A Heuristic Method for Job-Shop Scheduling with an Infinite Wait Buffer
- From One-Machine to Multi-Machine Problems

Z. J. Zhao * J. Kim M. Luo H. C. Lau S. S. Ge J.B. Zhang

Abstract—Through empirical comparison of classical Job Shop Problems (JSP) with multi-machine consideration, we find that the objective to minimize the sum of weighted tardiness has a better wait property compared with the objective to minimize the makespan. Further, we test the proposed Iterative Minimization Micro-model (IMM) heuristic method with the Mixed Integer Programming (MIP) solution by CPLEX. For multi-machine problems, the IMM heuristic method is faster and achieves a better solution. Finally, for a large problem instance with 409 jobs and 30 types of machines, IMM-heuristic method is compared with ProModel and we find that the heuristic method is slightly better.

Index Terms—Job Shop Scheduling, Multiple Machine, infinite in-process wait buffer

I. INTRODUCTION

Multi-machine shop scheduling problems (job-shop, flow-shop and etc.) receive more and more attention ever since 10 years ago, as it can greatly improve the production rate and reliability. Its application can be found in logistics, semiconductor manufacturing, robotics, and etc.

In [3], there is a broad classification for classical one-machine problem. In [6], the same notation follows and further, an extensive review for multi-machine problems is given. Multi-machine problems are usually studied in one of the following cases:

Case1: There are multiple exchangeable and renewable machines $C_k$ for each machine type $k$, and each specific machine could perform no more than one operation at any time;

Case2: There is exactly one machine for each type $k$, which can perform as many as $C_k$ operations at any time, and $C_k \geq 1$;

Case3: There are cases with combination of Case 1 and Case 2.

Z. J. Zhao is with School of Information System, Singapore Management University, He is currently a PhD candidate at the Electrical and Computer Engineering Department of the National University of Singapore, 117576 Singapore. Tel.: +65-68518564. Fax: +65-67791103. Email: zzytgx@yahoo.com.cn

* Corresponding author

J. Kim is with the Institute of Transportation & Logistics, UTAC, Korea.

M. Luo and J.B. Zhang are with the Singapore Institute of Manufacturing Technology.

H.C. Lau is with School of Information Systems, Singapore Management University.

S.S. Ge is with the Electrical and Computer Engineering Department of the National University of Singapore.

This research is partially funded by the A∗STAR SERC TSRP GRANT P0520104 and IMSS TSRP 052 116 0075.

From Formulation’s point of view, the Differences between one-machine problem and multi-machine problem are noted as following:

FD-1: One-machine problem can be formulated as disjunctive graph [2], which could be easily mapped to MIP model and the processing time can be non-integers.

FD-2: For multi-machine problem, it is so far not yet nicely formulated in our opinion especially when processing time is non-integer. In [8], a model is proposed with $\delta()$ function, which can only be solved by heuristic methods and the optimality is not guaranteed for all cases. For integer processing time, in [4], the Pritsker’s formulation [7] was applied and a 0-1 Integer Problem (IP) is solved.

With respect to the Solutions, there are following Differences, which will be further explained later.

SD-1: One-machine problem has a solution as the task schedule, while the multi-machine problem’s solution includes task schedule and machine dispatching.

SD-2: The solution of task schedule for both one-machine problem and multi-machine problem is unique, while solution for machine dispatching in multi-machine problem is not unique.

In this paper, our work is focused on the multi-machine JSP under Case 1. In order to make benchmark with optimal solutions, we use Pritsker’s 0-1 formulation and all processing time are integers. Although we assume infinite wait buffer in the problem formulation, we can see that the wait-in-process is reduced when machine capacity is increased. An iterative minimization micro-model is implemented to solve the overall problem iteratively. The solution performance is bench-marked with the CPLEX solution for small-size problems and compared with ProModel solution for large-size problems. The notation of this paper is summarized in Table I.

II. SAMPLE PROBLEM: MT6, $C_k = 2, \forall k$

For a clear demonstration, we use the MT6 problem [5], where totally there are 6 jobs and each job has 6 tasks on 6 different types of machines. The original MT6 problem in [5] is one-machine problem $C_k = 1, k \in \{1,2,...,6\}$, while our sample problem differs in $C_k = 2, k \in \{1,2,...,6\}$. The solution Gantt chart is shown in Fig. 1 where Fig. 1(a) is represented by job-grouping and Fig. 1(b) is by machine-grouping. Both Fig. 1(a) and Fig. 1(b) are actually from the same solution. Y-axis in Fig. 1(a) is job-id, while in Fig. 1(b), integer represents machine type and fraction represents machine-id of
TABLE I
NOTATION FOR JOBS, PROCESSES, MACHINES

<table>
<thead>
<tr>
<th>Notation for Jobs and Processes</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>Total number of jobs</td>
</tr>
<tr>
<td>$\alpha_i$</td>
<td>Total number of operations (i.e., processes) for job $i$</td>
</tr>
<tr>
<td>$p_{ij}$</td>
<td>Processing time of job $i$ process $j$</td>
</tr>
<tr>
<td>$d_i$</td>
<td>Due time of job $i$</td>
</tr>
<tr>
<td>$W_i$</td>
<td>Delay penalty per unit time if job $i$ is completed after its due time</td>
</tr>
</tbody>
</table>

Notation for Machines

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>Total number of machine types</td>
</tr>
<tr>
<td>$k$</td>
<td>Machine type index, $k \in {1, \ldots, K}$</td>
</tr>
<tr>
<td>$T$</td>
<td>Total time slot in discrete time formulation</td>
</tr>
<tr>
<td>$C_k$</td>
<td>Capacity of machine $k$, i.e., number of operations that can be executed at anytime on machine type $k$</td>
</tr>
<tr>
<td>$M_{ij}$</td>
<td>Mapping from job $i$ and process $j$ to machine type</td>
</tr>
</tbody>
</table>

Variables in Process Scheduling

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{ijt}$</td>
<td>Discrete decision variable for $i^{th}$ job, $j^{th}$ process at time slot $t$</td>
</tr>
<tr>
<td>$T_{ij}$</td>
<td>Start time of process $j$ in job $i$</td>
</tr>
<tr>
<td>$T_{ij}$</td>
<td>Completion time of process $j$ in job $i$</td>
</tr>
<tr>
<td>$M_k(t)$</td>
<td>Machine utilization function in continuous time</td>
</tr>
<tr>
<td></td>
<td>Each $k$ indicates a specific machine type</td>
</tr>
</tbody>
</table>

Notation in the IMM heuristic method

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>Iterative number</td>
</tr>
<tr>
<td>$L_k(R)$</td>
<td>Total number of executable tasks on machine type $k$ at iteration $R$; executable in the sense of precedence constraints</td>
</tr>
<tr>
<td>$t_l^j$</td>
<td>Arrival time for task $l$, $1 \leq l \leq L_k(R)$</td>
</tr>
<tr>
<td>$t_{ml}$</td>
<td>Release time for machine-id $m$ of type $k$</td>
</tr>
<tr>
<td>$1 \leq m \leq C_k$</td>
<td></td>
</tr>
<tr>
<td>$mi(l)$</td>
<td>A 1-to-1 mapping from task-id $l \in {1, 2, \ldots, L_k(R)}$ to machine-id $mi \in {1, 2, \ldots, C_k}$</td>
</tr>
</tbody>
</table>

$RTAM$ | Mean Release Time of all machines after scheduling |

$CTAJ$ | Mean Completion Time of all jobs |

$TW$ | Total Wait count |

$WT$ | Mean Wait Time |

$WT^M$ | Max Wait Time |

$AMU$ | Mean percentage of Machine Utilization of all |

that type. For the multiple solutions in machine dispatching, which is mentioned previously in SD-2, it can be explained that in Fig. 1(b), exchanging the machine-id (fractional part of machine scheduling) will not affect the task scheduling.

We further group the Gantt chart in colors, for the detailed readers to verify that Fig. 1(a) and Fig. 1(b) are actually the same solution. In Fig. 1(a), the same color represents the same machine type, while in Fig. 1(b), the same color represents the same job-id.

III. A 0-1 FORMULATION

A 0-1 formulation is proposed by [7], which might be the only solvable model for rigorous multi-machine problems.

A. Decision variable

Before formulating the problem in discrete time domain, one has to estimate a proper time unit $U_T \in \mathbb{R}$, with which the continuous time domain is discretized to $\bar{T} \in \mathbb{N}$ time slots in total. This estimation is introduced [4] [8], and the total time period $U_T \bar{T}$ could be available by heuristic solution. In discrete time domain, $t \in \{1, 2, \ldots, \bar{T}\}$ is used to represent each time slot. A binary decision variable is $X_{ijt}$, where

$$X_{ijt} = \begin{cases} 1 & \text{if process } j \text{ in job } i \text{ starts by time } t \text{ inclusively;} \\ 0 & \text{if process } j \text{ in job } i \text{ has not yet started time at } t \end{cases}$$

There are following constraints by assumption:

$$X_{ijt} - X_{ij,t+1} \leq 0, \quad \forall i,j,t \in \{1, 2, \ldots, \bar{T} - 1\}. \quad (1)$$

$$X_{ijt} \in \{0, 1\}, \quad \forall i,j,t \in \{1, 2, \ldots, \bar{T}\}. \quad (2)$$

Inequality (1) assumes that every task is non-preemptive.
B. Constraints on task precedence and machine capacity

Task precedence constraint is specified as inequality (3), which means that in-process wait is allowed. We name it to be an Infinite Wait Buffer because the model does not specify on:

- wait buffer capacity, which is usually related with machine settings,
- upper and lower bounds on wait time, which is usually from specifications of task schedule.

Machine capacity constraint is stated as the inequality (4), which is well studied in [8].

C. Objective functions

There are usually two types of criterions CRI-1 and CRI-2:

CRI-1: The objective is to minimize the makespan, which by nature, is to minimize the time length from the starting time of very first task to completion time of the very last task.

CRI-2: The objective is to minimize the sum of weighted tardiness, which actually, is to minimize some areas. Their formulations are following:

CRI-1: For a general job-shop problem, two types of dummy variables \( \{X_{S,t}, X_{E,t} : t = 1, 2, ..., T \} \) are introduced to formulate the Start time of very first task, and the End time of the very last task. See (5) and the inequality constraints (6) and (7).

\[
\text{minimize} \sum_t (1 - X_{E,t}) - \sum_t (1 - X_{S,t}) \quad (5)
\]

\[
X_{i,j,t} - X_{S,t} \leq 0, \forall i, j, t \quad (6)
\]

\[
X_{E,t} - X_{i,j,t} \leq 0, \forall i, j, t \quad (7)
\]

CRI-2: The most generalized formulation introduces two lists of parameters, \( W_i : i = 1, 2, ..., N \) for each job’s weight (or importance) and \( d_i : i = 1, 2, ..., N \) for each job’s due time. For a special kind of flow shop with some sequencing constraints like [8], by adjusting the parameter \( W_i \), (8) represents either minimizing the makespan or minimizing the sum of weighted tardiness.

\[
\text{minimize} \sum_i \left( W_i \sum_{t \geq d_i - p_{i,0}} (1 - X_{i,o_i,t}) \right) \quad (8)
\]

D. An empirical comparison of above two criterions

Classical jobshop problems are usually to minimize the makespan [1] [2]. However, we observe that the overall final machine release time, which is more important in our opinion especially for multi-machine problems and it is related with the sum of weighted completion time. Some experiments are given to compare the two criterions, the sample problems are chosen to be \( \{MT6, MT10\} \) at machine capacity \( C_k \in \{1, 2, 3, ...\} \).

Some terms are defined at first. There is wait between \( i^{th} \) job’s \( j^{th} \) process and its precedence process iff. there is \( T_{i,j-1}^s \neq T_{i,j}^s, i = 1, ..., N, j = 2, ..., o_i \).

- Total Wait: \( TW \) - the total element count in \( \{ (i,j)|T_{i,j-1}^s \neq T_{i,j}^s, i = 1, ..., N, j = 2, ..., o_i \} \);
- Max Wait Time: \( WT^M \) = max \{ \( T_{i,j}^s - T_{i,j-1}^s \mid i = 1, ..., N, j = 2, ..., o_i \) \};
- Mean Wait Time: \( \text{WT} \) = \( \frac{T_{\text{TotalWaitTime}}}{\text{TotalWait}} \).

Relation of continuous variable \( T_{i,j}^s \) and discrete variable \( X_{i,j,t} \) are studied in [8] and listed as follows:

\[
X_{i,j,t} = 1 \leq 0 \quad \land \quad X_{i,j,t^*} = 1 \\
\begin{align*}
T_{i,j}^s &= t^* - 1, i^{th} \text{ job's } j^{th} \text{ process starts at the time slot } t^*.
\end{align*}
\]

For the terms related with the machine utilizations, our definition is based on previous work [8] w.r.t. the construction of machine utilization function \( M_k(t) \). Their definitions are shown as following.

- Mean Machine Utilization for type \( k \): \( \mu_k(t) = \frac{\int M_k(t) dt}{t} \)
- Mean Utilization All Machines: \( \mu_{\text{All}} = \frac{\sum_k \int M_k(t) dt}{T_{\text{Max}}} \)

where the integration is from the very first time when the machine type \( k \) is put to use to the very last time when the machine type \( k \) stops using.

- Mean Complete Time is defined as \( \mu_{\text{CTA}} = \frac{T_{\text{Total}}}{N} \)

For the discrete formulation under assumption that all jobs’ first process is arrived at time 0, it is equivalent to \( \sum_{i,j} (1 - X_{i,o_i,t}) \).

![Fig. 2. Criterion Comparison: One-machine MT6 problem, grouping by machine](image-url)
Fig. 2 demonstrates a one-machine MT6 optimal solution by CPLEX. Fig. 2(a) is to minimize the sum of completion time (or criterion-2 with \( d_i = 0, W_i = 1, \forall i \)), while Fig. 2(b) is to minimize the makespan as criterion-1. The authors suggest to focus on only comparison of the length (makespan) and the area of last task with x-axis (sum of weighted tardiness). Here, we use grey figure because it is a 1-machine MT6 problem, in contrast with Fig. 1 which is 2-machine MT6 problem. It should be noticed that Fig. 1 is for two different problems (criterion-2 in Fig. 2(a) and criterion-1 in Fig. 2(b)).

Precisely, Table IV is used to compare \{1, 2\}-machine by CRI-\{1, 2\} on the MT6 and the MT10 problems. All schedule solutions are done by CPLEX.

From Table IV, we have following observations:

- When the total number of machines is increasing, both the makespan and the sum of weighted tardiness are dropping. When it reaches some level, both will keep the same value. This is similar to the Infinite Resource Model in [8].
- The sum of weighted tardiness (criterion-2) model is consistently better in the wait property (i.e. less in total wait count and the wait time) and the mean complete time than makespan model (criterion-2)
- For the mean machine release time, the 1-machine problem and the multi-machine problem are contrary to each other, while for the makespan and the machine utilization, there is no consistency.

E. Complete model for comparison with heuristic method

Based on above observation, we will further study our heuristic method for multi-machine job-shop with CRI-2 as objective to minimize. For simplicity without loss of generality, \( d_i = 0, W_i = 1, \forall i \). The complete model to benchmark our IMM heuristic method is shown as follows.

\[
\text{JSP} \quad \text{MultiMach} \quad \text{InfWait} \quad \text{SumCompleteTime}(8)
\]
\[
\text{subject to:} \quad \text{Inequality}(1) \quad \text{Inequality}(3) \quad \text{Inequality}(4) \quad 0-1 \ \text{Variable as}(2)
\]

This is the formulation for CPLEX solution of a 2-machine MT6 problem and shown in Fig. 1.

IV. HEURISTIC SOLUTION: ITERATIVE MINIMIZATION OF A MICRO-MODEL

The micro-model is defined for one particular machine type \( k \) with capacity \( C_k \), on which, totally \( L_k(R) \) processes are to be done. This model is solved and updated iteratively for the overall multi-machine JSP. \( R \) is a counter for iterations, and \( L_k(R) \) is the total number of executable tasks on machine type \( k \). "Executable" means satisfying sequencing constraints.

A. Minimization of a micro-model

In the Micro-model, there are \( C_k \) exchangeable and renewable machines, serving \( L_k(R) \) tasks. Each machine has a release time \( t_m^r : m \in \{1, 2, ..., C_k\} \) while each task has an arrival time \( t_l^a : l \in \{1, 2, ..., L_k(R)\} \) and processing time \( p_l : l \in \{1, 2, ..., L_k(R)\} \).

- \( t_m^r \) is a time for a particular machine \( m \) to be ready to start a new task.
- \( t_l^a \) is a time for task \( l \) to be ready to start. It is the completion time of task \( l \)'s precedence task which is specified in sequencing constraint.
- \( p_l, 1 \leq l \leq L_k(R) \), processing time for task \( l \).

The problem is to minimize the sum of completion (or end) time. The summation is done for \text{min}(C_k, L_k(R)) \) tasks.

Minimize: \text{min}(l) \sum_{l \in \{1, ..., C_k\}} \sum_{l \in \{1, 2, ..., L_k(R)\}} t_l^f \quad (9)

subject to: \text{Inequality}(1) \quad \text{Inequality}(3) \quad \text{Inequality}(4) \quad 0-1 \ \text{Variable as}(2)

This micro-model is to minimize the sum of end time in schedulable and executable task schedule as in (9) under the constraints of (10) and (11) with variable mapping (12), and the time variable lists (13).

From (9), total number of items in the summation is \text{min}(C_k, L_k(R)) \) where \( C_k \) is for schedulable constraints by the machine availability and \( L_k(R) \) is for executable constraints of precedence. In (9), \text{mi}(l) \) is a l-to-1 mapping from the task-id \( l \in \{1, 2, ..., L_k(R)\} \) to the machine-id \( m = \text{mi}(l) \in \{1, 2, ..., C_k\} \).

There is a post-process after solving above minimization micro-model, which is to update the machine release time for next group of tasks, as shown in (14).

\[
t_m^r = t_l^f, \forall l \ \text{executable & schedulable} \
\]

B. Three scenarios and computational complexity

According to the relative value of \( C_k \) and \( L_k(R) \), there are totally 3 scenarios for consideration:

- **IMM1:** \( L_k(R) \geq C_k = 1 \), this is actually a one-machine problem. This can be solved optimally by enumeration, which is shown in Proposition-1.
- **IMM2:** \( C_k \geq L_k(R) \geq 1 \), the total number of machines is not less than the total number of tasks. The optimal solution for this case is following in Proposition-2.
- **IMM3:** \( L_k(R) > C_k > 1 \), the total number of machines is less than the total number of tasks.

**Proposition-1:** For IMM1, the single machine is \( m_1 \) and the task list is \( l \in \{1, 2, ..., L_k(R)\} \). The optimal objective value is following:

\[
\min \left( \min_{l : t_l^a \leq t_m^r} \{p_l + t_m^r\} \min_{l : t_l^a > t_m^r} \{p_l + t_l^a\} \right) \
\]

**Proof:** Done by enumeration.

\]
**Proposition-2:** For IMM2, the optimal solution is to assign earliest available $L_k(R)$ machines for tasks by first come first serve rule.

**Proof:** We begin with $2$ tasks $l_1, l_2$ on $2$ parallel machines $m_1, m_2$. The processing time and arrival time of the $2$ tasks are $\{p_{l_1}, t_{l_1}^a\}$ and $\{p_{l_2}, t_{l_2}^a\}$, while the machine release time of the $2$ machines are $t_{m_1}^r, t_{m_2}^r$. Without loss of generality, we assume that $t_{l_1}^a \leq t_{l_2}^a$ and $t_{m_1}^r \leq t_{m_2}^r$. The optimal objective value is $p_{l_1} + \max\{t_{l_1}^a, t_{m_1}^r\} + p_{l_2} + \max\{t_{l_2}^a, t_{m_2}^r\}$ because it can be verified by all the possible cases as following:

\[
\begin{align*}
& t_{l_1}^a \leq t_{l_2}^a \leq t_{m_1}^r \leq t_{m_2}^r \\
& t_{l_2}^a \leq t_{l_1}^a \leq t_{m_1}^r \leq t_{m_2}^r \\
& \vdots \\
& t_{m_1}^r \leq t_{l_1}^a \leq t_{m_2}^r \leq t_{l_2}^a \\
& t_{m_1}^r \leq t_{l_2}^a \leq t_{m_2}^r \leq t_{l_1}^a 
\end{align*}
\]

The first come first serve rule will result as following mapping $mi(l_1) = m_1, mi(l_2) = m_2$, so it is true for 2-task-2-machine case.

For general cases, $C_k \geq 2, L_k(R) \geq 2$, for any mapping solution $mi(l)$, if $\exists\{l_1, l_2, m_1, m_2\}$, s.t. $mi(l_1) = m_1, mi(l_2) = m_2$, and $(t_{l_1}^a - t_{l_2}^a)(t_{m_1}^r - t_{m_2}^r) < 0$. By swapping the mapping to first come first serve order will always achieve a better solution.

For IMM3, one solution strategy is to schedule one machine id after another and repeatedly apply Proposition-1 to be solved. This is not an optimal solution, but a fast feasible solution. The optimality depends on machine id sequencing.

Computational complexity for Proposition-1 is $O(L_k(R))$, and for Proposition-2 is $O(L_k(R) \cdot \ln(L_k(R)))$, while the fast feasible solution of IMM3 is $O(L_k(R) \cdot C_k)$

**V. REPORT FOR THE EXPERIMENTS**

**A. Benchmark on classical problems {MT6, MT10} – a multi-machine version**

There are 6 pairs of experiments to test our proposed IMM-heuristic method with CPLEX solution. The job-process-machine setting is exactly same as the MT6 and MT10 problems. First 3 pairs are the MT6 with $\{1, 2, 3\}$-machine(s) for each type. Last 3 pairs are the MT10 with $\{1, 2, 3\}$-machine(s) for each type. The results are shown in Table V. There are following points for notice:

- CPLEX solver engine is stopped after one hour and return the best feasible solution;
- both CPLEX and IMM-heuristics solver runs on Intel(R) Xeon(R) CPU of X3220 2.4 GHz with 2.4GHz & 4.0 GB RAM;
- IMM-heuristic method runs about 3 seconds for MT10 problems or 1 second for MT6 problems.

From the table, we can see that the optimality of IMM-heuristic method becomes better when total number of machines are increased. Except for MT10 one-machine problem, IMM-heuristic method achieves better solution than CPLEX in other problems.

We observe that for one-machine MT10 problem, CPLEX [1] failed to have a feasible solution within one hour. It is because the 0-1 formulation results as a very large size problem [8].

**VI. CONCLUSION AND FURTHER RESEARCH SUGGESTIONS**

We empirically evaluated the impact of multi-machine consideration on solution of several classical JSPs with an infinite wait buffer. Statistics show that the objective of sum weighted tardiness is more suitable for multi-machine JSP. Furthermore, an IMM heuristic method has been proposed for a fast feasible solution and the performance is bench-marked with CPLEX solutions. Especially, our IMM heuristic method has following features:

- It supports fractional processing time;
- The micro-model considers the machine release time and the task arrival time, which can be extended as an input from the problem of multi-machine JSP;
- It outputs the machine dispatching solution and the machine release time, so the micro-model can be easily applied to construct a huge multi-machine problems;
The overall computational complexity of IMM-heuristic method is not more than $O((N \cdot K \cdot \max_i=1,2,\ldots,N \{o_i\})^2)$

- It achieves local optimality for micro-model in 2 scenarios, which are IMM1 and IMM2.
- The optimality of the overall solution is affected by the sequencing in selection of machine type $k$ at iteration $R$ for solving the Minimization Micro-model and the sequencing of solving IMM3 during a specific iteration. This, to our opinion, deserves further research.

In the future, multi-period consideration should come into picture for better machine utilization and dynamics handler. What’s more, a fast solution towards optimality of IMM3 also deserves further study.

Intuitively, the classical model of traveling-salesman is well and deep studied. Afterward, the problems of traveling group with multi agents (or robots), and the case when some parts of the group are ill (or not in best performance) will be further studied.

### VII. Acknowledgement

Much appreciation goes to Ireneus Wior for his excellent simulation work which makes ProModel a really efficient tool.

### REFERENCES


