Scientia Iranica, Vol. 9, No. 3, pp 283–288

SCIENTIA
IRANICA

Research Note

Supply Chain Production Planning Using Linear Optimization

J. Ahmadi* and R. Benson[1]

In this paper, the formulation and creation of an enterprise-wide global production planning system for the semiconductor industry are described, using linear optimization as the core planning technology. The objective is to generate production plans on a weekly or daily basis which are optimized to maximize customer service and cash flow in response to a prioritized statement of demand. When evaluating production alternatives, the system observes defined product and process structure relationships, as well as capacity and materials limitations within the various stages of manufacturing.

INTRODUCTION

While the basic large scale production planning problem has already drawn significant attention from academics and practitioners [1], the incorporation of semiconductor specific supply chain issues and the use of optimization technology for practical solution to these problems is a relatively recent development. As opposed to mere extensions to the traditional planning problem, semiconductor specific supply chain issues render traditional MRPII (enhanced materials requirements planning) solution strategies inappropriate for the industry. Some of these issues include re-entrant manufacturing routes, where a lot visits the same workstation along its process flow, co-product binning and downgrade substitution, the availability of alternative yet non-homogenous workgroups at various processing steps and a partitioned demand statement into multiple priority classes. Earlier work which addresses some of these items is found in [2].

Although the inclusion of the above issues is essential for a planning system at a semiconductor firm, it results in a substantial increase in the size and complexity of the resulting models when compared to generic production planning approaches. Various commercial vendors and corporations have responded to this increase in complexity by developing sub-optimal heuristic algorithms to generate production plans. It has been found that linear programming provides both an efficient search engine to produce optimal solutions as well as providing an appropriate framework to model the semiconductor planning problem.

BUILDING BLOCKS FOR THE SUPPLY CHAIN MODEL

The scale of the problem precludes modeling each processing step as a separate planning decision. Instead, stages of manufacturing are modeled and production plans are generated which specify the input quantities by individual item at each stage of manufacturing. Shop floor scheduling within each manufacturing stage is beyond the scope of this work and it is assumed that each manufacturing area possesses its own shop floor scheduling capabilities.

There are generally six major stages of manufacturing defined in the planning model. They are: Wafer fabrication part 1, wafer fabrication part 2, wafer sort, assembly, test/mark/pack part 1 and test/mark/pack part 2. This division of the supply chain is represented graphically in Figure 1. These six stages were chosen to divide the manufacturing supply chain along typical organizational, product structure and facility boundaries. Additional stages can be added to extend the supply chain vertically into card or board assembly, or each major stage may further be subdivided into various stages. Problem size is the primary limitation on the number of manufacturing areas, since the number of decision variables and constraints grow linearly by

*. *Corresponding Author, Information Technology Management, Advanced Micro Devices, Austin, Texas 78741.*

1. *Information Technology Management, Advanced Micro Devices, Austin, Texas 78741.*

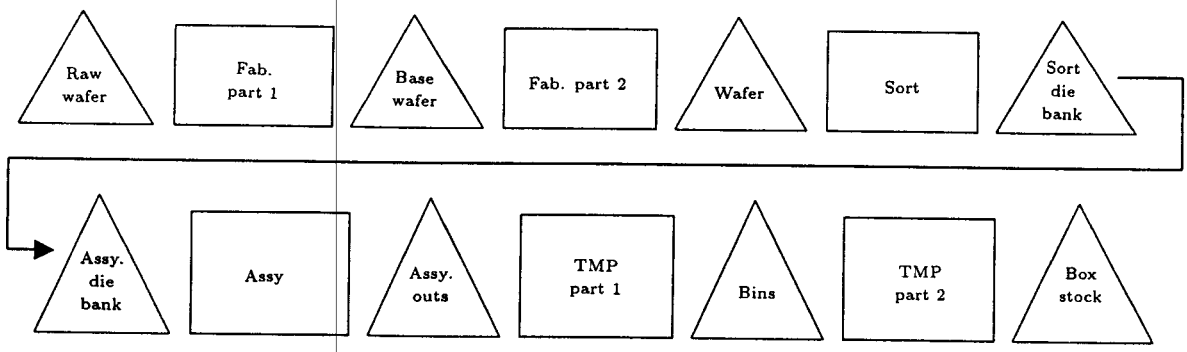

Figure 1. Manufacturing stages.

number of stages and time periods in the planning horizon.

The rectangles represent the manufacturing stages, while the triangles represent Corporate Inventory Points (CIPs). The CIPs are points in the supply chain where the system has the option to plan the accumulation of inventory. Since the logistics within a manufacturing area is delegated to shop floor systems, the planning system does not dictate inventory levels at individual processing steps within a manufacturing stage. Instead, it regulates the flow of material into a manufacturing stage, which, in turn, limits the total inventory or Work-In-Process (WIP) within the stage. The ability to plan inventory only at CIPs is the primary motivation for dividing both wafer fabrication and Test/Mark/Pack (TMP) into two manufacturing stages. In wafer fabrication, the split is added to handle situations such as ROM code products where a generic base wafer is manufactured up to a step in the process and then held in inventory awaiting a signal to continue processing using specific metal masks. Similarly, the TMP area is split into two manufacturing stages to allow for the planning of accumulated inventory and for product substitution at the finished unmarked (FUM) inventory point. Although Figure 1 depicts a single facility at each manufacturing stage, the planning model allows for multiple separate facilities, both internal and foundries/subcontractors, at each stage of manufacture.

By dividing up the manufacturing supply chain into various segments, the planning model imposes an allowable Bill Of Materials (BOM) for the purpose of planning. Since shop floor logistics are beyond the scope of the planning model, the planning system can only recognize product structure alternatives between manufacturing stages that encompass multiple parallel or serial facilities typical of an extended supply chain. This is exploited to define the notion of a connection:

$$cnxn(parent\_product\_area, parent\_product, route,$$

$$child\_product\_area, child\_product, time\_period),$$

which takes on the value of 1 if it is legal to use the parent_product manufactured in the parent_product_area as the input to manufacture the child_product in the child_product_area using the defined route in the specified time_period. It takes on the value 0 if the combination is not valid. The connection concept is applied between adjacent stages of manufacturing and it serves to define the possible production transformations along the supply chain. This data is constructed from BOM definitions. A connection, therefore, describes a valid transformation of a product to another using a specific method, that is also inclusive of the change of location of the resulting child product. Associated with each connection is a process time, a yield factor and resources requirements, that may be time period dependent. At any time, multiple alternative methods may be available for such transformations.

PRODUCT BINNING AND COMPONENT KITTING

While connections provide a standardized way of representing production transformations between each stage of manufacturing, it assumes that one unit of a parent product is consumed in the making of a single child product. In semiconductor manufacturing, there exist situations where one unit of a parent product is transformed into one of many child products. This is the co-product or binning process. Alternatively, multiple parent products can be required in the manufacture of a single child product. This is the kitting process. To address these aspects of the semiconductor supply chain, the notion of a transformation link has been developed to represent the numerical conversions necessary for kitting and binning. A link is represented by:

$$link(parent\_product, child\_product\_area,$$

$$child\_product, time\_period),$$

where the arguments represent the valid transformations and the numerical value is the transformation multiplier, as in the bin split for the case of binning.

The discrete nature of the physical process is ignored and it is assumed that links can take on real values. It is important to note that link data contains no route argument. In case of kitting, the link data describe the set and quantities of the kit elements required to initiate a transformation (fusion). In case of binning, the link data specifies the ratio of the by-products created per unit of transformation activity (fission). Both instances of fusion and fission may be simultaneously associated with a given transformation. In presence of kitting or binning, the parent or child products defined in the connection data are "logical" products representing the kit set, or binning siblings. By combining connections and links, therenow exists a robust set of building blocks for a model of the supply chain.

CONNECTING THE SUPPLY CHAIN

The two basic sets of decision variables found in the planning model represent the planned starts of a product into a manufacturing stage and the planned output of a product from a manufacturing stage. Production start decision variables at each CIP are represented by:

$$S(parent\_product\_area, parent\_product, route,$$

$$child\_product\_area, child\_product, time\_period),$$

where the value of the variable represents the start quantity for the combination of arguments. A production start decision will then result in reduction of inventory of the parent product (or kit of parent products) and a future increase of child product (or set of bins). Start variables are defined consistent with the connection data.

To connect the supply chain and form a basic planning model, the manner must be defined in which starts into a manufacturing stage relate to starts of products in the adjacent upstream manufacturing stage. Inventory balance equations (flow conservation equations) provide this connection. In its basic form, an inventory balance equation for product j in period t is: $inv(j,t) = inv(j,t-1) + upstream\_outs(j,t) - downstream\_starts(j,t)$. The expansion of this basic equation to account for multiple manufacturing areas and for transformation links results in:

$$Inv(uc\_area, uc, t) = Inv(uc\_area, uc, t-1)$$

$$+ upstream\_outs(uc\_area, uc, t)$$

$$- [\sum_{\substack{\forall link(uc,dc\_area,dc,t)>0.0 \\ \forall cnxn(uc\_area,uc,route,dc\_area,dc,t)=1}}$$

$$link(uc, dc\_area, dc, t)^*$$

$$S(uc\_area, uc, rout, dc\_area, dc, t)], \qquad (1)$$

where the convention is used that *uc* means upstream child and *dc* mans downstream child. By enforcing this equation for all products and periods between all manufacturing areas and by forcing $Inv()$ to remain non-negative, a conservation of mass is forced at the CIP points. Upstream outs in the initial periods of the planning horizon represent WIP output from the manufacturing area and are assumed to be provided to the planning model by the shop floor control system or by some other WIP out prediction algorithm. A model to predict WIP output is found in [1]. Notice that the inventory has been tied to the upstream area and to the upstream child. The convention will also be adopted that the CIP point is logically located immediately following the upstream manufacturing area and, therefore, the shipping time to the downstream area is added to the cycle time of the downstream route. A special case addresses the situation where inventory of the parent product is present at the beginning of the child manufacturing area at the start of the planning horizon [3].

Once the inventory balance equations are defined, the connection of the entire supply chain reduces to defining the relationship between the start and output quantities within a manufacturing area. It is assumed that planning process yields and cycle times have been defined for each route and that the planning cycle times will be adhered to if the plan does not exceed the capacity limitations described in the next section. By assuming that the starts of a product will be made uniformly over a planning period, these starts can be mapped to their resulting output in future periods by offsetting the starts by the planning cycle time and decreasing them by the process yields. This is shown in Figure 2.

Starts in one planning period may result in output in more than one period due to cycle times which are non-integers of the planning periods or due to a time varying working calendar for the manufacturing area. By assuming uniform starts across a period, the resulting output in various periods is simply the fraction of the total output area which falls into the period of interest. In Figure 2, the output in period 3 is $(m/n)^*O$. Since the output area O is defined as (S^* route process yield), the output in period s is given by $(m/n)^*$(process yield)$^*S$. The quantity $(m/n)^*$

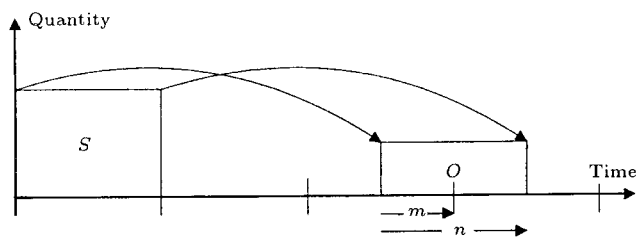

Figure 2. Start to outs transformation.

(process yield) is defined as the output map coefficients and provide a linear representation of output in terms of starts. A more rigorous development of this concept is found in [2]. Taken together with the inventory balance equations, the stages of the supply chain are linked to form a connected planning model. Instead of carrying two sets of equations, the expression for upstream_outs in the inventory balance equations are replaced with the equivalent expression in terms of starts and the entire supply chain is linked in a single set of equations:

$$Inv(uc\_area, uc, t) = Inv(uc\_area, uc, t-1)$$

$$+ [\sum_{\substack{\forall link(up,uc\_area,uc,t)>0.0 \\ \forall out\_map(up\_area,up,route,uc\_area,uc,s,t,)>0.0 \\ \forall cnxn(up\_area,up,route,uc\_area,uc,t)=1}}$$

$$link(up, uc\_area, uc, t)^* out$$

$$\_map(up\_area, up, route, uc\_area, uc, s, t)^*$$

$$S(up\_area, up, route, uc\_area, uc, s)]$$

$$- [\sum_{\substack{\forall link(uc,dc\_area,dc,t)>0.0 \\ \forall cnxn(uc\_area,uc,route,dc\_area,dc,t)=1}}$$

$$link(uc, dc\_area, dc, t)^*$$

$$S(uc\_area, uc, route, dc\_area, dc, t)]$$

$$\forall uc\_area, uc, t = 1, \ldots, T. \qquad (2)$$

DEMANDS

At various points along the supply chain, the downstream activity is an allocation of production volume against demand instead of production starts into a manufacturing area. While the typical semiconductor supply chain applies demand as wafer sales, die sales and completed IC sales, the model is structured so that demand can be applied at any CIP point. Demand is restricted to CIP points to simplify the model. In order to make this applicable for die sales, a route is defined in assembly which contains only the saw and some packaging steps. Die sales are then modeled to occur from the assembly outs CIP point. By limiting the application of demand to CIPs, distribution variables, $D()$ can be defined, to represent the planned physical flow of product. Equation 2 then becomes:

$$(2) - \sum_{\forall link(uc,uc\_area,opn,t)=1} link(uc, uc\_area,$$

$$opn, t)^* D(uc\_area, uc, opn, t), \qquad (3)$$

where the term *opn* (ordering part number) represents the demanded product name and link concept is used to define the allowable products to allocate towards meeting the demand.

Since the shortfall of distribution versus demand requires to be explicitly tracked, and since the distribution requires limiting so that it does not exceed demand, backorder variables are defined and used to express backorder balance equations. They are defined as:

$$Back(opn, t) = Back(opn, t-1) + dem(opn, t)$$

$$- \sum_{\forall link(uc,uc\_area,opn,t)=1} link(uc, uc\_area, opn, t)^*$$

$$D(uc\_area, ucopn, t) \forall opn, t = 1, \ldots, T. \qquad (4)$$

PRODUCTION LIMITATIONS

Various types of production limitations are supported in the planning model. They represent both user defined behavioral limitations as well as physical limitations. Planned inventory minimums and maximums are modeled by placing bounds directly on inventory variables.

For production starts, the model supports minimum and maximum limits placed on subsets of products within an area. In addition, it supports limiting the maximum rate of change on subsets of products from one period to the next. The definitions are found in [3]. Limitations on starts are used to observe contractual limitations at foundries and subcontractors, as well as to observe product line allocations within the wafer fabrication areas. The lower bounds in both the inventory and starts limits constraint sets may render the entire model infeasible if set too high. Safeguards have been put in place to avoid such a situation.

Through linkage with starts activities, materials limitations are enforced by ensuring that the cumulative availability of a material up to a planning period exceeds the cumulative consumption of that material up to the same planning period. These constraints are enforced up to a user defined re-order lead time. Beyond the lead time, it is assumed that the material is infinitely available.

Capacity constraints serve to limit the total load on a resource in a planning period. Both the re-entrant nature of the process routing and the presence of alternative workgroups need to be comprehended in developing an appropriate capacity model for this industry. The re-entrant nature of the process is addressed by first creating load map coefficients in similar fashion to the way the output map coefficients were described in pervious section. The load map coefficients represent the yielded fraction of a start quantity which results in load at a step in a planning period. The planning process yield and cycle time up to the step are used to compute these coefficients. When considering the load resulting on a particular resource, r, in a period, the load map is multiplied by (1/UPH)

where UPH is the units per hour for the resource applied to that step. The result of this multiplication will be reffered to as the resource map, res_map.

In addition to the resource maps, the allowable resource configurations are required at each potential bottleneck step. Defining configurations begins with assigning resources to workgroups, wg. Resources represent a group of homogenous machines and the same resource can be defined in multiple workgroups. In an effort to generate a compact model, resources should only be assigned to workgroups if they represent potential bottlenecks. If a production plan uses a workgroup, it is assumed that it will use all the resources within the workgroup. An example is found in the test area, where a workgroup may be defined as a test system and a handler. Workgroups are then attached to individual processing steps along a route. If multiple workgroups are attached to the same step, it is assumed that these workgroups represent capacity alternatives at that step. Workgroups can be attached and detached to processing steps over the planning horizon to represent changes in the capacity alternatives over time. Based on these modeling constructs, the capacity constraints take the form:

$$\sum_{\forall res\_map()>0.0} res\_map(uc\_area, uc, route,$$

$$dc\_area, dc, t, step, wg, r, t)^*$$

$$A(uc\_area, uc, route, dc\_area, dc, t, step, wg)$$

$$\leq capacity(dc\_area, r, t) \forall r, t = 1, \ldots, T, \qquad (5)$$

$$\sum_{wg} A(uc\_area, uc, route, dc\_area, dc, t, step, wg) =$$

$$S(uc\_area, uc, route, dc\_area\_dc, t)$$

$$\forall dc\_area, dc, t = 1, \ldots, T, \qquad (6)$$

where $A()$ represents the allocation of load to a particular workgroup and $capacity()$ represents the maximum hours of productive work available. Much detailed formulation of the problem for the interested reader can be found in [3].

OBJECTIVE FUNCTIONS AND PRIORITY CLASSES

While the previous sections describe ways to model the supply chain and its relationship to demands, objectives for the planning model still require definition. Typical objectives are: Increase revenue or profit, decrease backorders or inventory, increase throughput, or decrease cycle time. This planning model includes provisions for minimizing backorder penalties (customer service objective) and for maximizing cash flow (economic objective). The customer service objective assigns costs to both the backorder and inventory variables where backorder costs are much higher than the inventory costs. This drives the system to search for a plan which minimizes the lateness cost while attempting to manufacture in a just-in-time fashion. The economic objective applies a revenue benefit to each distribution made towards the satisfaction of demand based on average selling prices. From that revenue, it subtracts the variable cost of manufacturing the item plus an inventory holding cost.

The model supports multiple objectives in order to plan the various types of demand found in the semiconductor industry. For example, booked orders, quotes, demand contracts and certain types of forecasts are demands where AMD already has a delivery obligation to their customer. These demand types require a customer service type objective to reflect this obligation. For other demand forecasts or risky bookings, the economic objective is more relevant. In order to apply these two objectives, the total demand is divided into priority classes. There are no restrictions on mapping demand to classes, however the planning model generates plans for each class in strict priority order. Support for higher priority classes cannot be sacrificed to support demand in a lower priority class. Each class is planned using one of the two objective functions. Typically the customer service class is used for the higher priority classes and then the economic objective is used for the remainder.

APPLICATION DEVELOPMENT AND PERFORMANCE

The current version of the application dubbed MPP for Micro Processor Production Planning, is a web based application written in Java and C++ using an Oracle data base system and IBM Optimization Subroutine Library Version 2.0. The web based nature of the application allows various geographically distributed facilities to rapidly update relevant data components such as latest work-in-process, yield and capacity information. Model exercises are managed by the application as what-if scenarios with user controlled security provisions. Significant data integrity and completeness checking is provided at the data base level. A graphing interface is provided to allow users to draw the products BOM structure as a general graph, with point and click options to convey data and attributes with the structures defined. A typical planning exercise covers 90 planning periods mostly in increments of days followed by weekly periods. Consisting of 23 CIP's, 156 products, 40 opn's, with 12-14 resource types. The associated model sizes have roughly 260,000 rows and

340,000 columns and are super sparse. Since the model structure has a Dantzing Wolf block structure, the solution process uses a modified decomposition strategy for a fast and very high quality initial solution generation. The solve time of the problem to optimality is approximately 5 minutes on a 130 Mhz IBM Power2 machine running under AIX operating system. Data extraction and model generation time for a new scenario may take up to 15 minutes and the report generation about 45 seconds. The solve time is 40% faster on an AMD 750 Mhz Athlon processor from Advanced Micro Devices under Linux operating system. Faster solve times are expected using the CPLEX solver, since the CPLEX solver has been optimized for the Athlon CPU, offering a significant price performance advantage with Ghz capability.

To our knowledge there are no competitive vendor solutions using pure optimization methods matching this capability. Most commercial based solutions, while faster and more scaleable, are heuristic based with an unknown measure of optimization and lacking the flexibility to capture untypical business rules. With mathematical modeling techniques, it is more readily possible to manage such complexities. This is an issue that is important in terms of worth of the degree of optimization possible and conformance to the business process. Certainly with the recent more expensive products such as high performance CPU's that have much more complex fabrication and testing structures, it is a worthwhile consideration. The generalized notion of connections and links is very consistent with modern multi-facility product structures that encompass various geographies, readily encompassing non-manufacturing activities such as transportation, with full time dependency of parameters and structures, yet requiring the minimum number of decision variables for problem formulation

ACKNOWLEDGMENT

The authors would like to thank Advanced Micro Devices Information Technology Organization, as well as the entire plan supply team for their continual support and partnership.

REFERENCES

1. Elsayed, E. and Boucher, T., *Analysis and Control of Production Systems*, Prentice Hall (1985).
2. Leachman, R., Benson R., Liu, C. and Raar, D. "IMPReSS: An automated production planning and delivery quotation system at Harris corporation - semiconductor sector", *Interfaces*, **26**, pp 6-37 (1996).
3. Ahmadi, J., Benson, R. and Supernaw-Issen, D. "Planning of production in semiconductor manufacturing - Chapter 10", in T. Ciriani, Ed., *Operational Research in Industry*, Macmillan Press Ltd. (in press).