International Journal of Production Research

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713696255

Optimal supply chain configurations in semiconductor manufacturing

To link to this article: DOI: 10.1080/00207540600792499
URL: http://dx.doi.org/10.1080/00207540600792499

International Journal of Production Research,
Vol. 45, No. 3, 1 February 2007, 631–651

Optimal supply chain configurations in semiconductor manufacturing

DAVID CHIANG†‡, RUEY-SHAN GUO†‡, ARGON CHEN*†,
MENG-TSE CHENG† and CHENG-BANG CHEN†

†Graduate Institute of Industrial Engineering, National Taiwan University,
‡Graduate Institute of Business Administration, National Taiwan University

(*Revision received March 2006*)

As a semiconductor supply chain becomes widespread and the competition pressure is very fierce, the detrimental effects of increasing varieties and variations are magnified in the supply chain. But many important issues, such as different service priorities, adaptability, controllability and scalability of performance metrics, have not been addressed in the literature. Conventional modelling techniques of supply chain operations are no longer effective for supply chain configuration. Therefore, a proposed empirical model was first built to catch up the relationship between supply-chain configuration and metrics under the influence of the variability sources. Next, an optimal supply chain configuration model is formulated as a polynomial goal programming model to accommodate different goal objectives. Finally, an efficient solution methodology is further developed to find out the optimal supply chain configuration. Our results show that our proposed approach can easily be adapted to the practices in semiconductor supply chain, and the solution methodology developed in this paper is truly promising.

Keywords: Semiconductor supply chain configuration; Polynomial goal programming model; Heuristics

1. Motivations and objectives

Semiconductor fabrication itself is a very complicated manufacturing process. Its global, cross-company supply chain operations (figure 1) are even more complicated and dynamic such that usual planning and scheduling solutions sometimes have become impossible to employ. Many semiconductor companies in Taiwan tried to implement the most sophisticated advanced planning and scheduling (APS) solutions provided by notable supply chain planning (SCP) vendors, but the results of implementation are less satisfactory.

As reported by Maiorana and Iuliano (1997), manufacturing variability is a critical factor with great impact on work-in-process (WIP) level, cycle time, and throughput. We use coefficient of variation (CV) as a measure of the manufacturing variability. Figure 2 shows that Q-curves describe the relationship between cycle time

*Corresponding author. Email: achen@ntu.edu.tw

International Journal of Production Research
ISSN 0020–7543 print/ISSN 1366–588X online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00207540600792499

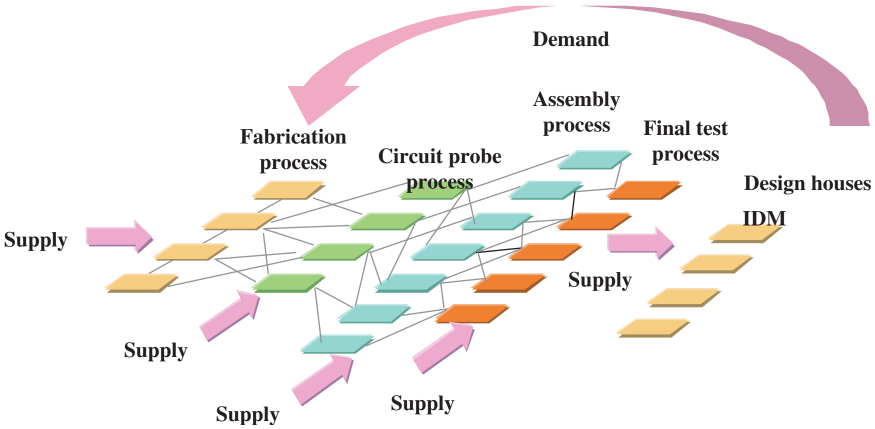

Figure 1. Semiconductor supply chain.

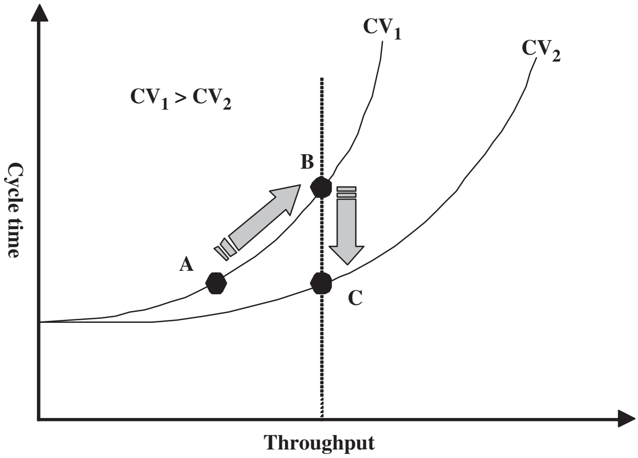

Figure 2. Q-curves under different CVs.

and throughput under different levels of CV. Q-curves in figure 2 are observed based on a queueing system that can be used to mimic a semiconductor supply chain. Under constant system variability CV$_1$, manufacturing planners usually pursue a higher throughput and move point A to point B in figure 2. A higher throughput cannot be achieved without paying a price. As shown in figure 2, the cycle time is rising exponentially as the throughput increases. Based on the Little’s law, it also leads to an exponentially increased WIP level and causes a lot of material handling problems both on the shop floor and in the supply chain. In fact, it is not impossible to achieve a higher throughout while keeping or even shortening the cycle time. The solution is to keep the operational variability low, i.e. moving from point B to point C to lower the coefficient of variation from CV$_1$ to CV$_2$.

Hopp and Spearman (2000) had reported several causes of variability including physical dimensions, process times, machine failure/repair times, quality measures, temperatures, material hardness, setup times and so on. There are two major sources existing: process time variability and flow variability. Reducing variability is a subject intensively discussed and investigated in the fields of stochastic control and quality engineering. It has been long recognized in the discipline of statistical process control (SPC) that unnecessary manual interferences will only result in bigger process variability. Contrary to many engineering approaches demanding variability elimination, the philosophy of statistical control is to tolerate, while closely monitor, a certain level of variability. In addition to off-line system design to minimize the variability, on-line monitoring of variability also helps to seek out causes resulting in unusually large variability and hence the opportunity to further improve the system performance. Recently, such a philosophy has been adopted by production control researchers (Chen *et al.* 2000) and practitioners (Croft and Hand 2003, Devlin *et al.* 2003, Segal and Kalir 2003) alike. In (Chen *et al.* 2000), both statistical optimization and control techniques have been proposed and applied to semiconductor manufacturing systems.

On the other hand, there are several supply chain researchers focusing more on developing a better plan under a complicated and dynamic environment in semiconductor industry. Padillo *et al.* (1995) presented a strategic decision support system (DSS), which has been conceptualized and designed by SEMATECH to assist the large semiconductor manufacturer in managing its extensive supply chain network. This DSS is intended to be used for evaluating future factory concepts and to assist business planners in strategic decision about product allocation and major resource planning.

Toktay and Uzsoy (1998) addressed a capacity allocation problem for a semiconductor wafer fabrication facility. They formulated the problem as a maximum flow problem on bipartite network with integer side constraints and developing efficient heuristics which obtain near-optimal solutions in negligible computation time based on the objectives of maximizing throughput and minimizing deviation from predetermined production goals. Rupp and Ristic (2000) presented a distributed planning methodology for semiconductor manufacturing supply chains. The developed system is based on an approach that leaves as much responsibility and expertise for optimization as possible to the local planning systems while a global coordinating entity ensures best performance and efficiency of the whole supply chain. They used the Factory Planner (FP), one module of the X-CITTIC system as a tool of implementation that can optimize both local and global order flow through the Virtual Enterprise concept.

Frederix (2001) developed a methodology, more flexible and efficient than the traditional time-bucket-based techniques and dynamic dispatching heuristics, to plan the extended semiconductor enterprise and schedule work at the different production entities. The methodology uses stepwise search procedures to improve plans and make-or-buy decision processes to solve resources constraints. Hung and Cheng (2002) proposed a hybrid capacity modelling for an alternative machine types production planning problem, and justified their model that would build a more accurate model at the expense of a small loss in speed.

Christie and Wu (2002) presented a multistage stochastic programming model for strategic capacity planning at a major US semiconductor manufacturer. Their model's objective is to minimize the gaps between products' demands and the capacity allocated to the technology specific for each product. They constructed two scenario-analysis independent versus arbitrary, and showed that independent model allows a varying degree of scenario aggregation without significant prior information, while the arbitrary model allows planners to play out specific scenarios given prior information. Karabuk and Wu (2002) examined the main issues of a decentralized coordination scheme in a setting observed at a major US semiconductor manufacturer: marketing managers reserve capacity based on product demands, while attempting to maximize profit; manufacturer managers allocate capacity to competing marketing managers so as to minimize operating cost while ensuring efficient resource utilization. They formulated the local marketing and manufacturing decision problem as a separate stochastic program, then formulated a centralized stochastic programming model to maximize the firm overall demands. Jo Min (1995) constructed and analysed an economically efficient way of pricing and allocating semiconductor chips of which production technology is characterized by persistent quality variations and of which production capacity is exceeded by potential demand. In his model, specification levels and allocation priorities of competing orders from customers are systematically determined for a single profit maximizing producer. He formulated the producer's profit maximization problem as a nonlinear programming program and investigated the optimality of the proposed allocation rule. Harker (2004) presented an integrated model of incentive problems arising in forecasting and capacity allocation. He proposed a game theoretic model and designed a mechanism (a bonus scheme for all managers and an allocation rule to allocate realized capacity to the product managers) that elicits truthful reporting by all managers. He also showed that large classes of allocation rules, including the current allocation practice of the firm, are manipulable. Helal and Jones (2004) proposed hybrid simulation environment to provide the practical framework to achieve the needed integration. They also studied the impact of the top decisions in developing a production schedule as well as in rescheduling the production facility as needed.

Although many researchers have worked on the topics related to semiconductor supply chain, most of the studies still focus on discussion of internal dispatching rules, resource allocation, and real-time production scheduling in a centralized or distributed supply chain with a single objective (for minimizing cost or maximizing throughput or profit). However, many important issues in the semiconductor supply chain, such as different service priorities, adaptability to process varieties and engineering changes, controllability and scalability of performance metrics, have not been addressed in the literature. Conventional modelling techniques of supply chain operations are no longer effective for supply chain configuration, performance diagnoses and improvement. Thus, the objective of this research is to focus on the strategic decisions related to allocate orders to different routes at different service priorities and control points given the fixed product mix and capacity mix by considering different performance measures simultaneously.

In order to catch the relationship between the supply chain configuration and performance matrices, an empirical supply chain model is developed first to describe how the supply chain configuration affects the chosen performance metrics and their variability. With such models, an optimal supply chain configuration can be formulated for different types of products, priorities, routes, and control points. With this configuration, a possible monitoring mechanism can be developed in the future to quickly find out the adjustment solution once any configuration change is detected.

This paper is organized as follows: initially, the research motivation and objectives are provided; next, an empirical supply chain model using both simulation and response surface method is constructed; after the empirical supply chain model is developed, an optimal supply chain configuration goal programming model intending to determine the best allocation mechanism in the supply chain is constructed, and a heuristic of solving a polynomial nonlinear goal programming is provided based on the characteristics of the goal programming model; an example consistent with the practices in Taiwan semiconductor industry is also provided to illustrate our approach and validate our model; finally, brief conclusions and future related researches are presented.

2. An empirical model for semiconductor supply chain

There are several performance metrics which can be identified for the needs of semiconductor supply chain. These includes cycle time, level of WIP, product throughput, delivery performance, capacity utilization, yield rates, and so on. However, from the entire supply chain point of view, the most valuable performance metrics are cycle time, WIP, product throughput and delivery performance because other measures such as capacity utilization and yield rates are more likely internal performance driven. Therefore, we choose the mean and variability (standard deviation) of cycle time, WIP, and demand fulfillment rate as the metrics to evaluate the supply chain performance in semiconductor manufacturing.

Because lots of factors in semiconductor manufacturing may affect the supply chain performance metrics, it is thus necessary to consider these factors before the supply chain model is built up. These factors include: demand and product mix, number of tiers in the supply chain, number of facilities in each tier, production capability of each facility, cycle time of each product at each facility, production capacity of each facility, dispatching rules setting for each control point, and the length of planning horizon. For instance, choosing different routes and control points or setting different priority for customer demands will have great impact on the performance metrics such as cycle time and on-time delivery. Therefore, in this study, we intend to explode the impacts of allocation variables on the performance metrics in semiconductor supply chain. These allocation decision variables investigated in this study are defined as follows: π_{kq} the percentage of product k assigned to priority group (q); ρ_{kr} the percentage of product k assigned to supply chain route (r), and τ_{ri} the percentage of route r assigned to supply chain control point (i). Since these variables are measured in percentages, it can be easily implemented as customer demands fluctuate over the planning horizon. The relationship among these allocation decision variables is shown in figure 3. Besides, we assume that the priority mix is independent of the supply chain route mix without loss of generality.

In order to describe the interrelationships between chosen performance metrics and these supply chain allocation decision variables in semiconductor manufacturing, an empirical supply chain simulation model is developed initially. In this simulation model, major input parameters includes the presetting value of allocation variables (priority mix percentage, route mix percentage, control point mix percentage), demand arrival time interval and quantity, total demand, product mix

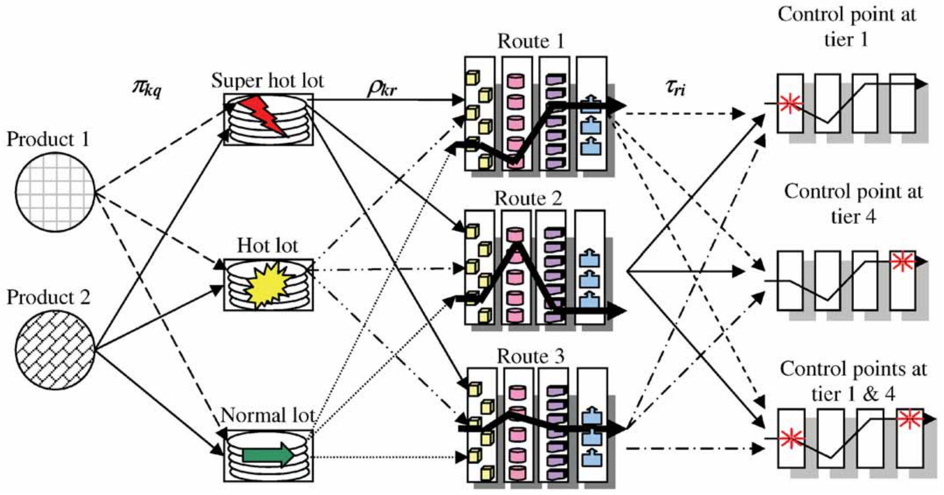

Figure 3. Supply chain allocation decision variables.

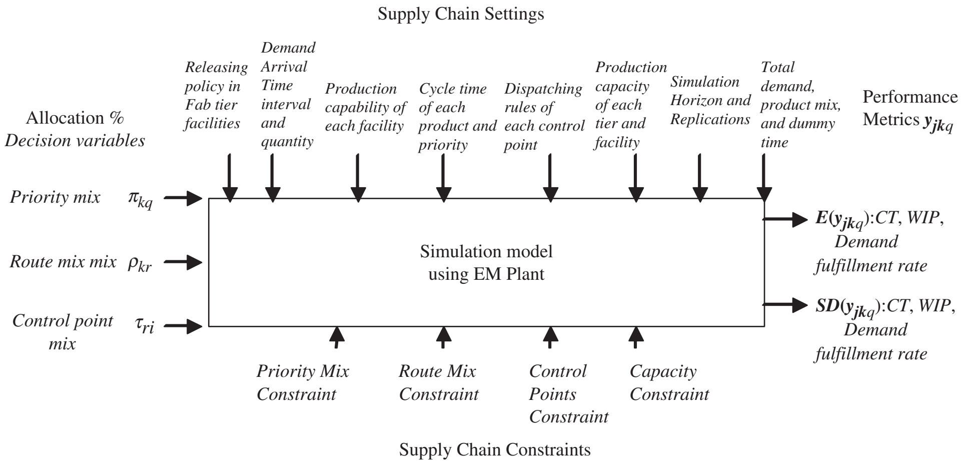
Supply Chain Settings
Allocation % Decision variables
Releasing policy in Fab tier facilities
Demand Arrival Time interval and quantity
Production capability of each facility
Cycle time of each product and priority
Dispatching rules of each control point
Production capacity of each tier and facility
Simulation Horizon and Replications
Total demand, product mix, and dummy time
Performance Metrics y_{jkq}
Priority mix π_{kq}
Route mix mix ρ_{kr}
Control point mix τ_{ri}
Simulation model using EM Plant
$E(y_{jkq})$:CT, WIP, Demand fulfillment rate
$SD(y_{jkq})$:CT, WIP, Demand fulfillment rate
Priority Mix Constraint
Route Mix Constraint
Control Points Constraint
Capacity Constraint
Supply Chain Constraints

Figure 4. Supply chain simulation model.

and dummy time, simulation horizon and replications, number of facilities in each tier, production capability of each facility, cycle time of each product at each priority, production capacity of each tier and facility, dispatching rules of each control point and releasing policy in Fabrication tier facilities. The supply chain constraints are priority mix constraints, route mixed constraints, control points' constraints and capacity constraints. (These constraints will be discussed further in the next section). Finally, the performance metrics could be the means and standard deviations of cycle time, WIP, and demand fulfillment rate. The supply chain simulation model can be illustrated in figure 4. Before simulations are run, some basic setting, demand setting, capacity setting, due date setting and cycle time estimation are predetermined based on empirical data collected from the semiconductor industry in Taiwan. Next, the D-optimal approach of experimental design is adopted (Atkinson and Donev 1992), and the corresponding performance metrics for each run were collected and recorded. After that, a response surface function of each customer priority group is generated from the results of these runs to indicate the interrelationships between each performance metric and these supply chain allocation decision variables. Of course, once the product mix changes significantly, the corresponding response surface functions need to be regenerated to accommodate the changes.

3. An optimal configuration model

Once the response surface functions of each set of performance metrics and supply chain allocation variables for different customer priority groups are found, an optimal configuration model in a semiconductor manufacturing is ready to be constructed. Before the model is presented, related indices and variables are defined:

Indices

i threads (service control points), $i = 1, \dots, I$
j performance metrics, $j = 1, \dots, J$
k type of products, $k = 1, \dots, K$
q service priority, $q = 1, \dots, Q$
r service routes, $r = 1, \dots, R$
t tiers of supply chain, $t = 1, \dots, T$
ϕ factory in each tier, $\phi = 1, \dots, \Psi$

Parameters

E_t utilization rate of tier t in semiconductor supply chain,

$C_{\psi t}$ capacity ratio at factory ψ of tier t in semiconductor supply chain,

p_k the percentage of product k in product mix,

$PT_{\psi t}$ average production cycle time of single product at factory ψ of tier t in semiconductor supply chain,

$PT_{k\psi t}$ average production cycle time of product k at factory ψ of tier t in semiconductor supply chain,

η_q the maximum percentage of products produced at the priority q,

α_r the maximum percentage of products produced at production route r,

y_{jk} the performance metric j of product k at the priority q,

$\mathrm{E}(y_{jkq})$ the mean of the performance metric j of product k at the priority q which is determined by the response surface method. That is, $\mathrm{E}(y_{jkq})$ is the function of π_{kq}, ρ_{kr}, and τ_{ri},

$\mathrm{SD}(y_{jkq})$ the standard deviation of the performance metric j of product k at the qth priority which is determined by the response surface method. That is, $\mathrm{SD}(y_{jkq})$ is the function of π_{kq}, ρ_{kr}, and τ_{ri},

$W_{E(yjkq)}/W_{SD(yjkq)}$ the weights of the *j*th performance metric (mean or standard deviation) of product *k* at the priority *q*.

Decision variables

π_{kq} the percentage of product *k* assigned to be produced at the priority *q*,
ρ_{kr} the percentage of product *k* assigned to be produced at the route *r*,
τ_{ri} the percentage of route *r* assigned to be controlled by different control point *i*.

Since this model must consider several performance metrics simultaneously, the weights of performance metrics for different products at different priorities must be determined before a goal programming model is employed. In general, the weights of performance matrices can be determined by using multi-criteria decision approaches such as AHP (Analytical Hierarchical Process) (Saaty 1988) or simply given by subjective weights which are made jointly by the supply chain planners who are responsible for allocating customers' order in supply chain. In order to reflect different levels of importance of the customer demands, the weights for different service priorities should be different, too. The generalized mathematical formulation of goal programming model is listed as follows:

Minimize

$$Z = \sum_{q} \sum_{j,k} (W_{E(y_{jkq})} \times E(y_{jkq}) + (W_{SD(y_{jkq})} \times SD(y_{jkq})) \tag{1}$$

where

$$E(y_{jkq}) = f_{jkq}(\pi_{kq}, \rho_{kr}, \tau_{ri}) \tag{2}$$

$$\mathrm{SD}(y_{jkq}) = g_{jkq}(\pi_{kq}, \rho_{kr}, \tau_{ri}) \tag{3}$$

Subject to

$$\sum_{k} p_k = 1 \tag{4}$$

$$\sum_{k} p_k \cdot \pi_{kq} \leq \eta_q \quad \forall q \tag{5}$$

$$\sum_{q} \pi_{kq} = 1 \quad \forall k \tag{6}$$

$$\sum_{k} p_k \cdot \rho_{rk} \leq \alpha_r \quad \forall r \tag{7}$$

$$\sum_{r} \rho_{rk} = 1 \quad \forall k \tag{8}$$

$$\sum_{i} \tau_{ri} = 1 \quad \forall r \tag{9}$$

$$E_t \sum_{r:\phi \in r} \sum_{k} (p_k \rho_{rk}) \frac{PT_{k\phi t}}{PT_{\phi t}} \leq C_{\phi t} \quad \forall \phi, t \tag{10}$$

The objective function in (1) is to minimize the weighted mean value and the variation of the performance metrics. (Note: if the mean (e.g. for throughput) is to be maximized while minimizing its standard deviation, the sign in front of mean should be changed to be negative.) The mean value and standard deviation of performance metrics are expressed as the function of decision variables π_{kq}, ρ_{kr}, τ_{ri} in (2) and (3).

The meanings of seven sets of constraints from (4) to (10) are explained as follows:

1. Constraint sets in (4) are product mix constraints; i.e. the sum of proportion of all the products must add up to 1. These proportions are inputted from the practices.
2. Constraint sets in (5) are the priority mix constraints; i.e. the total proportion of demands to be produced at the priority level q must not exceed a predetermined upper limit based on the pre-setting policy in practice.
3. Constraint sets in (6) are also related to the priority mix; i.e. the proportion of a product assigned to be produced at different priority levels should be added up to 1.
4. Constraint sets in (7) are route mix constraints; i.e. the total proportion of demands to be produced at the specific route must not exceed a predetermined upper limit based on the pre-setting policy in practice.
5. Constraint sets in (8) also related to the route mix; i.e. the proportion of a product assigned to be produced at different routes should be added up to 1.
6. Constraint sets in (9) are the constraints related to control points; i.e. the proportion of a product assigned to be controlled by different control points should be added up to 1.
7. Constraint sets in (10) are the capacity constraints of facility ψ in supply chain tier t. The constraint sets are derived from the following inequalities:

$$\sum_{r:\phi\in r}\sum_{k}(p_k\rho_{rk})D_tPT_{k\phi t} \leq \frac{C_{\phi t}PT_{\phi t}D_t}{E_t} \quad \forall\phi, t \tag{11}$$

where D_t is the average total demand during a planning horizon. Since the model considers that all the customer demands must be manufactured through all the tiers of semiconductor supply chain, D_t must be the same for all the tiers. Thus, the meaning of (11) is that the total production time at factory ϕ of tier t needed under the optimal configuration must be less than total available capacity at factory ϕ of tier t by considering the current utilization rate at tier t. Because both sides of the inequality have D_t, D_t can be canceled out. Finally, we multiply both sides of the inequality by E_t to obtain the inequality of (10).

Once the model is formulated, it can be solved by applying different methodologies. We will recommend a solution methodology in the next section. The solution of our optimal configuration model can be straightforwardly adapted to the real decisions. When a new order of product k arrives, a service priority of that order may be pre-assigned by the supply chain planner initially. But the status of the new order needs to be further confirmed due to the current availability at that priority, i.e. the new order may keep the same original priority or lower its priority level depending on the availability at that priority by comparing the optimal percentage of product k assigned at that priority with current percentage of product k already assigned at that priority. Once the priority level of that order is assigned, current percentage of different routes and control points combinations at that particular service priority level based on the existing scheduled orders will be reexamined, and the combination with maximum difference between the optimal percentage and the current percentage will be assigned to that order. Of course, the current percentage needs to be updated if an order is completed and removed from the production list.

4. The solution methodology

The solution methodology becomes very difficult to develop when a large variety of products and routes are involved in semiconductor supply chain and the response surface function has a nonlinear property. In order to reduce the complexity of solution methodology, we develop an approximation method to solve a polynomial goal programming model. This method is named as semi-definite quadratic approximation method (SQAM) (Arjan *et al.* 1996, Auslender and Coutat 1996). The details of our algorithm are shown in figure 5, and explained as follows:

Step 1: Build up the model: an optimal configuration model is built up based on response surface functions at each priority group and constraint sets defined in section 3.

Step 2: Test semi-definite quadratic property: if the objective function has a quadratic form, the semi-definite quadratic property needs to be tested on that objective function by examining whether the quadratic form is semi-definite, i.e. all the eigenvalues of quadratic matrix must be no less than 0. If the objective function is a semi-definite quadratic function, then go to step 8. Otherwise, go to step 3.

Step 3: Approximate objective function toward a semi-definite quadratic function: the objective function will be approximated toward semi-definite quadratic function by applying the least square method to all range of decision variables.

Step 4: Check the errors within the tolerance limits. The errors of the least square method in step 3 are computed and compared to the allowable tolerance limits predetermined by decision maker. If the errors are within the allowable tolerance limits, then go to step 8.

Step 5: Determine the number of hyper-planes to cut through the feasible region: appropriate cut points or hyper-planes are searched for cutting through the feasible range of decision variables to form several disjoint regions. In this paper, we first set the second-order partial derivatives of objective function with respect to each decision variable to be zero to determine the inflation function of each decision variable. Since the inflation function of one decision variable is the function of other decision variables, the inflation points of one decision variable will be changed if the hyper-plane jointly determined by other inflation functions changes. In order to determine the hyper-planes to cut through the feasible region, we use the results from step 3 by choosing the global optimal obtained at step 3 to be a reference point. To determine the cut-through hyper-plane of the *i*th decision variable, we input all the values of decision variables at the reference point except the value of the *i*th decision variable into the inflation function of the *i*th decision variable to determine

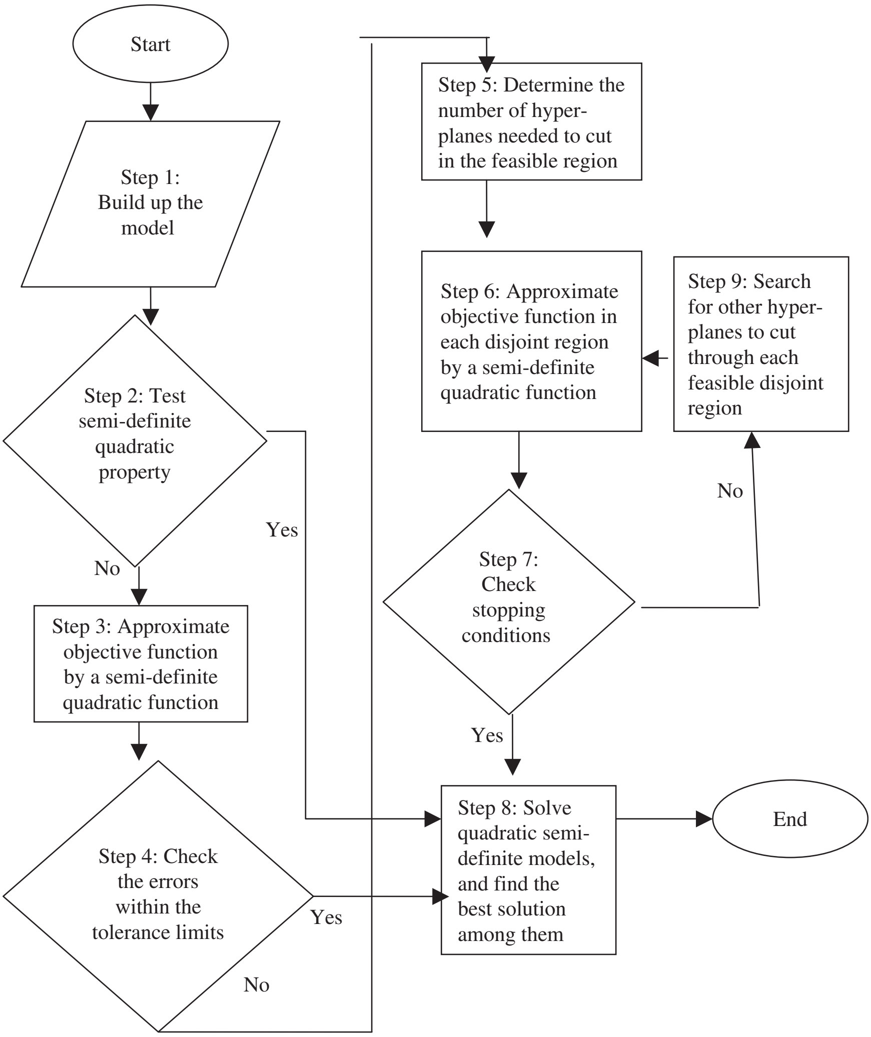
Start
Step 1: Build up the model
Step 2: Test semi-definite quadratic property
Yes
No
Step 3: Approximate objective function by a semi-definite quadratic function
Step 4: Check the errors within the tolerance limits
Yes
No
Step 5: Determine the number of hyper-planes needed to cut in the feasible region
Step 6: Approximate objective function in each disjoint region by a semi-definite quadratic function
Step 9: Search for other hyper-planes to cut through each feasible disjoint region
No
Step 7: Check stopping conditions
Yes
Step 8: Solve quadratic semi-definite models, and find the best solution among them
End

Figure 5. Semi-definite quadratic approximation method (SQAM).

the reflection points of the *i*th decision variable. These reflection points of the *i*th decision variable will cut the entire solution space into several disjoint solution space. By performing the procedure repeatedly for each decision variable, we may cut the entire solution space into huge numbers of disjoint solution spaces. To avoid huge computational efforts, we only consider the reflection points of important decision variables to cut through the solution space to tradeoff the computation time with solution quality. The rules of choosing important decision variables are stated as follows: the first priority of choosing one decision variable is to choose the one that has more reflection points. If there are two variables that have the same number of reflection points, we choose the one which has the minimum sum of distance between reflection points and the centre of its feasible region. In this paper, we set the maximum number of important decision variables to be a finite number, U.

Step 6: Approximate the objective function toward semi-definite quadratic function of decision variables in each disjoint region. For each disjoint region, the objective function is approximated toward semi-definite quadratic function by applying the least square method to each disjoint feasible region.

Step 7: Check stopping conditions: if the approximation is good enough to meet the terminated conditions, then go to step 8. Otherwise, go to step 9. These terminated conditions include: the errors are within a predefined tolerance range, the maximal number of regions is obtained, and no other hyper-plans can be found to cut through the regions.

Step 8: Solve a quadratic semi-definite model in each disjoint region by adding the constraints (4) to (10) in the goal programming formulation, and find out the optimal solution by applying Wolfe-dual method (Wolfe 1959, Rusin 1971) and stop. Since only limited cuts are generated which is constrained by U, the algorithm will terminate in a finite number of iterations.

Step 9: Search for other hyper-planes to cut through each feasible disjoint region: if the approximation fails to meet the terminated condition, other appropriate cut points or hyper-planes need to be searched for cutting through the feasible range of decision variables to form more sub-regions.

5. An example

In order to demonstrate our methodology, an example based on industrial practices in Taiwan semiconductor industry is illustrated. By collecting the data from research papers and personal interviews from the semiconductor industry, we are able to build up a simulation model. The production environmental setting in our simulation is described as follows:

1. There are four tiers in our simulation model including FAB, Circuit Probe (CP), Assembly and Final test. Also, six FAB plants exist in the first tier, two Circuit Probe plants exist in the second tier two assembly plants exist in the third tier, and two final test facilities are available in the fourth tier. The capacity in each facility of tiers 1, 3, and 4 is listed in table 1. The capacity

Table 1. The capacity at each facility of each tier.

FAB	Capacity	Assem.	Capacity	FT	Capacity
FAB1	1468 K	Assem1	3265	FT1	3200
FAB2	1376 K	Assem2	3200	FT2	3265
FAB3	922 K	Total	6465	Total	6465
FAB4	1133 K	• Capacity in 200 m wafers per year			
FAB5	1202 K	• Fab 1: TSMC 5, 6, or 8			
FAB6	689 K	• Fab 2: UMC 8C, 8D, 8E, or 8F			
Total	6465 K	• Fab 3: TSMC 2			
		• Fab 4: TSMC 3, 4, or 7			
		• Fab 5: UMC 6A, or 8AB			
		• Fab 6: WaferTech, VIS or SSMC			

in the second tier, CP, is assumed to be infinite since it is not a bottleneck stage in the practice of Taiwan semiconductor industry.

2. Three major product groups are produced in supply chain network: product A has the simplest production process, product B has more complex production process, and product C can only be produced at some facilities with certain technology capability. There are capacity contentions between three different products at the factory level. The ratio of these three types of products produced in our supply chain network is close to 2:1.7:1. Besides, there are three priorities for each product: regular, hot lot and super hot lot. Based on different priorities from the practice of Taiwan semiconductor industry, the customers' required delivery duration are very different; for example, the required delivery duration for super hot lot is set to be 1.3 times the raw process time of each product; the required delivery duration for hot lot is set to be 2.1 times the raw process time of each product; and the required delivery duration for regular lot is set to be 3 times the raw process time of each product in our simulation model. The details of possible routes for each product and the definitions of each route are shown in figure 6 and table 2, respectively.
3. Since, in most of the cases, the control point is located at FAB tier in Taiwan's semiconductor supply chain, thus, we do not include the control

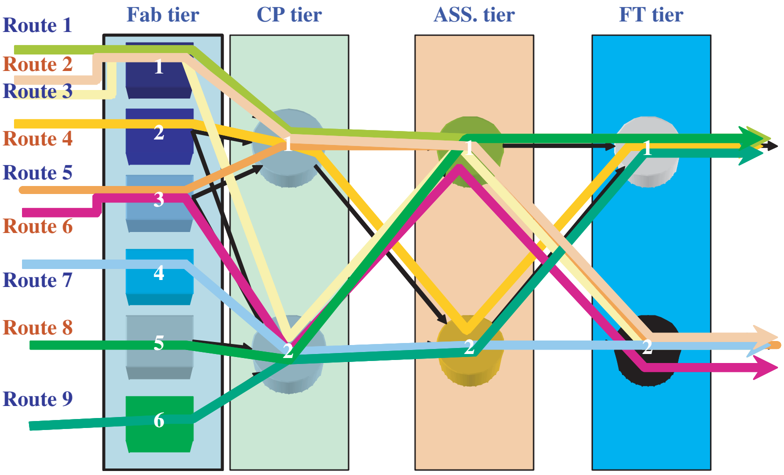

Figure 6. Semiconductor supply chain routes.

Table 2. Product mix versus route mix.

Product	Route no.
A	145 789
B	2467
C	34

point level related decision variables and parameters in our example. Also, the dispatching rule used at each of the facilities in the example is FCFS (First come first serve).

4. The results of the current studies shows that the product cycle time is related to the product types, equipments used in different facilities, service priorities and capacity utilization at each facility. In order to simulate the effect of cycle time of single product in a complex network while other products sharing the same facility, we first estimate the average product cycle time of each product at each facility with different service priorities under different capacity utilization rates by assuming that each product has an assigned percentage of the capacity. Once the average product cycle times are estimated, we are able to obtain the real product cycle times by taking these values into our simulation model to simulate the joint effect of different products sharing the same facility. For a single product, the production cycle time is assumed to be infinite if capacity utilization rate at one facility approaches 100% and the production cycle time is raw processing time if capacity utilization rate is 0%. By following these assumptions, the product cycle time for each product at each plant in different priorities can be estimated based on different utilization rates for each product at each plant in different priorities. Many researchers have done lots of studies on how to estimate the product cycle time in the Fab by applying queueing model. In this example, an M/M/1 model for each facility in semiconductor supply chain is assumed based on some researchers' studies on Fab product cycle time (Raddon and Grigsby 1997, Wood 1997, Chung and Huang 2002). The general function of cycle time with capacity utilization rate based on M/M/1 model can be approximated as $a * \mathrm{e}^{(b/(2-x))}$, where a and b are constants, and x is capacity utilization rate. Our simulation model is based on 80% capacity utilization rate which is average utilization rate at the Taiwan semiconductor industry in the past several years, and the parameters a and b are estimated from the data collected from the personal interview at the industry by using a model fitting package, Best-fit. The expected cycle times and raw process time for each product at each plant of different priorities in FAB, Assembly, and Final test are listed in tables 3, 4, and 5, respectively. The expected cycle time in CP is assumed to be equal to raw process time since we assume that there is infinite capacity in each facility at the stage of CP.

In this simulation example, we choose X-factor and standard deviation of cycle time as our performance metrics. And the factors of experimental design are the decision variables π_{kq} and ρ_{kr}. We design five levels for each factor, but total levels in this experimental design have only 15 levels because the sum of decision variables π_{kq} and ρ_{kr} must add up to 1. These level settings are shown in table 6. Next, a D-Optimal method is adopted such that 180 simulation runs are generated. Finally, each run is performed 20 replicates, and the duration of each replicate lasts for 90 days.

By performing the response surface method, we are able to obtain the response surface functions at different priority levels. For instance, the resulting response surface functions at priority 1 group are shown below.

$$\begin{aligned}
\text{X-factor}_1 = {} & 3.04028 - 0.76932 \times \rho_{18} - 5.32407 \times \pi_{11} \times \rho_{18} - 1.70306 \times \rho_{22} \\
& + 2.55627 \times \pi_{11} \times \rho_{22} - 3.21206 \times \rho_{24} + 2.0617 \times \pi_{11} \times \rho_{24} \\
& + 3.35593 \times \rho_{22} \times \rho_{24} + 6.45807 \times \rho_{24}^2 - 6.01907 \times \rho_{26} \\
& + 2.64128 \times \pi_{11} \times \rho_{26} + 2.43013 \times \rho_{18} \times \rho_{26} + 3.95107 \times \rho_{22} \times \rho_{26} \\
& + 3.59718 \times \rho_{24} \times \rho_{26} + 6.27775 \times \rho_{26}^2 - 2.08025 \times \rho_{24} \times \rho_{33} \\
& + 0.95507 \times \rho_{33}^2 - 0.89682 \times \rho_{22} \times \pi_{12} - 2.73278 \times \rho_{22} \times \pi_{21} \\
& - 2.94563 \times \rho_{24} \times \pi_{21} + 4.33951 \times \pi_{11} \times \pi_{31} + 2.20954 \times \rho_{22} \times \pi_{31} \\
& + 2.26427 \times \rho_{26} \times \pi_{31} - 3.90953 \times \rho_{33} \times \pi_{31} + 3.88601 \times \pi_{11} \times \pi_{32} \\
& - 0.69994 \times \rho_{33} \times \pi_{32} - 3.96577 \times \pi_{11} \times \rho_{11} - 1.57582 \times \rho_{22} \times \rho_{14} \\
& - 1.4345 \times \rho_{15} - 6.01892 \times \pi_{11} \times \rho_{15} + 4.19858 \times \rho_{26} \times \rho_{15} \\
& + 6.76398 \times \pi_{21} \times \rho_{15}
\end{aligned}$$

$$\begin{aligned}
\text{CT-STD}_1 = {} & 10.76424 - 30.39664 \times \rho_{18} + 20.91593 \times \rho_{17} \times \rho_{18} + 44.38031 \times \rho_{18}^2 \\
& - 7.04376 \times \rho_{22} + 3.75327 \times \pi_{11} \times \rho_{22} - 19.49143 \times \rho_{17} \times \rho_{22} \\
& - 14.41669 \times \rho_{24} - 19.90259 \times \rho_{17} \times \rho_{24} + 17.13541 \times \rho_{22} \times \rho_{24} \\
& + 19.52764 \times \rho_{24}^2 - 16.44176 \times \rho_{26} - 20.93092 \times \rho_{17} \times \rho_{26} \\
& + 15.80926 \times \rho_{22} \times \rho_{26} + 16.97536 \times \rho_{24} \times \rho_{26} + 23.21307 \times \rho_{26}^2 \\
& + 5.01638 \times \rho_{22} \times \pi_{22} - 6.09423 \times \pi_{31} + 9.03262 \times \rho_{22} \times \pi_{31} \\
& + 9.70458 \times \rho_{24} \times \pi_{31} + 7.29345 \times \pi_{12} \times \pi_{31} - 16.26179 \times \rho_{11} \\
& + 18.03107 \times \rho_{17} \times \rho_{11} + 27.58228 \times \rho_{18} \times \rho_{11} - 5.73555 \times \pi_{12} \times \rho_{11} \\
& + 28.52365 \times \rho_{17} \times \rho_{14} + 29.47225 \times \rho_{18} \times \rho_{14} - 5.43344 \times \rho_{22} \times \rho_{14} \\
& + 23.20821 \times \rho_{11} \times \rho_{14} - 55.31215 \times \rho_{14}^2 - 17.66627 \times \rho_{15} \\
& + 17.71866 \times \rho_{17} \times \rho_{15} + 22.28432 \times \rho_{18} \times \rho_{15} + 11.24435 \times \rho_{26} \times \rho_{15} \\
& - 5.38951 \times \pi_{22} \times \rho_{15} + 32.58414 \times \rho_{11} \times \rho_{15} + 25.65426 \times \rho_{14} \times \rho_{15}
\end{aligned}$$

Before the polynomial goal programming model can be formulated, the response surface functions of X-factor and standard deviation of cycle time at each priority are validated by R^2 value shown in table 7. Since all the R^2 of response surface function are greater than 0.75, we believe that our simulation model can really catch up the characteristics of the data.

After that, the objective function of a polynomial goal programming model is formed by adding up the product of response surface function for each performance measure at each priority and the pre-assigned weight corresponding to each priority group. In this paper, we assume that subjective weights for priorities 1, 2, and 3 are 15, 5, and 1, respectively. Once the model is formulated, our proposed semi-definite quadratic approximation method is used to solve the problem. The results of polynomial goal programming are obtained. The optimal results for priority mix allocation and for route mix of product 1, 2, and 3 are shown in tables 8, 9, 10, and 11.

Table 3. Estimated cycle time with raw process time (RPT) for products, plants and priorities in FAB.

Product		Product 1		Product 2		Product 3	
FAB	Priority	CT_{Expect}	RPT	CT_{Expect}	RPT	CT_{Expect}	RPT
FAB1	Priority 1	65145.79	43545.6	72207.19	55065.6	81200.68	65491.2
(Min)	Priority 2	74119.98	43545.6	86415.72	55065.6	97630.54	65491.2
	Priority 3	88554.68	43545.6	101967.7	55065.6	108735.7	65491.2
FAB2	Priority 1	67332.99	46569.6	75223.14	55209.6	84255.93	66974.4
	Priority 2	77308.54	46569.6	85297.14	55209.6	99159.08	66974.4
	Priority 3	92718.25	46569.6	101261.4	55209.6	111180.8	66974.4
FAB3	Priority 1	69076.43	45576	77143.15	57096	Not	Not
	Priority 2	75997.7	45576	88053.38	57096	Not	Not
	Priority 3	91623.63	45576	103922.3	57096	Not	Not
FAB4	Priority 1	64217.57	40000.4	77806.07	55886.4	Not	Not
	Priority 2	72970.5	40000.4	86780.05	55886.4	Not	Not
	Priority 3	86251.17	40000.4	104554.1	55886.4	Not	Not
FAB5	Priority 1	68337.91	45748.8	Not	Not	Not	Not
	Priority 2	76444.73	45748.8	Not	Not	Not	Not
	Priority 3	90985.53	45748.8	Not	Not	Not	Not
FAB6	Priority 1	71811.93	43027.2	Not	Not	Not	Not
	Priority 2	73571.61	43027.2	Not	Not	Not	Not
	Priority 3	89618.61	43027.2	Not	Not	Not	not

Table 4. Estimated cycle time with raw process time for products, plants and priorities in assembly.

Product		Product 1		Product 2		Product 3	
Fab	Priority	CT_{Expect}	RPT	CT_{Expect}	RPT	CT_{Expect}	RPT
Asse 1	Priority 1	11529.42	8523.07	12016.27	9001.94	12423.72	9403.24
(Min)	Priority 2	13382.98	8523.07	13971.16	9001.94	14387.28	9403.24
	Priority 3	15578.61	8523.07	15697.45	9001.94	17025.96	9403.24
Asse 2	Priority 1	11567.02	8560.02	12162.65	9146.06	12569.84	9547.2
	Priority 2	12979.55	8560.02	13584.52	9146.06	13997.58	9547.2
	Priority 3	15200.21	8560.02	15823.9	9146.06	16248.83	9547.2

Thus, Product 1 should be produced 5% at the first priority (super hot lot), 25% at the second priority (hot lot), and 70% at the third priority (normal); Product 2 should be produced 5% at the first priority (super hot lot), 10% at the second priority (hot lot), and 85% at the third priority (normal); Product 3 should be produced 15% at the first priority (super hot lot), 10% at the second priority (hot lot), and 75% at the third priority (normal) Also, although there are six alternative routes available to produce product 1, all six routes 1, 4, 5, 7, 8, and 9 are chosen to use and the proportion of production at these routes are 10, 10, 20, 16.46, 20 and

Table 5. Estimated cycle time with raw process time for products, plants and priorities in final test.

Product		Product 1		Product 2		Product 3	
Fab	Priority	$\mathrm{CT}_{\mathrm{Expect}}$	RPT	$\mathrm{CT}_{\mathrm{Expect}}$	RPT	$\mathrm{CT}_{\mathrm{Expect}}$	RPT
FT1	Priority 1	18126.78	15051.3	19459.48	16376.1	19583.39	16499.34
(Min)	Priority 2	20075.95	15051.3	21420.23	16376.1	21545.15	16499.34
	Priority 3	24223.82	15051.3	25604.66	16376.1	25732.75	16499.34
FT2	Priority 1	11593.87	11593.87	19450.56	16713.24	19286.78	16550.16
	Priority 2	20197.47	15170.94	21761.9	16713.24	21596.65	16550.16
	Priority 3	24036.97	15170.94	25639.42	16713.24	25470.4	16550.16

Table 6. The levels of experimental design in the example.

	Lv.1	Lv.2	Lv.3	Lv.4	Lv.5
π_{kq}					
π_{11}	0.05	0.075	0.1	0.125	0.15
π_{12}	0.1	0.15	0.2	0.225	0.25
π_{13}	–	–	–	–	–
π_{21}	0.05	0.075	0.1	0.125	0.15
π_{22}	0.1	0.15	0.2	0.225	0.25
π_{23}	–	–	–	–	–
π_{31}	0.05	0.075	0.1	0.125	0.15
π_{32}	0.1	0.15	0.2	0.225	0.25
π_{33}	–	–	–	–	–
$\rho_{\kappa\rho}$					
ρ_{11}	0.1	0.13	0.167	0.18	0.2
ρ_{14}	0.1	0.13	0.167	0.18	0.2
ρ_{15}	0.1	0.13	0.166	0.18	0.2
ρ_{17}	0.1	0.13	0.167	0.18	0.2
ρ_{18}	0.1	0.13	0.166	0.18	0.2
ρ_{19}	–	–	–	–	–
ρ_{22}	0.1	0.175	0.25	0.275	0.3
ρ_{24}	0.1	0.175	0.25	0.275	0.3
ρ_{26}	0.1	0.175	0.25	0.175	0.3
ρ_{27}	–	–	–	–	–
ρ_{33}	0.4	0.45	0.5	0.55	0.6
ρ_{34}	–	–	–	–	–

Table 7. R^2 of each response surface function.

R^2	X-factor	Standard deviation of cycle time
Priority 1	0.8641	0.9562
Priority 2	0.7534	0.7828
Priority 3	0.9206	0.9175

Table 8. An optimal priority mix allocation.

	Product 1	Product 2	Product 3
Priority 1	0.05	0.05	0.15
Priority 2	0.25	0.1	0.1
Priority 3	0.7	0.85	0.75

Table 9. An optimal route mix allocation of product 1.

Route 1	Route 4	Route 5	Route 7	Route 8	Route 9
0.1	0.1	0.2	0.16463	0.2	0.23537

Table 10. An optimal route mix allocation of product 2.

Route 2	Route 4	Route 6	Route 7
0.269471	0.215127	0.178696	0.336706

Table 11. An optimal route mix allocation of product 3.

Route 3	Route 4
0.497767	0.502233

23.54%, respectively. Product 2 should be manufactured 26.95% by route 2, 21.51% by route 4, 17.87% by route 6 and 33.67% by route 7. Product 3 should be produced 49.78% by route 3 and 50.22% by route 4. Our optimal solution is compared with the results from Lingo. The results show that our method has a lower objective function value in this example.

In this example, it takes average 1.5 min for each run (20 replicates) and 4.5 h for total 180 runs to generate the response surface model, and takes 40 min to solve the model. The computational results are reported from running on a PC with Intel Pentium M processor 1.5 GHz and 768 Ram.

Finally, in order to validate our results, we compare the pre-180 runs of simulation results, goal programming results, and the simulation result by implementing our best configuration. These results of validation of X-factor and cycle-time variability are shown in tables 12 and 13, respectively. From the results, it can be found that average X-factor of priority 1, 2, and 3 by applying our optimal configuration model are 1.4033, 1.6198, and 2.0961, compared to the results that average X-factor of priority 1, 2, and 3 from our original simulation model are 1.5897, 1.8307, and 2.1226. Also, a 95% confidence interval for the X-factor at each different priority is constructed to show the statistical significance of our approach. From the results of table 12, it shows that our goal programming results and

Table 12. Validation of X-factor.

X-factor		Priority 1	Priority 2	Priority 3
180 Runs of simulation experiments	Min	1.403319	1.47294	1.703227
	Lower bound of confidence interval (95%)	1.56803	1.81195	2.10761
	Average	1.5897	1.8307	2.12257
	Upper bound of confidence interval (95%)	1.61137	1.84945	2.13754
	Max	2.195282	2.193816	2.338799
Goal programming results		*1.4034*	*1.6198*	*2.0961*
Simulation validation of goal programming		*1.3873*	*1.6251*	*1.9941*

Table 13. Validation of cycle time variability.

Cycle time STD (month)		Priority 1	Priority 2	Priority 3
180 Runs of simulation experiments	Min	0.485356	0.815753	1.245471
	Lower bound of confidence interval (95%)	1.00998	1.37429	1.77755
	Average	1.09511	1.43061	1.8152
	Upper bound of confidence interval (95%)	1.18024	1.48694	1.85285
	Max	3.531987	2.48982	2.604576
Goal programming results		*0.1928*	*0.4684*	*1.6018*
Simulation validation of goal programming		*0.2383*	*0.4429*	*1.7613*

simulation validation results for all three priorities are significantly better than pre-optimization cases. Therefore, our approach has greatly improved the performance of supply chain for all three priority groups. The same results are found for validation of cycle-time variability in table 13. The above result is also validated by the simulation result by implementing our best configuration.

6. Conclusions and remarks

In summary, in this research we first build up an empirical model to describe the relationship between supply-chain configuration and metrics under the influence of the variability sources. Next, an optimal supply chain configuration model is formulated as a polynomial goal programming model to accommodate different goal objectives. Finally, an effective solution methodology is developed further to find out the most robust supply chain configuration. Our simulation results show that our proposed model and methodology are really promising. In the future, more intensive experimental tests are needed to justify the effectiveness and the efficiency of our algorithm. Also, a possible monitoring mechanism can be further developed to quickly find out the adjustment solution once an abnormal situation is detected.

Acknowledgement

We thank Bou-Chia Chung for implementing the simulation model. This research was supported by SRC (SRC/ISMT 2004 OJ-1214), NSC (NSC93-2218-E-002-048), and NSC (NSC93-2416-H-002-012).

References

Atkinson, A.C. and Donev, A.N., *Optimum Experimental Design*, 1992 (Oxford University Express: New York).

Arjan, B.B., Jansen, B. and Terlaky, T., The optimal set and optimal partition approach to linear and quadratic programming. *Econometric Institute Report* 30, Econometric Institute, Erasmus University, Rotterdam, The Netherlands, 1996.

Auslender, A. and Coutat, P., Sensitivity analysis for generalized linear-quadratic problems. *J. Optimization Theory and Algorithm*, 1996, **88**(3), 541–559.

Chen, A., Guo, R.S. and Lin, P., Statistical analysis and design of semiconductor manufacturing systems. *International Symposium on Semiconductor Manufacturing* (*ISSM'2000*), Tokyo, Japan, 2000.

Christie, R.M. and Wu, S.D., Semiconductor capacity planning: stochastic modeling and computational studies. *IIE Trans.*, 2002, **34**, 131–143.

Chung, S.H. and Huang, H.W., Cycle time estimation for wafer fab with engineering lots. *IIE Trans.*, 2002, **34**(2), 105–118.

Croft, T. and Hand, B., Using real-time data in reducing idle time due to avoidable interrupptions, *International Symposium on Semiconductor Manufacturing* (*ISSM'2003*), pp. 63–66, San Jose, CA, 2003.

Devlin, P., Thielemier, C. and Wehrlin, D.E., Lost utilization – constraint performance management. *International Symposium on Semiconductor Manufacturing* (*ISSM'2003*), pp. 51–54, San Jose, CA, 2003.

Frederix, F., An extended enterprise planning methodology for the discrete manufacturing industry. *Eur. J. Oper. Res.*, 2001, **129**, 317–325.

Harker, P.T., Coordinating supply chain with competition: capacity allocation in semiconductor manufacturing. *Eur. J. Oper. Res.*, 2004, **159**(2), 330–347.

Helal, M. and Jones, A., a study of the impact of production scheduling using enterprise simulation. *IIE Annual Conference and Exhibition*, pp. 839–840, 2004.

Hopp, W. and Spearman, M., *Factory Physics*, 2nd ed., 2000 (McGraw-Hill: New York).

Hung, Y. and Cheng, G., Hybrid capacity modeling for alternative machine types in linear programming production planning. *IIE Trans.*, 2002, **34**, 157–165.

Jo Min, K., Economic determination of specification levels and allocation priorities of semiconductor products. *IIE Trans.*, 1995, **27**(3), 321–331.

Karabuk, S. and Wu, S.D., Decentralizing semiconductor capacity planning via internal market coordination. *IIE Trans.*, 2002, **34**, 743–759.

Maiorana, A. and Iuliano, G., Improving Cycle Time through managing variability in a DRAM production line, in *Proceedings of IEEE International Symposium on Semiconductor Manufacturing*, 1997, pp. 29–32.

Padillo, J.M., Ingalls, R. and Brown, S., A strategic decision support system for supply network design and management in the semiconductor industry. *Computers Industry Eng.*, 1995, **29**, 443–447.

Raddon, A. and Grigsby, B., Throughput time forecasting model. *Proceedings of the 1997 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 430–433, 1997.

Rupp, T.M. and Ristic, M., Fine planning for supply chains in semiconductor manufacture. *J. Mater. Processing Technol.*, 2000, **107**, 390–397.

Rusin, M.H., A revised simplex method for quadratic programming. *SIAM J. Appl. Math.*, 1971, **20**(2), 143–160.

Saaty, T., *The Analytic Hierarchy Process*, 1988 (McGraw-Hill: New York).

Segal, T. and Kalir, A., A breakthrough in utilization maximization via real-time tool performance feedback, *International Symposium on Semiconductor Manufacturing* (*ISSM'2003*), pp. 39–41, San Jose, CA, 2003.

Toktay, L.B. and Uzsoy, R., A capacity allocation problem with integer side constraints. *Eur. J. Oper. Res.*, 1998, **109**, 170–182.

Wolfe, P., A simplex method for quadratic programming. *Econometrics*, 1959, **37**, 382–398.

Wood, S.C., Cost and cycle time performance of Fabs based on integrated single-wafer processing. *IEEE Trans. Semiconductor Manuf.*, 1997, **10**(1), 98–111.