ON CAPACITY MODELING FOR PRODUCTION PLANNING WITH ALTERNATIVE MACHINE TYPES

ROBERT C. LEACHMAN, TALI F. CARMON

Analyzing the capacity of production facilities in which manufacturing operations may be performed by alternative machine types presents a seemingly complicated task. In typical enterprise-level production planning models, capacity limitations of alternative machine types are approximated in terms of some single artificial capacitated resource. In this paper we propose procedures for generating compact models that accurately characterize capacity limitations of alternative machine types. Assuming that processing times among alternative machine types are identical or proportional across operations they can perform, capacity limitations of the alternative machine types can be precisely expressed using a formulation that is typically not much larger than the basic linear programming formulation that does not admit alternative resource types. These results have important implications for industrial practice, suggesting that in the case that processing times are nearly proportional among alternatives, the prevalent approximation that involves using a single, capacitated, artificial resource may be dropped in favor of our formulation incorporating the approximation that processing times among the alternatives are proportional. Another advantage is that the set of capacity constraints we formulate can be used to check the feasibility of suggested production schedules or demands simply by plugging them into the constraints, without need to develop values for allocation variables.

■ In this paper we present an efficient approach for formulating models of the capacity limitations of alternative machine types for use in corporate-level capacity analysis. In accordance with the hierarchical approach to production planning and control proposed by Gershwin [7], the corporate-level production planner is primarily interested in determining the optimal, capacity-feasible product mix for each planning period over some planning horizon, i.e., the optimal rates of production of each product type in periods such as weeks or months. One or more lower level planners (e.g., factory-floor schedulers) then determines the best way to produce this mix in a much shorter time frame such as a day, taking into consideration the detailed current status of the factory.

Linear Programming (LP) is often proposed as a tool for production planning and scheduling [3], [4], [9], [11], [13]. However, LP formulations can be very large for large organizations with complex production environments, and care must be exercised to develop the most compact model possible achieving a desired level of accuracy. The microelectronics manufacturing industry is an example of such a complex manufacturing environment [1], [13], [15]. A large semiconductor firm must plan the production of thousands of products subject to capacity limitations imposed by hundreds of equipment types. In addition to the complexity caused by the variety of processing requirements of the different product types, frequently there are alternative machine types suitable for performing manufacturing operations. (Throughout this paper we use the phrase "alternative machine types for an operation" to refer to machine types that are functionally different but nevertheless all suitable for performing the manufacturing operation. Such machine types have partial overlap in terms of the operations they can perform. We view multiple identical machines as a single machine type with given capacity.) Due to this equipment flexibility, higher throughput and capacity utilization can be achieved through balancing the workloads among the alternative machine types. However, analyzing the available capacity of this flexible environment in order to determine the optimal production quantities presents a seemingly complicated task.

If the high-level capacity analysis were to precisely characterize available capacity, one could imagine developing a detailed LP formulation in which the variables are defined as the amount of each product type taking each route through the factory (i.e., the workload assigned to a set of machines that perform a complete sequence of operations on the product type). The total number of variables in each planning period would then be the number of all possible routes, for all product types. For example, if a product requires a single processing operation that has 4 alternative machine types, then 4 variables are required to represent it. However, if the route for the product incorporates this processing operation twice, then 16 variables are required to represent all alternatives. Thus, in a re-entrant process flow the number of variables grows as the number of alternative machine types for each operation to the power of the number of re-entries to that operation.

The semiconductor industry is characterized by re-entrant routes [1], [13], [15]. In the Lithography stage, for example, there are on the order of 10-20 re-entries, with several alternative machine types suitable for performing most lithography operations, but perhaps only one machine type suitable for critical operations. This is a typical case in which the number of variables in an LP formulation of this type would grow unattractively large.

In general, it is computationally cumbersome to carry out detailed assignments of operations to machines when the purpose at hand is to adjust company-wide demands to achieve capacity-feasibility. In typical enterprise-level planning applications, capacity limitations of alternative machine types are approximated using some sort of average resource capacity to constrain volumes of each product type to be produced, with varying degrees of success. For example, Spence and Welter [14] estimate a semiconductor manufacturing work cell's capacity using a cycle time—throughput trade off curve.

The problem of accurately modeling the capacity of alternative, non-identical resources arises in almost all manufacturing environments where process technology is evolving, as well as in numerous other application areas. Surprisingly, published research on this topic is scant.

Semiconductor shop floor scheduling on alternative machine types has been addressed by Bitran and Tirupati [2]. Their goals were to minimize the schedule makespan and total tardiness. However, they considered a single stage process, and assigned jobs to machine types based on heuristic rules. Their approach is not suitable for high level planning of a multi-stage process. Federgruen and Groenevelt [5] find capacity-feasible schedules for n jobs with given release times and due dates. They use network flow techniques similar to ones we rely on, in that maximal throughput of jobs corresponds to a maximal flow in the network. However, they allow for preemption, and assume that all jobs can be processed on any of the machine types, albeit at different speeds.

In the next sections we develop alternative formulation approaches that are much more compact than the route-variable approach briefly discussed above. First, we develop a formulation that replaces variables representing entire production routes with variables representing production activity at each operation in a product's multi-stage production process. For each operation on each product, there are allocation variables to spread the product-operation production quantities among alternative machine types, and between consecutive operations there are constraints guaranteeing consistency of production volumes. This formulation, a natural extension of an approach suggested in [11], is termed the "Step-Separated Formulation" and is presented in the next section. This formulation is quite general but has the disadvantage of relatively large dimensions due to the explicit scheduling of machine assignments to product-operation volumes in order to determine capacity-feasibility.

We then introduce more compact models applicable under a uniformity assumption concerning the machines types. This assumption requires the processing times among alternative machines types to be either identical or else proportional across all operations performed by the machines. Exploiting this uniformity assumption, one of these formulations utilizes allocation variables for the total workload in planning periods rather than for the workloads of individual process- steps. We refer to such a formulation as the "Workload Allocation Formulation." We next show how an exact formulation may be generated without use of any allocation variables, i.e., including only variables for the production of each product. We term such a formulation the "Direct Product Mix Formulation."

We also compare the different formulation techniques for a couple of examples that illustrate the range of dimensions that may result. We discuss the technological reasons underlying the existence of alternative machine types as they arise in industrial practice. We illustrate that for the most commonly arising sets of alternative machine types, the Direct Product Mix formulation is the most compact formulation by a substantial margin, and is typically not much larger than the conventional LP planning formulation that does not admit alternative resources. The restrictiveness of the uniformity assumption also is discussed. We explain the underlying technological factors that make this assumption a good approximation in semiconductor manufacturing, suggesting that the compactness of the proposed formulation is worth the loss of generality.

In all formulations we assume that the set of machine types suitable for performing a particular processing step is independent of the machine types selected to perform other steps on the same product. This assumption is typically realistic for semiconductor manufacturing, as well as for many other manufacturing environments.

We also assume that the production cost is independent of the product's processing route. This assumption is reasonable for semiconductor manufacturing since the incremental costs of raw material and labor are relatively small, and are typically considered identical or almost identical for all alternative machine types. Revenue is much larger than variable production cost, and

therefore the major issue the production planner is concerned with is the utilization of production capacity.

The Step-Separated Formulation

We introduce the following notation for the data concerning the production planning problem:

$t = 1,\ldots,T$ is the time period index, where period t denotes the time interval $(t-1, t]$,

$i = 1,\ldots,n$ is the product type index,

$j = 1,\ldots,J_i$ is the index of process-steps (operations) for product type i,

L_{ij} denotes the average flow time for product type i from the start of its production process until initiation of step j (the "lead time" up to step j),

L_i denotes the average flow time for product type i from start to finish of its entire production process (the "lead time" for the process),

$k = 1,\ldots,K$ is the machine type index,

$k \in P(i,j)$ denotes that machine type k is suitable for performing step j on product type i,

a_{ijk} is the time required to process one unit of product type i in step j on machine type k,

C_{kt} is the available capacity (in time units) of machine type k in time period t,

D_{it} is the maximum cumulative number of units of product type i that can be sold by time t (i.e., the total market forecast, including both committed and potential future orders),

d_{it} is the minimum cumulative number of units of product type i that must be supplied by time t (i.e., the committed orders),

p_{it} is the estimated net discounted cash flow from producing and selling one unit of product type i in time period t, and

h_{it} is the estimated cost of holding one unit of inventory of product type i at time t.

For simplicity of exposition, we shall omit the effect on the formulations of various factors such as yields, initial status of work in process, and infeasibilities of given minimum demands. Extensions to account for such phenomena are addressed in detail in [13]. Also for simplicity, we assume that all lead times are integer; in the case they are fractional, a formulation may be developed in which step workloads and process outputs are split between adjacent periods. The lead times also can be time-varying, represented as parameters specific to epochs marking the end points of the planning periods. See [13] for details. All these extensions preserve the special structure developed here for coping with alternative resources.

All formulations we shall present include the following basic production and inventory variables:

X_{it} is the number of units of product type i to be started in time period t,

I_{it} is the number of units of product type i held in inventory at time t, and

B_{it} is the shortfall of the cumulative production vs. the cumulative max demand for product type i at time t.

In the Step-Separated formulation, additional variables are introduced to represent the portion of the workload of each product type in each processing step in each time period that is assigned to each suitable machine type. Formally, we let W_{ijkt} denote the workload of product type i in step j in time period t that is assigned to machine type k, defined only for machine types $k \in P(i,j)$. The formulation may be expressed as follows:

$$\text{Max} \sum_{i=1}^{n} \sum_{t=1}^{T} p_{it} X_{it-L_i} - h_{it} I_{it}$$

subject to

$$X_{it-L_{ij}} = \sum_{k \in P(i,j)} \frac{W_{ijkt}}{a_{ijk}} \quad \forall\ i=1,\ldots,n;$$

$$\forall\ j=1,\ldots,J_i;\ \forall\ t=1,\ldots,T$$

(for each i, one of these expressions could be substituted to replace $X_{it-L_{ij}}$)

$$\sum_{\substack{i=1 \\ \{i,j \mid k \in P(i,j)\}}}^{n} \sum_{j=1}^{J_i} W_{ijkt} \leq C_{kt}\ \forall k=1,\ldots,K;\ \forall t=1,\ldots,T$$

$$\sum_{\tau=1}^{t} X_{i\tau-L_i} - I_{it} + B_{it} = D_{it}\ \forall\ i=1,\ldots,n;\ \forall\ t=1,\ldots,T-1$$

$$\sum_{\tau=1}^{T} X_{i\tau-L_i} + B_{iT} = D_{iT}\ \forall\ i=1,\ldots,n$$

$$B_{it} \leq D_{it} - d_{it}\ \forall i=1,\ldots,n;\ \forall t=1,\ldots,T$$

$$X_{it} \geq 0,\ I_{it} \geq 0,\ B_{it} \geq 0,\ W_{ijkt} \geq 0, \text{ for all } i,j,k,t.$$

Note that the formulation assumes no intermediate accumulation of inventory is allowed, i.e., only finished goods may be stored. The formulation allows output

illing the portion of market forecasts in excess of nmitted orders ($D_{it}-d_{it}$) to be delivered late, albeit less discounted revenue. In the objective function this and all subsequent formulations, discounted net h flow is credited against production, even though s production might not be immediately sold. How-er, we assume the inventory holding costs include : difference in discounted sales revenue from one pe-d to the next. Since ending inventory is prohibited, rrect discounted net cash flow is assessed overall. ote that in the Step-Separated formulation, an addi-nal set of constraints is required to guarantee that r each product type, the number of units processed each step is consistent with the start quantity. The pacity constraints then force the total workload on ch machine to be less than or equal to the machine apacity.

umber of Variables and Constraints

Let M_{ij} be the number of alternative machine types at can perform step j on product type i. Excluding nventory and backorder variables that will be common o all formulations we shall present, the number of var-ables per time period in the Step-Separated formulation s

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i} M_{ij},$$

hile the number of capacity-related constraints per time eriod is

$$K+\sum_{i=1}^{n}(J_i-1) = K-n+\sum_{i=1}^{n} J_i$$

since in addition to the K machine capacity constraints, J_i-1 equality constraints are required for each product ype i to guarantee that the number of units processed t each step is consistent with the start quantities. These imensions are somewhat reduced in the degenerate case here some M_{ij}'s are unity, i.e., in the case that there re some operations with only a single suitable machine ype for which allocation variables are not needed. (We mit discussion of the dimensions of the demand con-raints since these constraints are identical for all for-nulations to be presented.)

In the above formulation, the allocation variables rep-esent assignments of volumes to machines. The for-nulation is inefficient in the sense that it actually solves he detailed problem of allocating process operation orkloads among alternative machine types, yet the roduct start rates (X_{it}'s) are the only variables truly esired at the highest level of the planning and sched-ling hierarchy. At this level of planning, it is not re-listic to develop a detailed, operation-by-operation enterprise-wide schedule that one expects to be precisely followed; but it is nevertheless important to develop starts rates that are capacity-feasible based on average rates of available capacity. In typical practice, the operation-level scheduling problem would be solved much more frequently and with a shorter time horizon, whereby its formulation would reflect more refined knowledge about the state of the factory (e.g., machine availability and work in process) than can be reflected in the average rates assumed in high-level planning. That is, the corporate level production planner views shop floor activities as rates, and, under certain assumptions, as long as the shop floor data used by the high level planner is accurate, a feasible high level production plan can always be translated into a feasible schedule on the shop floor [7].

In the next section we show that it is possible to formulate the production planning problem in a more efficient manner. In lieu of variables representing allocations of the workload of individual process-steps, variables are defined to represent the allocation among machine types of the total workload that can be performed by a set of machine types. This formulation requires a uniformity assumption, specifically, that processing times among alternative machine types are either identical or else proportional across all operations performed by the machine types.

The Workload Allocation Formulation

If all alternative machine types have identical processing times, then the workload in a given time period for the set of process-steps that can utilize any of a set of alternative machine types may be summed to express the total workload in the time period on the set of machine types. If instead the processing times on alternative machine types are proportional across all operations the alternatives perform, the processing times and capacities for the machine types may be scaled in terms of some "standard" machine type to achieve identical processing times and thereby obtain the same result. This assumption allows a substantial simplification of the Step-Separated formulation, replacing allocation variables for individual process-steps with allocation variables for the total workload in a time period. We develop this formulation as follows. (We are indebted to Professor Ilan Adler for suggesting this formulation.)

When the uniformity assumption holds, the k index of the a_{ijk} time coefficients can be eliminated since these coefficients no longer depend on the machine type selected for processing step j of product type i. The following additional notation is required for this formula-

tion:

S_m, $m=1,\ldots,M$, denote the unique sets of alternative machine types among the sets $P(i,j)$ appearing in the problem data,

Z^m_{kt} is a variable representing the workload on set S_m assigned to machine type k in period t, defined for each $k \in S_m$.

The Workload Allocation formulation is then expressed as follows:

$$\text{Max} \sum_{i=1}^{n} \sum_{t=1}^{T} p_{it} X_{it-L_i} - h_{it} I_{it}$$

subject to

$$\sum_{\substack{i=1 \\ \{i,j \mid P(i,j)=S_m\}}}^{n} \sum_{j=1}^{J_i} a_{ij} X_{it-L_{ij}} = \sum_{k \in S_m} Z^m_{kt} \quad \forall m;\ \forall t$$

$$\sum_{m=1}^{M} Z^m_{kt} \le C_{kt} \quad \forall k;\ \forall t$$

$$\sum_{\tau=1}^{t} X_{i\tau-L_i} - I_{it} + B_{it} = D_{it} \ \forall\ i=1,\ldots,n;\ \forall\ t=1,\ldots,T-1$$

$$\sum_{\tau=1}^{T} X_{i\tau-L_i} + B_{iT} = D_{iT} \quad \forall\ i=1,\ldots,n$$

$$B_{it} \le D_{it} - d_{it} \quad \forall\ i=1,\ldots,n;\ \forall\ t=1,\ldots,T$$

$X_{it} \ge 0$, $I_{it} \ge 0$, $B_{it} \ge 0$, $Z^m_{kt} \ge 0$, for all i,j,k,m,t.

Number of Variables and Constraints

The number of capacity constraints per time period in the Workload Allocation formulation depends on the number of sets of alternative machine types that appear in the problem data. Similarly, the number of variables per time period appearing in the capacity constraints depends on the cardinality of each of these sets. To express capacity limitations in each time period, the formulation includes K inequality constraints, one for the capacity of each machine type, plus M equality constraints relating the allocation variables for each machine set to the workload for the set. The number of equality constraints may total to slightly less than M when there are one or more sets S_m that are singletons. Examples may be constructed for which M can range from $K-1$ to 2^K-K-1, but for most industrial cases we have examined it is $O(K)$ or $O(K^2)$. As for the number of variables, let

$$M_t = \sum_{\substack{m=1 \\ |S_m| \ge 2}}^{M} |S_m|$$

The Workload Allocation formulation includes $n + M_t$ variables appearing in capacity constraints in each time period. We remark that $|S_m|$ is typically $O(K)$; hence M_t is typically $O(K^2)$ or $O(K^3)$.

The Direct Product Mix Formulation

For the assumptions of the previous section, it is possible to construct an exact formulation of the capacity limitations without using any allocation variables at all. We call such a formulation the Direct Product Mix formulation. This formulation is developed by introducing capacity constraints for *sets* of alternative machine types. The particular sets that need to be included depend on the machine usage patterns appearing in the problem data, whereby the total number of sets required may turn out to be as small as the number of different machine types or as large as the power set of all machine types. The justification for this formulation is based on an examination of the allocation problem in terms of Duality Theory for Network Flows. Before plunging into this theoretical justification, we summarize the construction of the Direct Product Mix formulation as follows. First, a procedure we shall term the *Capacity Set Generation Procedure* is used to identify the sets of machine types to be represented with corresponding capacity constraints in the formulation. Basically, the specification of such a capacity constraint involves the definition of two sets: a set of machine types whose capacities are summed to form the right hand side of the constraint, and a set of operations (i.e., product-steps) whose loads are summed to form the left-hand side of the constraint. The procedure develops these sets as follows:

Capacity Set Generation Procedure

Step 1. Identify all sets of alternative machine types appearing in the problem data that are suitable for performing one or more product-steps. Capacity constraints will be formulated for each of these identified sets. The set of operations for this constraint will include all operations that load the identified set of machine types or any proper subset. The capacity of the set of machine types for this constraint is the sum of the scaled capacities of the machine types in the identified set.

Step 2. For all sets of machine types identified in Step 1 that have elements in common, form unions of these sets of machine types and unions of the operations that would appear in the corresponding constraints. Constraints are also formulated for all new, larger sets so formed.

Step 3. Continue this process of forming unions of intersecting sets of machine types so as to combine sets created in Step 2 with each other or with sets identified in Step 1, terminating when no new distinct sets of machine types can be generated.

Let S denote an arbitrary set of machine types generated by this procedure. We write $(i,j) \in S$ to denote that the load from performing j for product i is included in the left hand side of the capacity constraint for set S, and we write $(k) \in S$ to denote that the capacity of machine type k is included in the right hand side of the capacity constraint for set S. The resulting Direct Product Mix formulation can then be written as follows:

$$\text{Max} \sum_{i=1}^{n} \sum_{t=1}^{T} p_{it} X_{it-L_i} - h_{it} I_{it}$$

subject to

$$\sum_{i=1}^{n} \sum_{(i,j) \in S} a_{ij} X_{it-L_{ij}} \leq \sum_{(k) \in S} C_{kt} \text{ for all generated sets } S;\ \forall t$$

$$\sum_{\tau=1}^{t} X_{i\tau-L_i} - I_{it} + B_{it} = D_{it} \ \forall\ i=1,\ldots,n;\ \forall\ t=1,\ldots,T-1$$

$$\sum_{\tau=1}^{T} X_{i\tau-L_i} + B_{iT} = D_{iT} \quad \forall\ i=1,\ldots,n$$

$$B_{it} \leq D_{it} - d_{it} \quad \forall\ i=1,\ldots,n; \quad \forall\ t=1,\ldots,T$$

$$X_{it} \geq 0,\ I_{it} \geq 0,\ B_{it} \geq 0, \text{ all } i, \text{ all } t.$$

Number of Variables and Constraints

Defined this way, the number of variables appearing in capacity constraints each time period is equal to n, the number of product types, which is always smaller than the number of variables in the previous approaches. The number of capacity constraints per period is the number of generated sets S of machine types, which depends on the usage patterns appearing in the problem data. Two simple examples discussed in the next section illustrate that this number could be as small as K, or as large as $2^K - 1$. As will be discussed in a later section, for most practical cases, this number is close to the number of machine types K, making this formulation very attractive. In fact, the formulation is usually only slightly larger (in terms of the numbers of variables and constraints) than the standard LP planning formulation that does not admit alternative resource types.

Theoretical Justification

The capacity feasibility conditions used in the Direct Product Mix formulation are justified using a modification and application of a feasibility theorem for Transshipment [12] Network Flows. We now transform the capacity limitations of the production planning problem into a Transshipment (more specifically, a Transportation) network representation, whereupon we shall state the feasibility theorem and formally develop its application to the construction of capacity constraints in the Direct Product Mix formulation.

Network Representation of the Capacity Limitations

A transportation—type network representing the capacity limitations of the production planning problem in an arbitrary time period t in the case of alternative machine types is depicted in Figure 1. Let $(i-j)$ be a node representing step j of product type i, and let (k) be a node representing machine type k. An arc $(k,i-j)$ from node (k) to node $(i-j)$ indicates that step j of product type i can be processed on machine type k. These two sets of nodes and the arcs connecting them form a bipartite graph G which may have one or more components (a component of G is a connected subgraph of G that is not connected to other subgraphs of G, i.e., the components of G determine a unique partition of its nodes and arcs [12]). Let A be the set of arcs in the network. The labels (L_{ab}, U_{ab}) indicate the minimal and maximal flows allowed on arc (a,b). The labels on the source nodes on the left denote available inflows, while the labels on the sink nodes on the right denote required outflows to be explained below. Note that the condition that the processing time coefficients a_{ijk} are independent of k is necessary for the network representation of Figure 1. Otherwise, multiple nodes are required to describe each step of each product type—a node for each of the alternative machine types.

The exogenous outflow V_{ij} requirements in the network can be identified as the total workload in time period t for step j of product type i (which in the formulations is expressed in terms of the total production of product type i, i.e., $V_{ij} = a_{ij} X_{it-L_{ij}}$). The C_k inflows to the network represent the (known and limited)

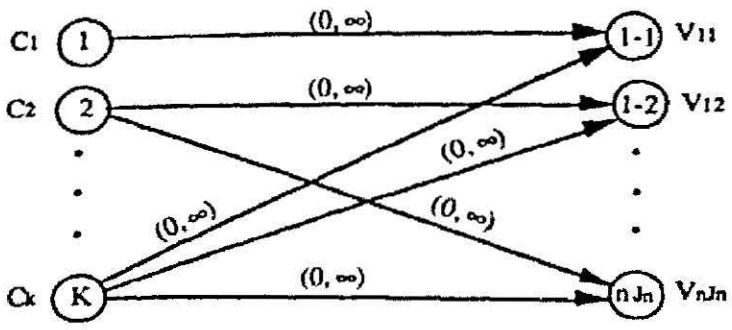

Figure 1: Network representation of the capacity limitations

capacities of the machine types in period t. (C_{kt}'s in the formulation). The W_{ijkt} variable of the Step-Separated formulation, which represents the workload of step j for product type i assigned to machine k, can be identified as the flow on the arc $(k,i-j)$, and the capacity constraints of the Step-Separated formulation, which guarantee that the total workload assigned to each machine does not exceed the machine capacity, correspond to the flow conservation constraints at the machine type nodes. (We omit from the network a dummy sink node for absorbing excess capacities.)

Note also that if the X_{it}'s were known, the V_{ij}'s would be known, and the decision on assignment of workload to machines would become the Transportation [9] problem. Thus the capacity—feasibility of an arbitrary set of X_{it}'s may be evaluated in terms of the conditions for a feasible solution of the underlying Transportation problem.

Applying Gale's Flow Feasibility Theorem

There are several flow feasibility theorems that can be applied to the foregoing network. All of these theorems are specialized versions of the general max-flow min-cut theorem. Hoffman's Circulation Theorem [10], [12] expresses feasibility conditions for general networks. Hall's Theorem [8] is a special case that applies to bipartite graphs for the 0-1 assignment problem, and perhaps could be extended to apply to the problem at hand. However, it is most convenient to make use of Gale's Flow Feasibility Theorem for Transshipment Networks [6], of which Transportation Networks are a special case.

Gale's Feasibility Theorem states (in our notation) that a Transshipment problem is feasible if and only if

$$\sum_{(i-j)\in S} V_{ij} \le \sum_{(k)\in S} C_k + \sum_{(k,i-j)\in(T,S)} U_{k,i-j}$$

for all cutsets (S,T). Here, a cutset is defined as a partition of the set of all nodes into two nonempty, mutually exclusive subsets S and T. We now apply this theorem to generate capacity constraints on the X_{it} variables. We shall start by identifying what we shall term the *dominant cutsets* of the network which are cutsets that correspond to the required capacity constraints. We then demonstrate the redundancy of all other cutsets. Formally, we define the dominant (S,T) cutsets as those cutsets of the set of product-step and machine nodes that satisfy the following conditions:

1. If $(i-j)\in S$ then $(k)\in S \quad \forall\ (k,i-j)\in A$
2. If $\exists\ (i-j^*)$ such that whenever $(k,i-j^*)\in A$ we have $(k)\in S$, then $(i-j^*)\in S$
3. S is connected.

Proposition (Elimination of Dominated Cutset

Cutsets for which one or more of the conditions a are violated are dominated in the sense that the c tions of the flow feasibility theorem are automati satisfied.

Proof:

1. If $(i-j)\in S$ and $\exists\ (k)\in T$ s.t. $(k,i-j)\in A$ then s $U_{k,i-j} = \infty$ we get

$$\sum_{i-j\in S} V_{ij} \le \infty$$

which is always satisfied.

2. If $\exists\ (i-j^*)\notin S$ such that $(k)\in S\ \forall (k)$ s.t. $(k,i-j^*)$ then the resulting constraint is

$$\sum_{(i-j)\in S} V_{ij} \le \sum_{(k)\in S} C_k$$

which is dominated by the constraint resulting fi adding $(i-j^*)$ to S, i.e.,

$$\sum_{(i-j)\in S} V_{ij} + V_{ij^*} \le \sum_{(k)\in S} C_k .$$

3. If S is not connected then without loss of genera let S' and S'' represent two subsets of S s.t. $S'+$ $= S$, S' is connected and S'' is connected. The sulting constraint is

$$\sum_{(i-j)\in S'} V_{ij} + \sum_{(i-j)\in S''} V_{ij} \le \sum_{(k)\in S'} C_k + \sum_{(k)\in S''} C_k$$

which is dominated by the constraints

$$\sum_{(i-j)\in S'} V_{ij} \le \sum_{(k)\in S''} C_k$$

$$\sum_{(i-j)\in S''} V_{ij} \le \sum_{(k)\in S''} C_k$$

Applying the flow feasibility theorem to the domi cutsets, we obtain a set of inequalities of the for

$$\sum_{(i-j)\in S} V_{ij} \le \sum_{(k)\in S} C_k$$

i.e., in the appropriate form to serve as LP cap constraints, where the V_{ij}'s are the workloads inc by the LP variables (the X_{it} variables defined ea representing the feasible amounts of product type can be processed, and the C_k's are the machine capacities (the C_{kt}'s in the formulation). Interp the definition of a dominant cutset, note that it d a set of machine types and a set of product-steps actly those product-steps that can be performed b or more of the defined machine types are include

connectedness of S indicates that S is defined in terms of a set of machine types that with respect to one another are all alternative types for some particular product-step, or else S is defined as a union of such sets that have one or more elements in common. It should now be clear that the Capacity Set Generation Procedure introduced at the beginning of this section is precisely the procedure for generating the sets of machine types and product-steps that correspond to the dominant cutsets.

Assuming the input data is sorted by product-steps and by machine types, the theoretical complexity of the Capacity Set Generation Procedure is $O(PN + P2^K)$, where P is the maximal cardinality of the alternative machine sets, K is the number of the machine nodes of the maximal (in terms of number of the machine nodes) connected component of the bipartite graph of product-steps and machine types, and N is the number of product-steps in the problem data. The first step of the procedure requires an $O(PN)$ sorting operation to eliminate multiple identical machine sets and identify supersets and their subsets. The second and third steps of the procedure require a number of comparisons which is $O(P2^K)$. Note that although the Procedure is exponential, it is practically very useful since K, the maximal number of machine types in a connected component of the bipartite graph, is typically a relatively small number, as explained in the next section.

Comparison of the Formulations

At the outset it should be noted that the constraints of the Direct Product Mix formulation have much wider application than just inclusion in linear programming models. The capacity-feasibility of any proposed set of start rates (X_{it}'s) may be checked by simply substituting the proposed schedule into these constraints. In contrast, the Step-Separated and Workload Allocation formulations require one to develop optimal values for allocation variables to assess the capacity-feasibility of proposed start rates.

It is also important to note that in most industrial enterprises, the overall data set describing capacity limitations can be broken down into separate data sets describing several independent groups of machine types. For example, a semiconductor wafer fabrication facility might have capacity data concerning photolithography machines, ion implant machines, plasma etch machines, etc., whereby each group performs different kinds of processing steps. The capacity constraints for each group may be constructed separately, and then combined to form the complete set of constraints. For each independent group, K will be typically a relatively small number, e.g., 1-4. Thus, we are primarily interested in studying the size of formulations with small values of K. The number of products and the number of process-steps per product is perhaps more variable from enterprise to enterprise. For semiconductor wafer fabrication facilities we have studied, the number of product types n ranges from 10-300, while the average number of process steps per product per independent group of machine types ranges from as low as one for certain kinds of deposition equipment to a high of 18 for the photolithography group. An enterprise-level planning model may involve dozens of manufacturing facilities, thousands of products and hundreds of machine types, but nevertheless the overall formulation of capacity limitations may be developed by combining formulations for relatively small, independent groups of machine types.

We now compare the dimensions of the Step-Separated, Workload Allocation and Direct Product Mix formulations. As noted before, the precise dimensions depend on the particular machine usage pattern evident in the problem data. We shall focus on a couple of simple examples that concisely illustrate the range of dimensions of these formulations that may be encountered. For each example, we shall note the number of constraints, variables and nonzero elements in each type of formulation. We make the assumption that all products have one or more process-steps that load each set of alternative machine types appearing in the data. For simplicity of exposition of the results for the Step-Separated formulation, we further assume that for each product there are exactly J process-steps (operations) loading one or more of the K machine types, and that the assignment of process-steps to sets of machines is uniformly distributed.

In the first example, suppose the only sets of alternative machine types appearing in the problem data are the singleton machine type A, the pair of machine types A and B, the set of three machine types, A, B, and C, and so on, up through the set of all K machine types. That is, the sets of alternative machine types for the various process-steps can be arranged as a series of *nested subsets*. As will be discussed in the next section, this is the most common pattern of alternative machine sets in semiconductor manufacturing. The middle column of Table 1 displays dimensions of the formulations of this example. As can be seen, the Step-Separated formulation pays a heavy price for the generality of non-uniform processing times, with the number of capacity-related constraints per time period $O(nJ + K)$, and the number of variables appearing in these constraints

FORMULATION TYPE	PROBLEM TYPE	
	Nested Sets	All Singletons and All Pairs
Step Separated		
Constraints	$\left(1 - \frac{1}{K}\right)nJ + K = O(nJ + K)$	$nJ \frac{\binom{K}{2}}{K + \binom{K}{2}} + K = O(nJ + K)$
Variables	$\frac{nJ}{K}\left\{\frac{K(K+1)}{2} - 1\right\} + n = O(nJK)$	$2nJ \frac{\binom{K}{2}}{K + \binom{K}{2}} + n = O(nJ)$
Nonzeros	$nJ\left(K - \frac{3}{K} + 2\right) + n = O(nJK)$	$nJ\,(n+4) \frac{\binom{K}{2}}{K + \binom{K}{2}} + nK = O(n^2J + nK)$
Workload Allocation		
Constraints	$2K - 1$	$K + \binom{K}{2} = O(K^2)$
Variables	$n + \{\frac{K(K+1)}{2} - 1\} = O(n + K^2)$	$n + 2\binom{K}{2} = O(n + K^2)$
Nonzeros	$nK + 2\{\frac{K(K+1)}{2} - 1\} = O(nK + K^2)$	$n\left[K + \binom{K}{2}\right] + 4\binom{K}{2} = O(nK^2)$
Direct Product Mix		
Constraints	K	$2^K - 1$
Variables	n	n
Nonzeros	nK	$n\,(2^K - 1)$

Table 1. Comparison of Formulation Dimensions for the Case of Nested Sets and the Case of All Singletons and All Pairs (Number of Capacity-Related Constraints per Time Period, Number of Variables, and Number of Non-Zero Elements Appearing in these Constraints)

$O(nJK)$. Capacity-related constraints dominate demand constraints in this formulation by a factor of J. For large values of n and J, this formulation is prohibitively large.

In contrast, the Direct Product Mix formulation in each time period has the bare minimum number of capacity constraints (the number of machine types K) and the bare minimum number of variables (the number of product types n). It is essentially no different in its matrix dimensions than the basic LP formulation that does not admit alternative machine types. (This result is obtained because Steps 2 and 3 of the Capacity Set Generation Procedure do not generate any new distinct sets for the nested-subset structure.)

Compared to the Direct Product Mix formulation, the Workload Allocation formulation for this case includes almost double the number of capacity-related constraints per time period (specifically, $2K-1$), with an additional $O(K^2)$ variables appearing in these constraints. The number of nonzero elements also is larger by an additive factor of $O(K^2)$.

The last column of Table 1 displays results for an example that yields the worst-case machine usage pattern for the Direct Product Mix formulation, i.e., a usage pattern that leads to the power set when the Capacity Set Generation Procedure is applied. To obtain this result requires a somewhat pathological example in which every machine type and every combination of two machine types each appear as a suitable set of machine types for at least one operation (the case of "All Singletons and All Pairs"). In such a case, Steps 2 and 3 of the Set Generation Procedure will generate all unions of 3, 4,...,K machine types, and the resulting number of capacity constraints per time period is

$$\sum_{i=1}^{K}\binom{K}{i} = 2^K - 1 .$$

Employing the Workload Allocation formulation to develop capacity constraints for this example, this formulation generates $O(K^2)$ capacity constraints per time period on $O(n + K^2)$ variables. For small K in this problem type, the two formulations are comparable, but the Workload Allocation formulation is preferred when K is 4 or more. With respect to translating these comparisons into optimization run-time comparisons, it should be remembered that the capacity constraints in these two formulations typically account for a much smaller portion of the overall constraint matrix than do the demand constraints, which are identical among the alternative formulations.

The Step-Separated formulation for this example still has the very unattractive features whereby the number of variables and the number of capacity-related constraints are proportional to nJ.

Alternative Machine Types Arising in Semiconductor Manufacturing

In a company-wide capacity model for semiconductor manufacturing that the first author developed, the resulting number of capacity constraints per time period when the Direct Product Mix formulation was applied turned out only slightly larger than the total number of resources K. To understand this result requires an explanation of the underlying technological factors giving rise to alternative machine types in this industry. The theme underlying the vast majority of cases of alternative machine types is technological progress, whereby the alternative machine types can be ordered in terms of capability. The most common attribute of technological progress is machine precision. For example, the newest machine type may be capable of precision to very fine geometries, making it eligible to perform operations with the most critical geometries, as well as all other operations. Older machine types may be unable to perform the most critical operations. After several generations of progressively improved machine types have been installed, the factory finds itself in a state such that only the highest precision machine type can be used for the most critical operations, the best two types can be used for the next most critical set of operations, and so forth, down to the set of least critical operations which may be performed on any of the alternative machine types. As we have seen, the number of capacity constraints per time period for the Direct Product Mix formulation in these cases is exactly K, with only the n planning variables appearing in them.

For nested set patterns of machine usage, it is clear that the Direct Product Mix formulation is preferred from the point of view of formulation size. In applying the Direct Product Mix formulation to more than 100 independent groups of alternative machine types, pathological examples leading to the complete power set explosion of constraints, such as the all-singletons-and-all-pairs-of-types example discussed in the previous section, were never encountered for values of K larger than 2.

The remarkable economy of the Direct Product Mix formulation relies on the assumption of uniformity of processing times among alternative machine types, so it is of interest to know how reasonable this assumption is in industrial practice. Reviewing processing time data from many semiconductor factories, we find that the proportionalities of processing times for alternative machine types are typically very similar but not exactly constant across different processing steps. Again, there are technological reasons underlying this empirical result. The overall processing time of an operation consists of two parts: a true machine processing portion, and a generally much smaller portion devoted to material handling just before and just after true processing activity. A new generation of equipment occasionally will proportionally improve the speed of true processing, but continue to use the same material handling technology. Since the handling activity is only a small portion of total processing time, the overall processing times for old and new machine types are approximately proportional.

A good approximation of this situation may be developed by computing weighted-average ratios of the processing times among alternative machine types, and then converting the processing time and capacity data into equivalents for a machine type arbitrarily selected as the "standard." Thus the Direct Product Mix formulation of the planning problem may be employed with only relatively minor loss of accuracy.

Conclusions and Further Research

In this paper we have analyzed techniques for formulating capacity constraints describing the limitations of alternative machine types in corporate-level production planning models. We have proposed the Direct Product Mix formulation for this purpose, justified using a network representation of the process and a flow feasibility condition that determines the required set of capacity constraints. The main advantage of this approach is that under a uniformity assumption concerning processing times, capacity limitations of alternative machine types can be precisely expressed using a formulation that is typically comparable in size to the basic

LP formulation that does not admit alternative resource types. These results have important implications for industrial practice, suggesting that in the case that processing times are nearly proportional, the prevalent, crude approximation that represents capacity of alternative machine types in terms of an artificial, average resource may be replaced by the refined approximation that processing times among the alternatives are proportional. Another advantage is that the resulting set of capacity constraints can be used to check the feasibility of suggested production schedules or demands simply by plugging them into the constraints, without need to develop values for allocation variables.

For large numbers of alternative machine types that do not appear as nested sets in the problem data, the Workload Allocation formulation may be preferred. Cases where this formulation is preferred seem to be rare in practice.

The problem of formulating the capacity constraints for alternative resources that are not independent (e.g., alternative tools and alternative machines used at the same time) is not treated in this article. This problem is very common in semiconductor testing operations and is the subject of current research.

Acknowledgements

This research was supported by grants to the University of California at Berkeley from Harris Corporation—Semiconductor Sector and from Semiconductor Research Corporation. We are grateful to Harris Corporation—Semiconductor Sector for the opportunity to study capacity data from more than 20 semiconductor manufacturing facilities.

REFERENCES

[1] Bai, X., and Gershwin, S. B., "A Manufacturing Scheduler's Perspective on Semiconductor Fabrication," Laboratory for Manufacturing and Productivity, MIT (1989).

[2] Bitran, G. R., and Tirupati, D., "Planning and Scheduling for Epitaxial Wafer Production Facilities," *Operations Research*, Vol. 36, No. 1, pp 34-49 (1988).

[3] Burman, D. Y., Gurrola-Gal, F. J., Nozari, A., Sathaye, S., and Sitarik, J. P., "Performance Analysis Techniques for IC Manufacturing Lines," *AT&T Technical Journal*, Vol. 65, No. 4, pp 46-56 (1986).

[4] Dantzig, G. B., *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ (1967).

[5] Federgruen, A., and Groenevelt, H., "Preemptive Scheduling of Uniform Machines by Ordinary Network Flow Techniques," *Management Science*, Vol. 32, No. 3, pp 341-349 (1986).

[6] Gale, D., *The Theory of Linear Economic Models*, McGraw-Hill Publishing Co., New York, NY (1960).

[7] Gershwin, S. B., "Hierarchical Flow Control: a Framework for Scheduling and Planning Discrete Events in Manufacturing Systems," *IEEE Proceedings of Special Issue on Discrete Event Systems* (1988).

[8] Hall, P., "On Representatives of Subsets," *J. London Math Society*, Vol. 10, pp 26-30, 1935.

[9] Hillier, F. S., and Lieberman G. J., *Introduction to Operations Research*, McGraw-Hill Publishing Co., New York, NY (1986).

[10] Hoffman, A. J., "Some Recent Applications of the Theory of Linear Inequalities to Extremal Combinatorial Analysis," *Proceedings of Symposium on Applied Mathematics*, Vol. 10 (1960).

[11] Johnson, L. A., and Montgomery, D. C., *Operations Research in Production Planning, Scheduling, and Inventory Control*, John Wiley & Sons, New York, NY (1973).

[12] Lawler, E. L., *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart, and Winston, New York, NY (1976).

[13] Leachman, R. C., "Modeling Techniques for Global Production Planning in the Semiconductor Industry," ESRC Report 92-7, Engineering Systems Research Center, University of California, Berkeley (1992), to appear as a chapter in *Optimization in Industrial Environments*, John Wiley & Sons, New York, NY (1992).

[14] Spence, A. M., and Welter, D. J., "Capacity Planning of a Photolithography Work Cell in a Wafer Manufacturing Line," *Proceedings of the IEEE International Conference on Robotics and Automation* (1987).

[15] Uzsoy, R., Lee, C. Y., and Martin-Vega, L., "A Review of Production Planning and Scheduling Models in the Semiconductor Industry," Research Memorandum No. 90-11, Purdue University (1991).