Feasibility Study of Real-Time Scheduling Using the Lagrangean Relaxation Method Under an APS Environment*

Kaikou SHIN**, Mitsuru KURODA***
and Kouichi NATSUYAMA****

Advanced Planning and Scheduling (APS) has been widely recognized as a promising method for solving real production planning and scheduling problems. Based on the proposal of a real-time job shop scheduling mechanism under an APS environment, which adopts the Lagrangean relaxation method as the optimization logic, the present paper describes a feasibility study of this mechanism by evaluating its calculation speed and re-scheduling quality. Numerical experiments have been carried out for various models having different scales, as well as different densities and strengths of random events, such as the arrival of new jobs or changes to the due dates for existing jobs. The results of experiments show that the proposed scheduling mechanism has the potential to satisfy the real-time scheduling requirements, not only in terms of calculation speed and solution quality, but also with respect to predictability of the calculation load. Finally, an improvement to the Lagrangean relaxation method is proposed to improve re-scheduling quality.

Key Words: Lagrangean Relaxation Method, Advanced Planning and Scheduling, Re-scheduling

1. Background

The total optimization planning framework described by Shin^(*) has become increasingly feasible. The goal of it is the realization of consistent and efficient planning from the strategic level to the tactical and operational levels over the entire supply chain network. In this framework, the Advanced Planning and Scheduling (APS) mechanism is critical. It attempts to determine optimized and feasible solutions by which to deal with procurement, production and transportation problems while considering various constraints throughout the supply chain. APS has attracted a great deal of attention and has higher expectations due to both the use of various advanced mathematic tools and the integrated nature of this

mechanism, which has brought a number of long-researched functions closer to practice realization.

One such function is real-time scheduling, which is expected to maintain the optimized schedule even when dealing with unpredictable events, e.g. new orders or frequent changes to orders or machine breakdowns. Based on the real-time scheduling capability, a comprehensive real-time *Availability-To-Promise* (ATP) logic has emerged. Figure 1 shows the procedure of this logic, which performs multi-location inventory checks, as well as checks of the existing production schedule and component availability by re-scheduling in order to provide better responsiveness to customer requests.

Two key requirements in realizing the above-mentioned real-time ATP logic are calculation speed and re-scheduling quality. The re-scheduling quality depends on how closely the new schedule corresponds to the original schedule. The calculation speed should be predictable considering the constructing of such mechanisms in the real production environment. Due to the exponential growth of calculation time and the setting of impractical objective functions, application of most theoretical methods is generally difficult in

* Received 2nd September, 2002 (No. 02-4242)

** SCM Solutions, Professional Services Organization, SAP Japan Co., Ltd., Tokyo 100-0004, Japan. E-mail: kaikou.shin@sap.com

*** Information and Systems Engineering, Aoyama Gakuin University, Tokyo 157-8572, Japan

**** RF Planning Group, TU-KA Cellular Tokyo Inc., Tokyo 105-8540, Japan

Fig. 1 Real-time ATP logic

actual practice.

The present paper describes a feasibility study of a real-time job shop scheduling mechanism, which adopts the Lagrangean relaxation method as the optimization logic. The objective is to explore the applicability of this type of mechanism in order to realize the above ATP logic under an APS environment.

Following a short review of relevant research concerning the application of the Lagrangean relaxation method to scheduling problems, a mathematical formulation of job shop scheduling problems and a framework of Lagrangean relaxation method are presented. Then, the experimental models and the simulation workflow which use the formulation and the framework are explained. After the initial experimental results are presented and analyzed, an improvement to the Lagrangean relaxation method is proposed and its effectiveness is proven through further experimentation.

2. Literature Review

Luh et al.⁽⁵⁾ first reported that one of the motivations for developing the Lagrangean relaxation method for scheduling problems is to close the gap between *theoretical methods*, which produce good schedules without the ability to react to dynamic changes, and *practical methods*, which react to dynamic changes without the ability to reliably produce a good schedule. Hoitomt et al.⁽⁴⁾ explored the use of the Lagrangean relaxation method to schedule in job shops which contain multiple machine types, generic precedence constraints and simple routing considerations, and found that near-optimal and feasible schedules could be obtained efficiently. In addition, a procedure by which to evaluate the quality of the schedule by generating a lower bound on the optimal cost was developed. Extensive researches by Luh and his colleagues concerning both of the method and its application prove that the Lagrangean relaxation method has good generality to deal with various scheduling problems, and the obtained solutions can

be evaluated theoretically using their lower bound values.

A smart approach to solving relaxed job level sub-problems using a pseudo-polynomial algorithm based on the dynamic programming approach is proposed by Chen et al.⁽¹⁾ In this algorithm, not only the relaxation of the precedence constraints becomes unnecessary, and thus the oscillation problems existing in operation level sub-problems vanish, but also the application of Lagrangean relaxation method has been expanded to problems which were regarded as indecomposable. As an example, a very complex scheduling problem, in which lot sizing and lot sequencing are to be optimized simultaneously, has been solved by Muramatsu⁽⁶⁾ by adopting the pseudo-polynomial algorithm as part of Lagrangean relaxation method.

In addition to studies concentrating on the Lagrangean relaxation method, Kuroda et al.⁽²⁾ and Kuroda⁽³⁾ explore the application of the Lagrangean relaxation method to dynamic scheduling problems. In order to simulate an actual dynamically changing environment, a concept referred to as the occurrence density of random events is introduced and the behavior of the Lagrangean relaxation method has been widely investigated according to various occurrence densities. The scheduling results are compared to those obtained by widely used dispatching rules, and the Lagrangean relaxation method is proved to be able to generate better schedules for the specific model used for the experiments.

According to Yoneda⁽⁸⁾, two features are believed to provide the Lagrangean relaxation method with the potential for practical application: one is scalability and the other is simplicity of logic. Theoretical analysis of its application to general static scheduling problems provides evidence for the former, whereas for the latter, the logic of the method is perfectly matched to the rules and phenomena of common social and business activity, referred to as *auction*.

Scalability is extremely important for any scheduling method considering practical application. In the present study, scalability is regarded as one of the conditions when evaluating the feasibility of the Lagrangean relaxation method under an APS environment. An explicit relation between calculation speed and the model scale is expected to be clarified. Furthermore, the present study attempts to evaluate scheduling with respect to practical application rather than using the lower bounds, as is the case for most of the above-mentioned studies. This leads to an improvement in the algorithm, which will be explained before the Summary Section.

3. A Lagrangean Relaxation Framework for Job Shop Scheduling Problems

3.1 Problem formulation

A typical job shop scheduling problem for N jobs on M machines is considered in the present research. The objective is to minimize the total cost incurred due to tardiness of the jobs, while considering the constraints arising from operation precedence and machine availability. The following is the notation used in the formulation of the scheduling problem:

J : objective function to be optimized;

J_i : objective function for job i ;

ω_i : penalty cost for job i if tardiness occurred;

C_i : calculated completion time of job i ;

D_i : due date of job i ;

δ_{ijhk} : =1, if operation (i, j) occupies workcenter h at time k ; =0, if operation (i, j) does not occupy workcenter h at time k ;

t_{ij} : processing time of operation (i, j) ;

M_{hk} : available machine number for workcenter h at time k ;

K : planning horizon;

H : type of workcenter;

W_i : a set of operations for job i ;

P_j : a set of operations preceding immediately operation j ;

c_{ij} : completion time of operation (i, j) ;

where δ_{ijhk} and c_{ij} are the decision variables.

The scheduling problem can be formulated as follows:

$$\text{Minimize } J = \sum_{i=1}^N J_i = \sum_{i=1}^N \omega_i \times \max(0, C_i - D_i) \quad (1)$$

subject to the machine available constraint:

$$\sum_{ij} \delta_{ijhk} \leq M_{hk}, \quad 1 \leq i \leq N, j \in W_i, h \in H, k \in K \quad (2)$$

and the precedence constraint:

$$c_{il} + t_{ij} \leq c_{ij}, \quad 1 \leq i \leq N, l \in P_j, j \in W_i \quad (3)$$

3.2 The Lagrangean relaxation method

The Lagrangean relaxation method adopted in the present study is summarized as follows:

(a) Formulation of the original scheduling problem P . The objective is to form the problem that can be decomposed to sub-problems at the job level. Formula (1), (2) and (3) represents P .

(b) Relaxation of a number of constraints using non-negative Lagrangean multipliers. In the present research, we adopt λ_{hk} ($h \in H, k \in K$) to relax the machine availability constraints whose initial values are decided randomly. The relaxed problem RP becomes:

$$RP = \text{Min} \left\{ \sum_{i=1}^N J_i + \sum_{hk} \lambda_{hk} \times [\sum_{ij} \delta_{ijhk} - M_{hk}] \right\}$$

$$= \text{Min} \left\{ \sum_{i=1}^N \omega_i \times \max(0, C_i - D_i) + \sum_{hk} \lambda_{hk} \times \sum_{ij} \delta_{ijhk} \right\} - \sum_{hk} \lambda_{hk} \times M_{hk} \quad (4)$$

$$= \sum_{i=1}^N \text{Min} \{ \omega_i \times \max(0, C_i, D_i) + \sum_{hk} \lambda_{hk} \times \sum_j \delta_{ijhk} \} - \sum_{hk} \lambda_{hk} \times M_{hk} \quad (5)$$

subject to (3). The transformation from (4) to (5) is possible based on the assumption that there is no precedence constraint between any two operations of different jobs, and therefore constraint (3) has functional reparability in terms of job i .

(c) Construction of a dual problem DP to the problem P , as follows. This is an optimization problem to obtain maximum value by regarding Lagrangean multipliers as parameters.

Maximum J

$$= - \sum_{hk} \lambda_{hk} \times M_{hk} + \text{Min} \left\{ \sum_{i=1}^N \omega_i \times \max(0, C_i, D_i) + \sum_{hk} \lambda_{hk} \times \sum_{ij} \delta_{ijhk} \right\} \quad (6)$$

(d) Decomposition of the RP problem (5) into independent sub-problems at the job level. The sub-problem SP_i is:

$$SP_i = \text{Min} \{ \omega_i \times \max(0, C_i, D_i) + \sum_{hk} \lambda_{hk} \times \sum_j \delta_{ijhk} \} \quad (7)$$

(e) Adoption of the pseudo-polynomial dynamic programming algorithm developed by Chen and *et al.*⁽¹⁾ to solve the job-level sub-problems. The algorithm requires to construct the cost calculation π_{ik}^{ij} for any operation (i, j) finished at time k ($k \in K$),

$$\pi_{ik}^{ij} = \begin{cases} +\infty, & \text{if } k < t_{ij} \\ \sum_{\theta=k-t_{ij}}^{k-1} \lambda_{h\theta}, & \text{if } j \neq W_i^* \text{ and } k \geq t_{ij} \\ J_i + \sum_{\theta=k-t_{ij}}^{k-1} \lambda_{h\theta}, & \text{if } j = W_i^* \text{ and } k \geq t_{ij} \end{cases} \quad (8)$$

where W_i^* is the last operation of i . The results of this step are optimized completion times for each job, although these results will not produce a feasible schedule for the workshop.

(f) Construction of a feasible schedule based on (e) by List scheduling. The corresponding value of the objective function J , which can be obtained by (1) is regarded as an upper bound J^* of the optimized solution.

(g) Adoption of the subgradient algorithm to obtain optimized multipliers λ_{hk} corresponding to the feasible schedule. These optimized values can be used to calculate the lower bound J_* of the original problem.

(h) Evaluation of the feasible schedule. The relative duality gap can be calculated by $(J^* - J_*)/J_*$, and if the gap is found to be smaller than the pre-defined value, the entire calculation is finished. If not, the method will continue the calculation from (e).

using the latest optimized Lagrangean multipliers' values.

4. Experiments and Results

Based on the above framework, the following experimental models and simulation approach are adopted to clarify the feasibility of the Lagrangean relaxation method under an APS environment and evaluate the results of the Lagrangean relaxation method with respect to practical application.

4.1 Experimental model

Table 1 shows the *basic* experimental model adopted in the present study. Three work centers achieve different operations in this workshop; each of which contains two machines whose function and capability are regarded to be identical. A new job, which is regarded as a new customer order, arrives randomly at the workshop and requires three operation steps at different work centers. The operation routine, operation time and due date for the job are generated randomly. Each simulation run is terminated when the 150th job leaves the workshop.

The above basic model is defined as the level 1 model. In order to enlarge the scale of the model, 10 levels are designed for experiments. The level n model is composed of $n \times 10$ initial jobs, where n new jobs arrive together at a random time, and $n \times 2$ machines are available at each work center. The simulation for the level n model is terminated when the $(n \times 150)$ th job leaves the workshop. Based on this design, the load factor of the workshop can be maintained constant for different level models.

In addition to the arrival of new jobs, changes in the due dates of existing jobs are adopted as random events during simulation experiments. In the present study, two characteristics of the random events are introduced, and their effects on the calculation efficiency are clarified.

The first characteristic, referred to as density D , represents the frequency of the occurrence of random events. D can be calculated as:

$$D = (f + 1)/t, \quad (9)$$

where t represents the cycle time of normal schedul-

ing and f represents the number of random events occurring during time t . Both the arrival of new jobs and changes in the due dates of existing jobs are included in f . Since the normal scheduling can be regarded as an event happened during time t , it is also included in the numerator. In the present study, t is set to be 100. Three levels are set for f : 4, 8 and 12. According to (9), the corresponding values of density D are 0.05, 0.09 and 0.13.

The second characteristic, referred to as *strength*, is used to represent the ratio of jobs for which the due date has changed. In the present study, strength is designed to three levels, in which jobs for which the due date has changed make up 10%, 20% or 30%, respectively, of all jobs. The new due date is calculated from the original due date:

$$\text{New due date} = \text{Original due date} + CO \quad (10)$$

where the coefficient CO is selected randomly from as -8 , -4 , $+4$, or $+8$.

This model implies the features of component makers in high-tech industries. The essential scheduling cycle for these component makers is a daily cycle. However, for key customers, component makers have to perform real-time ATP checks so as to provide precise commitments. Therefore, schedules must require adjustment several times a day.

4.2 Simulation approach

Figure 2 is the experimental workflow used to simulate the workshop's planning and execution processes using a real-time scheduling mechanism. The simulation is executed on an NEC PC equipped with a Pentium 233-MHz processor and 32-Mbyte Memory.

4.3 Experiment results

Simulation experiments are categorized into three types, which attempt to determine:

Type 1: The relation between calculation speed and model scale without considering random events;

Type 2: The effect of D on the calculation speed for simulation models having different scales;

Type 3: The effect of the *strength* of random events on the calculation speed of simulation models having different scales.

For each type, the number R of Lagrangean Multiplier's updates or iterative calculations for searching a solution and the relevant calculation time T are recorded. Both R and T are measured separately for initial solutions and continuous solutions after re-scheduling, and average values of R and T for re-scheduling are calculated accordingly.

Because the experimental results for Types 1 and 3 show the same trends, only results of 20 Type-2 simulation experiments are shown in Figs. 3 and 4. The following conclusions were obtained:

Table 1 Basic model (Level 1 model)

Number of initial jobs	10
Finished jobs as simulation terminal condition	150
Work Center (WC)	3
Number of machines per work center	2
Number of operations per job (Fixed number)	3
Operation routine, randomly selected from the following (the No. refers to WC type)	
(1-2-3) (1-3-2) (2-1-3)	
(2-3-1) (3-1-2) (3-2-1)	
Operation time (UT)	U[1, 35]
Due date (UT)	Arrival time + α * total operation time
Due date factor α	U[1, 3]

Note: UT: Unit time during simulation run

U[a, b]: Unified distribution between a and b.

Fig. 2 Simulation workflow

1. For the same level of experimental model, R and T increase with both initial and average values for increasing D , and R and T also increase with the expansion of the model scale.

Although the values of R for the *initial* and the *average* solutions are not explicitly different, the relevant values of T are very different, as shown in Figs. 3 and 4. This means that the update of Lagrangean Multipliers after the initial calculation requires much less calculation time for all model levels, because the Lagrangean Multipliers for the initial solution are used for the progressing scheduling. These results prove explicitly that, under the condition that the workshop has a very similar status, the use of initialized Lagrangean Multipliers can greatly improve the scheduling efficiency. This important feature also provides the Lagrangean Relaxation method with the ability to satisfy the real-time scheduling requirement of calculation speed, even for relatively large production models.

2. From Fig. 4, the calculation time T is found to increase linearly with the expansion of the model scale. In order to clarify this result, calculation load L is introduced as:

$$L = T * D \quad (11)$$

Fig. 3 Simulation results of experiment type 2 (1)

Fig. 4 Simulation results of experiment type 2 (2)

Fig. 5 Relation between calculation load L and model level for different density D

The unit for L is the calculation time per simulation unit time, which indicates the amount of re-scheduling calculation effort required under a specific changing environment.

Figure 5 shows the relation between calculation load and model scale, which is extremely important for realizing real-time ATP check mechanism. In order to more closely examine the relations between model level and calculation load, level line fit plots for each density as obtained by regression analysis are shown in Fig. 6. Values of Significance F and Multiple X are also listed as in the figures. The linear relation between the model level and the calculation load are confidently proved, because the threshold value of significance F is generally set to <0.05 , and the closer the multiple X approaches to 1, the closer of the correlation between the two parameters becomes.

By this linear relation, the calculation load for an even larger model can be estimated. Furthermore, based on the experience concerning the frequencies of new customer orders and changes to orders per

Fig. 6 Regression analysis of the relation between calculation load L and model level for different density D

scheduling cycle, the possibility of executing real-time ATP can be evaluated for a specific production environment. In addition, the appropriate hardware specification can be estimated as well.

As an example, if real-time ATP is applied to a real environment of a component maker, the daily scheduling cycle is equivalent to the time required for 100 simulation units. The calculation load for the PC used in the present experiment is less than 1.2 seconds per 4.8 minutes for a relative large workshop. The result is quite practicable if only the re-scheduling efficiency is considered.

3. In order to evaluate the quality of re-scheduling, the present study investigates the difference in the start/end times of operations and the total lead-time of jobs. Relative values for differences in start/end times of operations and lead-time of jobs are adopted as:

$$\begin{aligned} \text{Relative Operation Difference } RO \\ = (\text{Absolute Difference} / \text{Original Operation End Time}) * 100\% \end{aligned} \quad (12)$$

$$\begin{aligned} \text{Relative Lead-time Difference } RL \\ = (\text{Absolute Lead-time Difference} / \text{Original Lead-time}) * 100\% \end{aligned} \quad (13)$$

Table 2 shows the average values of RO and RL after

Table 2 Difference of operations' start/end time and lead-time after re-scheduling runs

	RO (%)			RL (%)
	Operation 1	Operation 2	Operation 3	
Job1	0.0	16.7	6.2	6.2
Job2	0.0	3.7	3.0	3.0
Job3	0.0	1.1	25.1	25.1
Job4	0.0	14.2	14.1	14.1
Job5	59.9	52.1	20.7	4.4
Job6	0.0	1.3	25.0	25.0
Job7	0.0	6.4	32.6	32.6
Job8	0.0	6.1	3.9	4.0
Job9	0.0	0.9	36.1	39.3
Job10	0.0	1.6	12.8	14.0
Job11	6.4	92.1	82.9	119.6
Job12	0.0	2.8	69.9	85.3
Average	5.5	16.6	27.7	31.0

two re-scheduling runs, which occurred due to new job arrivals and due date changes in one experiment. The average values in the last row are calculated based on the absolute values of RO or RL for all jobs. Most data are not equal to zero, which indicates that the start/end time of most operations are changed from the original schedule and that the lead-time becomes longer after re-scheduling.

5. Improvement

Generally, considering real production, the difference between the new schedule and the original schedule due to re-scheduling leads to extra preparation time, which can affect the feasibility of the re-scheduling results. Therefore, the re-scheduling results shown in Table 2 require further improvement.

In the present study, we attempted to modify the cost function used in the Lagrangean relaxation algorithm by embedding the operation time difference. As explained in Step (e) of the method, the cost calculation defined by formula (8) then becomes

$$\pi_{ijk}^{ij} = \begin{cases} +\infty, & \text{if } k < t_{ij} \\ \alpha_1 T_j + \alpha_2 \sum_{\theta=k-t_{ij}}^{k-1} \lambda_{h\theta}, & \text{if } j \neq W_i^* \text{ and } k \geq t_{ij} \\ \alpha_1 T_j + \alpha_2 T_i + \alpha_2 \sum_{\theta=k-t_{ij}}^{k-1} \lambda_{h\theta}, & \text{if } j = W_i^* \text{ and } k \geq t_{ij} \end{cases} \quad (14)$$

where T_i is the absolute difference between time k and the original operation completion time, and α_1 and α_2 are coefficients used to adjust the weights between this new item and the original Lagrange cost.

Three different patterns obtained by applying this improvement have been tested:

Pattern 1: Apply the new cost functions to *all* operations;

Pattern 2: Apply the new cost functions to the *immediate succeeding* operations only;

Pattern 3: Apply the new cost functions to the *last* operations only.

The results for 10 experiments for each pattern

Table 3 Comparison of scheduling results among different patterns and the original results

Approach	RO(%)			RL (%)	Immediate Succeeding Operation's Difference	Calculation Time
	Operation 1	Operation 2	Operation 3			
Original	5.3	28.7	32.9	38.2	23.4	1.2
Pattern 1	4.4	15.2	18.0	21.3	13.2	1.2
Pattern 2	8.1	27.3	33.1	36.8	16.0	1.5
Pattern 3	5.1	28.7	29.6	32.5	20.9	1.1

and the original scheduling are shown in Table 3, whereas α_1 and α_2 are set to be 1. In addition, the difference and calculation time for the immediate succeeding operations *RO* and *RL* are recorded, because the difference for the immediate succeeding operations has the greatest effect on the feasibility of rescheduling results with respect to real production preparation.

The experimental results reveal that all three patterns show improvements compared to the original results for the immediate succeeding operations. Among these three patterns, Pattern 1 produces the best results for all criteria shown in Table 3, while the calculation speed is kept to be the same. However, further researches are needed to clarify the generality of this result.

6. Summary

The feasibility of the proposed real-time scheduling mechanism has been investigated based on aspects of calculation load and re-scheduling quality. The experimental environment is extracted from a very real APS environment, in which real-time ATP and scheduling are required in order to maintain high customer satisfaction. Two characteristics of random occurrence, *density* and *strength*, are investigated with respect to their effects on the feasibility of the scheduling mechanism. Lagrangean relaxation, which is central to the proposed scheduling mechanism, is proven to have great potential for being adapted for use in such an environment.

The present paper focuses on the scheduling aspect of the APS environment, and future studies are expected to evaluate the method under the entire planning structure, which implies integration between planning and scheduling.

References

- (1) Chen, H., Chu, C. and Proth, J., An Improvement of the Lagrangean Relaxation Approach for Job Shop Scheduling: A Dynamic Programming Method, *IEEE Transactions on Robotics and Automation*, Vol. 14, No. 5 (1998), pp. 786-795.
- (2) Kuroda, M., Tookuni, Y. and Enomoto, M., Quasi-optimization of Realtime Scheduling with Lagrangean Relaxation Method, (in Japanese), *Proceedings of Production Scheduling Symposium*, (1998), pp. 65-70.
- (3) Kuroda, M., Lagrangean Decomposition and Coordination Method and Dynamic Scheduling, (in Japanese), *Operations Research*, Vol. 45, No. 6 (2000), pp. 263-269.
- (4) Hoiomt, D., Luh, P. and Pattipati, K., A Practical Approach to Job-Shop Scheduling Problems, *IEEE Transactions on Robotics and Automation*, Vol. 9, No. 1 (1993), pp. 1-13.
- (5) Luh, P., Hoiomt, D., Max, E. and Pattipati, K., Schedule Generation and Reconfiguration for Parallel Machines, *IEEE Transactions on Robotics and Automation*, Vol. 6, No. 6 (1990), pp. 687-696.
- (6) Muramatsu, K., Simultaneous Optimization for Lot Sizing and Lot Sequencing: Multi-item Single Machine Lot Size Scheduling with Setup Time, (in Japanese), *Journal of Japan Industrial Management Association*, Vol. 52, No. 5 (2001), pp. 313-321.
- (7) Shin, K., Total Optimization Model of Supply Chain Management, *Proceeding of APIEMS '99*, (1999), pp. 677-680.
- (8) Yoneda, K., Lagrangean Relaxation Method and Production Scheduling, (in Japanese), *Production Scheduling*, (2002), pp. 141-161, Asakura Shoten Pub. Co.