



MULTIPLE CONSTRAINT SYNCHRONIZATION IN A HIGH-MIX, LOW-VOLUME MANUFACTURING ENVIRONMENT

Author: R. Michael Mahoney



MULTIPLE CONSTRAINT SYNCHRONIZATION IN A HIGH-MIX, LOW-VOLUME MANUFACTURING ENVIRONMENT

INTRODUCTION

Customers are demanding products that offer solutions to their particular problems. Manufacturers are increasingly working with customers to provide the solutions know-how and capabilities to provide a custom customer solution for their customer's specific problem. Manufacturer's who pursue such a strategy are known as agile competitors. The only constraint in this environment will be the constantly increasing variety of products. Product design is becoming more modular to facilitate the ease of configuration required to economically satisfy the requirements of a particular customer solution. A proliferation in the diversity of products demanded by customers, along with a consequent reduction in volumes produced, has resulted in an inescapable evolution to high-mix, low-volume manufacturing. High-mix, low-volume manufacturing environments are complex. For high-mix, low-volume manufacturing environments, mix can be as high as 600 different products or greater, and incoming order quantities may range anywhere from 0-1,000 or greater for individual products within the mix over a particular time horizon.

Although quality and cost are order winners (i.e., differentiators) for lean production environments that exist today, they are marketplace qualifiers for high-mix, low-volume manufacturers. It is important to note that customers are often willing to pay a premium for information and services loaded solutions. Responsiveness and delivery performance are competitive differentiators for a high-mix, low-volume manufacturing environment.

To facilitate excellence in responsiveness and delivery performance, flexibility is a must. It is essential to thoroughly understand the relevant dimensions of flexibility (i.e., mix, volume, and workforce flexibility), and the extent to which flexibility enables a high-mix, low-volume manufacturer to produce a high variety of high-quality, low-cost products. Flexible manufacturing environments are characterized by automated processes that have low sequence-independent setup times, defocused process configurations, and manual processes where the workforce is trained, cross-trained, creative, and empowered to effect process improvements.

Proper planning and scheduling decisions are fundamental to successfully managing a high-mix, low-volume manufacturing environment. The magnitude of the difficulty for solving the high-mix, low-volume manufacturing problem will be presented first. Secondly, a manufacturing operations model will be presented that can be used to facilitate the management decision-making process for properly planning, and accurately predicting results in a high-mix, low-volume manufacturing environment. Finally, the groundbreaking Multiple Constraint Synchronization(MCS™) scheduling algorithm will be presented.

INTRACTABILITY

High-mix, low-volume scheduling problems are intractable (i.e., np-complete where "np" is nondeterministic polynomial). An optimal makespan solution cannot be ascertained without explicitly evaluating each and every scheduling possibility. Difficulties encountered when searching for optimality are referred to as complexity, and are based on what is referred to as the time complexity function. Think of time as the computer time required to explicitly evaluate all scheduling alternatives. Manufacturing scheduling problems involve polynomial and exponential time complexity functions. Polynomial time is computer processing time which grows as a function of n , where: n, n^2, n^3, \dots Exponential time is computer processing time which grows as a function of n , where: $2n, 3n, \dots$ For obvious reasons, exponential time complexity functions have explosive growth rates. Thus, polynomial time complexity functions are generally regarded as more desirable.

High-mix, low-volume manufacturing scheduling problems are referred to as np-hard. Np-hard problems are a subset of np-complete problems.

The concept of np-hard is best demonstrated by example. Consider the case where five products are to be produced by two parallel machines (machine 1 and machine 2) and the processing time for each product is 5, 2, 6, 3, and 4 minutes for products A, B, C, D, and E respectively. The objective is to associate each job with a machine (machine 1 or machine 2) such that makespan (i.e. the total time to process all jobs) is minimal. The solution is to assign products A, B, and D to one of the two machines and assign



products C and E to the remaining machine. The optimal makespan is 10 minutes.

No direct solution to the parallel machine problem has been offered when lot preemption (i.e., lot splitting) is not permitted. The parallel machine makespan optimization problem is np-hard. Np-hard defines the situation where the simplest case is just as hard as the more complex problems one can encounter for a particular problem under consideration. Consider the parallel machine problem where 10 products are to be processed by two machines in the minimal makespan. The parallel machine sequencing problem proves to be formidable. There are $10!2$ possible sequences, and each sequence must be explicitly evaluated in order to ascertain the optimal makespan. Plainly and simply, high-mix manufacturing scheduling problems defy explicit computer enumeration techniques.

Branch and bound methods have been developed which use a strategy where the universe of feasible solutions is eliminated without having to explicitly evaluate them all (a.k.a., implicit computer enumeration). Although optimal solutions are not guaranteed, the resulting solutions are good. Heuristic methods have also been developed to obtain good suboptimal solutions. Heuristic solutions are based on inductive inferences. A heuristic has been developed for the parallel machine problem. This can be accomplished by successively scheduling the longest remaining job to the machine where it will be completed the earliest. If shortest processing time (SPT) sequencing is subsequently performed for the jobs associated with each machine, the mean number of jobs in the system will also be reduced. The complex game of chess is another example where heuristics are used extensively. Chess books abound with such rules.

Complex high-mix manufacturing problems have no suitable analytic solution available and do not have optimal solutions in the strict sense. When planning and scheduling customer orders for a high-mix, low-volume manufacturing environment, the overriding goal is to develop schedule sequences that are in alignment with the overall organizational objectives established at the highest level of a manufacturing organization. The proper modeling of operations is critical in this regard.

MODELING OPERATIONS

It is impractical, if not impossible, to solve high-mix, low-volume manufacturing scheduling problems using a direct analytical approach. Rather than mutilate

a high-mix, low-volume manufacturing scheduling problem until it conforms to a model, the best approach is to modify the solution procedure to fit the problem so as to obtain a directed search of the solution space.

Consider Fig. 1. Sequencing and lot sizing rules have a significant effect on the overall financial position for a high-mix, low-volume manufacturer. The minimum lot size, L_{Min} , is equal to one unit for each product under consideration, and the maximum possible lot size, L_{Max} , is equal to the total quantity ordered for a particular product over a particular time horizon under consideration. Increased lot sizes increase the costs associated with batch error occurrences and inventory while setup time and idle capacity costs are reduced. Conversely, if lot sizes are reduced, the costs associated with batch error occurrences and inventory are reduced while setup time and idle capacity costs are increased. Given the assumption of fixed setup times for the mix of products produced, decreased lot sizes will decrease available capacity and increase responsiveness. For the case of increased lot sizes, available capacity is increased and responsiveness is decreased.

The cost of quality will adversely affect the financial position of a high-mix, low-volume manufacturer across the entire spectrum of lot sizes that can be selected. This is also true for the case where disruptions occur due to machine breakdowns, unscheduled worker absenteeism, stockouts, defective material, etc.

Clearly, the reduction of sequence-independent setup time will positively affect the financial position of any manufacturer over the entire range of lot-sizes that can be selected. It is important to note that group technology methods will not achieve the desired effect in a high-mix, low-volume manufacturing environment. Group technology is based on the notion that a particular ordering of products to be processed can result in the reduction or avoidance altogether of setup time for particular consecutively run products. It will be shown that the sequencing of products through a high-mix, low-volume manufacturing environment must be based on constraint considerations. Group technology methods used for the purpose of local machine setup time avoidance criteria will damage a high-mix, low-volume manufacturer's competitiveness in the marketplace.

By obtaining the monetary values for Work-in-process (WIP), the total cost of quality, the cost of disruptions, and sequence-independent setup cost, the loss function can be derived. Clearly, reductions in sequence-independent setup time, the cost of qual-



ity, and disruptions will have the greatest impact on the financial performance for any manufacturer. The derivative of the loss function with respect to WIP represents the minimum point of the loss function and is referred to as the pareto optimum.

There is a degree of robustness to changing lot sizes that occurs when the operating point is at or near the pareto optimum. It is essential to note that the pareto optimum is preferred only under the condition that customer delivery and responsiveness performance criteria are being satisfied. Competitive differentiators in a high-mix, low-volume manufacturing environment are delivery and responsiveness performance. Thus, the operating point will be shifted toward LMin the degree required to not damage the cost competitiveness of a high-mix, low-volume manufacturer. Typically, customers are willing to pay a premium for superior delivery and responsiveness performance. The costs associated with reductions in lot size are ultimately based on the price customer's are willing to pay for a desired level of responsiveness and delivery performance. Those manufacturer's that only compete based on price will cease to exist in this marketplace.

A situation can also occur where incoming customer order quantities increase to the extent that the current lot size method selected results in the reduction of available capacity to the extent that the makespan required to deliver the quantity ordered is inordinately long. Such a condition can result in the loss of customers if alternative means of raising capacity (e.g., overtime and hiring additional temporary workers) are insufficient to make up the shortfall in available capacity. Lot sizes may have to be increased (i.e., shifted toward LMax) to increase available capacity. This is particularly true if cosourcing or outsourcing are not options. For a high-mix, low-volume manufacturer, increasing and decreasing incoming customer order volumes versus finite manufacturing resources is the dilemma that must continually be addressed. Lot sizing is a strategic driver for satisfying this requirement.

The cross-functional relationship that exists between responsiveness and capacity utilization establishes the parabolic characteristic of the loss function as WIP (i.e., lot size) is varied. Responsiveness is defined as the time required to negotiate (i.e., produce) at least a quantity of one unit of each product produced over the particular time horizon under consideration (e.g., one month). For a high-mix, low-volume manufacturer, capacity utilization is a measure based on the critical constraint resource(s), and is expressed as a percentage of total available constraint capacity.

The model presented identifies and defines those factors affecting a manufacturer's objectives. This will greatly assist management in gaining a better understanding of problems and serve as a focal point for systematic discussion of objectives and alternatives. Clearly, no model can be expected to exactly predict real-world manufacturing conditions.

WIP levels and capacity constraint utilization are not exclusively controlled by lot sizing methods in a high-mix, low-volume manufacturing environment. Sequencing decisions are critical in a high-mix, low-volume manufacturing environment due to the dissimilarities of processing times for particular products within the mix of products to be produced.

MULTIPLE CONSTRAINT SYNCHRONIZATION (MCS™)

To understand, in the simplest terms, the importance of sequencing decisions in a high-mix, low-volume manufacturing environment, we can analyze a single machine process step (see Fig. 2).

Repetitive Just-In-Time manufacturing environments are characterized by products that have relatively equal processing times for each process step. Such a condition facilitates what is known as line balancing. Although line balancing is impossible to achieve in a high-mix, low-volume manufacturing environment, the degree of imbalance can be minimized through the use of balance delay equations. Consider Fig. 2a. It is clear that all sequencing choices will result in the same makespan performance (5 time units). What is also apparent is the mean number of units (3) and mean time (3 time units) that all units spend in the system are constants, irrespective of any sequence chosen. Consider Fig. 2b. High-mix, low-volume manufacturing environments are characterized by products that have unequal processing times for each associated process step. Although all sequencing choices will result in the same makespan performance (15 time units), the mean number of units and mean time that all units spend in the system as WIP are significantly affected by the sequence chosen. For the case of Fig. 2b, the optimal sequence is obtained by running the products in order of non-decreasing processing time (SPT: Shortest Processing Time). The worst possible sequence is obtained by running the products in order of non-increasing processing time (LPT: Longest Processing Time). Despite the clear advantages associated with sequencing, it is astonishing that many manufacturers today pay as little attention as they do to sequencing considerations.



High-mix, low-volume manufacturing environments are characterized by different products that use a common set of manufacturing resources. For dissimilar products, the processing times associated with each process step will most likely differ. One or more capacity constraints will exist and “shift” based on the product mix and volumes being produced over a particular time horizon. Consider Fig. 3. If the processing times for the mix of products produced are constrained at process steps X and Y respectively, and the cumulative total processing time at process step Y is greater than that of X, process step Y will manifest itself as the bottleneck process step. Alternatively, if a shift in the mix and volumes occurs such that the cumulative total constraint processing time at process step X is greater than that of process step Y, process step X will manifest itself as the bottleneck process step. Such occurrences are not atypical of high-mix, low-volume manufacturing environments. The Multiple Constraint Synchronization (MCS™) algorithm was developed to alleviate the problems associated with multiple moving constraints and bottlenecks. Consider the 3-machine constraint problem depicted in Fig. 4. The optimal sequence is determined by using the following algorithm:

MULTIPLE CONSTRAINT SYNCHRONIZATION (MCS™) ALGORITHM

OBJECTIVE: Minimize makespan.

Step 1: Determine the job with the maximum processing time at the most downstream machine. Place this job in the first available position of the sequence.

Step 2: Remove this job and its associated machine from consideration and return to step 1.

If jobs or machines = 0, then stop.

This algorithm is remarkably simple to use and will always find an optimal sequence for minimizing makespan for the multistage, serial machine, multiple constraint synchronization problem. Although the MCS algorithm will find an optimal sequence, it will not find them all (see sequences I and II of Fig. 5). There are two optimal solutions to the 4-machine multiple constraint problem. The MCS™ sequence has the attribute that the inverse of the MCS™ sequence represents the worst possible makespan (see sequence III of Fig. 5). This result can be used to quantify the benefit associated with using the MCS™ sequence. Note that the constraints are aligned (i.e. synchronized) as closely as possible with respect to time for the optimal MCS™ sequences (see Figs 4

and 5). There are advantages in reduced WIP and improved constraint utilization if the processing times for the constraint process steps are equal. Although WIP reduction and improved constraint utilization are highly desirable, equating constraint times is not a necessary condition for obtaining an optimal sequence.

In order to minimize WIP and obtain maximum utilization of all constraint operations, the equating of time at the constraint operations is based on the notion of the system least common denominator (SLCD). The SLCD is selected such that it is not less than the greatest constraint time. This concept is depicted in Fig. 6. The ability to equate constraint times based on the SLCD is based on the lot sizes selected and the granularity of the processing times of the products to be produced.

Consider Fig. 7. All products are listed apriori in SPT sequence under their associated constraint operations. Additional information such as customer order quantity, forecast quantity, work content processing time, and setup time are included for each product to be produced. The aggregate median processing time and setup time is derived for the overall system. The setup time is amortized based on the production rate established during the production planning process. This will ensure that the resulting schedule sequences are in alignment with the overall objectives developed at the highest level of the manufacturing organization. The derived quantity is subsequently multiplied by the aggregate median processing time to obtain the system least common denominator (SLCD). The lot-size for each product is determined by dividing the SLCD by the processing time for the particular product under consideration. Each product at each constraint operation will be iteratively produced in SPT sequence until a condition arises where an insufficient customer order quantity of a particular product is available to satisfy the SLCD requirement. The shortfall can be made up by using the required forecast order quantity for this particular product. Actual customer orders should take precedence over forecast orders. Dissimilar products can be combined to satisfy the SLCD by using the customer order quantity of the next product in SPT sequence required to satisfy the shortfall. If all orders associated with a particular constraint are insufficient to satisfy the SLCD, the remnants will be produced without disrupting the overall objectives of the schedule. All production lots based on the SLCD for each constraint operation are sequenced based on the MCS™ algorithm.



The responsiveness of the overall system is based on the total time required to negotiate at least one product for the entire mix of products produced.

The MCS™ tableau is used to establish the master plan. As actual customer orders are received, the forecast order quantity is decremented by the same amount. In this manner, the forecast order quantity is consumed by actual customer orders. Control of the timing of the execution of the MCS™ sequence is based on the use of generic Kanban. Daily available capacity is established through the use of balance delay calculations.

Fig. 8 depicts WIP simulation results based on real-world data for a manufacturing system before and after the manufacturing system is based on the MCS™ algorithm. The simulation was performed using Witness, a simulation product of AT&T™, Istel.

The simulation was free-run using deterministic data, and unit transfer lot sizes were used.

MCS™ planning alleviates many of the problems inherent to manufacturing resource planning (MRP II) systems by:

- Alleviating problems associated with moving constraints and bottlenecks
- Improving flexibility to changing customer demand
- Improving capacity constraint utilization
- Recognizing changes in lead time that occur based on changes in lot size

Avoiding random start times by forward-scheduling production operations.