# Automated Planning Tool (APT): A mixed integer non-linear programming problem solver for Workorder scheduling

**Anupam Purwar (RME Science, Amazon)** [*1]

## Abstract

Workorder scheduling is a type of Resource allocation problem which is NP-complete. Workorder scheduling related to maintenance work becomes more challenging because of the inherent complexity of e-commerce facilities owing to various local requirements and variety of equipments viz.electrical, mechanical and electro-mechanical etc. Hence, we formulated the problem by using multiple decision variables to accommodate all these requirements. Next, we decomposed the problem into two smaller sub-problems to ensure a fast yet accurate solution with load balancing of workorder hours as the main objective function. Through this objective function, we are addressing the business requirement of reducing overtime and ensuring uniformity in workorders to be executed across the planning horizon. We have a mix of continuous and binary decision variables with multiple linear constraints, hence, this problem has been formulated as a Mixed Integer Nonlinear programming problem (MINLP) with nonlinear objection function. Being a NP-hard hard problem the runtime increases as the problem size increases (problem size is proportional to number of planning days and number of workorders). However, through problem decomposition, code vectorization and use of AWS EC2 instances, we have achieved a run time in the range of 5-20 minutes. Through this article, we have covered the vital details of optimization model, cloud compute infrastructure used to solve this NP-hard problem and salient points of MINLP model formulation.

## 1. Introduction

Workorder scheduling is a popular problem which has been solved using optimization models for production planning [1], facility maintenance [2] and online scheduling in smart factories [3]. In this paper, we have discussed our findings from development of an optimization model for scheduling maintenance activities in e-commerce facilities. Present day e-commerce facility process large volume of customer orders everyday, only possible because of high reliability of equipment supporting these facilities. The health of these equipments is taken care by site maintenance teams by executing and planning/scheduling workorders (WOs) for thousands of equipments of different types every day. Planning and scheduling carried out by the sites is key to maximize equipment uptime through load-balancing and prioritization of critical maintenance activities at these facilities/sites. Generally, Workorder scheduling is done manually using excel spreadsheets requiring manually updating information on ERP system to produce the daily plan. This process is time consuming, sub-optimal and done for short time windows( 5-7 days). This takes away any scope of centralization as well as associated opportunities related to centralized planning, cluster level work aggregation and realize associated efficiencies. Centralized planning using a data driven solution can help create multiple plans for different scenarios by minimizing variance in workorder allocation across planning period. Also, it makes workorder scheduling more accurate, less time consuming, easily explainable to the site maintenance teams and bring uniformity in scheduling across all the sites. Hence, we propose a data driven solution namely Automated Planning Tool (APT) to serve as a recommendation engine for the e-commerce facilities to efficiently and effectively execute day to day scheduling as well as long-term planning.

## 2. Problem Statement

The problem at hand is to make sure all the open workorders are scheduled to the available technicians in a uniform manner. Every Workorder (WO) needs to be mapped to a technician/technicians depending who carry the right qualification keeping in mind their availability. The availability of a technician is governed by their availability information and technician-shift mapping information. The purpose of the model is to schedule every WO to a given shift/day and a technician pool, while achieving the following objectives :

- The sum total of hours assigned to technicians at facility per day across planning horizon shall be as uniform as possible
- The percentage of workorders getting scheduled by the

model shall be as high as possible

- Model shall abide by site level planning requirements viz. geographic co-location of equipment and technicians, equipment nature (high criticality equipments are serviced in Downtime only), shift timings of facility as well as variable shift-technician mapping

To meet the first objective, we need to arrive at a distribution of WOs in a way that the number of hours allocated across the "n" day planning horizon (PH) remains as uniform as possible. (n is a hyper parameter entered by planner). Ideally, total hours allocated per day (y-axis) when plotted with respect to PH (x-axis), the plot should represent WOs being spread out as evenly as possible across the planning horizon. To achieve this, we have set minimization of variance of Workorder hour distribution as the objective function.

## 3. Model Development

In the present article, we have solved a resource allocation problem inspired from the use case of workorder scheduling done in e-commerce facilities. We have developed a tool around this model called Automated Planning Tool (or APT). Briefly, the problem is to map the technicians to a set of tasks based on certain requirements modelled as constraints. The complexity of APT is governed by the nature of its constraints viz. the time window within which a task is to be completed, eligibility/availability of technicians to work on a given task and the physical construct of the facility for which the schedule is being created. Last but not the least, the schedule/plan is to be created at a click of few buttons within few minutes (5-20 mins per run). Keeping all these challenges in mind, we divided our research objectives into multiple categories viz. model creation, model testing, runtime optimization and deployment.

### 3.1. WO scheduling: Resource allocation problem

WO scheduling can be modelled as a Resource allocation problem(RAP). Generally, the complexity of a RAP determines the complexity of a scheduling problem (SP). A difficult RAP would result in an equally hard scheduling problem. In mathematical terms, optimally solving a scheduling problem to determine an optimized network topology is an NP-complete problem. NP-complete problem is a problem which is both NP (verifiable in non-deterministic polynomial time) and NP-hard (any NP-problem can be translated into this problem) [4]. NP-complete problems is a class of computational problems for which no efficient solution algorithm has been found. Some examples are Travelling Salesman problem, satisfiability problems and graph covering problems [5]. These problems become more challenging when RAP/SP has time bound tasks. One has to explore entire search space to find the optimal solution of the SP in a reasonable amount of time, therefore, several heuristics are deployed to find a near-optimal solution. As a precursor to understand a RAP, first problem formulating and modelling needs to be done. Modelling is the most critical step towards solving RAP and is at the core of machine learning/Artificial Intelligence. Robust model creation is the first step of devising an efficient problem-solving strategy. However, modelling NP-complete real world problems is not straightforward because there are a host of factors at play consisting of both known unknowns and unknown unknowns, where the former is still easier to anticipate but the latter are intractable. Besides, no mathematical guidelines or algorithms exist which can guide about the factors which need to be accounted for during the modelling process. Through this work, we have carefully balanced the scientific rigour of innovation with ever evolving requirements and expectations of the customers to develop a model which is modular and modifiable.

The model requires input data from demand (e.g. predicted workorders) and supply sides (e.g., technician-shift, technician-trade, technician-zone mapping and shift-zone mapping for technicians). Besides, there are some other inputs which user needs to enter viz. 1) Planning horizon: no of days for which one wants to pick and schedule workorders for, 2) percentage utilization of technician per shift, 3) No of days beyond planning horizon which user wants to expand the scheduling window to. The key modelling output is the name of technician/technicians as well as shift/date and scheduling block to which a WO is mapped.

### 3.2. Workorder scheduling: A non-linear MIP

Work order assignment has been formulated as a Mixed integer non-linear optimization problem with linear constraints and non-linear objective function. The objective function is to minimize the variance of WO hours allocated across planning horizon. Mixed Integer-Programming or MIP problems are formulated when the decision variables are to be restricted to either integer or binary values. For example, in present case one of the decision variable represents allocation of a technician to a workorder which has to be a binary value and another decision variable represents allocation of a workorder to a shift which again has to be a binary value. On the other hand, we have decision variables which assume continuous values like how many hours during a given day will a technician work on the assigned workorder. Thus, a mix of binary and continuous decision variables make this problem also a Mixed Integer Programming (MIP) problem, and in generally MIPs are harder to solve than LP problems [6, 7, 8]. In the current problem, objective function selected is minimization of deviation of hours allocated per day from the mean of this distribution of allocated hours. Deviation being a non-linear quantity makes the objective function also non-linear. Hence, the

WO scheduling problem is Mixed Integer Non-linear programming problem or MINLP. This MINLP has been solved using Branch and Bound technique [9].

### 3.3. Algorithm

Routine 1 consumes the Demand side (all open WOs of PM/PDM/PR type) and supply side: technician-trade, technician-shift, technician availability information to match the WO demand hours to supply of technician hours at a trade and shift level. In order to accommodate, the zone level assignment of WOs, another constraint is written which makes sure only those shifts are selected which map to the zone mentioned against the WO.

#### 3.3.1. ROUTINE 1: MAP WORKORDER TO A SHIFT

We have introduced slack/surplus variables to convert inequality constraints into equality constraints. If a constraint is binding, then the corresponding slack variable will assume zero value else a non-zero value. Thus, routine 1 does the mapping of WOs to zone-shift (date and site's zone-shift combination on that day) with deviation minimization of "WO allocated hours distribution" as the objective function.

**Index**

I: Set of all work orders
O: Set of work orders to be completed in ORT
D: Set of work orders to be completed in DTW
J: Set of days in the planning horizon
T: Set of all trades (work types)
A: Set of all technicians
S: Set of all shifts in the planning horizon
PH: Set of all days in the planning horizon
J(s): The day which shift s belongs to

**Input data (after pre-processing)**

$n_{it}$: Number of technicians required for WO i of trade t
$h_i$: Estimated working hours required for work order i
$z_i$: Zone required for work order i
$L_s$: Length of shift s in hours
$L_{so}$: Number of ORT hours available in shift s
$L_{sd}$: Number of DTW hours available in shift s
$p_i$: A penalty cost for NOT completing work order i during the planning horizon
$r_i$: It will be zero if no eligible shift is found for the WO i and 1 if there is atleast one shift eligible to complete WO i

**Decision variables**

Binary variable $x_{is}$ = 1 if work order **i** is assigned to shift **s**
Continuous variable $yo_{ast} \in$ [0,1] indicates fraction of time technician **a** is assigned to trade **t** on shift **s**
Continuous variable $yd_{ast} \in$ [0,1] indicates fraction of time technician **a** is assigned to trade **t** on shift **s**
Binary variable $z_i$=1 if work order is NOT allocated
Subscript: "o" for ORT orders and "d" for DTW orders

**Constraints: Routine 1**

Resource allocation constraint: maps WO to a shift based on trade, number of technicians and number of hours required by WO. This constraint makes sure the workorders are not allocated in a way which exceeds the maximum allowed hours in the shift s across all technicians

For ORT Workorders,

$$\sum_{i=1}^{o} min(L_{so}, h_i) \cdot n_{it} \cdot x_{is} = \sum_{a=1}^{n} L_s \cdot yo_{ast} \forall s \in S, t \in T$$

For DTW Workorders,

$$\sum_{i=1}^{o} min(L_{so}, h_i) \cdot n_{it} \cdot x_{is} = \sum_{a=1}^{n} L_s \cdot yd_{ast} \forall s \in S, t \in T$$

Tech schedule/eligibility constraint:

$yd_{ast} = 0 \, \forall s \in S$, if technician a is not eligible to do trade t
$yd_{ast} = 0 \, \forall t \in T$, if technician a does not work on shift s
$yo_{ast} = 0 \, \forall s \in S$, if technician a is not eligible to do trade t
$yo_{ast} = 0 \, \forall t \in T$, if technician a does not work on shift s

This constraint makes sure the technician 'a' does not work beyond the maximum allowed hours in the shift 's'

$$\sum_{t \in T} yd_{ast} \leq \text{DTW hrs ratio in shift s}, \forall a \in A, s \in S$$

$$\sum_{t \in T} yo_{ast} \leq \text{ORT hrs ratio in shift s}, \forall a \in A, s \in S$$

$x_{is}$=0 if shift **s** does not have the minimal number of techs with the required trades for WO **i**

Zone-shift combination constraint:

$x_{is} = 0$ if shift s is not mapped to the zone z of WO i

Shift available hours constraint: Workorder needs to be assigned with sufficient working hours for completion

$$\sum_{s \in S} min(L_{so}, h_i) \cdot x_{is} + short_i - over_i = h_i \forall i \in O$$

$$\sum_{s \in S} min(L_{sd}, h_i) \cdot x_{is} + short_i - over_i = h_i \forall i \in D$$

$$z_i \geq \frac{1}{M} * shortage_i, \forall i \in IWorkorders$$

$x_{is}$=0, if WO **i** is a DTW WO but shift **s** has no DTW
$x_{is}$=0, if WO **i** is an ORT WO but shift **s** has no ORT

Compliance window constraint:

$x_{is} = 0$ if shift s is NOT within compliance window of WO i

Define daily workload deviation:

$$dev_j \geq \frac{1}{J} \sum_{s \in S}^{i \in I} min(L_s, h_i) \cdot x_{is} - \sum_{s \in S}^{i \in I} min(L_{so}, h_i), \forall j \in J$$

$$dev_j \geq \sum_{s \in S}^{i \in I} min(L_s, h_i) - \frac{1}{J} \sum_{s \in S}^{i \in I} min(L_{so}, h_i) \cdot x_{is}, \forall j \in J$$

Objective function:

$$\textbf{Minimize} : \sum_{j \in J} dev_j + p_i \sum_{i \in I} z_i + p_{over} \sum_{i \in I} over_i + \sum_{i \in I} r_i \cdot p_i \cdot (\sum_{i \in I} (1 - x_{is})) + \sum_{i \in I} (ps_i \cdot z_i + po_i \cdot over_i)$$

$ps_i$: Penalty cost for short time variable ($short_i$) for $WO_i$
$po_i$: Penalty cost for surplus time variable ($over_i$) for $WO_i$

### 3.3.2. Routine 2: Map workorder to a technician

Routine 2 performs the next step of finding technician for a given WO by consuming the shift estimated by Routine 1. While doing so, the WO would be assigned preferentially to a technician who is mapped to zone estimated for this WO by Routine 1. However, in case estimated zone-shift combination does not exist, then the WO can be assigned to any technician who is available and eligible for the WO. To model this, costs in ascending order are defined for A) technician who is mapped to estimated zone (lowest cost), B) a technician who is not mapped to any zone and C) highest cost for a technician who is not mapped to the estimated zone. In case, there is no zone associated with the WO, any eligible/available technician in the estimated shift can be assigned to WO. Thus, routine 2 finds out the most appropriate technician for a WO by minimizing cost associated with its allocation( see sample model output in Table 1).

**Index**

I: Set of all work orders assigned to the shift
T: Set of all trades (workorder types)
A: Set of all technicians who work during a given shift
S: shift-day combination

**Input data**

$n_{it}$: Number of technicians required for WO **i** in trade **t**
$h_i$: Working hours assigned to the shift for WO **i** L: Length of the shift

**Decision variables**

Binary variable $x_{iat}$ = 1 if workorder **i** is assigned to tech **a** to perform trade **t**
Continuous variable $c_{iat}$ which denotes the cost of performing workorder **i** by tech **a** for trade **t**

**Constraints**

Following constraint makes sure that WOs ($i \in$I) remain unscheduled by assigning a non-zero value to $q_i$ which in turn gets penalized for being non-zero using a very high penalty (M) in objective function.

$$\sum_{a \in A} x_{iat} + M \cdot q_i \geq n_{it} \forall i \in I$$

Following constraint makes sure that across all WOs ($i \in$I) mapped to a given shift (s), total number of hours allocated does not exceed the maximum hours (L) available in that shift for ORT and DTW WOs respectively.

For ORT Work orders, L = $L_o$

$$\sum_{i \in I, t \in T} min(h_i, L) \leq L \; \forall a \in A$$

For DTW Work orders, L = $L_d$

$$\sum_{i \in I, t \in T} min(h_i, L) \leq L \; \forall a \in A$$

$$\sum_{t \in T} x_{iat} \leq 1 \; \forall i \in I, a \in A$$

$x_{iat} = 0$ if technician a is not eligible to do trade t

Constraint which defines the cost $c_i$ of allocating $WO_i$

$$\sum_{a \in A, t \in T} c_{iat} \cdot x_{iat} \leq c_i, \; \forall i \in I$$

*Table 1.* Overall model output for a given WO

| WO | SHIFT | DATE | ZONE | TECHNICIAN |
|---|---|---|---|---|
| 123456 | ABC1 | 2020-12-18 | ABC1-18 | XXYZ |

Objective function:

$$\textbf{Minimize}: \sum_{i \in I} c_i + M \sum_{i \in I} q_i + p_i \sum_{i \in I} w_i$$

## 4. Current state

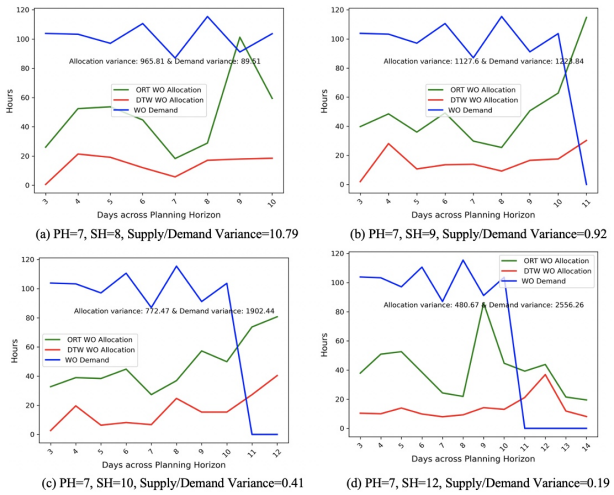
Currently, APT model has two components viz. the User Interface (UI) and the backend compute capacity. Each module of APT is written in Python and utilizes AWS infrastructure for data storage/handling. It's solver module uses Xpress package which is a commercial solver for optimization problems [10]. APT users can use the model by accessing it's Web UI and run the tool themselves, download the output and verify it. The Web UI requires the user to enter the site name and the number of days for which they want to plan the WOs. On the output side, the UI provides several metrics including the number of WOs picked/scheduled, model run time, visual plots of demand hours with respect to allocated hours as a function of dates across which WOs are due and a downloadable .csv file which contains details of scheduled WOs.

## 5. Results

The essence of having solver module is to spread out the WO hour allocation as uniformly as possible across the planning horizon. We can see that the demand hours as shown in blue have a lot of fluctuation with respect to time, however, the allocated hours across DTW and ORT are more uniformly spread out as scheduled by APT.

### 5.1. Results for Site 1

The plots here compares how the demand contrasts with the allocation of hours across Downtime window (DTW) and Operational runtime (ORT). In each of the 4 model runs, the number of workorders picked is the same, however, the number of days in planning horizon has been varied. Workorders are picked for 7 days (planning horizon: PH) in all 4 cases, however, the number of days across which they could be spread is varied. In Fig.1(a), workorders as spread across only the 8 days (scheduling horizon: SH) for which they have picked up, however, in Fig.1 (b, c, d), workorders have been spread across 9, 10, 12 days (scheduling horizon). As we can see here, the variance of allocated hours distribution is lower than variance of demand hours distribution across all the days. However, the ratio of variance of allocated hours and demand hours distribution does not essentially go down by just adding more number of days to the scheduling horizon. This can be explained by the fact that ability to schedule workorders is limited by various constraints like workorder trade, zone, compliance window and availability of downtime window for DTW workorders.




(a) PH=7, SH=8, Supply/Demand Variance=10.79
(b) PH=7, SH=9, Supply/Demand Variance=0.92
(c) PH=7, SH=10, Supply/Demand Variance=0.41
(d) PH=7, SH=12, Supply/Demand Variance=0.19

*Figure 1.* Results for site 1: Demand vs. Allocated Hours

### 5.2. Results for Site 2

Similar to previous section, the plots here compares how the demand contrasts with the allocation of hours across Downtime window (DTW) and Operational runtime (ORT). In each of the 4 model runs, the number of workorders picked is the same, however, the number of days in planning horizon has been varied. Workorders are picked for 7 days in all 4 cases, however, the number of days across which they could be spread is varied. In Fig.2 (a), workorders as spread across only the 8 days (scheduling horizon) for which they have picked up, however, in Fig.2 (b, c, d), workorders have been spread across 9, 10, 12 days (scheduling horizon). In this case, we observe an interesting thing here, the variance of allocated hours distribution is higher than the variance of demand hours distribution in Fig. 2(a). This may indicate that the model is not able to meet its objective of minimizing variance. However, a more careful investigation tells that the model could not allocate enough workorders on days: 3,7 because most workorders falling due on these days are mapped to zone/shifts which are not available on days: 3,7. This is the direct outcome of the zone constraint in our model, which is a hard constraint and specifies that a workorder can only be done in the shift mapped to the zone of that workorder. Similar to the case for site 2, the ratio of

variance of allocated hours and demand hours distribution does not essentially go down by just adding more number of days to the scheduling horizon. However, it is very evident that model does reduce the variance of allocated hours distributed as compared to demand hours. Thus, APT's model achieves load-balancing as much as possible by complying the requirements of each workorder and the associated constraints with the available resources/technician hours.

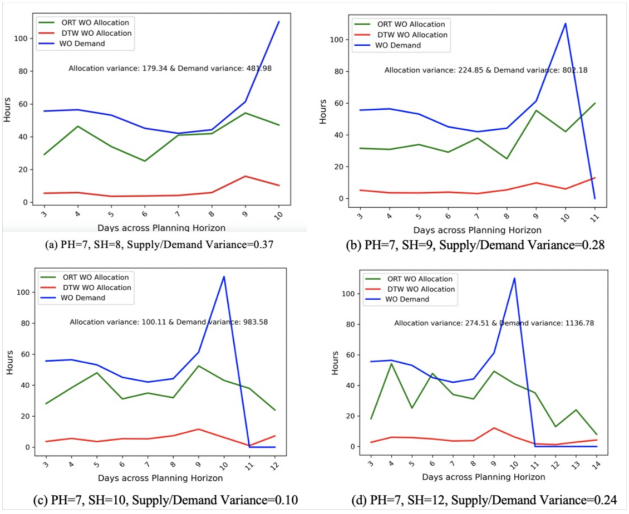



*Figure 2.* Results for site 2: Demand vs. Allocated Hours

### 5.3. Results: Model runtime

APT model solves a NP-hard problem and as expected the runtime increases as the number of workorders increases. The runtime of APT is defined by the time taken in a) Definition of decision variables and constraints, b) Addition of constraints, c) Solution of the model. Pre-processing and post-processing consume only 15% of overall runtime. Several time studies were carried to evaluate the runtime of the code developed around the algorithm and our deep dive helped identify following:

- Number of constraints increase as the number of workorders and scheduling horizon or number of shifts increase (Resource allocation constraint: Sec 3.3.1)
- Number of constraints increase as number of technicians increase (Available hour constraint: Sec 3.3.1)

Among the above 2 reasons, the former needs more attention because it is directly proportional to number of workorders and definition of this constraint requires one to iterate over all the possible shift-workorder combinations. Thus, the definition time of resource allocation constraint is directly proportional to problem size. On the other hand, definition time for shift available hours constraint does not grow as strongly with increase in problem size because its right hand side is only proportional to number of workorders. Besides, other constraints like workorder compliance window, trade eligibility and zone eligibility are also proportional to number of shift-workorder combinations. It was observed that solving for 1000 workorders would require up to 4 hours in a 30GB RAM ECS Fargate container. Hence, optimization of code was carried, several iterative loops were combined into one to define the constraints together into one by using conditional statements alongwith abstraction of nested iterative loops. These code optimization efforts helped bring down APT run time with p90 value of 20 minutes.

### 5.4. Results: Percentage of workorder scheduled

The model has 6 important constraints which drastically reduce the possible workorder-shift- technician mappings. However, with all these hard constraints in place, APT model has been found to schedule 95% of the workorders on an average. This number goes up when the scheduling horizon is increased keeping the planning horizon constant. Also, this number goes up when the percentage utilization of technician per shift is increased. This is quite obvious because increasing shift utilization increases the number of hours available in a given shift, thereby allowing more workorders to get scheduled. Most of the workorders which were unscheduled belonged to downtime window which is not available in every shift and even if available has just a 2 hour window per work day. Hence, we have allowed APT user to configure the shift utilization percentage and scheduling horizon on their own while using the model. This degree of freedom provided to APT users empowers them to utilize the model as per their requirement during times where the facilities have more than the usual volume of work.

## 6. Conclusion

This paper reports the development of an optimization model for a scheduling problem. Development of the model involved understanding the workorder scheduling process of e-commerce facility maintenance teams besides the mathematical formulation of problem. Workorder scheduling has been formulated as a Mixed-Integer Non-linear programming problem with a mix of binary/continuous decision variables and a non-linear objective function. An algorithm has been formulated keeping in mind various requirements related to workorders modelled as constraints. Some important facts about the MINLP model are as follows:

1. MINLP models uses branch-and-bound method and is able to find a feasible solution in a finite amount of time (p85 for run time is 15 mins).

2. MINLP model has 6 important constraints which drastically reduce the possible workorder-shift-technician mappings. However, with all these hard constraints in place, the model has been found to schedule 90 percent of the workorders on an average. This number goes up when the scheduling horizon is increased keeping the planning horizon constant.

3. Strong attention is required on standardizing the ingestion of data into such complex optimization models. We created an entire data engineering pipeline using various AWS services like S3 buckets, Redshift to get accurate data with zero latency.

4. Resource allocation problems like scheduling once decomposed into smaller sub-problems makes them more tractable.

## References


1. Herrmann, Jeffrey W. "Information flow and decision-making in production scheduling." IIE Annual Conference. Proceedings. Institute of Industrial and Systems Engineers (IISE), 2004. [Accessed 2020].

2. Weiwei Chen, Keyu Chen, Jack C.P. Cheng, Qian Wang, Vincent J.L. Gan, BIM-based framework for automatic scheduling of facility maintenance work orders, Automation in Construction,Volume 91,2018, Pages 15-30, ISSN 0926-5805 [Accessed 2020].

3. Tong Zhou, Dunbing Tang, Haihua Zhu, Zequn Zhang, Multi-agent reinforcement learning for online scheduling in smart factories, Robotics and Computer-Integrated Manufacturing,Volume 72, 2021,102202, ISSN 0736-5845 [Accessed 2021].

4. E. W. Weisstein, "Wolfram- NP Complete Problem," Wolfram, [Online]. Available: https://mathworld.wolfram.com/NP-CompleteProblem.html. [Accessed 2021].

5. T. E. o. E. Britannica, "Britannica, The Editors of Encyclopaedia. "NP-complete problem". Encyclopedia Britannica, 16 Jan. 2018. Accessed 24 January 2021.," [Online]. Available: https://www.britannica.com/science/NP-complete-problem. [Accessed 2021].

6. M. F. Tompkins, Optimization Techniques for Task Allocation and Scheduling in Distributed Multi-Agent Operations, Massachusetts: Massachusetts Institute of Technology, 2003.

7. Module_C on MIPs, [Online]. Available: http://web.tecnico.ulisboa.pt/mcasquilho/compute/_linpro/TaylorB_module_c.pdf. [Accessed 1 10 2020].

8. D. P. Bertsekas, Non-Linear Programming, Athena Scientific, 1995.

9. Discrete Optimization, Coursera, Available: https://www.coursera.org/learn/discrete-optimization

10. F. Xpress, Branch and Bound, Fico Xpress, [Online]. Available:https://www.fico.com/fico-xpress-optimization/docs/dms2019-03/solver/optimizer/HTML/chapter4_sec_section4003.html. [Accessed 08 2020].