Journal of the Operational Research Society (2012) 63, 1566–1577

www.palgrave-journals.com/jors/

Optimized material requirements planning for semiconductor manufacturing

RJ Milne[1,*], C-T Wang[2], C-KA Yen[3] and K Fordyce[4]

[1]*Clarkson University, New York, USA;* [2]*National Central University, Jhongli City, Taiwan;* [3]*Wolverine Decision Technology Inc., California, USA; and* [4]*IBM Corporation, Essex Junction, Vermont, USA*

This paper describes a custom operational research algorithm, which is run nightly by IBM to create a material requirements plan for its semiconductor fabrication facility in Vermont, USA. To model alternative manufacturing processes and part substitutions, this application interweaves linear programming and heuristic methods to reap the benefits of each decision technology. At each level of the bills of materials supply chain with complex decision choices to be made, parallel linear programmes are invoked and their results are fed into a material requirements planning (MRP) heuristic, which processes parts through multiple iterations. The results from processing one level of the bills of materials supply chain are exploded to create demand for the next level and the interweaving of the two decision technologies continues. The algorithm creates recommended manufacturing releases and work-in-process priorities. These outputs point out opportunities for improvement in order to satisfy all demands on time. The output can be interpreted with well-known MRP assumptions.

Journal of the Operational Research Society (2012) **63,** 1566–1577. doi:10.1057/jors.2012.1
published online 22 February 2012

Keywords: practice of OR; production; material requirements planning (MRP); heuristics; linear programming

Introduction

Material requirements planning (MRP) systems have been in use for decades (Orlicky, 1975) and remain popular. According to a survey by Jonsson and Mattsson (2006), 75% of manufacturing companies use MRP as a main method of material planning. As observed by Pandey *et al* (2000), Ornek and Cengiz (2006), and Taal and Wortmann (1997), the relative simplicity of MRP systems makes them preferred by many over math programming approaches. Users understand MRP logic and develop a good understanding of the relationships between MRP inputs and outputs. This understanding helps the users identify which input data are in error and which inputs need to be improved to obtain better outputs. MRP users can be confident in their production plans in contrast to those output by 'black box' math programmes. Ornek and Cengiz (2006) go so far as to say, 'MRP planners require decision-support rather than decision-making software'. Clearly, MRP systems are important in practice. It is within this context that we present an optimized MRP system, which is being used by IBM to create material requirements plans at its semiconductor fabrication facility in Vermont, USA. This system contains an original custom algorithm, which interweaves linear programming and traditional MRP technologies. Our approach allows the system to retain the key MRP input/output relationships, which are easily understood by planners. Furthermore, the integrated optimization capability provides smart decisions for complex situations. The contribution of this paper is the original methodology of this novel hybrid algorithm we created.

MRP systems operate at a part number/plant level of detail. They explode end-item demands through a bills of materials supply chain to determine planned manufacturing releases and planned purchase orders. In addition, they determine need date priorities for existing purchase order receipts and work-in-progress (WIP) jobs. These priorities are determined by matching the independent and dependent demands for a part with existing receipts (eg, projected WIP and firm purchase orders) being used to support these demands. After these existing receipts have been netted from the demands, the remaining uncovered demands are adjusted for yields and lead times to create new planned receipts (planned manufacturing releases and planned purchase orders). The planned manufacturing releases are exploded to create dependent demands on the part's components. Because of the timing of the demands and the changing availability of covering receipts—due to stochastic yields and lead times—often MRP will suggest that some activities should have occurred in the past, for

*Correspondence: RJ Milne, School of Business, Clarkson University, 107 B.H. Snell Hall, PO Box 5790, Potsdam, NY 13699-5790, USA.
E-mail: jmilne@clarkson.edu

instance that a planned manufacturing release should have been released 4 days ago to meet demand on time or that a WIP job should have been processed through an additional 6 days of operations already.

Harrison and Lewis (1996) and Billington *et al* (1983) have suggested modifying MRP systems so that their outputs are made feasible automatically. This approach ignores crucial business processes called 'chase' or 'recovery'. Prior to running MRP calculations by site, IBM conducts division planning to establish a centralized supply chain plan and applies this plan to all sites within its Microelectronics Division. Each site in this Division thus obtains a shipment plan, which was capacity and lead time feasible at the time of its creation. Each site is also obligated to deliver shipments according to the central plan. Those shipments are then fed as independent demands into the MRP systems used by the IBM Vermont semiconductor fabrication facility and other sites. When the site MRP runs create infeasible outputs, this suggests areas where improvements need to be made. For instance, when capacity infeasibilities result, either more capacity must be obtained (eg, set up equipment to run a different product) or workload must be satisfied in a different manner (eg, small shift in timing of preventive maintenance). Or, when lead time infeasibilities occur, this suggests job priorities need to be adjusted so that late jobs are completed in faster than normal remaining lead time. Because of IBM's business processes, the infeasibilities resulting from MRP plans are helpful as they point to areas that need improvement, and human judgement will determine which opportunities to pursue (on the other hand, an automatically generated feasible plan would rarely satisfy all demands on time).

The semiconductor industry is the epitome of high-tech manufacturing and requires intensive capital investments. IBM's flagship semiconductor facility cost over US$4 billion for the building and equipment. The manufacturing process has hundreds of processing steps resulting in lead times of several months. The process begins by cutting raw wafers out of silicon ingots. Then, on the surface of the wafer, four essential manufacturing steps are repeated to build circuit components (such as transistors, resistors, etc) and interconnect these components to form integrated circuits, one layer at a time. These four essential steps are *deposition*, *photolithography*, *etching*, and *ion implantation/wiring*. Depending on the design, these four steps may be repeated dozens or hundreds of times, and the result is a three-dimensional, layered circuit structure built on the surface of the wafer. Finished wafers are cut (diced) into *devices*, each of which may contain millions of circuit components with lines tens of nanometers in width. Devices are tested to separate the working units from those that do not function. Working devices are further tested and sorted into broad functional categories (such as speed or power consumption), followed by a sequence of assembly steps to mount sorted devices onto a substrate to make a *module*. Modules go through a series of testing. The semiconductors produced by the IBM Vermont facility are used for a variety of purposes including cell phones, global positioning systems, digital cameras, and government applications. For more detail on semiconductor manufacturing, please refer to Denton *et al* (2006), Mönch *et al* (2011), and Lyon *et al* (2001).

Figure 1 is a simplified representation of bills of materials product flows at the IBM Vermont site. Although simplified, this figure contains the major complexities that the site needs to address. The figure shows two product lines. Upon the completion of wafer fabrication in the new product line, untested devices (UD) on the wafer (WAF) are tested for speed: 60% of the UD will be characterized as fast devices (FD), 30% as medium devices (MD), and 10% as slow devices (SD). This characterization process is called *binning* and can be described as, 'UD bins into FD, MD and SD with proportions 60, 30, and 10%, respectively'. These proportions are the result of statistical fluctuations in the semiconductor manufacturing process. A second set of tests bins the unsorted modules (UFM, UMM, and USM) into fast, medium, and slow modules (FM, MM, and SM). Because module speed sorting follows device speed sorting, this is referred to as *sequential binning*. The same scenario can also be observed in the old product line, which is shown on the right side of Figure 1 (the 'O' prefix distinguishes parts of the old product line from parts of the new product line).

Binning is often associated with *substitutions*. In Figure 1, for example, the fast module of the old product line OFM can be used as a substitute for slow module OSM if a shortage occurs for OSM. Substitutions can also occur across different product lines, such as the slow module of the new product line SM substituting for both fast and slow modules of the old product line OFM and OSM. We call this type of substitution *complex substitution* because the substituting part has a different component than the part being substituted. Also seen in Figure 1, there are parts that can be made with different component parts, such as the fast module of the old product line OFM, which can be made from OUFM or OUSM. These are represented as *alternative processes* and we use P1, P2 and so on to distinguish these different manufacturing processes making the same part. For instance, the processes of making OFM using OUFM or OUSM are denoted as P1 and P2, respectively.

Because traditional MRP systems plan one part at a time, they are not well suited for making good choices when planning the parts of Figure 1. There are many alternative paths for satisfying the end-item demands of Figure 1. For instance, the demand for fast module of the old product line OFM can be satisfied through OUD binning to OFD resulting in OUFM, which is binned to

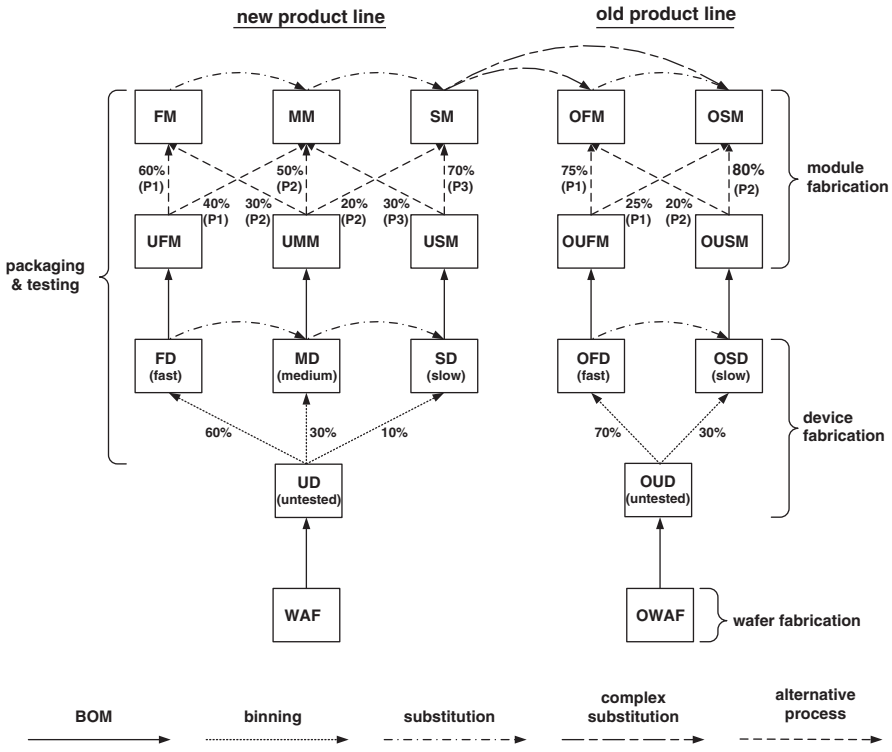

Figure 1 Double speed-sort semiconductor manufacturing processes.

OFM; this demand can also be satisfied through OUD binning to OSD resulting in OUSM, which also bins to OFM; or, the same demand can also be satisfied through a complex substitution by the slow module of the new product line SM. The optimal supply paths through these product flows depend upon the relative demands for products over time, the time-varying binning percentages, the finished and in-process inventory of each part and so on. These interdependencies suggest that the entire product structure of Figure 1 should be planned simultaneously. Linear programmes (LPs) have been widely adopted to make good decisions in such contexts where there are many alternative paths to choose. By solving a set of constraints simultaneously with respect to an objective function, LP technology is superior to traditional MRP heuristics in handling the complexities of Figure 1.

In the context of IBM's business processes, however, LPs have drawbacks not held by traditional MRP logic. First and foremost, LPs create feasible plans, but IBM prefers infeasible plans. Infeasibilities point out areas where improvements are needed, as elaborated above in the third paragraph of this Introduction. Second, because of run-time considerations, LPs will aggregate supply chain activities into multi-day time periods, whereas MRP—being a fast heuristic—can process in daily granularity. This accuracy of MRP is advantageous in the context of site planning. Another drawback of LPs is their lack of integral considerations such as lot sizing. LPs already run much slower than MRP, and the incorporation of lot sizing would only make run times even longer. Conversely, lot sizing does not create much burden on MRP run times. In addition to LPs having longer run times, they also consume more memory and disk space than MRP. Finally, an LP-only solution would not map well to the established approaches used by the analysts and familiar to management at IBM.

So how can we choose alternatives wisely for building parts with complex product structures (as can be done with an LP) while driving to meet all demands on time with a reasonably fast run time (as can be done with a traditional MRP)? Our initial approach was to model infeasibilities within an LP and penalize those infeasibilities. This would have increased the problem size beyond the capacity of available computers to deliver acceptable run times. Consequently, we developed a solution called Optimized Material Requirements Planning (OMRP) by blending LP decision technology with traditional MRP decision technology. At some levels of the bills of materials supply chain, OMRP first runs an LP to make optimal choices of supply paths (eg, when to substitute and when to use alternative processes and their corresponding bills of materials). OMRP feeds those LP choices into an MRP heuristic, which adjusts the material plans so that all demands are met on time (Milne *et al*, 1999). Because the OMRP output looks as if it was created by a traditional MRP (but smarter), many of the input/output relationships

are well understood by the user who can comprehend which input data should be changed to result in the desired output. Once the LP has determined the supply paths to take, the input/output relationships along a path are straightforward to understand. This blending of the LP and MRP decision technologies provides the understandability, intelligence, and speed that IBM needs.

Observe how our approach contrasts with another approach, which also blends optimization with heuristic methods for material planning: Tang *et al* (2008) use optimization (via Lagrangian relaxation) to find an infeasible solution and subsequently apply a heuristic algorithm to their optimization's output to obtain a feasible solution. In contrast, our method first optimizes to create a feasible solution and then applies a heuristic that typically makes the solution infeasible.

Ornek and Cengiz (2006) interweave mixed integer programmes (MIPs) and MRP to create capacity-feasible material requirements plans. They do this by first running a lot size-relaxed MIP to establish capacity-feasible gross requirements for end-item demand. These gross requirements are fed as inputs to MRP. At each level of the bills of materials, the MRP lot sized planned manufacturing releases are fed into another MIP that adjusts them to restore capacity feasibility. The resulting planned manufacturing releases are exploded to create gross requirements for their component parts. Observe that their method feeds MRP results to an MIP at each level, whereas our method proceeds in the opposite direction feeding LP results to MRP at each required level. A greater difference is that Ornek and Cengiz do not consider substitutions or alternative bills of materials, whereas consideration of both of these possibilities is a key aspect of our method.

Lin *et al* (2008) propose heuristics for lot release times with a focus on minimizing set-up times and timely delivery for the thin film transistor-liquid crystal display production industry. For the same industry, Lin *et al* (2009) propose math models for the planning of critical components used in the module assembly process. Each module can be assembled from alternative configurations, that is, types and quantities, of components that are purchased from the suppliers. Customers prefer certain configurations, and the firm has a desired ratio for each component of how much to acquire from each supplier. Their math models determine purchase quantities for each component, and thus all customer demands are satisfied while minimizing the differences between the preferred and the actual supply ratios. Their models consider only a single level of components (they do not consider components of components). This contrasts with our model where choices among alternatives at one level are made based on material inventories, yields, substitution possibilities, and binning percentages across multiple levels of the bills of materials supply chain.

Geunes (2003) developed a heuristic method for solving the single assembly level of an MRP problem in which there are alternative component parts to support each product assembly. Geunes (2003) does not handle the complex substitutions or multiple product levels illustrated, for example, in our Figure 1. These same comments apply to an earlier paper by Balakrishnan and Geunes (2000). Similarly, Hung and Wang (1997) handle single-level alternative material planning for semiconductor bin allocation but not the complicated multi-level product structures illustrated in our Figure 1.

Leachman *et al* (2002) solved a materials planning problem for a DRAM semiconductor facility; that facility has orders of magnitude fewer products than the IBM Vermont fabrication facility. Leachman *et al* (1996) solved a large-scale materials planning problem for a different semiconductor facility using linear programming at some levels and MRP at other levels. Neither of the Leachman-related literature addresses alternative bills of materials or complex substitutions.

Carmon and Nahmias (1994), Bitran and Leong (1992, 1995), and Bitran and Gilbert (1994) propose models and solution methods for planning binning and substitutions. These four works protect against the uncertainty of binning output. Because they apply only at a single component level of the bills of materials, their solutions are limited in the context of the IBM environment.

Ram *et al* (2006) assume that the bill of materials component quantities of an assembly part is flexible. They use an LP to determine the component quantities based on target quantities and scheduled receipts of a single level of components.

To conclude this introduction, OMRP addresses the following key features, which are observed at the IBM Vermont semiconductor facility:

- complex substitutions;
- alternative processes that can have different bills of materials;
- more than two levels of bills of materials;
- outputs that point out improvements needed for satisfying all demands on time (ie, infeasible plans) and which can be interpreted with well-known MRP assumptions.

None of the above related papers model more than two of these four features. This paper describes how we did this by integrating the best aspects of linear programming with the best aspects of MRP methodology to determine planned manufacturing releases and purchase order releases, as well as need dates for existing purchase orders and WIP jobs. The remainder of this paper describes our LP model, the key concepts of blending linear programming and MRP decision technologies, how we improved run-time performance, the OMRP algorithm, and finally

a numerical example to illustrate OMRP and its advantages over pure linear programming and regular MRP approaches.

Linear programming model

We present in the following an LP model that is suitable for handling complex product structures discussed in the previous section. This model is a simplified version of the LP model that is being used in the production OMRP algorithm. See Denton *et al* (2006) for details on the complete LP model. We refer to the following model as the 'many-path LP model' to differentiate it from a simpler LP model—which handles only a single product flow path. We discuss the simpler LP model late in the paper.

Indices

j	Time period
m, n	Parts in the bills of materials supply chain
SA_m	Set of assembly parts, which consume component m
e	Manufacturing process or purchase process
SP_{mej}	Set of periods in which if part m is started using process e, the manufacturing will be completed in period j

Objective function coefficients

PRC_{mej}	Cost to build one piece of part m using process e during period j
$SUBC_{mnj}$	Cost to substitute each piece of part n using part m during period j
$INVC_{mj}$	Cost to hold each piece of part m in inventory at the end of period j
BOC_{mj}	Cost to backorder each piece of part m at the end of period j

Parameters

$DEMAND_{mj}$	Quantity requested for part m during period j
$RECEIPT_{mj}$	Quantity of projected WIP and/or purchase order for part m expected to arrive during period j
$QTYPER_{menj}$	Quantity of component part m needed for building each piece of part n during period j using process e
$YIELD_{mej}$	Output for each piece of part m released (started) during period j using process e
CT_{mex}	Cycle time (ie, number of periods from the start to the finish of a part's manufacturing) of starting part m using process e during period x; note that $SP_{mej} = \{x \mid x + CT_{mex} = j\}$

Decision variables

I_{mj}	Inventory of part m at the end of period j
P_{mej}	Start quantity of part m using process e during period j
L_{mnj}	Quantity of part n being substituted by part m during period j
F_{mj}	Quantity of customer shipment during period j to satisfy demand for part m
B_{mj}	Backorder for demand for part m at the end of period j

Objective function

$$\begin{aligned}\text{Minimize} & \sum_m \sum_e \sum_j PRC_{mej} P_{mej} \\ & + \sum_m \sum_n \sum_j SUBC_{mnj} L_{mnj} \\ & + \sum_m \sum_j INVC_{mj} I_{mj} + \sum_m \sum_j BOC_{mj} B_{mj} \end{aligned} \quad (1)$$

Constraints

Material balance:

$$\begin{aligned} I_{mj} = I_{m(j-1)} + RECEIPT_{mj} + \sum_e \sum_{x \in SP_{mej}} YIELD_{mex} P_{mex} \\ + \sum_n L_{nmj} - \sum_n L_{mnj} - F_{mj} \\ - \sum_{n \in SA_m} \sum_e QTYPER_{menj} P_{nej}, \quad \forall m, j \end{aligned} \quad (2)$$

Backorder conservation:

$$B_{mj} = B_{m(j-1)} + DEMAND_{mj} - F_{mj}, \quad \forall m, j \quad (3)$$

Non-negativity:

$$I_{mj}, F_{mj}, B_{mj} \geqslant 0, \quad \forall m, j \quad (4)$$

$$P_{mej} \geqslant 0, \quad \forall m, e, j \quad (5)$$

$$L_{mnj} \geqslant 0, \quad \forall m, n, j \quad (6)$$

The objective function of Equation (1) minimizes the total of production, substitution, inventory holding, and backorder costs across the planning horizon. Note that the coefficients in Equation (1) are not accounting costs; rather, they have been chosen so that the model will yield results that the business desires. For instance, meeting demand on time is the most important business objective, and therefore the backorder cost coefficients (BOC_{mj}) have higher values than the others. The material balance

constraints in Equation (2) ensure that parts going into inventory (in the form of WIP or purchase orders, manufacturing release, or other parts' substitutions) stay there until coming out to substitute for other parts' fabrication, satisfy demand, or be used in the making of their assembly parts. The backorder conservation constraints in Equation (3) ensure that demand not satisfied in one time period is backordered to the next. Finally, non-negativity constraints in Equations (4), (5), and (6) require that all the decision variables in the model be non-negative.

Blending linear programming and heuristics

Although a simplified representation of the real situation, Figure 1 highlights the major complexities the IBM Vermont facility needs to address so that it can deliver the shipments determined by centralized division planning. These major complexities are alternative processes (which may be formed by binning processes producing the same parts, as illustrated in Figure 2) and complex substitutions (as illustrated in Figure 3). To generate optimal material plans for complexities such as these, all parts in Figure 2 should be planned simultaneously, as the solution would depend on these parts' relative demands, inventory assets, yields and so on. Similarly, all parts in Figure 3 should be planned together.

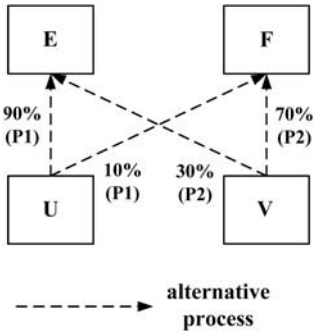

Figure 2 Alternative processes.

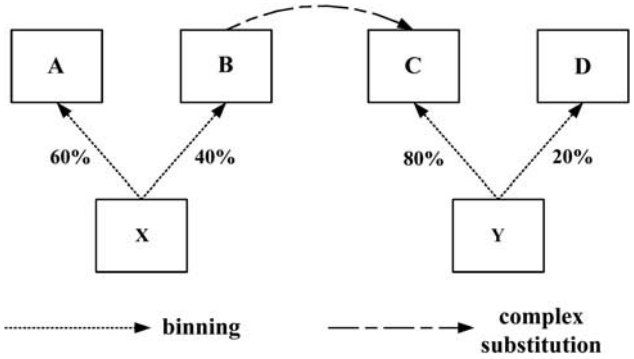

Figure 3 Complex substitutions.

Linear programming is an ideal tool for this task, and we will show how we blended linear programming with traditional MRP heuristics to meet IBM's planning needs.

First, we discuss the scenario of alternative processes, as illustrated in Figure 4: part R can be made via process P1 or P2 using component part S or T, respectively. Observe that Figure 4 is mathematically equivalent to Figure 5 where we have instituted a new, 'dummy' part R' and allow R' to substitute for R at zero cost. This suggests that an LP processing Figure 5 data would result in an equally good solution as an LP processing Figure 4 data. Therefore, we will handle the alternative processes of Figure 4 (a 'multiple process per part' representation) using the substitution possibility of Figure 5 (a 'single process per part' representation). In this modelling approach, if a part has n alternative processes, $n-1$ dummy parts will be created in the new representation, with each dummy part possessing only a single process. Furthermore, all data related to making the part using a particular alternative process—such as yields, cycle times, and processing costs—will be applied to the corresponding new dummy part. In the new representation, there are no longer choices as to which process should be used to make the part—they have all been

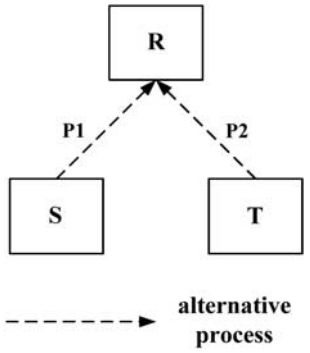

Figure 4 Two alternative processes to make part R (multiple process per part representation).

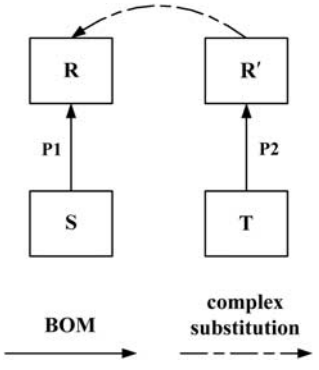

Figure 5 Equivalent single process per part representation for Figure 4.

converted to choices of substitutions (which will be discussed shortly). The modelling approach remains the same when alternative processes are formed because of binning (eg, Figure 2), as binning only affects how many pieces will be available for the substitution but does not change the structure of the new representation (eg, Figure 5).

The above discussion points out that once the 'multiple process per part' representation has been transformed into a 'single process per part' representation, part substitution is the remaining core complexity for creating material plans for the IBM Vermont fabrication facility. To understand the meaning of substitution from an MRP point of view, consider Figure 6, which shows the result of an LP run recommending a substitution of 100 pieces of part *C* using part *B* in some time period. This substitution means that 100 pieces are removed from *B*'s inventory and placed into *C*'s inventory. Mathematically, this is equivalent to the illustration of Figure 7, where a satisfied demand of 100 pieces is placed on part *B* and a scheduled receipt of 100 pieces are received by part *C*. Therefore, this is how OMRP solves Figure 3 substitution complexity: run the many-path LP model described in the previous section to determine optimal substitutions (eg, Figure 6); convert those LP-determined substitutions into demand and receipt records (eg, Figure 7); run an MRP heuristic, which will take the new LP-converted demands and receipts into account. That MRP run will drive the material plans to meet all demands on time. The MRP run may require receipts to be moved to a time period earlier than when the LP determined they would be available. Consider the receipt of 100 pieces of part *C* in Figure 7, which were created as a result of an LP-determined substitution. If this receipt must be delivered, say, 4 days earlier in time to meet the demand of *C*, then the corresponding 'demand' of 100 pieces on part *B* must also be rescheduled 4 days earlier in time. The rescheduling of part *B*'s demand means that *B* needs to be re-planned. This necessitates a rerun of MRP with the revised demand. Consequently, following the LP run, which recommends an initial set of substitutions, MRP is run multiple times until either no more rescheduling is required (ie, no more receipts need to be pulled earlier in time) or an MRP iteration limit is reached. In practice, a few iterations of MRP are enough to converge or nearly converge. An iteration limit of about 10 was instituted to ensure robustness and reasonable run times for rare cases.

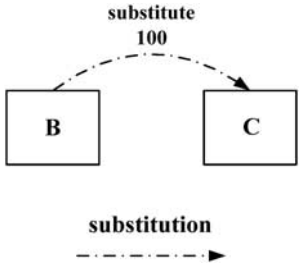

Figure 6 Substitution recommended by an LP run.

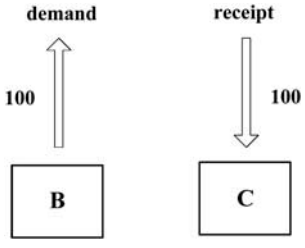

Figure 7 Equivalent representation of substitution.

The above logic of running a many-path LP followed by iterations of MRP occurs at each level of the bills of materials supply chain when complexity such as that of Figure 3 occurs due to substitutions. Once material plans have been finalized at one level of the supply chain, the planned manufacturing releases of that level are exploded through MRP logic, creating demands for parts at the next level of the supply chain. Then the process of running an LP and iterations of MRP repeats for the next level. When the LP runs, it plans parts at that level in the supply chain and parts at all lower levels. (The parts at higher levels have already been planned.) Referring to Figure 1, an LP would run considering all parts of Figure 1 (after all alternative processes have been converted to substitution possibilities). An MRP would then be run—potentially multiple times—for module parts FM, MM, SM, OFM, and OSM. The planned manufacturing releases for these modules would then be exploded to create dependent demands on the untested modules UFM, UMM, USM, OUFM, and OUSM. Because these untested modules are each made in a single way, they are planned by an MRP running a single time and their releases exploded to create dependent demands on the sorted devices FD, MD, SD, OFD, and OSD. The process continues level by level until all parts in the figure have been planned.

This OMRP logic requires that the calculations proceed through the supply chain level by level. Traditional MRP systems have ensured level-by-level processing by first calculating for each part its *low-level code* (LLC). An LLC indicates the stage of a part within the entire manufacturing process. Typically, a part at the end of the manufacturing process is assigned an LLC of one, and the component of any assembly is assigned an LLC one unit higher than the highest LLC of any of its assemblies. In the case of OMRP, it is also required that parts have the same LLC as parts for which they can substitute and parts that are produced from a same binning process. For example, sorted devices FD, MD, and SD in Figure 1 must all have the same LLC. For further details on the LLC calculations of OMRP, refer to Milne *et al* (1999).

Performance improvement through decomposition

To improve the run-time performance of the OMRP method, we apply two ideas: the first is to ensure that each product flow structure is solved with the most computationally efficient method required to create a good solution; the second idea is to apply parallelization—where possible—to those activities that are computationally intensive.

To demonstrate the two ideas, observe that there are four types of product flow structures in Figure 8. First of all, part A2 can substitute for part A3 and it is a complex substitution. The many-path LP model described above must be used on these two parts to make the best supply path choices. The same applies to those parts that are related to A2 and A3 in the bills of materials product structure, namely all parts in Figure 8 beginning with the letter 'A'. Next, parts Bl and B2 have multiple supply path choices because of their alternative processes. This also requires an LP run to make good choices. Consequently, all parts starting with the letter 'B' will be processed with the many-path LP model. In contrast, parts D1, D2, and D3 have a simple product flow and can be solved with MRP heuristics; this would be faster than if they were to be solved with linear programming. Finally, because part C1 can substitute for part C2 and both are resulted from a same binning process, a small 'single-path LP model' can be used to create good material plans for these parts. This single-path LP is described in Lyon *et al* (2001) along with the custom MRP method, which imbeds this LP. This single-path LP can be applied to just C1, C2, and C3 (they are all the parts involved in the binning process). Part C4 will be planned with an MRP heuristic. The single-path LP is slower than the MRP heuristic, but faster than the (larger) many-path LP model. By applying these different decision technologies selectively to product flows with different complexities, the OMRP method is faster than if all parts of Figure 8 were to be run through a single LP model.

Further, observe that the product flows of parts Al through A7 and those of parts Bl through B5 are independent (they do not intersect). Therefore, these parts can be solved in parallel to further reduce run times by solving Al through A7 on one processor and Bl through B5 on another. There are diminishing returns from applying parallelization to many independent product flows that require to be processed by the many-path LP model. As a result, we segregate these product flows into just a few groups and then solve these groups simultaneously.

The OMRP algorithm

Putting together the above conceptual building blocks of a solution results in the below algorithm. When 'MRP' is mentioned below, this refers to the custom MRP method, which embeds the single-path LP model described above to make planning decisions for the single-path parts stemming from a common binning process (Lyon *et al*, 2001).

Step 1: Transform the data so that each part has a single manufacturing process.
Step 2: Calculate LLCs as described above.
Step 3: Repeat the following processing for each LLC (beginning at the top of the supply chain moving downwards):
If there is at least one complex substitution possibility at the current LLC,
Then
Identify the sets of parts that have the current or a higher LLC and require (many-path) LP processing.
Solve these sets of parts using parallel many-path LPs.
Create demand and receipt records corresponding to each substitution created by the LP.
3.1. Run MRP for the parts having the current LLC.
3.2. If substitution receipts that stem from the many-path LP need to be rescheduled earlier and the predetermined MRP iteration limit has not been reached, adjust the

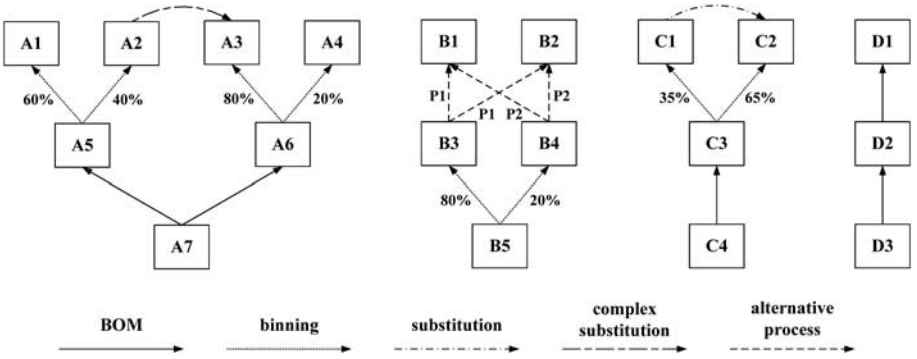

Figure 8 Sample product flows in semiconductor manufacturing.

receipts and their corresponding demands and go to Step 3.1.

Else (ie, LP processing is not required for parts at the current LLC)

Run MRP for parts at the current LLC.

Step 4: Post process MRP outputs by replacing the dummy parts created in Step 1 with their corresponding original parts.

The purpose of the last step is to reverse-convert data resulted from Step 1, so that outputs that are created by the Step 3 core calculations of the OMRP algorithm are displayed in the original, multiple process per part representation. For instance, if there are manufacturing releases created for 'dummy' part *R*' of Figure 5, Step 4 will convert them back to manufacturing releases of original part *R* of Figure 4.

Numerical illustration

We illustrate the advantages of OMRP by showing example material plans created by three methods. As mentioned towards the end of this paper's Introduction, none of the methods identified in the referenced literature address all these aspects that are present at IBM: complex substitutions, alternative processes with different bills of materials, and applicability to more than two levels of bills of materials. All these aspects can be considered by a stand-alone LP, which can determine the best feasible plan. However, a feasible plan is not desired, as it does not indicate opportunities for improvement the way that MRP plans do. We show how material plans would be created using traditional MRP, stand-alone LP, and OMRP.

Our example uses the old product line structure and binning percentages contained within Figure 1. The data are simpler than normal for ease of illustration. The time periods are daily buckets. The lead times are zero for the parts resulting from a binning process (OFM, OSM, OFD, and OSD), 4 days for OUFM and OUSM, and 2 days for OUD and OWAF. The yields are all 100%. There are demands of 750 and 450 pieces of OFM in Day 3 and Day 8, respectively. The WIP (referred to sometimes in the MRP literature as 'scheduled receipts') is 940 pieces of OUFM and 400 pieces of OUSM all expected to arrive to stock in Day 4. Because of lot sizing, OWAF parts must be released in multiples of 100.

The material plans created by traditional MRP, a stand-alone LP, and OMRP are shown in Figures 9, 10, and 11,

Low level codes								
1	2	3	4	5	Day	Gross requirements	WIP arriving at stock	Planned manufacturing releases
OFM					8	450		600 (P1)
					3	750		1000 (P1)
	OUFM				8	600		
					4		940[a]	600
					3	1000		
					1			60[b]
		OFD (fast)			4	600		857
					1	60[c]		86[b]
	OUSM				4		400[d]	
		OSD (slow)						
			OUD (untested)		4	857		
					2			857
					1	86[c]		86[e]
				OWAF	2	857		
					1	86[f]		100[g], 900[h]

[a]Expedite by 1 day to arrive on time on day 3.

[b]Expedite by 2 days to arrive on time.

[c]Due 2 days before day 1.

[d]MRP does not have the intelligence to utilize the WIP on the slower part supply path.

[e]Expedite by 4 days to arrive on time.

[f]Due 4 days before day 1.

[g]Expedite by 6 days to arrive on time.

[h]Expedite by 1 day to arrive on time.

Figure 9 Material plans created by traditional MRP.

Low level codes					Day	Gross requirements	WIP arriving at stock	Planned manufacturing releases
1	2	3	4	5				
OFM					9			497 (P1), 213 (P2)
					8	450[a]		175 (P2)
					4			940 (P1), 225 (P2)
					3	750[b]		
	OUFM				9	497		
					5			497
					4	940	940	
		OFD (fast)			5	497		710
	OUSM				9	213		
					8	175		
					5			213
					4	225	400	
		OSD (slow)			5	213		710
			OUD (untested)		5	710		
					3			710
				OWAF	3	710		
					1			710

[a]415 of the 450 pieces are shipped one day late.
[b]All 750 pieces are shipped one day late.

Figure 10 Material plans created by stand-alone LP.

Low level codes					Day	Gross requirements	WIP arriving at stock	Planned manufacturing releases
1	2	3	4	5				
OFM					8	450		497 (P1), 388 (P2)
					3	750		940 (P1), 225 (P2)
	OUFM				8	497		
					4		940[a]	497
					3	940		
		OFD (fast)			4	497		710
	OUSM				8	388		
					4		400[b]	213
					3	225		
		OSD (slow)			4	213		710
			OUD (untested)		4	710		
					2			710
				OWAF	2	710		
					1			800[c]

[a]Expedite by 1 day to arrive on time on day 3.
[b]Expedite at least 225 pieces of the WIP to arrive by day 3.
[c]Expedite by 1 day to arrive on time.

Figure 11 Material plans created by OMRP.

respectively. Traditional MRP explodes the gross requirements of OFM backward through the bills of materials using the faster parts. Gross requirements are netted of WIP. These netted requirements are offset by lead time and divided by any binning percentages and yields (and when applicable lot sized) to create planned manufacturing releases. The planned manufacturing releases are exploded using the bills of materials to create dependent gross requirements for components. MRP recommends expediting the 940 pieces of WIP of OUFM to arrive by Day 3 to satisfy that day's gross requirement on time and also recommends expediting some of the planned manufacturing releases, as indicated by footnotes in Figure 9. Expediting involves giving jobs higher than normal priority so that they can move through the manufacturing line fast. MRP does not have the intelligence to use inventories created through the slower part supply line on the right side of Figure 1. Consequently, MRP creates 1000 ($=100+900$) pieces to release for OWAF. These releases need to have happened before the current day (Day 1) to meet the demand on time using standard lead times; this suggests expediting actions should be taken on these planned manufacturing releases.

In contrast to traditional MRP, both stand-alone LP and OMRP are smart enough to use all supply paths. The stand-alone LP creates feasible material plans, as shown in Figure 10. These plans would result in late shipments for most of the demand quantities; moreover, the LP ignores lot sizing and thus releases 710 pieces of OWAF.

OMRP applies lot sizing and releases 800 pieces of OWAF (Figure 11). This is 200 pieces less than traditional MRP. If there were no lot sizing considerations, then OMRP would release 710 pieces—the same as stand-alone LP. OMRP suggests expediting WIP and planned manufacturing releases when required to meet all demand on time. This advantageous OMRP function is similar to that of MRP from the user perspective.

Figures 9, 10, and 11 reflect fundamental differences between the three methods. The traditional MRP points out where in the bills of materials supply chain improvements are needed to satisfy demand on time. But MRP lacks the intelligence required to handle any of the complexities discussed earlier in the paper, and thus created the largest release quantities of OWAF. The stand-alone LP results in an intelligent allocation of inventory and planned manufacturing releases. No other method could have created better feasible material plans. Because of feasibility issues, the stand-alone LP creates material plans that result in late shipments. The stand-alone LP does not suggest to the planner what actions to take to satisfy all the gross requirements on time. In contrast, OMRP indicates exactly where and how to make improvements: the two WIP jobs (OUFM and OUSM) and one planned manufacturing release (OWAF) must be expedited to arrive 1 day earlier than normal. The OMRP plans in Figure 11 have a total of 2140 ($=940+400+800$) pieces expedited 1 day (2140 pieces-days). In contrast, the traditional MRP plans in Figure 9 expedite 1840 pieces 1 day and 100 pieces (for the difficult to achieve) 6 days for a total of 2440 ($=1840+6^{*}100$) pieces-days.

In summary, as illustrated in this problem context where on-time delivery is paramount, OMRP is better than stand-alone LP. OMRP and traditional MRP each support on-time delivery, but OMRP does so with less expediting and less quantities released than traditional MRP.

Conclusions

OMRP is implemented in C++ and AIX scripts (AIX is IBM's version of the UNIX operating system). Presently, OMRP runs nightly for about 10 000 parts, with an average of 1.3 processes per part and over 1000 part substitution possibilities. The entire run completes within 15 min. (An LP-only run against the same problem size ran out of disk space after 1.5 h during the generation of the LP matrix to feed the solver.) By blending linear programming with MRP heuristics, OMRP provides both speed and intelligence for the matching process. It runs faster than a stand-alone LP planning tool, accommodates daily granularity requirements, and allows the IBM Vermont semiconductor manufacturing facility to make good business decisions while suggesting specific actions planners should take to meet all demand on time.

Future research opportunities include the development of near real-time solutions that can handle large-scale MRP problems in the presence of many alternative supply paths. The authors wonder whether concepts of this paper could be combined in a creative way with ideas of Lawrynowicz (2008) to achieve this.

References

Balakrishnan A and Geunes J (2000). Requirements planning with substitutions: Exploiting bill-of-materials flexibility in production planning. *Manufacturing & Service Operations Management* **2**: 166–185.

Billington PJ, McClain JO and Thomas LJ (1983). Mathematical programming approaches to capacity-constrained MRP systems: Review, formulation and problem reduction. *Management Science* **29**: 1126–1141.

Bitran GR and Gilbert SM (1994). Co-production processes with random yields in the semiconductor industry. *Operations Research* **42**: 476–491.

Bitran GR and Leong T-Y (1992). Deterministic approximations to co-production problems with service constraints and random yields. *Management Science* **38**: 724–742.

Bitran GR and Leong T-Y (1995). Co-production of substitutable products. *Production Planning & Control* **6**: 13–25.

Carmon TF and Nahmias S (1994). A preliminary model for lot sizing in semiconductor manufacturing. *International Journal of Production Economics* **35**: 259–264.

Denton B, Forrest J and Milne RJ (2006). IBM solves a mixed-integer program to optimize its semiconductor supply chain. *Interfaces* **36**: 386–399.

Geunes J (2003). Solving large-scale requirements planning problems with component substitution options. *Computers & Industrial Engineering* **44**: 475–491.

Harrison TP and Lewis HS (1996). Lot sizing in serial assembly systems with multiple constrained resources. *Management Science* **42**: 19–36.

Hung Y-F and Wang Q-Z (1997). A new formulation technique for alternative material planning—An approach for semiconductor bin allocation planning. *Computers & Industrial Engineering* **32**: 281–297.

Jonsson P and Mattsson S-A (2006). A longitudinal study of material planning applications in manufacturing companies. *International Journal of Operations & Production Management* **26**: 971–995.

Lawrynowicz A (2008). Integration of production planning and scheduling using an expert system and a genetic algorithm. *Journal of the Operational Research Society* **59**: 455–463.

Leachman RC, Benson RF, Liu C and Raar DJ (1996). IMPReSS: An automated production-planning and delivery-quotation system at Harris Corporation—Semiconductor sector. *Interfaces* **26**: 6–37.

Leachman RC, Kang J and Lin V (2002). SLIM: Short cycle time and low inventory in manufacturing at Samsung Electronics. *Interfaces* **32**: 61–77.

Lin JT, Chen T-L and Lin Y-T (2009). Critical material planning for TFT-LCD production industry. *International Journal of Production Economics* **122**: 639–655.

Lin JT, Wang FK and Peng CC (2008). Lot release times and dispatching rule for a TFT-LCD cell process. *Robotics and Computer-Integrated Manufacturing* **24**: 228–238.

Lyon P, Milne RJ, Orzell R and Rice R (2001). Matching assets with demand in supply-chain management at IBM Microelectronics. *Interfaces* **31**: 108–124.

Milne RJ, Orzell RA and Yen C (1999). *Advanced material requirements planning in microelectronics manufacturing*. United States Patent 5 943 484.

Mönch L, Fowler JW, Dauzere-Peres S, Mason SJ and Rose O (2011). A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations. *Journal of Scheduling* **14**: 583–599.

Orlicky JA (1975). *Material Requirements Planning*. McGraw-Hill: New York.

Ornek AM and Cengiz O (2006). Capacitated lot sizing with alternative routings and overtime decisions. *International Journal of Production Research* **44**: 5363–5389.

Pandey PC, Yenradee P and Archariyapruek S (2000). A finite capacity material requirements planning system. *Production Planning & Control* **11**: 113–121.

Ram B, Naghshineh-Pour MR and Yu X (2006). Material requirements planning with flexible bills-of-material. *International Journal of Production Research* **44**: 399–415.

Taal M and Wortmann JC (1997). Integrating MRP and finite capacity planning. *Production Planning & Control* **8**: 245–254.

Tang L, Liu G and Liu J (2008). Raw material inventory solution in iron and steel industry using Lagrangian relaxation. *Journal of the Operational Research Society* **59**: 44–53.

Received September 2010;
accepted November 2011 after one revision