drive the overall design, and are responsible for the core operations research techniques that do the actual matching. The core decision technology challenge the team faced was to successfully interweave the power of linear programming to search for optimal matches with the power of traditional material-resource-planning algorithms to handle large volumes of parts with detail granularity. The team built and successfully deployed applications to support three matching areas: BCD, OMRP, and ATP.

Technical Overview of the PROFIT BCD and OMRP Matching Applications

The heart of the OMRP and BCD applications is moving work units (WIP or starts) either forward to project completed parts or backwards to determine starts required across the bill-of-material (BOM) chain following the appropriate manufacturing movement rules, such as cycle time, yield, capacity, and product structure. We typically use implosion to estimate what finished goods will be available to meet demand and explosion to estimate what starts are needed at what due dates to insure meeting the existing demand on time.

To review implosion and explosion, we will describe a simple production or bill of material (BOM) flow (Figure 1). The first manufacturing activity is the production of Wafer 2. This manufacturing activity

Continuous incremental improvement is the new paradigm.

has a cycle time of 30 days. That is, it takes on average 30 days to take a raw wafer and create a completed wafer with the part id Wafer 2. The second activity is device production. To create one unit of Device 2 requires 10 days of cycle time and the consumption of one unit of Wafer 2. Module 2 consumes one unit of Device 2 and takes three days to produce. Card 2 consumes two units of Module 2 and takes four days to produce.

We illustrate implosion with the following example (Figure 1). Manufacturing estimates that four units of Device 2 will be available or completed on day 10. This is

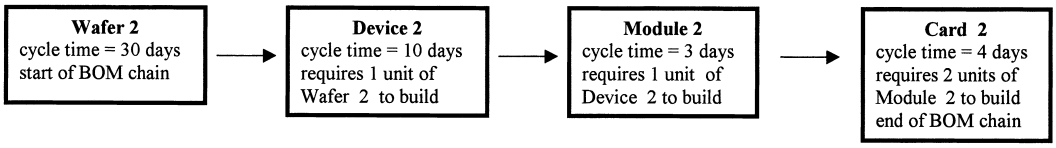

Figure 1: The first manufacturing activity is wafer production. A wafer becomes a device, which becomes a module, which becomes a card. In implosion, WIP is projected forward to its final good form. If four device units will be available on day 10, these four devices are projected to be four modules on day 13 and two cards on day 17. In explosion, a demand is projected backward to determine the starts needed to make this demand. If one card is required on day 20, then two modules are required on day 16, two devices on day 13, and two wafers on day 3.

called a projected receipt. If manufacturing immediately uses these four units to produce Module 2, then on day 13 (10 + Module 2 cycle time = 10 + 3 = 13) four units of Module 2 will be completed. Continuing the projection process, the four units of Module 2 are immediately used to create two units of Card 2, which will be available on day 17 (13 + cycle time for Card 2 = 13 + 4 = 17). The implosion process enables manufacturing to estimate the future supply of finished goods.

We can also illustrate explosion with an example (Figure 1). To meet demand for one unit of Card 2 on day 20, the plant must have two completed units of Module 2 available on day 16 (20 minus the cycle time for Card 2 = 20 − 4 =16). This generates an exploded demand of two units of module 3 with a due date of day 16. To continue the explosion process, to produce the two units of Module 2, the plant must have two units of Device 2 available on day 13 (16 minus the cycle time for Module 2 = 16 − 3 = 13). Next, the device demand is exploded creating a demand for two units of Wafer 2 on day 3 (13 − 10). This exploded information creates the guidelines for manufacturing to meet existing demand. For example, the device department must start production of two units of Device 2 no later than day 3 to meet the demand for one unit of Card 2 on day 20.

Optimized Material Resource Planning Application

The optimized material resource planning (OMRP) application provides detailed guidelines to manufacturing and a detailed estimate of supply. It contains such traditional MRP features as lot-level details, lot sizing, daily time buckets, and the ability to handle partial-day cycle times. In addition, it has the ability to optimally allocate an asset when there are competing demands for this asset. It consists of three core components: the binning-material-resource-planning (BMRP) module, an alternative BOM material-resource-planning (AMRP) module, and the partitioning module (PARTITIONER).

OMRP was designed to handle five core decision technology challenges: (1) simple binning with downgrade substitution; (2) alternative production processes for the same part (alternative BOM structure), complex binning, and general substitution; (3) granularity at the individual manufacturing lots with cycle times that have partial days; (4) handling production-specification information, such as yields,

cycle times, and availability and capacity that change regularly during the planning horizon (date effectivity); and (5) determining the optimal match between assets and demand in a reasonable amount of time (performance).

Simple Binning and BMRP

In the production of devices, simple binning (Figure 2) with downgrade substitution is a common. Binning or sorting refers to assigning a part a specific identity only after testing it. After the wafer is completed and tested, there is a 50 percent chance it will be classified as Device A, a 30 percent chance it will be classified as Device B, and a 20 percent chance it will be classified as Device C. These values are called the binning percentages, and the devices are referred to as coproducts. Also, Device A can be used (substituted) to meet demand for Devices B and C.; Device B can be used to meet demand for Device C. This is called downgrade substitution.

The challenge is to make optimal use of

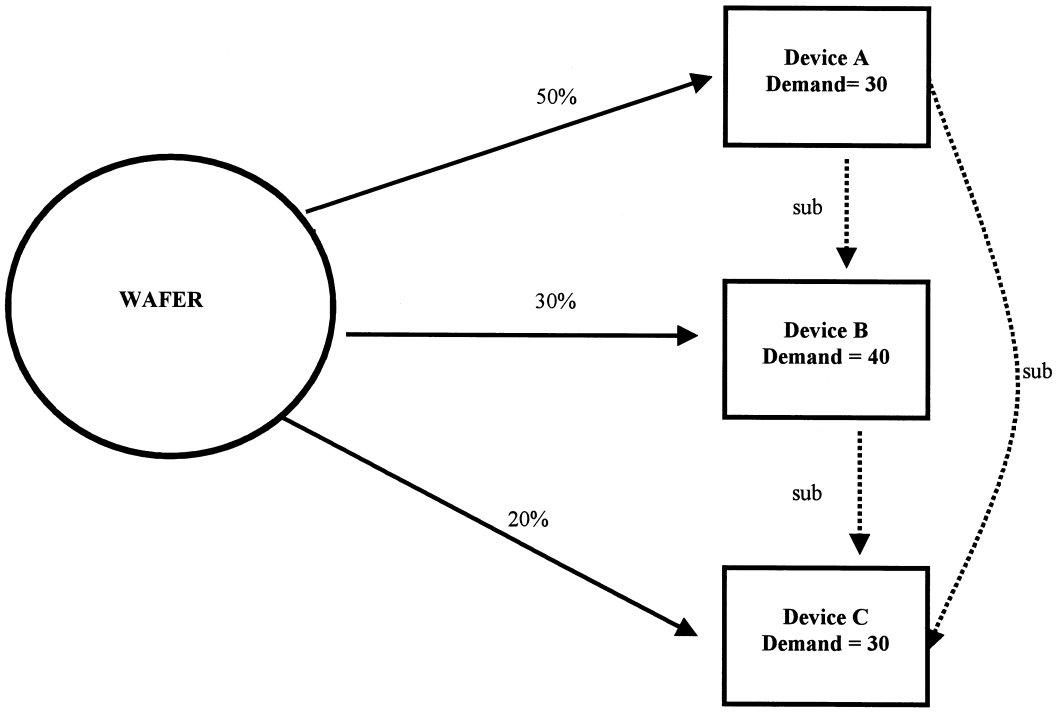

Figure 2: Binning or sorting refers to assigning parts specific identities only after testing them. After the wafer is completed and tested, there is a 50 percent chance it will be classified as Device A, a 30 percent chance it will be classified as Device B, and a 20 percent chance it will be classified as Device C. These values are called the binning percentages and the devices are referred to as co-products. In addition, Device A can be used (substituted) to meet demand for Devices B and C; Device B can be used to meet demand for Device C. This is called downgrade substitution. If the demand for devices is 30 for Device A, 40 for Device B, and 30 for Device C, the challenge is to determine the minimum number of wafers that must be produced to meet all of the demand. If we optimally account for coproducts and substitutions, the minimum number of wafers required to meet this demand is 100. Starting 100 wafers creates 50 of A, 30 of B, and 20 of C. The extra 20 of A are used to cover the shortfall of 10 of B and 10 of A.

coproducts and substitution to avoid overstating the required starts needed to meet demand. If the demand for devices is 30 for Device A, 40 for Device B, and 30 for Device C, the challenge is to determine the minimum number of wafers that must be produced to meet all of the demand. If we optimally account for coproducts and substitutions, the minimum number of wafers required to meet this demand is 100. Starting 100 wafers creates 50 of A, 30 of B, and 20 of C. The extra 20 of A are used to cover the shortfall of 10 of B and 10 of A.

Other factors complicate the determination of the minimum number of starts required to meet demand in simple binning production structures: demands for devices spread throughout the planning horizon, existing inventory, projected completion of WIP, and binning percentages and allowable substitutions that change during the planning horizon (date effectivity).

The PROFIT team developed the BMRP module of OMRP to handle simple demand, which includes simple binning. Parts that are classified as *simple demand* have no alternative BOM structures, general substitution, or complex binning in their production path. These demands may have one or more simple binning procedures in their production paths. The BMRP module is similar to a traditional MRP explosion algorithm except at simple binning points within the BOM chain. At these manufacturing activities, a small LP is invoked to determine the optimal required starts. The PROFIT team developed an LP formulation that calculates the minimum number of required starts each day spread across a planning horizon for a specific binning activity, algorithms and data structures to dynamically identify each instance of simple binning within a traditional MRP explosion process, and linkages to dynamically execute the binning LP model as needed. We provide details of the binning LP formulation in the appendix.

Alternative BOM Structures, General Substitution, and Complex Binning

Within the production of modules, an increasingly common manufacturing characteristic is alternative production options (called alternative bill of material (BOM) structures), general substitution, and complex binning. Complex binning refers to a situation in which one binning activity immediately invokes another or in which substitutions are permitted across binning activities. When alternative methods are available to produce a part, the tool to handle explosion must select one of two or more paths in propagating demand back through the BOM structure. With alternatives comes a need for a search to find the best alternative.

For example, two processes (P1 and P2) can be used to build Module 8 (Figure 3). The P1 process consumes Device 8A, and the P2 process consumes Device 8B. The explosion engine must determine how to explode demand for Module 8 back to the device level: half to P1 and half to P2, two-thirds to P1 and one-third to P2, or all to P1? The objective is to divide the demand for Module 8 across P1 and P2 to make best use of existing inventory and WIP, to minimize new starts, and to meet other relevant guidelines (for example, sharing percentages). Determining the best result requires an extended search through

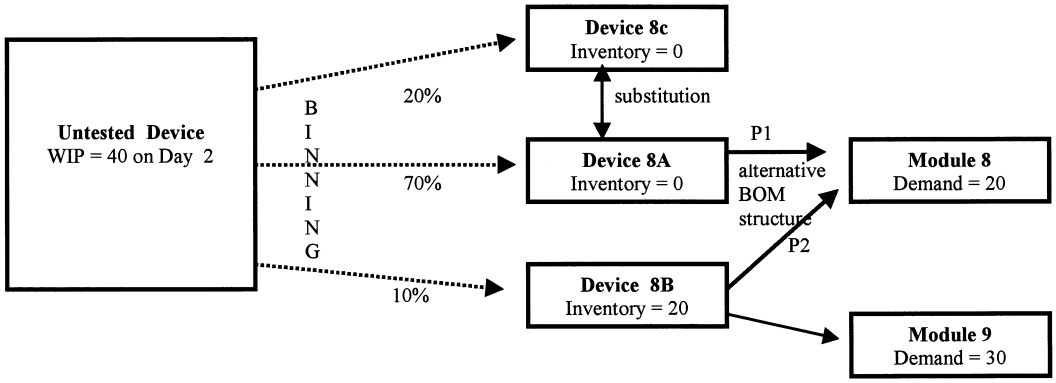

Figure 3: In this complex BOM chain, 40 units of WIP at untested device will be available on day 2. The binning percentages for Devices 8C, 8A, and 8B are 20 percent, 70 percent, and 10 percent, respectively. Module 8 can be made by two different manufacturing processes. Process 1 consumes Device 8A and Process 2 consumes Device 8B. Device 8C can be substituted for Device 8A. Module 9 can be built only with Device 8B. There are 20 units of demand for Module 8 and 30 units for Module 9. The objective for an intelligent explosion algorithm is allocating demand for Modules 8 and 9 back across production to make best use of existing inventory and WIP, to minimize new starts, and to meet other relevant guidelines.

the entire BOM structure.

The demand for Module 8 is 20 units. Twenty units of Device 8B are in inventory, which can be used to make Module 8 with process P2. Also Device 8A can be used to build Module 8 with process P1. There is no current inventory of Device 8A. Most of the search engines in heuristics that guide explosion through alternative BOM structures would explode the 20 units of demand for Module 8 down the P2 leg or process. However, a broader search would uncover available options among untested devices and avoid the conflict for Device 8B between Module 8 and Module 9.

There are 30 units of demand for Module 9, and Module 9 can be made only from Device 8B. Are there other options to meet the demand for Module 8? There are 40 units of projected WIP among the untested devices; after binning or sorting, eight will become Device 8C, 28 will become Device 8A, and four will become Device 8C. Since Device 8C can generally be substituted for Device 8B, there are 36 (8 + 28) future devices that can be used to produce Module 8 but not Module 9. It is probably not optimal to explode the demand for Module 8 down the P2 leg.

When alternative BOM structures and general substitutions are part of the manufacturing process, explosion is challenging. Because so many factors are relevant (inventory, WIP, demand, binning percentages, permissible substitutions) at multiple bill-of-material levels and because these factors vary across time, we thought that linear programming was the best way to solve such problems.

Demands for part numbers that have alternative BOM structures, general substitution, or complex binning as part of their production process are called *complex demands*. To optimize the explosion-process decisions for complex demands, we devel-

oped the large LP. We represented the entire explosion process in the large LP equations, and the implementation is completely data driven. The core decision variable of the large LP is the quantity of starts at each manufacturing activity during each time bucket. The objective is to minimize a weighted average tardiness in meeting demand on the date requested. Each demand is placed in one of multiple demand classes. The material balance equations represent the entire BOM chain and asset (starts and WIP) movement in all of their complexity. These equations accommodate binning, alternative BOM structures, substitution, different shop calendars, date effectivity, capacity, and sourcing. They insure temporal feasibility. The large LP supports variable time buckets. The large LP can be used as an MRP tool or a BCD tool with only minor adjustments.

Lot Sizing, Lot Identity, and Daily Granularity

The large LP does a superb job optimizing across the complexities created by binning, substitution, and alternative BOM structures. However, it lacks three key features of traditional MRP explosion algorithms: lot sizing, maintaining lot identity, and daily granularity. An application to provide daily manufacturing guidelines without these three features is worthless.

The key challenge in traditional explosion is allocating exploded demand among the alternative paths generated when binning, substitution, or alternative BOM structures occur in the production process. The key challenge in linear programming is maintaining detail granularity. Traditionally, linear programming models aggregate production information into time buckets and part buckets. As a result, critical lot-level and daily detail information is lost. The PROFIT team developed the advanced material resource planning (AMRP) module, which contains a unique method to interweave these two decision technologies to gain the best of each. AMRP invokes, when necessary, the large LP at each low-level code iteration of a traditional MRP explosion process to identify how to optimally allocate substitutions, select from alternative BOM paths, and calculate starts at binning. Low-level code refers to assigning each manufacturing process a number indicating its level in the manufacturing process. Manufacturing activities that produce parts used to meet customers' demands and never consumed by any other manufacturing activity are assigned a low-level code value of 1. In the example in Figure 1, the manufacturing activity Card 2 has a low-level code of 1. Manufacturing activities that produce parts that are consumed only by manufacturing activities assigned a low-level code of 1 are low-level code 2. Low-level code 2 activities may produce parts that are shipped directly to customers in some cases, but in at least one instance, the parts are consumed by a low-level code 1 manufacturing activity. In the example in Figure 1, the manufacturing activity for Module 2 has a low-level code of 2.

PROFIT interweaves traditional MRP explosion and the large LP in AMRP (Figure 4). First AMRP converts all alternative BOM structures to equivalent substitution structures. It then runs the large LP on all low-level codes (all parts from module

through wafer), posting the optimal use of substitutions as receipts (expected completion dates of parts). AMRP then invokes the MRP explosion engine and explodes demands from low-level code 1 (LLC-1) to LLC-2 using the posted receipts to guide its allocations across alternative BOM structures and its consumption of the existing substitutable supply. In the example in Figure 4, the large LP will instruct the traditional MRP explosion how to allocate exploded demand for Modules 1 and 5 between the two alternative process available for each (P1 and P2). At LLC-2, there are no complex substitutions or alternate processes for these raw modules. Consequently, traditional MRP logic is sufficient and there is no value to running the more time consuming large LP at LLC-2. The LLC-2 demand is the exploded demand from the low-level code 1 explosion plus any independent demand for LLC-2 parts. MRP logic then explodes this demand to LLC-3. At LLC-3, complex substitutions are permissible, so it is necessary to run the large LP from LLC-3 (devices, wafers, and common wafer) to optimize this level. The large LP will provide guidelines at LLC-3 on how to use substitutions between Devices 1A2 and 1B2 optimally.

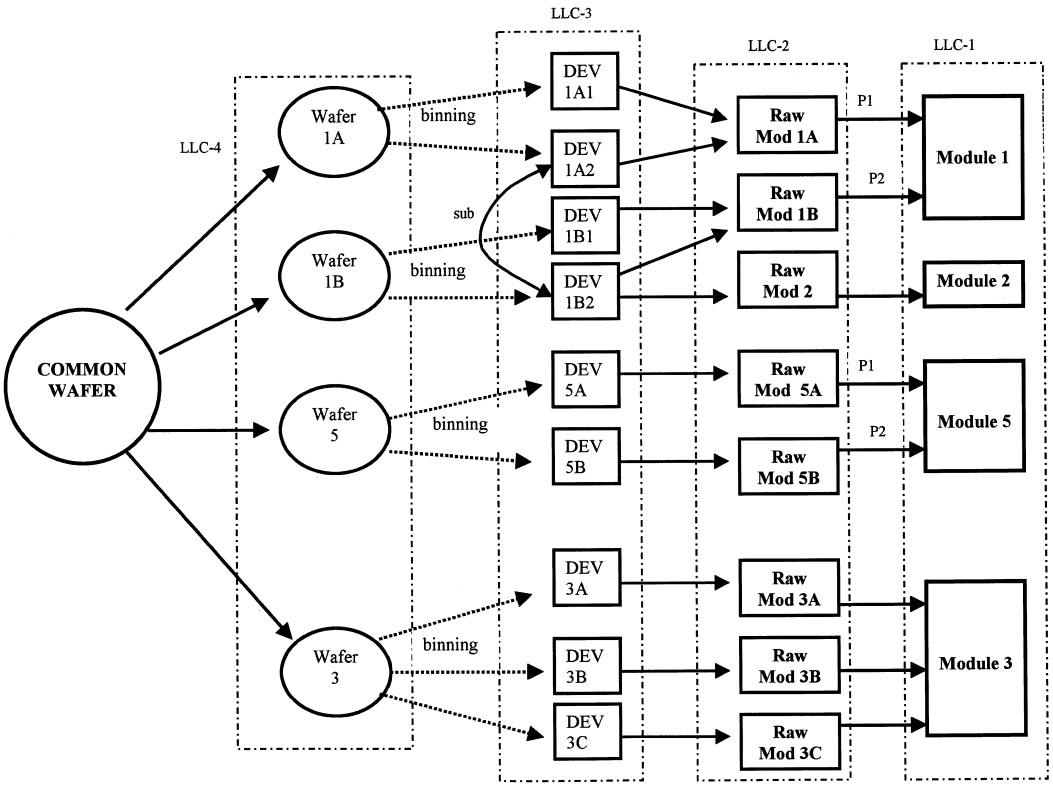

Figure 4: This complex BOM flow has binning, alternative BOM, and substitution. Each group of parts has a low-level code (LLC). The AMRP interweaves LP and a traditional MRP explosion algorithm by running the large LP at each level when needed to select the optimal use of substitutions and alternative BOM paths. The explosion algorithm then uses this information in a traditional one-level explosion.

Subsequent LLC iterations are not complex, so the large LP is not used again.

Performance

To resolve the performance challenge, the PROFIT team deployed two dynamic partitioning strategies and implemented parallel computation on an IBM supercomputer. In the first partition, the PARTITIONER module divides demands and their BOM parts and structures into two groups: complex and simple. The complex group is solved with AMRP, and the simple group is solved with BMRP. In the second partition, this module divides the complex group into logically independent groups that can be solved in parallel by separate LP runs.

The Division-Wide Best-Can-Do (BCD) Application

OMRP and BCD support different but related business functions. Both need the ability to move WIP and starts across the BOM chain and search for optimal matches. The BCD application determines which demand can be met when. Its focus is on the overall allocation of manufacturing assets among competing demands. The OMRP assumes that the demand it is given by the BCD application can be met, and it focuses on the optimal use of existing assets and providing detailed guidelines to manufacturing. The BCD does not need to accommodate such details as lot sizing, lot identity, and daily granularity. Daily granularity means identifying the arrival and departure of an asset from a manufacturing activity on a specific date (and time if needed). The BCD is able to operate with time buckets, for example, identifying the arrival or departure of an asset to a specific week instead of a specific day. The core decision challenge to the BCD application is intelligently allocating manufacturing assets to a prioritized demand spread across a planning horizon to minimize total tardiness when different demands have different priorities and therefore different tardiness weights

Module 1 and Module 2 are both made from Device 12 (Figures 5 and 6). Demand for Module 1 is 10 units on day 10 (demand C) and 15 units on day 12 (demand D). The cycle time to build Module 1 is 10 days. The demand for Module 2 is eight units on day 5 (demand A) and two units on day 6 (demand B). The projected supply for Device 12 is 10 units on hand now (0 days), 30 units on day 2, and 20 units on day 10. The problem is how to allocate the anticipated supply of Device 12 parts between Module 1 and Module 2.

Option 1 (Figure 5) might be: (1) immediately allocate eight of the 10 units of Device 12 on hand to meet demand A (eight units of Module 2 on day 5) one day early (on day 4 = 0 + 4); (2) to immediately allocate the remaining two units of Device 12 on hand to meet demand B (two units of Module 2 on day 6) two days early (on day 4 = 0 + 4); (3) on day 2, to allocate 10 units of the projected supply of 30 units of Device 12 to demand C (10 units of Module 1 on day 10) two days late on day 12 (12 = 2 + 10); (4) on day 2, to allocate 15 units of the projected supply of 30 units of Device 12 to demand D (15 units of Module 1 on day 12) on time (12 = 2 + 10). The score card for this solution is demand A early, demand B early, demand C late by two days, and demand D on time.

Figure 6 shows a second option. By searching, we can find alternative solu-

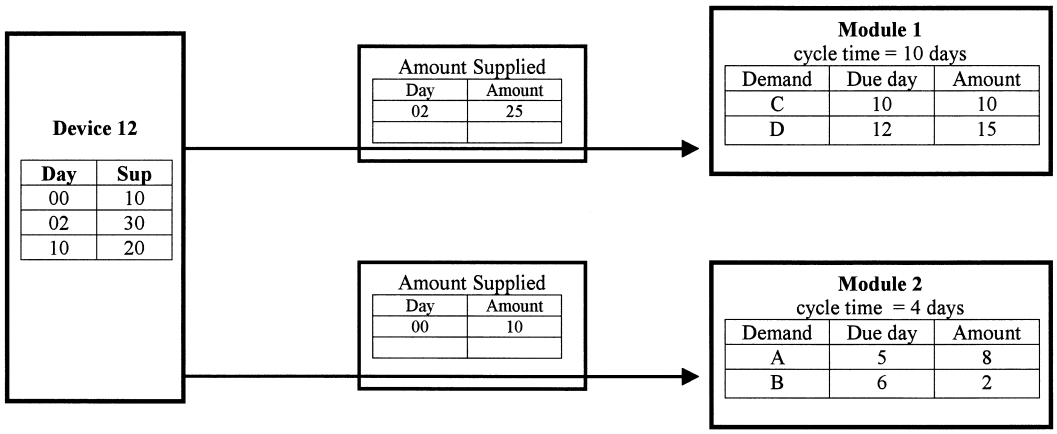

Figure 5: In this on-time delivery example showing option 1, Module 1 and Module 2 are both made from Device 12. The cycle time is 10 days for Module 1 and four days for Module 2. The demand for Module 1 is 10 units on day 10 and 15 units on day 12. The demand for Module 2 is eight units on day 5 and two units on day 6. In this solution, 10 units of Device 12 are allocated to Module 2 on day 0 and 25 units of Device 12 are allocated to Module 1 on day 2. With this allocation, demand A is met one day early, demand B is met two days early, demand C is met two days late, and demand D is met on time.

tions to the problem. The key questions for the search engine are when should it truncate the search and how should it evaluate the relative merits of the alternatives. For example, if the priority on demand A (eight units of Module 2 on day 5) was much higher than the priority on demand C (10 units of Module 1 on day

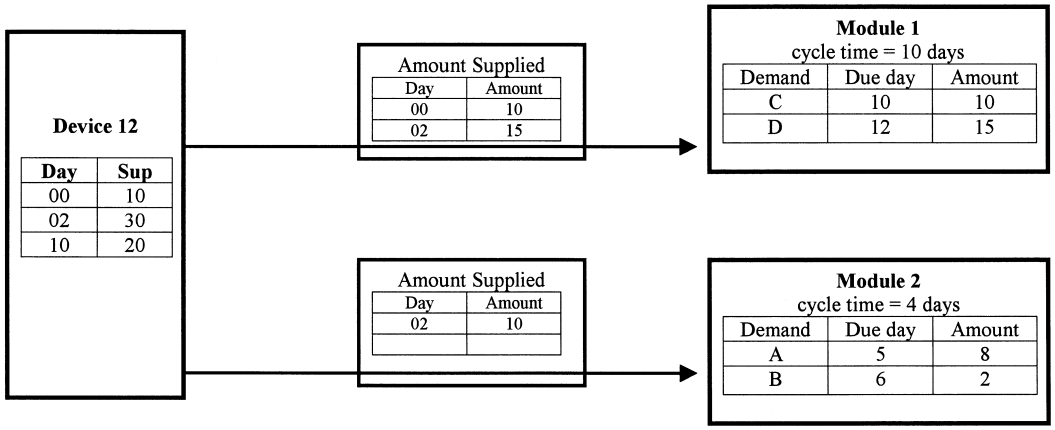

Figure 6: In this on-time delivery example showing option 2, Module 1 and Module 2 are both made from Device 12. The cycle time is 10 days for Module 1 and 4 days for Module 2. The demand for Module 1 is 10 units on day 10 and 15 units on day 12. The demand for Module 2 is 8 units on day 5 and 2 two units on day 6. In this solution 10 units of Device 12 are allocated to Module 1 on day 0, 15 units of Device 12 are allocated to Module 1 on day 2, and 10 units of Device 12 are allocated to Module 2 on day 2. With this allocation demand A is met 1 day late, demand B is met on time, demand C is met on time, and demand D is met on time.

10), option 1 would probably be preferred over option 2.

The PROFIT team relied on three core components in designing the BCD application: (1) the large LP to handle complex demand, (2) the IMEX BCD heuristic to handle simple demand, and (3) a dynamic partitioning algorithm to split demand between the two BCD solvers and to create parallel computational opportunities.

The IMEX BCD heuristic is a high-speed heuristic relying on an explode/implode paradigm to match assets with demand in the best way possible. This MRP-based heuristic has three parts. We use a special variation of the BMRP to explode demand across the entire BOM chain. During the explosion portion, the heuristic optimizes simple binning situations and analyzes key resultants against constraints to establish guidelines for the subsequent implosion. After completing the explosion, the application posts three files: capacity required (at each manufacturing point where capacity is measured), starts required, and need dates for all projected receipts (need dates for anticipated completions of WIP). The starts file lists demand priorities for the entries based on the original demands. In the second step, a user or another program can modify the starts file, the projected-receipts file, and the capacity-available file. Third, an implosion-based heuristic creates a projected supply of finished goods and estimated commitment dates for demands that meet all constraints (temporal, asset based, and business policy). IMEX runs substantially faster than the large LP for BCD but is not as intelligent. IMEX's major weakness is handling complex demands.

Business Value of the PROFIT Matching Applications

The matching applications (1) improved manufacturing productivity or tool utilization and increased inventory turns, (2) improved customer relationships, and (3) improved IBM's ability to respond to emerging market opportunities.

IBM Microelectronics substantially improved its use of assets: (1) In 1999, it improved its use of assets by $80 million, and (2) in 2000, it anticipates improvement of over $200 million.

It has also improved its responsiveness to its customers: (1) The average order-response time (elapsed time between the initial customer-order request and IBM's commitment to deliver) has dropped from almost four days to 0.6 days as of March 2000. For 70 percent of the orders, the response time is under 0.3 days, and for 90 percent, the response time is under one day. For many of the orders that come in electronically, the order commitment occurs within a few minutes. (2) It has improved its ability to commit to the initial customer-request date from 10 to 40 percent. (3) It has increased on-time delivery from 90 percent in 1998 to 97 percent in 1999.

Conclusion

The challenge the supply-chain management team faced was to reengineer the SCM or COF system to support IBM Microelectronics Division's transformation into a world-class provider of leading-edge semiconductor and packaged solutions for the networked world. To support this effort, the PROFIT team developed a suite of decision technology tools to match

assets with demand that advanced decision technology in our industry. It is the first tool kit to both handle daily material resource planning and determine the best the division can do to meet its demands. It uses a unique interweaving of linear programming and traditional MRP logic and takes full advantage of commercially available supercomputing. This work has made substantive savings in asset utilization and dramatic improvements in order response. Without the reduction of order-response time from multiple days to hours, the IBM Microelectronics Division would not have a customer-order-fulfillment process that was competitive.

Although we are pleased with the work done to date, e-business and the continued drive to improve customer responsiveness create some significant challenges to IBM Microelectronics and opportunities for the PROFIT team to continue to push the envelope of decision technology in support of SCM.

The Web, the Internet, and deep computing are viewed as the core tools to meet the expectation of improved customer responsiveness. The challenge is how to organize and apply these tools and eliminate disconnects in business processes to meet this expectation. This challenge of the disconnects is the opportunity window for PROFIT and the decision-technology community in general. Until this challenge is met, most of the e-business work is limited to moving input and output data from one computing device to another. Imagine what the world would be like if the deployment of the phone had stopped at simply providing the telegraph operator in each town with a voice alternative to Morse code.

In the past, at IBM and at other major organizations, including Franz Edelman Award finalists, the development and deployment of applications like those built by the PROFIT team were viewed as temporary big-bang efforts. Firms put together teams, established target completion dates, built and deployed applications, put the applications into maintenance mode, and dispersed the teams. This big-bang paradigm no longer works in the e-business world—continuous incremental improvement is the new paradigm.

We at IBM Microelectronics and IBM in general have no intention of just resting on our laurels. The next set of enhancements is already on the drawing board. As good as the current applications are, there is room for improvement. We intend to capture this opportunity (Sullivan 1999). What a great time to be an operations research professional.

APPENDIX
Detailed Description of Small or Binning Linear-Programming Model

The purpose of this small LP is to determine the minimum production of the binned part required so that the demand for all output parts is satisfied on time. The output parts ($j = 1 \ldots J$) are the parts that result when the binned part is produced. Usually, these output parts are the same as the parts with demand ($k = 1 \ldots K$). In Figure 2, for instance, the wafer is the binned part and devices A, B, and C are the output parts, each of which has demand.

Definition of Subscripts

j = output part number that results from the binning process ($j = 1 \ldots J$).

k = part number with demand ($k = 1 \ldots K$).
t = time period ($t = 1 \ldots T$).
Decision Variables
P_t = production of the binned part in period t.
S_{jkt} = quantity of output part j used to satisfy demand of part k in period t (j may equal k).
I_{jt} = inventory of output product j at the end of period t.
Z_{kt} = unsatisfied demand of part k during period t.
Constants
D_{kt} = demand for part k in period t.
I_{j0} = inventory of output part j available at the beginning of the horizon ($t = 0$).
R_{jt} = fixed receipts of part j in period t.
B_{jt} = binning percentage, i.e. percentage of output part j resulting per piece of production of the binned part in period t (includes yield as well as binning distribution).
Objective Function

Minimize

$$\sum_t \left[P_t + \sum_j \left[10^9 \, Z_{jt} + 0.0001 \, I_{jt} + \sum_k 0.00001 \, S_{jkt} \right] \right] (0.9)^t.$$

Constraints

$$I_{jt} = I_{jt-1} + B_{jt} \, P_t + R_{jt} - \sum_k S_{jkt}.$$

$$D_{kt} = Z_{kt} + \sum_j S_{jkt}.$$

$$I_{jt} >= 0, \; P_t >= 0, \; Z_{kt} >= 0,$$

$$S_{jkt} >= 0, \ldots j, \; t, \; k.$$

References

Sullivan, G. 1989, "IBM Burlington's logistics management system (LMS)," *Interfaces*, Vol. 20, No. 1, pp. 43–61.

Sullivan, G. 1999, "Supply chain management, decision technology, and e-business information technology at IBM Microelectronics," *MicroNews*, a publication of the IBM Microelectronics Division, Vol. 5, No. 4, pp. 18–21, www.chips.ibm.com/micronews/vol5no4/fordyce.html

---

Nick Donofrio, IBM Senior Vice President, Technology and Manufacturing, made the following remarks on videotape for the Edelman 2000 competition: ". . .The fundamental challenge was to reengineer the supply-chain management system to make our Microelectronics team the world-class provider of semiconductor and packaged solutions for the networked world. The suite of models that the PROFIT team developed really pushed the envelope of applied decision technology in our industry. The result has been substantive savings in asset utilization as well as notable improvements in order response.

"Crucial elements of IBM's revitalization over the years have been precision teamwork and breakthrough thinking. The PROFIT team has consistently demonstrated both qualities. I am deeply proud of their success and equally proud of the entire IBM Microelectronics supply-chain team. . .

"We're all familiar with the Internet and the Worldwide Web. Based on open standards, those technologies have erased the barriers to information flow among organizations and, most important, have transformed the way people approach computing.

"With e-business, customer responsiveness takes center stage. The ability to simultaneously respond to customers' needs and emerging business opportunities in an intelligent, orderly manner is a survival requirement for today's marketplace. Our customers continue to tell us that the quality of our responsiveness is as important

to them as the quality of our products. . .

"E-business creates the expectation of improved customer responsiveness. . .The challenge becomes the opportunity for the decision technology community. The work done by the innovators on our PROFIT team, as well as their colleagues, has enabled us to meet the customer responsiveness challenge. And I'm convinced that the monetary savings associated with the PROFIT work—though significant in their own right—understate the work's true value to IBM.

"Operations research professionals are the key to harnessing the opportunity created by e-business and deep computing. I'm convinced that organizations that make the best use of decision technology are those that will be the most successful from now on."

Stu Reed, IBM Vice President, Integrated Supply-Chain Development and Deployment, made the following remarks at the Edelman 2000 competition: "My job is to ensure excellence across IBM's supply chains and communicate IBM's intent to become a premier e-business. What we have here is a process transformation that was hardened by the application of information technology that was set upon the fundamentals of key operations research theory and—most importantly from my boss's perspective—we created a lot of business value.

"We had to capture customers that had lifelong relationships with other suppliers. To do that, albeit semiconductor technology is one differentiator, the other key differentiator was going to be our ability to respond. . .We knew the supply-chain design we had—which we affectionately called a coagulation of warring nations—would not work. We knew an integrated centrally run demand/supply, demand management, and available-to-promise process was the only way we were going to survive. . .

"What used to take a week was now two to two and a half hours of validation within a weekly run, enterprise wide for this division. . .This is the quickest, most comprehensive development and deployment of process and systems function ever done in the semiconductor industry unequivocally and, to the best of our knowledge, within the total high tech industry.

"Prior to the deployment and during the deployment, we benchmarked. . .we confirmed. . .we chose to go best-in-class. . .

"The hard science is the basis for the trust in the data. Without people trusting the data that came out of the central engine, the entire transformation would have been halted. . .We combined process reengineering cemented by the application of I.T. based on operations research hard science to bring true business value to the IBM Microelectronics Division."