Performance analysis and optimization of queueing network production systems considering non-conforming products rework and departure

Changjun LI*, Jinbin ZHANG* and Bo LI*

* School of Aeronautics and Astronautics, University of Electronic Science and Technology of China
2006 Xiyuan Ave, West Hi-Tech Zone, Chengdu 611731, China
E-mail: libo@uestc.edu.cn

Received: 19 January 2018; Revised: 21 March 2018; Accepted: 10 May 2018

Abstract
During the past decades, the production performance analysis and optimization have played a significant role in queueing network system production planning and operation design for the manufacturing plants, but how to model the optimized production system performance and solve it is facing great challenges. Since the more and more wide use of queueing network, the non-conforming product scenarios such as rework and departure, cause mass random interference. These problems when modelled in the conventional analysis and optimization models are extremely difficult to solve as they are coupled and NP-hard problems. This paper considers a modified M/M/n/m-first come first service (M/M/n/m-FCFS) queueing network approximation model that is developed to maximize utilization and profit by analysing the queue time and queue length of the products under the constraint conditions. A hybrid meta-heuristics is applied to solve the approximation model, both the service rate and buffer allocation of the queueing network system can be determined. Finally, to assess the effectiveness of the proposed method, extensive numerical experimental results from the approximation model are compared to the ARENA simulation, and sensitivity and validation analysis are also conducted, which prove it accurate, believable, and effective.

Keywords: Production planning, Operation design, Analysis and optimization, Queueing network, Non-conforming product, Hybrid meta-heuristics, Approximation model

1. Introduction
Most manufacturing companies are facing the challenge of quickly implementing flexible and complex production systems (Walid and Ali, 2002). But the production systems often experience worse performance for the unavoidable variability exists in reality. This phenomenon appears widely in queueing network and multi-product mixed production system. Performance analysis and optimization problems in queueing systems have received considerable attention in recent years, due to their wide applicability in the performance evaluation and optimal design of manufacturing systems and computer networks (Song et al., 1998). Now, the long serial lines analysis and optimization have been well studied using decomposition (Shi and Gershwin, 2014; Damodaran and Hulett, 2012; Cauffriez and Willaeys, 2006) or aggregation (Veeger et al., 2011; Belmansour and Nourelfath, 2008) methods for a long time, and extensive results are obtained. Although many studies have focused on solving the line optimization problems, the studies in which a large number of approaches are currently applied to research the performance of queueing network systems considering the non-conforming products are still extremely limited.

Research in the field of performance optimization tends to focus on issue models, such as cycle time (CT) optimization (Gilenson et al., 2015), throughput (TH) optimization (Law et al., 2008), work-in-process (WIP) optimization (Reyes et al., 2013; Qiu, 2005), and buffer allocation optimization (Shihabudin et al., 2015; Nahas, 2014). The published literature mainly focuses on production line methods (i.e., parameter model (Jimenez et al., 2012; Colledani and Tolio, 2011), queuing model (Millhiser and Burnetas, 2013; Manitz, 2008), Markov model (Wu et al., 2016; Kim and Morrison, 2015), and particularly on simulation model (Saidabad and Taghizadeh, 2015; Abdelsalam and Bao, 2006)) while widely disregarding the non-conforming product problems, such as products rework and departure, other than considering and solving them. This study is attributable to the almost all ignoring the
non-conforming product problems for performance analysis and optimization in academic research, as well as in practice, despite its increasing significance.

By using mathematical model, Marsudi and Shafeek (2013) applied queueing network theory to study the effect of throughput in optimizing resource utilization. Purnomo and Wei (2014) proposed a bi-objective mathematical model to maximize the TH. To decrease the CT, Lin et al. (2008) developed a formula for standard WIP to set the range of WIP level. Zarifoglou et al. (2013) also developed a queueing model to gain an insight into the relationship among CT and lot size. Yaghoubi et al. (2013) modelled multi-class multi-stage assembly systems with finite capacity as queueing networks to optimally control the service rates. Hajba and Horváth (2014) constructed mixed integer linear programming (MILP) models to optimize the real production systems without considering the product departure halfway. Perlman et al. (2015) proposed a cost-based optimization model that seeks to minimize the cost. However, most of the previous mathematical models consider that the production systems have infinite buffer capacities so that products’ queue time and queue length are not explicitly addressed (Dhouib et al., 2009).

Additionally, the hybrid model has been an increasing interest in studying the performance analysis and optimization in the recent years, which integrates the heuristics algorithm (such as particle swarm optimization (PSO) (Nearchou, 2011), cat swarm optimization (CSO) (Guo et al., 2015), ant colony optimization (ACO) (Yagmahan and Yenisey, 2010)). Olaian et al. (2013) investigated simulation-based optimization, and a genetic algorithm (GA) was utilized to optimize the parameters. Amiri and Mohtashami (2011) proposed a design of experiment, simulation, and GA for buffer allocation. Then, Khalaji and Zafraee (2016) used computer simulation, design of experiment and PSO to optimize the productivity. To minimize the total system cost, Li and Peng (2014) used Markov process to model the dynamic multi-state series-parallel system (MSSPS) and the GA was applied to solve the optimization model. Recently, Pehrsson and Bermedixen (2016) combined simulation-based multi-objective optimization and post optimality analysis methods with the conflicting objectives to maximize the TH and minimize the number of required improvement actions simultaneously.

Many relative works are published, but almost all the published models do not consider the treatment of non-conforming products scenarios that lead to products rework and departure in whole manufacturing process. They introduced simplifying assumptions by no non-conforming products existed during the whole process. The results obtained from these published methods are either too idealization or somewhat impractical. Namely, most of them are more conceptual than reality. For example, if products’ rework and departure are ignored and the infinite buffer capacities of a station are assumed, the models will generate more obvious error than the queueing network systems in practice, and the results are difficult to obtain. Furthermore, due to the complexity of this optimization problem, none of the aforementioned published literature have considered queueing network with finite buffer, products rework and departure properties simultaneously. To the best knowledge of the authors, there is no method in the state of the art that combining the queueing network method with hybrid meta-heuristics and is applicable for performance optimization to address the problems as described above.

In this paper, a mathematical model based on the M/M/n/m-FCFS is proposed to model the queueing network production systems considered the scenario of non-conforming products, then a hybridization of the GA and artificial neural network (ANN) is applied to solve the TH optimization and buffer allocation problems. The aim is to maximize station utilization and total profit by analysing the queue time and queue length of the products under the constrains. The reason for the hybridization is to combine the strengths of individual intelligent algorithm and achieve optimal results in solving the coupled and NP-hard problem in queueing network systems. While exploiting the power of GA to obtain candidates in different areas in the search space, ANN can provide better non-linear fitting properties that help to improve the solution in random interference. And its performance on solving much larger scale problems is extremely superior (Xiao et al., 2016).

The remainder of the paper is organized as follows: Section 2 presents the model and solution, including notations, assumptions, M/M/n/m-FCFS model, and the implementation procedure for model solution. Section 3 verifies and discusses the effectiveness of the model, and necessary sensitivity and validation analysis are implemented. Finally, the concluding remarks and future research directions are provided in Section 4.

2. Model and solution

In this section, the detailed model and solution are described. A queueing network production system which considers non-conforming products rework and departure. The objective is to specify and optimize the service rates and buffer capacities of each station to maximize the expected discounted profit or station utilization. To this end, a modified M/M/n/m-FCFS model with variability based on its probability distribution is first carried out. Then, the hybrid meta-heuristics approach is used to solve the modified model.

2.1 Notations

The following notations used in the proposed model are presented as follows:
Both subscript $i$ and $j$ indicate the station index.
Subscript $k$ indicates the product class index, where $k = 1, 2, ..., K$ ($K$ is the total number of product classes).

| $S_i$ | Station $i$. |
|-------|--------------|
| $\lambda_{ik}$ | Arrival rate of the product class $k$ before the $S_i$. |
| $\lambda_{0kj}$ | External arrival rate of product class $k$ before the $S_j$. |
| $\lambda_{ik}$ | Arrival rate to the $S_j$ that came from $S_i$ by product class $k$. |
| $\mu_{ik}$ | Service rate of product class $k$ at the $S_i$. |
| $\tau_i$ | Aggregate service time of the $S_i$. |
| $\tau_k$ | Service time of the $S_i$ by product class $k$. |
| $s_{kseq}$ | Set of stations visited by product class $k$. |
| $\rho_i$ | Service intensity of the $S_i$. |
| $L_{iq}$ | Mean queue length at the $S_i$. |
| $T_{iq}$ | Mean queue time at the $S_i$. |
| $T_{ii}$ | Mean sojourn time at the $S_i$. |
| $d_i$ | Rejection rate of the $S_i$ (namely non-conforming products departure rate). |
| $r_i$ | Rework rate of the product at the $S_i$. |
| $r_k$ | Rework rate of the product class $k$ at the $S_j$. |
| $T_s$ | Mean sojourn time of the product in the whole queueing network production line. |
| $n_i$ | Parallel machine quantity in the $S_i$. |
| $n_{ib}$ | Busy machine quantity in the $S_i$. |
| $m_i$ | Buffer capacities before the $S_i$. |
| $p(m_i)$ | Blocking probability of the $S_i$. |
| $p(0)$ | Starvation probability of the $S_i$. |
| $q_{0jk}$ | External arrival rate at $S_j$ by product class $k$. |
| $q_{jk}$ | Transition probability from $S_i$ to $S_j$ by product class $k$. |
| $TH_i$ | Throughput of the $S_i$. |

2.2 Assumptions

The basic assumptions considered in this paper are detailed as follows:

- After a long enough time, the queueing network production system can reach a steady state finally.
- External arrival time intervals of product class $k$ obey a Poisson distribution. And the service rates of stations are exponential distribution.
- The service discipline of each station is based on first come first service (FCFS).
- The buffer capacities before $S_i$ have a limited capacity with maximum $m_i$. Due to its finite capacities, products may be lost if the corresponding buffer capacities are full at the time of their arrival.
- Ignore the production setup time and the transmission time between the stations.
- All machines are reliable during a research period.
- The blocking mechanism is blocking-after-service (BAS).
- All products in the production process once detected unqualified (i.e., non-conforming products), they will be reworked at the previous station, or depart the queueing network production system, and not come back again.

2.3 Modified queue optimization model

This subsection proposes a modified queueing model of product arrival rate and station service rate for a station performance optimization problem by considering non-conforming products rework and departure. In general, the main performance indicators include queue time, queue length, station utilization, WIP level, sojourn time, and flow time. Therefore, the indicators of performance are chosen from the literature for the performance optimization. In this model, the station state is directly impacted by number of the parallel machines $n_i$ and indirectly impacted by the rework rate $r_i$. Moreover, the station state is updated through the product arrival rate $\lambda_{ib}$, rework rate $r_i$, and the station service rate $n_i/\tau_i$. 

[DOI: 10.1299/jamdsm.2018jamdsm0052] © 2018 The Japan Society of Mechanical Engineers
simultaneously. In the process, if all the parallel machines in a station are busy simultaneously, the service rate of the station is invariable, i.e. $n_i / \tau_i (1-r_i)$. The state flow diagram of the queueing system of the $S_i$ shown in Figure 1 is detailedly described as follows:

Fig. 1 State transition relationship of M/M/n/m-FCFS queueing model

To optimize the queueing network production system, the queue parameters must be quantified. The product arrival rate before $S_j$ is $\lambda_{jk}$, which can be approximately calculated by Damodaran and Hulett (2012).

$$\lambda_{jk} = \lambda_{0jk} + \sum_{m \in I_k} q_{mk} \lambda_m (1 - d_m), \quad \forall j \in J_k, \forall i \in I_k, k \in K.$$  

(1)

where $\lambda_{jk}$ is the arrival rate of product class $k$ before $S_j$, $I_k$ is the upstream station of product class $k$ index set that connect the $S_j$, and $J_k$ is the downstream station of product class $k$ index set that connect the $S_j$. $r_i$ and $d_i$ can be estimated by sampling method and Pearson $\chi^2$ test, respectively. Define $p_{ijk}$ as the internal arrival ratio from all other station in the queueing network which is computed by

$$p_{ijk} = \frac{\lambda_{ijk}}{\lambda_{ik}}, \quad \forall j \in J_k, k \in K$$  

(2)

And

$$\lambda_{ijk} = \lambda_i q_{jk}, \quad \forall j \in J_k, k \in K$$  

(3)

So far, the product arrival rate before each station is obtained. Obviously, if the aggregate service time of each station can be obtained, then the queueing network model can be modelled. That is to say the queueing network model also relies on the aggregate service time of each station. The aggregate service time $\tau_j$ at station $S_j$ has been calculated by the previous studies (see (Damodaran and Hulett, 2012) and (Pradhan et al., 2008)). Eq. (4) calculates a weighted average of the service times taking into account the external arrival rate to $S_j$, and inputs from the other last stations in the queueing network system.

$$\tau_j = \frac{\sum_{i \in K} \lambda_{0ij} \tau_i + \sum_{i \in K} \sum_{i \in I_j} \lambda_{ijk} \tau_j v_{jk}}{\sum_{i \in K} \lambda_{0ij} + \sum_{i \in K} \sum_{i \in I_j} \lambda_{ijk} v_{jk}}, \forall j \in J$$  

(4)

where the mean number of visits to each station (i.e. node) $v_{jk}$ caused by the products rework is computed by

$$v_{jk} = q_{0jk} + \sum_{m \in I_k} q_{mk} v_{mk}, \quad \forall j \in J_k, k \in K$$  

(5)

According to Eq. (4), the station service intensity $\rho_j$ can be approximated by

$$\rho_j = \frac{\sum_{i \in K} \lambda_i \tau_j}{n_j (1-r_j)}, \quad \forall j \in J$$  

(6)

Then, the $\rho_j$ also can be approximated as the instantaneous utilization of the $S_i$.

The relations of the queueing model between the steady-state probabilities are as follows: The state space of the modified queueing network is

$$\{ \lambda_i = \lambda (i = 0, 1, \ldots, m - 1) \}
\{ \mu_i = \mu (i = 0, 1, \ldots, n - 1) \} \quad (\mu > 0)
\{ \mu_i = \mu (i = n, n + 1, \ldots, m) \}$$  

(7)

By utilizing Eq. (6), the starvation probability of the $S_i$ i.e., there is not any product at $S_i$ waiting for service, is written as follows.
where $z_i$ indicates one state of the $S_i$ in the queueing model, it is a nonnegative integer, and $z_i \leq n_i$.

Now, the average number of busy machines at $S_i$ can be computed as

$$n_{b_i} = \left[ \frac{n_i^x \times \rho_i^{n_i}}{(z_i-1)!} + n_i \sum_{z=0}^{n_i} \frac{n_i^x \times \rho_i^{n_i}}{n_i!} \right] \times P(0)$$

$$= n_i \rho_i \left[ \frac{n_i^x \times \rho_i^{n_i}}{(z_i-1)!} \times P(0) + \frac{(n_i \rho_i)^{n_i+1}}{(n_i-1)!} + \sum_{z=0}^{n_i} \frac{n_i^x \times \rho_i^{n_i}}{n_i!} \times P(0) \right]$$

$$= n_i \rho_i \left[ 1 - P(0) \right]$$

where $\phi_i = z_i - 1$.

The average loss quantity of products per unit time (minute) at $S_i$ can be calculated by

$$\lambda_{loss} = \lambda_i \left[ 1 - P(0) \right]$$

And the loss ratio $p_{loss}$ is defined by the blocking probability, which is computed based on the $M/M/n/m$-FCFS model and given by

$$P_{loss} = P_i(m_i) = \frac{n_i^x \times \rho_i^{n_i}}{n_i!} \times P(0)$$

The average waiting queue length $L_{wq}$ at $S_i$ can be written

$$L_{wq} = \left[ \frac{n_i^x \rho_i^{n_i+1} P_i(0)}{n_i! \left( 1 - \rho_i \right)} \right] \left[ 1 - (m_i - n_i + 1) \rho_i^{n_i+1} + (m_i - n_i) \rho_i^{n_i+1} \right] \times P(0)$$

$$= \frac{n_i^x \rho_i^{n_i+1} P_i(0)}{2n_i!} \left( m_i - n_i \right) \left( m_i - n_i + 1 \right) P_i(0)$$

Thus, the average queue length, namely average number of products at $S_i$ is

$$L_i = \sum_{z_i=0}^{n_i} z_i P_i(z_i)$$

$$= \sum_{z_i=0}^{n_i} z_i P_i(z_i) + \sum_{z_i=0}^{n_i} \sum_{z_i=0}^{n_i} \left( z_i - n_i \right) P_i(z_i) + \sum_{z_i=0}^{n_i} n_i P_i(z_i)$$

$$= L_{wq} + n_i - \sum_{z_i=0}^{n_i} \left( n_i - z_i \right) P_i(z_i)$$

Therefore, the number of WIP in the whole queueing network production system is

$$\text{WIP} = \sum_i L_i$$

And, the effective arrival rate of the products before $S_i$ can be written by

$$\lambda_i = \mu_i \left[ n_i - \sum_{z_i=0}^{n_i-1} \left( n_i - z_i \right) P_i(z_i) \right]$$

According to the Little’s law, the average sojourn time of products at $S_i$ can be got

$$T_{is} = \frac{L_i}{\lambda_i} = T_{wq} + \frac{1}{\mu_i}$$

Thus, the sojourn time of the products in the whole queueing network system can be written by

$$T_s = \sum_{i=1}^{n} T_{is} = \sum_{i=1}^{n} T_{wq} + \frac{1}{\mu_i}$$

And the queue time at $S_i$ is
Here, \( k \) and the coefficient of variation \( \rho \) are defined as
\[
\rho = \frac{\sigma}{\mu}
\]
flowing network system, the sequence of product class \( k \) is expressed as
\[
T_{\text{flow}} = \sum_{i \in k_{\text{flow}}} \left( T_i + \tau_i \right)
\]
And,
\[
T_{\text{flow}} = \max \{T_{\text{flow}} \mid s_{\text{flow}} = 1, 2, \ldots\}
\]
\[\text{CV}_p = \frac{\lambda_{\text{ok}}}{\sqrt{\lambda_{\text{ok}}}}\]
where \( a_i \) indicates the income unit, \( \beta_i \) and \( \gamma_i \) indicate the cost unit of service rate \( \mu_i \) and buffer capacities \( m_n \), respectively.

To sum up, the whole mathematical expression of the model, i.e. the cost objective function in steady state is as follows

\[
\text{max } O_{\text{pro}}(n_i, m_i) = \sum_i \left( \sum_j \frac{n_j^{-\lambda_j - \beta_j - \gamma_j}}{n_j!} \left( \sum_{k=0}^{\lambda_j} \frac{(n_j \times \rho_j)^k}{n_j!} \cdot \frac{(n_j \times \rho_j)^{m_j} \times (1 - \rho_j^{n_j})}{n_j!} \right) - \beta_i \mu_i - \gamma_i \right)
\]

Subject to

\[
\alpha_i > 0, \beta_i > 0, \gamma_i > 0
\]

\[
\rho = \frac{\lambda_i}{n_i \mu_i}, \quad 0 < \rho_i \leq 1, \quad \forall i
\]

\[
\lambda_i > 0, \quad \mu_i > 0, \quad \forall i
\]

\[
\text{int}(n_i), \quad \text{int}(z_i), \quad 0 \leq z_i \leq m_i, \quad \forall i
\]

\[
m_i > n_i > 1, \quad \forall i
\]

\[
\sum_{z=0}^{m_i} p_{z} = 1
\]

- Utilization maximization:

\[
\text{max } O_{\text{ut}}(n_i, m_i) = \sum_i (1 - P_0) - P(m_i)
\]

\[
= \sum_i \left( 1 + \frac{n_i^{-\lambda_i - \beta_i - \gamma_i}}{n_i!} \right) (1 - P_0)
\]

\[
= \sum_i \left( 1 + \frac{n_i^{-\lambda_i - \beta_i - \gamma_i}}{n_i!} \right) \left( \sum_{k=0}^{\lambda_i} \frac{(n_i \times \rho_i)^k}{n_i!} \cdot \frac{(n_i \times \rho_i)^{m_i} \times (1 - \rho_i^{n_i})}{n_i!} \right) - 1, \quad \rho_i = 1
\]

Subject to

\[
\lambda_i > 0, \quad \mu_i > 0, \quad i \in z^*, \quad \forall i
\]

\[
0 \leq P_0(0) \leq 1, \quad 0 \leq P(m_i) \leq 1
\]

\[
m_i, n_i \in z^* \land m_i > n_i
\]

The objective function \( O_{\text{pro}}(n_i, m_i) \) and \( O_{\text{ut}}(n_i, m_i) \) can be solved by the hybrid meta-heuristics programmed with MATLAB software, which is detailedly presented in Subsection 2.4.

### 2.4 Solution with hybrid meta-heuristics

For the solution of the optimization model, a hybrid meta-heuristics which integrates ANN and GA is proposed. The relationship diagram between the algorithms is illustrated in Fig. 2. In the hybrid meta-heuristics, the ANN is used to deal with the nonlinear relationship of the objective function. And the relationship between the input and output can be determined (there is a nonlinear relationship) by using the ANN self-learning (explicit input and output from a training dataset). After training weight value determination, the new input of the nonlinear objective function can be predicted. GA are used to solve the problems of constrained or unconstrained optimization, which is more flexible and without any restrictions. But the only difficulty is the choice of the chromosome encoding and the evaluation function, so the ANN can be used to can make up for this shortcoming. The basic steps of the hybrid meta-heuristics are: (1) construct a suitable ANN architecture based on the characteristics of the optimization function Eq. (26a) and Eq. (27a); (2) use the input and output data of the nonlinear objective functions to train the ANN; (3) predict the output of the function by the trained ANN, and take the prediction results of the trained ANN as the value of individual fitness for
GA; (4) search for the global optimal value and the corresponding input value of the objective function by selection, crossover and mutation operations.

The steps of the hybrid meta-heuristics algorithm are introduced as follows.

**Step 1:** Initialize ANN.

**Step 2:** Model ANN system, the architecture of ANN is 2 (input parameter, i.e., the service rate and buffer capacity)-5 (neuron)-1 (output parameter, i.e., the objective function).

**Step 3:** Train ANN until the termination condition is reached, i.e. the train epochs is 100, or the training error is 0.0000001.

**Step 4:** Predict the fitness of the objective function by the trained ANN.

**Step 5:** Create initial population randomly through two solutions (service rate, buffer capacities) within the prescribed range.

**Step 6:** Randomly select two individuals (solutions) from the population and obtain the new individuals by a crossover procedure. Then the feasibility of individuals is checked. If the two individuals are not feasible, then cross again.

**Step 7:** Mutate the new individual with a given slightly probability. Then the feasibility of individual is checked. If the individual is not feasible, then mutate again.

**Step 8:** Calculate the objective function (Fitness value) obtained by Step 4.

**Step 9:** Compare and select the best individual in the population. Join the better one of the two solutions to the population and abandon the other one.

**Step 10:** Randomly create new constructed solutions to add the current population.

**Step 11:** If the termination conditions (given number of generations) are met, then stop the hybrid meta-heuristics program. Otherwise, go to Step 6.

In the process of algorithm solution, the hybrid meta-heuristics function is employed with MATLAB software to produce the program solution.

By the optimization approach introduced above, the performance of single station is optimum, but the local optimization does not guarantee the global optimum. Therefore, the bottleneck stations are chosen to optimize in the next step. Since bottleneck stations lead to incoordination among all stations and impact on the utilization rate, the bottleneck stations can be found through the analysis and comparison of the queue length and queue time of products at all the stations in the queueing network system. If a station is a bottleneck station, adjust the utilization rate or profit by optimize the service rate and buffer capacities of the station to solve the bottleneck problem.
3. Numerical example

In this section, the numerical results of the proposed model apply to the queueing network system formed by 11 stations and 8 buffers is reported. 3 product classes with different arrival time intervals, service times, rework rates, departure rates, and different sequences of operations are chosen to examine the performance of the proposed method. Since a benchmark solution for comparison does not exist currently, two scenarios with the optimization method are performed in Subsection 3.2. Scenario I and Scenario II representing different optimization strategies respectively with maximizing profit and maximizing utilization introduced in Subsection 2.3 are modelled. Taking a multi-product mixed and rework production system with finite buffer capacities as an example, see Fig. 3. The service sequence of product class 1 is Station 1 → Station 2 → Station 3 → Station 4 → Station 5 or Station 1 → Station 7 → Station 8 → Station 5. The service sequence of product class 2 is Station 6 → Station 7 → Station 8. The service sequence of product class 3 is Station 9 → Station 10 → Station 11 or Station 9 → Station 8 → Station 11. The time intervals between arrivals of product class 1, product class 2, and product class 3 are Pois. (0.3571), Pois. (0.3846), and Pois. (0.3448), respectively.

As far as the hybrid meta-heuristics, the float-encoding (namely, real encoding) method is applied to encode the individuals in GA, and the length of each individual encoding is 2, because the nonlinear objective function has 2 input parameters. The individual value range is 1-10 and 0-50, respectively. The number of evolutionary times is 200. The basic parameters of GA are mostly based on empirical observation with respect to the issue, where population size is 20, crossover rate is 0.9, mutation rate is 0.1, selection function is tournament, both crossover and mutation function are uniform. According to the nonlinear objective function of the target model, there are 2 input parameters and 1 output parameters, so the structure of the ANN is 2-5-1. 10,000 groups of input and output data of nonlinear objective function randomly generated by MATLAB, of which 9,900 groups are randomly selected as the training dataset to train the ANN, and 100 groups are used to test the performance of the network fitting. After the ANN is well trained, it is used to predict the output of nonlinear objective functions.

Due to the novelty of theoretical foundations and method applications, there is no existing approach to compare the proposed optimization model as described above. In order to verify the capability of the proposed model, a computer simulation is used to provide a comparison because it is a popular method to model variability production system. However, when the number of alternative configurations is large, it is almost impossible to implement and evaluate all of the parameter configurations every time by simulation (Mehmet et al., 2016). Hence, only the computer simulation of the best obtained scenario is executed, and the results are compared. In this paper, ARENA software is chosen to simulate the scenarios due to its availability for wide range of researchers. In ARENA simulation, the initial simulation of system parameters setting used a chosen to simulate the scenarios due to its availability for wide range of researchers. In ARENA simulation, the initial simulation of system parameters setting used a

3.1 Model evaluation and verification

The goal of this subsection is to study the accuracy of the performance indicators obtained in Section 2.3. To implement accuracy validation, the relative error of performance value for the proposed approach and the simulation is defined by

$$
Relative\ error = \frac{|Value_{Simulation} - Value_{math}|}{Value_{Simulation}} \times 100\%
$$

(28)
Four main parameters of performance value, i.e., the busy number of parallel machines of the station \( (n_b) \), queue length \( (L_q) \), queue time \( (T_q) \), and the instantaneous utilization \( (U_i) \) are selected and compared. These parameters resulted from proposed method (indicated by Math.) compare with the ARENA simulation (indicated by Sim.) and their difference \( (Err.) \) namely, the relative error in Eq. (28) expressed in percentage) are presented in Table 1.

Table 1 Comparison of performance between mathematical model and simulation model of each station

| \( S_i \) | Sim. | Math. | \( n_b \) | Sim. | Math. | \( L_q \) | Sim. | Math. | \( T_q \) | Sim. | Math. | \( U_i \) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| \( S_1 \) | 2.54 | 2.69 | 6.04 | 4.04 | 4.08 | 1.22 | 1.59 | 1.52 | 4.63 | 0.85 | 0.93 | 9.85 |
| \( S_2 \) | 2.61 | 2.62 | 0.47 | 3.63 | 3.49 | 3.92 | 1.39 | 1.42 | 1.80 | 0.87 | 0.90 | 3.29 |
| \( S_3 \) | 2.51 | 2.51 | 0.15 | 2.53 | 2.50 | 0.94 | 1.01 | 1.05 | 4.43 | 0.84 | 0.86 | 2.32 |
| \( S_4 \) | 2.36 | 2.35 | 0.33 | 1.79 | 1.82 | 1.78 | 0.76 | 0.77 | 2.08 | 0.79 | 0.80 | 0.74 |
| \( S_5 \) | 2.47 | 2.47 | 0.06 | 2.71 | 2.51 | 2.73 | 1.10 | 1.02 | 7.62 | 0.82 | 0.83 | 1.75 |
| \( S_6 \) | 3.20 | 3.39 | 5.80 | 2.66 | 2.49 | 6.10 | 1.11 | 0.98 | 11.23 | 0.80 | 0.87 | 8.91 |
| \( S_7 \) | 2.70 | 2.62 | 3.03 | 3.91 | 3.48 | 11.02 | 1.43 | 1.37 | 4.07 | 0.90 | 0.90 | 0.24 |
| \( S_8 \) | 2.90 | 2.84 | 2.16 | 6.96 | 7.18 | 3.22 | 2.55 | 2.72 | 6.76 | 0.97 | 1.00 | 3.00 |
| \( S_9 \) | 2.52 | 2.71 | 7.47 | 3.45 | 3.79 | 9.67 | 1.37 | 1.40 | 2.18 | 0.84 | 0.96 | 14.58 |
| \( S_{10} \) | 2.59 | 2.59 | 0.23 | 3.37 | 3.29 | 2.32 | 1.30 | 1.34 | 2.71 | 0.86 | 0.89 | 3.05 |
| \( S_{11} \) | 1.95 | 1.95 | 1.22 | 5.34 | 5.28 | 1.14 | 2.74 | 2.70 | 1.31 | 0.97 | 1.00 | 3.09 |

To provide further insights, the results are also compared in terms of the whole accuracy of the proposed model in ARENA simulation, the whole performance of the queueing network system model and hybrid meta-heuristics are carried out. As shown in Table 2, the performance after optimization has the less number of WIP 61.25 and sojourn time 24.12 than its optimization before. Exactly, the proposed method can reduce the number of WIP of whole queueing network from 72.91 of optimization before to 61.25 of optimization after with the decreasing of 15.99%, and the sojourn time from 27.93 minutes before optimization to 24.12 minutes with the saving of 13.64%. From the table, the performance measure of the improved \( M/M/n/m-FCFS \) model is effective with a maximum error range 12.44%, which can be employed to further performance optimization.

Table 2 Comparison of performance between mathematical model and simulation model of whole queueing network production system

| WIP | \( \text{Sim.} \) | \( \text{Math.} \) | \( \text{Err.} \) (\%) | WIP | \( \text{Sim.} \) | \( \text{Math.} \) | \( \text{Err.} \) (\%) |
|---|---|---|---|---|---|---|---|
| Before | 68.71 | 72.91 | 6.10 | 24.84 | 27.93 | 12.44 |
| After | 59.16 | 61.25 | 3.53 | 23.15 | 24.12 | 4.19 |

3.2 Discussion and sensitive analysis

Fig. 4 summarizes the results of the comparison of performance with the different parameters, such as the service intensity, buffer capacity, service rate, and arrival rate of products. As seen in Fig. 4(a), with the service intensity increasing, both the queue time and number of WIP increased, especially when \( \rho \geq 0.6 \), they increased rapidly. That is to say, high service intensity keeps high queue time and WIP level, which is clearly noticeable in almost all results. Service intensity also mightly effects the queue time and makes it be exponential growth, especially when the parameter \( \rho \geq 0.6 \), see Fig. 4(b). According to Fig. 4(c), with respect to the buffer capacities levels \( (m \geq 15) \), it founds that there is no significant difference in developing the queue time. This is a side effect of the variability because buffer capacities indirectly diminish on the coefficient of variation of products arrival and station service rate. We precisely notice that optimize performance by increasing the buffer capacities is at the cost of high WIP level and long queue time when the buffer capacities are in the range from 0 to 15. Furthermore, it is well known that the buffer capacities have a limited contribution to the performance of the queueing network system beyond a certain level. According to Fig. 4(d), it is remarkable that the curves of queue time and queue length varying with their buffer capacities are tend to be similarly if ignore their units on figure. The queue time difference is relatively low compared to the queue length under the same buffer capacities practically. So reducing the buffer capacities can reduce the queue time and queue length, but its loss ratio defined in Eq. (11) increases rapidly. That is because reducing the buffer capacities also increases the station blocking probabilities which causes the effective service rate to decrease. As far as the buffer capacities before the stations increased, the average queue time of queueing network system grows significantly, although its average throughput grows slightly. To explore the influence of the parameters TH, m, and coefficient of variation on the performance of the system, the arrival rate and service rate are kept the same changes. The computational queue time and service rate are illustrated in Fig. 4(e). It can be observed that with the increasing service rate, both the queue time and queue length decrease rapidly, that is to say they are extremely sensitive to the parameter \( n \). But increasing the \( n \) also leads high production cost. So, as we know, increasing \( n \) to reduce the queue time and queue length is not a wise
choice. This has a negative influence on the objective function given by Eq. (26a) concerning the purchase cost and maintenance cost of the machines. The production managers may consider optimizing the buffer capacities to improve the performance in the case of limited resources. Finally, as seen in Fig. 4(f), it can be found that, with the growth of flow density, i.e., arrival rate, the average queue time and queue length are increased. That means increasing flow density can improve TH, but will result in worse queue time (sojourn time) (Zhang et al., 2016). As a conclusion, simply increasing the products arrival rate in this case is not a wise choice to be applied with performance optimization that contains coefficient of variation (see Eq. (24)).

![Fig. 4 Performance parameters relationship of single station](image)

Scenario I: for the maximizing profit, sum of all the classes products arrival rate $\lambda_a$ equals 2.9. And the departure ratio of non-conforming products $d = (0, 0, 0.02, 0, 0.02, 0, 0.02, 0, 0.02, 0, 0.02, 0)^T$, rework ratio $\pi = (0, 0.06, 0.05, 0, 0, 0, 0.07, 0, 0, 0.05, 0)^T$. In terms an optimized station, the cost parameters $\alpha = 50$, $\beta = 3$, and $\gamma = 0.2$. The elapsed time and error accumulative total of ANN obtained are 27.48 seconds and 3.49, respectively. And the best known fitness, TH,
and buffer capacities \( m \) to the GA are 135.17, 2.39 and 9, respectively. Fig. 5 exemplarily depicts the effect of the precision interval on ANN and GA in the process and result of utilization optimization. As seen in Fig. 5(a), the prediction output fits the expectation output very well. Fig. 5(b) confirms that the prediction results on performance indicators are generally minimal error, even if the maximum prediction error is no more than 0.2%. Fig. 5(c) represents that the convergence rate of the proposed hybrid meta-heuristics is satisfied. According to the curve, when the program runs to about the tenth generation, its maximum value can be reached.

![Fig. 5 Profit optimization sample for single station varies with service rate and buffer capacities](image)

Next, the proposed model is used to analyse different optimization scenarios. By comparing Table 1, the performance of the whole queueing network system runs not very well. Especially the Station 8 and Station 11 have the relative bad performance, such as the queue length (6.96) and queue time (2.55) of the Station 8, and the queue time (2.74) of the Station 11. So, Station 8 and Station 11 are the key optimization objects, and the optimization model and algorithm used as introduced above. The process and result of utilization optimization are shown in Fig. 6.

![Fig. 6 Utilization optimization sample for single station varies with service rate and buffer capacities](image)

According to Fig. 6(a), the prediction outputs almost are well predicted by ANN compare with the expectation output. The maximum prediction error of ANN is about 0.85%, which achieves a satisfactory prediction accuracy, see Fig. 6(b). Additionally, as depicted in Fig. 6(c), the fitness of GA reaches to the maximum value, when the algorithm program executes to no more than 40 generations. The corresponding parameters obtained such as, buffer capacities \( m = (12, 12, 10, 10, \infty, 10, 12, \infty, 10, 12, \infty) \) and number of parallel machines \( n = (3, 3, 3, 3, 4, 3, 4, 3, 3, 3) \).

In order to facilitate the discussion and sensitive analysis in the field of Factory Physics for the optimization results, and that the maximum utilization and profit are almost similar. Therefore, the following analysis are focused on the scenario of maximum utilization rate in detail.

As for the maximizing utilization in Scenario II: according to Eq. (27a), make both the parameters \( p_i(0) \) and \( p_i(m_i) \) or their sum be minimum. The best known optimized fitness is 0.8924, namely, the optimal utilization ratio of each station is 89.24%. And the corresponding service rate, buffer capacities obtained by GA are 1.66 and 18, respectively. Fig. 7 exemplarily depicts the optimal value of decision parameters and the corresponding performance under the proposed optimization model by using the same data. As the optimization results shown in Fig. 7, the performance optimized by the proposed method provides the better results than performance optimization before. Exactly, the proposed method can reduce the queue time at the Station 8 from 2.72 of Scenario II to 0.70 with the saving of 288.57%, and the queue length from 7.18 of Scenario II to 1.98 with the saving of 262.63%. Similarly, for the Station 11, the queue time from 2.59 to 0.92 with the saving 181.52%, and the queue length from 5.10 to 2.32 with the saving 119.83%. Furthermore, there is no doubt that the queue time and queue length at both the Station 8 and 11 reduce
obviously, but the other stations in the queueing network system also reduce slightly. It is because when the performance of a bottleneck station is developed, the whole performance of the queueing network system also can be improved. In other word, the local optimization cannot lead to the global optimization, but it can be close to the global optimization gradually.

Additionally, it is worth mentioning that profit maximization does not always improve the station utilization rate. As seen in Fig. 7(c), with the profit maximizing, the station utilization rate before the Station 8 from 0.96 to 0.73 with declining 31.51%. The reasons for constraints of costs in reality, sometimes suboptimal can also meet production requirements.

To analyse and compare the before and after optimization results of performance concerning a single class product, Table 3 summarizes the performance comparison of single class, which also confirms the findings in Fig. 7. After performance optimization, the number of WIP, TH, and queue time reach a better performance level. Especially, the number of WIP and queue time of product class 2 reduced by 28.85% and 48.79%, respectively. And the TH of product class 3 also increased by 29.23%. In addition, the flow time also is obviously optimized, in particular the product class 2 and 3. The flow time of product class 2 from 8.78 before (Bef.) optimization to 6.19 with the saving of 29.50%, and product class 3 from 8.38 after (Aft.) optimization to 6.31 with the saving of 24.70%. But at this optimization level of scenario, the parameters of service rate and buffer capacities at a station have no significant influence on the product class 1. In other words, although the proposed approach is applied to the scenario II for profit maximization, in terms of product class 1, its performance does not improve obviously.

In summary, due to the inevitable existence of variability, such as products arrival interval time variability and service time variability, resulting in the average utilization rate of each station cannot reach 100%, although the arrival rate of the product is always equal to the service rate of the station. Generally, higher coefficient of variation has a negative influence on all performance indicators. An effective way is try to make sure production process that keep the same service rate (or time) of each station, while reducing their various coefficient of variation likewise.

4. Conclusion

This paper has addressed a production performance analysis and optimization problem originated from the highly flexible manufacturing factory. A novel method, developing the latest advances of M/M/n/m-FCFS approximation queueing network with a hybrid meta-heuristics, has been presented to calculate the objective functions. The advantages of ANN system fitting and GA nonlinear optimization are integrated. It can overcome the limit of using others measures presented in some literature, that is hard to obtain the nonlinear value when the scale of the queueing network systems become extremely large. More importantly, from the mathematical derivation, it is easy to discover that although the product arrival parameter $\lambda$ and station service parameter $\mu$ implement the same changes, the corresponding performance is obvious difference. It has revealed that variability is more significant than the traditional view that the service rate and its coefficient of variation are the crucial factors for the modified M/M/n/m-FCFS model.
The main contribution of this paper is to develop an analysis and optimization approach for queueing network systems with finite buffer capacities, which considers the non-conforming products rework and departure comparing to previous published works. The paper presents an innovative perspective for practitioners to maximize the profit and utilization, which has a great potential for optimizing the performance of the large-scale queueing network systems. It also supports practitioners in achieving performance improvement in semiconductor manufacturing systems, and then the production planning and control can be well implemented.

It is also worth noting that this paper is limited. Although the coefficient of variation is involved in the proposed method, the transfer and convergence of variability among the stations and how their effect on the performance are not very clear, which still needs more work. Besides, the station (machine) failure has been ignored, and the Poisson, exponential distributions are assumed in the proposed model to make the performance analysis tractably. Such assumptions will be relaxed in future work.

Acknowledgment

The authors are thankful to the anonymous reviewers for their constructive and helpful comments that have led to this much improved manuscript. This work was supported by the National Natural Science Foundation of China under grant number 71671026, and also supported by the Sichuan Science & Technology Program under Grant No. 2018GZ0306 & 2017GZ0034.

References

Abdelsalam, H. M. E., and Bao, H. P., A simulation-based optimization framework for product development cycle time reduction, IEEE Transactions on Engineering Management, Vol. 53, No. 1 (2006), pp. 69-85.

Amiri, M., and Mohrshadi, A., Buffer allocation in unreliable production lines based on design of experiments, simulation, and genetic algorithm, The International Journal of Advanced Manufacturing Technology, Vol. 62, No. 1-4 (2011), pp. 371-383. DOI: 10.1007/s00170-011-3802-8.

Belmansour, A. T., and Nourzelfah, M., An aggregation method for performance evaluation of a tandem homogenous production line with machines having multiple failure modes, Reliability Engineering & System Safety, Vol. 95, No. 11 (2008), pp. 1193-1201. DOI: 10.1016/j.ress.2010.05.002.

Cauflfriez, L., and Willaeya, D., A predictive model for improving production line performance, International Journal of Advanced Manufacturing Technology, Vol. 29, No. 9-10 (2006), pp. 969-979. DOI: 10.1007/s00170-005-2583-3.

Colledani, M., and Tolio, T., Joint design of quality and production control in manufacturing systems, Cirk Journal of Manufacturing Science and Technology, Vol. 4, No. 3 (2011), pp. 281-289. DOI: 10.1016/j.cirpj.2011.06.008.

Damodaran, P., and Hulett, M., Analytical approximations to predict performance measures of manufacturing systems with general distributions, job failures and parallel processing, European Journal of Operational Research, Vol. 221, No. 1 (2012), pp. 74-86. DOI: 10.1016/j.ejor.2012.03.013.

Dhoubi, K., Gharbi, A., and Landolsi, N., Throughput assessment of mixed-model flexible transfer lines with unreliable machines, International Journal of Production Economics, Vol. 122, No. 122 (2009), pp. 619-627.

Gileson, M., Hassoun, M., and Yedidison, L., Setting defect charts control limits to balance cycle time and yield for a tandem production line, Computers & Operations Research, Vol. 53 (2015), pp. 301-308.

Guo, J., Sun, Z., Hong Tang, L. Y., and Zhang, Z., Improved cat swarm optimization algorithm for assembly sequence planning, Open Automation and Control Systems Journal, Vol. 7, No. 1 (2015), pp. 792-799.

Hajba, T., and Horváth, Z., MILP models for the optimization of real production lines, Central European Journal of Operations Research, Vol. 23, No. 4 (2014), pp. 899-912. DOI: 10.1007/s10100-014-0337-8.

Hopp, W. J., and Spearman, M. L., Factory physics: foundations of manufacturing management (2002), Tsinghua Press.

Jimenez. M. A. et al., Automation and parameters optimization in production line: a case of study, International Journal of Advanced Manufacturing Technology, Vol. 66, No. 9-12 (2012), pp. 1315-1318.

Khalaji Assadi, M., and Zahraee, S. M., Integration of computer simulation, design of experiments and particle swarm optimization to optimize the production line efficiency, International Journal of Swarm Intelligence and Evolutionary Computation, Vol. 5, No. 2 (2016), pp. 136. DOI: 10.4172/2090-4908.1000136.

Kim, W., and Morrison, J. R., The throughput rate of serial production lines with deterministic process times and random setups: Markovian models and applications to semiconductor manufacturing, Computers & Operations Research, Vol. 53 (2015), pp. 288-300. DOI: 10.1016/j.cor.2014.03.022.

Law, C. F., Gwee, B. H., and Chang, J. S., Asynchronous control network optimization using fast minimum-cycle-time analysis, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 27, No. 6 (2008), pp. 985-998. DOI: 10.1109/tcad.2008.923238.

Li, Y. F., and Peng, R., Availability modeling and optimization of dynamic multi-state series-parallel systems with random reconfiguration, Reliability Engineering and System Safety, Vol. 127, No. 4 (2014), pp. 47-57.

Lin, Y. C., Tsai, C. H., Li, R. K., Chen, C. P., and Chen, H. C., The study of the cycle time improvement by
work-in-process statistical process control method for IC foundry manufacturing, Asian Journal on Quality, Vol. 9, No. 3 (2008), pp. 71-91. DOI: 10.1108/1598268820800028.

Manitz, M., Queueing-model based analysis of assembly lines with finite buffers and general service times, Computers & Operations Research, Vol. 35, No. 8 (2008), pp. 2520-2536. DOI: 10.1016/j.cor.2006.12.016.

Marsudi, M., and Shafeek, H., Production line performance by using queuing model, Conference on Manufacturing Modelling, Management, and Control (2013), pp. 1152-1157.

Mehmet, P., Mustafa, Y., and Bilal, T., Variability modelling and balancing of stochastic assembly lines, International Journal of Production Research, Vol. 54, No. 19 (2016), pp. 5761-5782. DOI: 10.1080/00207543.2016.1177236.

Millhiser, W. P., and Burnetas, A. N., Optimal admission control in series production systems with blocking, IIE Transactions, Vol. 45, No. 10 (2013), pp. 1035-1047. DOI: 10.1080/0740817x.2012.706732.

Nahas, N., Buffer allocation and preventive maintenance optimization in unreliable production lines, Journal of Intelligent Manufacturing, Vol. 28, No. 1 (2014), pp. 85-93. DOI: 10.1007/s10845-014-0963-y.

Nearough, A. C., Maximizing production rate and workload smoothing in assembly lines using particle swarm optimization, International Journal of Production Economics, Vol. 129, No. 2 (2011), pp. 242-250.

Olaitan, O. A., Shaaban, S., and Geraghty, J., Evaluation of production control strategies for negligible-setup, multi-product, serial lines with consideration for robustness, Journal of Manufacturing Technology Management, Vol. 24, No. 3 (2013), pp. 331-357. DOI: 10.1108/17410381313118864.

Pehrsson, L., Ng, A. H. C., and Berendtzen, J., Automatic identification of constraints and improvement actions in production systems using multi-objective optimization and post-optimality analysis, Journal of Manufacturing Systems, Vol. 39 (2016), pp. 24-37. DOI: 10.1016/j.jmsy.2016.02.001.

Perlmutter, Y., Crispil, E., and Kalir, A., Production planning and WIP assignment for wafer fabrication tools with availability constraints, IEEE Transactions on Semiconductor Manufacturing, Vol. 28, No. 3 (2015), pp. 385-392.

Pradhan, S., Damodaran, P., and Srihari, K., Predicting performance measures for Markovian type of manufacturing systems with product failures, European Journal of Operational Research, Vol. 184, No. 2 (2008), pp. 725-744.

Purnomo, H. D., and Wee, H. M., Maximizing production rate and workload balancing in a two-sided assembly line using harmony search, Computers & Industrial Engineering, Vol. 76 (2014), pp. 222-230.

Qiu, R. G., Virtual production line based WIP control for semiconductor manufacturing systems, International Journal of Production Economics, Vol. 95, No. 2 (2005), pp. 165-178. DOI: 10.1016/j.ijpe.2003.12.004.

Reyes Levalle, R., Scavarda, M., and Nof, S. Y., Collaborative production line control: minimisation of throughput variability and WIP, International Journal of Production Research, Vol. 51, No. 23-24 (2013), pp. 7289-7307.

Saidabad, A. A., and Taghizadeh, H., Performance and improvement of production line function using computer simulation (case study: an iron foundry), American Journal of Computational Mathematics, Vol. 5, No. 4 (2015), pp. 431-446. DOI: 10.4236/ajcm.2015.54038.

Shi, C., and Gershwin, S. B., Improvement of the evaluation of closed-loop production systems with unreliable machines and finite buffers, Computers & Industrial Engineering, Vol. 75, No. 1 (2014), pp. 239-256.

Shihabudin Ismail, M., Hussain, M. I., Zain, Z. M., and Muhd, N., A Management guide for determination of optimal buffer allocation in un-paced production lines using 6 Steps OBA, Applied Mechanics and Materials, Vol. 793 (2015), pp. 637-641. DOI: 10.4028/www.scientific.net/AMM.793.637.

Song, D. P., Sun, Y. X., and Xing, W., Optimal control of a stochastic assembly production line, Journal of Optimization Theory and Applications, Vol. 98, No. 3 (1998), pp. 681-700.

Veeger, C. P. L., and Others, Predicting cycle time distributions for integrated processing workstations: an aggregate modeling approach, IEEE Transactions on Semiconductor Manufacturing, Vol. 24, No. 2 (2011), pp. 223-236. DOI: 10.1109/TSM.2010.2094630.

Walid, A. K., and Ali, G. B. Capacity estimation of a multi-product unreliable production line, International Journal of Production Research, Vol. 40, No. 18 (2002), pp. 4815-4834. DOI: 10.1080/0020754021000024148.

Wu, C., Chien, W., Chuang, Y., and Cheng, Y., Multiple product admission control in semiconductor manufacturing systems with process queue time (PQT) constraints, Computers & Industrial Engineering, Vol. 99 (2016), pp. 347-363. DOI: 10.1016/j.cie.2016.04.003.

Xiao, L., Song, S., Chen, X., and Coit, D. W., Joint optimization of production scheduling and machine group preventive maintenance, Reliability Engineering & System Safety, Vol. 146 (2016), pp. 68-78.

Yaghoubi, S., Noori, S., and Azaron, A., Lead time control in multi-class multi-stage assembly systems with finite capacity, Computers & Industrial Engineering, Vol. 66, No. 4 (2013), pp. 808-817.

Yagmahan, B., and Yenisey, M. M., A multi-objective ant colony system algorithm for flow shop scheduling problem, Expert Systems with Applications, Vol. 37, No. 2 (2010), pp. 1361-1368. DOI: 10.1016/j.eswa.2009.06.105.

Zarifoglu, E., Hasenbein, J. J., and Kutanoglu, E., Lot size management in the semiconductor industry: queueing analysis for cycle time optimization, IEEE Transactions on Semiconductor Manufacturing, Vol. 26, No. 1 (2013), pp. 92-99. DOI: 10.1109/TSM.2012.2230032.

Zhang, H. Y., Chen, Q. X., and Mao, N., System performance analysis of flexible flow shop with match processing constraint, International Journal of Production Research, Vol. 54, No. 20 (2016), pp. 6052-6070.