APEM journal

Advances in Production Engineering & Management
Volume 12 | Number 3 | September 2017 | pp 213–220
https://doi.org/10.14743/apem2017.3.252

ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Production lot-sizing decision making considering bottleneck drift in multi-stage manufacturing system

Zhou, F.L.[a,*], **Wang, X.**[a,b,*], **He, Y.D.**[a], **Goh, M.**[c]

[a]Department of Industrial Engineering, Chongqing University (CQU), Chongqing, P.R. China
[b]State Key Laboratory of Mechanical Transmission, Chongqing University (CQU), Chongqing, P.R. China
[c]Department of Analytics and Operations, NUS Business School and The Logistics Institute-Asia Pacific, National University of Singapore (NUS), Singapore

A B S T R A C T

Under a capacity constrained multi-product manufacturing system, the products are usually prepared and produced in lots. As a lot-sizing strategy is critical for effective production and high productivity, this encourages practical and research interest in the strategic batch sizing decision for a minimum procedure time in an order-to-delivery (OTD) operating environment. While the lot-sizing plan can be formed by studying the manufacturing parameters of the established bottleneck procedure, for a multi-stage manufacturing system, the bottleneck is not fixed and fluctuates with the production rate or batch size. This paper proposes a lot-sizing strategy to determine the optimal lot-size for each class of products taking bottleneck drifting into consideration. A queuing network analyser (QNA) method is employed to deal with the non-linear mixed integer programming model targeting at the total flow time minimization of the system. A practical case is presented and solved using the proposed method, and the results are validated with Flexsim, a simulation model.

A R T I C L E   I N F O

Keywords:
Manufacturing system
Multistage manufacturing system
Lot-sizing decision making
Lead time
Queuing network analyser (QNA)

**Corresponding authors:*
deepbreath@outlook.com
(Zhou, F.L.)
wx921@163.com
(Wang, X.)

Article history:
Received 18 January 2017
Revised 24 May 2017
Accepted 11 June 2017

1. Introduction

In the modern manufacturing era, enterprises are increasingly focusing on optimizing the total flow time under an order-to-delivery (OTD) environment [1]. With the proliferation in product variety, many products are usually processed and produced in the same production system with a view to improving the production efficiency and reducing cost. In the multi-product production system, the productivity and lead time of the manufacturing workshop is directly affected by the batch size of the product [2, 3]. With mass production, the productivity can be increased albeit with a longer lead time for the lot production; with low-volume production, the effect is reversed. In the literature, this phenomenon of large lot sizes will cause long lead times known as the batching effect. As the lot size decreases (the jobshop context), the lead time will also decrease, but once a minimal lot size is reached a further reduction in lot size will cause high traffic intensities resulting in longer lead times, called the saturation effect [4]. Therefore, to meet the shortest delivery lead time possible, it is often necessary to determine the right type and produce products in an appropriate quantity.

There are many models, in the literature, focusing on the complex relationship between batch, lead time, and work-in-process (WIP) [5-7]. Karmarkar [5] had studied the influences of batch (lot-size) on manufacturing lead times, and reported that the best lot sizes, high capacity utilization exact a high price in lead time and WIP. The relationship between batch size and lead-time variability was investigated by Kuik and Tielemans [6], who concluded that a minimization of the average queue delay or the average time in system performance measure, would not result in minimum lead-time variability. Vaughan [7] developed a comprehensive process lot sizing/order point model, and found that it was more efficient to adjust the safety stock and lead time than to change the shortage penalty parameter solely through adjusting the safety stock. Recently, some scholars argued that the lead time and productivity levels of the production system were determined by the bottleneck equipment, and the optimal processing of a single batch was studied and analysed [8-12]. Koo and Koh [8] proposed a batch decision optimization model that included a variety of product lines with the same preparation time by targeting maximum profit as the objective. Similarly, an extended model for optimizing the lot size in various production lines with different process and preparation times was formulated [9]. Liu [13] analysed the possibility of resources becoming the bottleneck through the bottleneck drift index, and calculated the optimal lot size and lead time of the bottleneck resource targeting the non-value added time and the difference in the manufacturing unit processing rate minimum as the optimization goal. However, they did not consider the bottleneck drifting caused by lot sizes.

In reality, in a relatively balanced production system with a large variety of products, the fluctuation of the product portfolio and the lot sizes can alter the load level of the resources, which might lead to bottleneck drifting. As a result, the lot size and lead time found in previous studies may not be optimal. Adacher and Cassandras [4] studied the optimal lot sizing problem with the example of two products and two processing operations, using the substitution method and the random comparison algorithm. Glock [14] solved the optimal lot size problem for a multi-stage manufacturing system by studying the relationship between the quantity and the total cost. Amy [15] proposed a mixed nonlinear integer programming model to solve the lot-sizing optimization problem with multiple suppliers in multiple periods considering quantity discounts, and a Genetic Algorithm (GA) is developed to minimize the total related cost. Therefore, this paper investigates the lot-sizing problem with bottleneck drifting. As the lot-sizing decision making is related to productivity, which influences the efficient output of the manufacturing system [1], this paper seeks to examine the lot-sizing problem for different experimental productivity scenarios.

In this paper, we apply QNA (queuing network analysis) [16] to establish a lot-size optimization model on a multi process production system, to decide the optimal lot sizes in the production of multiple products. The objective of the batch optimization model is to produce products with the shortest lead time according to the demand. QNA is an accurate queuing network analysis model developed by Bell Laboratories in the US [16]. QNA is used to analyze the queuing network by estimating the distribution and variation coefficient of each arrival process and each service time. QNA is suitable for solving complex queuing network problems, given its low computational complexity. Given the relationship between the processes of a multi operation system, we apply the QNA method to determine the waiting time of the tandem process queue, with the total process time as the shortest objective function, such that the changes caused by the bottleneck of the product batch will not affect the optimization objective.

The next section presents the studied problem and the assumptions thereof. In Section 3, the non-linear programing model considering bottleneck drifting is formulated and a four-step algorithm is designed by a traversal operation and implemented to deal with the lot-sizing optimization model. A practical case is presented in Section 4 to validate the proposed model. Lastly, we conclude with some remarks.

2. Problem description and assumptions

As shown in Fig. 1, the production line is a system with many work procedures. Each service station represents a processing or assembly operation. All kinds of products enter and leave independently with a minimum unit. Each batch of products follows a first-come-first-served (FCFS) queuing strategy.

The assumptions are as follows:

- The equipment utilization rate is less than 1, and productivity cannot exceed the production capacity;
- Each service station shows the same product capacity for the same products;
- The proportion of each product is known, and each class of products has a fixed manufacturing procedure.
- The processing times of various products are different and independent;
- Each lot of products arrives as a Poisson process;
- The preparation time before lot-size production, and the preparation time of each product is different and independent.

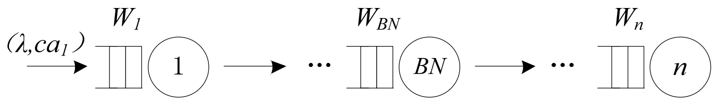

Fig. 1 Production system with multi-stage procedures

The parameters are denoted and described as follows:

Notation	Interpretation
F	Total processing time, units/period
$\bar{F}_j$	Average processing time in process *j*, units/period
$\bar{w}_j$	Average queuing time on equipment *j*, units/period
$\bar{s}_j$	Average processing time on equipment *j* , units/period
x_i	Productivity of product item *i*, (i = 1,2,3,...,m); $X = \sum_i x_i$
q_i	Lot size of product item *i*
r_i	Proportion of product item *i*, ($\sum_i r_i = 1$)
p_{ij}	Processing time on equipment *j* of product item *i*
τ_{ij}	Setup time on equipment *j* of product item *i*
s_{ij}	Total time required for product item *i* ($s_{ij} = q_i p_{ij} + \tau_{ij}$)
μ_{ij}	Service rate of a certain batch item *i* ($\mu_{ij} = 1/s_{ij}$)
ρ_j	Logistics intensity of equipment *j*
λ_i	Average number of arrivals of product *i* per time unit (arrival rate)
cs_j	Variation coefficient of processing time on equipment *j*
ca_j	Variation coefficient of product inter-arrival time on equipment *j*
cd_j	Variation coefficient of product inter-left time on equipment *j*

3. Model and algorithm

The non-linear mixed integer programming (MIP) model is formulated and constructed by minimizing the total flow time with the QNA technique. The algorithm procedures through the traversal operation are designed and implemented in this section.

3.1 Objective function

The optimization objective of the described problem is targeted as the total production time minimization considering production capacity and other common constraints. The model is formulated as follows.

$$\min F = \sum_{j=1}^{n} \bar{F}_j = \sum_{j=1}^{n} (\bar{w}_j + \bar{s}_j) \tag{1}$$

$$s.t. \sum_{i=1}^{m} (x_i p_{ij} + x_i \tau_{ij}/q_i) < 1 \tag{2}$$

$$u_j = \sum_i u_{ij}\,, i = 1,2, \dots, m \tag{3}$$

$$u_{ij} = x_i p_{ij} \tag{4}$$

$$v_j = \sum_i u_{ij}\,, i = 1,2, \dots, m \tag{5}$$

$$v_{ij} = x_i \tau_{ij}/q_i \tag{6}$$

$$q_i \le \bar{q_i}, q_i \epsilon Z^+ \tag{7}$$

$$x_i, q_i > 0 \tag{8}$$

where parameters u and v are the processing utilization rate and setup utilization rate respectively. Eq. 1 is the objective function for the total production time minimization. Eqs. 2 to 8 are the relevant constraints on productivity, order, and practical production.

3.2 Estimation of service time

In the production system, the arrival of a variety of products can be regarded as a Poisson process. It is reasonable to simulate the multi-process production system with the tandem queuing model, where $\bar{F_j}$ denotes the mean flow time of a product in processing j, which includes the waiting and processing times. The mean processing time of a unit product in process j is found from Eq. 9.

$$\bar{s}_j = \sum_i r_i \left(p_{ij} q_i + \tau_{ij}\right) \tag{9}$$

In a serial queuing system, the arrival processes are determined by the output of the previous process in addition to the first process. In this paper, the QNA method is used to estimate the arrival process variation coefficient of each node, which is regarded as an update process. Therefore, the queuing time in a steady state can be obtained [17]. Suresh and Whitt (1990) proposed the estimation method for the arrival process variation coefficient, and improved the accuracy of the estimation of the queuing time when the traffic intensity is 0.9 [18]. The improved coefficient of the arrival process is presented in Eq. 10.

$$\begin{aligned} cd_j^2 &= \left(1 - \rho_j^2\left(1 - \rho_{j+1}^{10}\right)\right) ca_j^2 + \rho_j^2\left(1 - \rho_{j+1}^{10}\right) cs_j^2 \\ ca_j^2 &= cd_{j-1}^2 \end{aligned} \tag{10}$$

The queuing time by the QNA technique is estimated from Eq. 11 and Eq. 12.

If $ca_j^2 < 1$

$$\bar{w}_j = \frac{\bar{s}_j (ca_j^2 + cs_j^2) \sum_i x_i (p_{ij} + \tau_{ij}/q_i)}{2[1 - \sum_i x_i (p_{ij} + x_i \tau_{ij}/q_i)]} \exp\left\{-\frac{2\left[1 - \sum_i x_i (p_{ij} + \tau_{ij}/q_i)\right]\left(1 - ca_j^2\right)}{3 \sum_i x_i (p_{ij} + \tau_{ij}/q_i)\left(ca_j^2 + cs_j^2\right)}\right\} \tag{11}$$

If $ca_j^2 \ge 1$

$$\bar{w}_j = \frac{\bar{s}_j p_j (ca_j^2 + cs_j^2)}{2(1 - p_j)} = \frac{\bar{s}_j (ca_j^2 + cs_j^2) \sum_i x_i (p_{ij} + \tau_{ij}/q_i)}{2[1 - \sum_i x_i (p_{ij} + x_i \tau_{ij}/q_i)]} \tag{12}$$

3.3 Productivity analysis

As the proportion of products r_i is known, the productivity x_i of an item i can be found from the total productivity X. The capacity constraints are illustrated in Eq. 13. From the upper bound on the lot-size of each class of products $\bar{q_i}$, we deduce the upper bound of X as shown in Eq. 14.

$$X \sum_i (r_i p_{ij} + r_i \tau_{ij}/q_i) < 1 \tag{13}$$

$$\bar{X} = \left\lfloor 1/\sum_i (r_i p_{ij} + r_i \tau_{ij}/q_i) \right\rfloor \tag{14}$$

where $\lfloor X \rfloor$ is the largest integer less than or equal to X.

3.4 Algorithm

To establish the optimized lot-sizing of each kind of product with different productivity levels, we targeted the total production time as the objective function and formulated a non-linear MIP model. To determine the optimal lot size, we introduce the algorithm for dealing with this issue. First, the upper bound on productivity is analysed based on the previous section. Then, the optimal solution is searched by a traversal operation. Fig. 2 shows the steps of the algorithm. As Fig. 2 shows, there are four steps in the proposed algorithm.

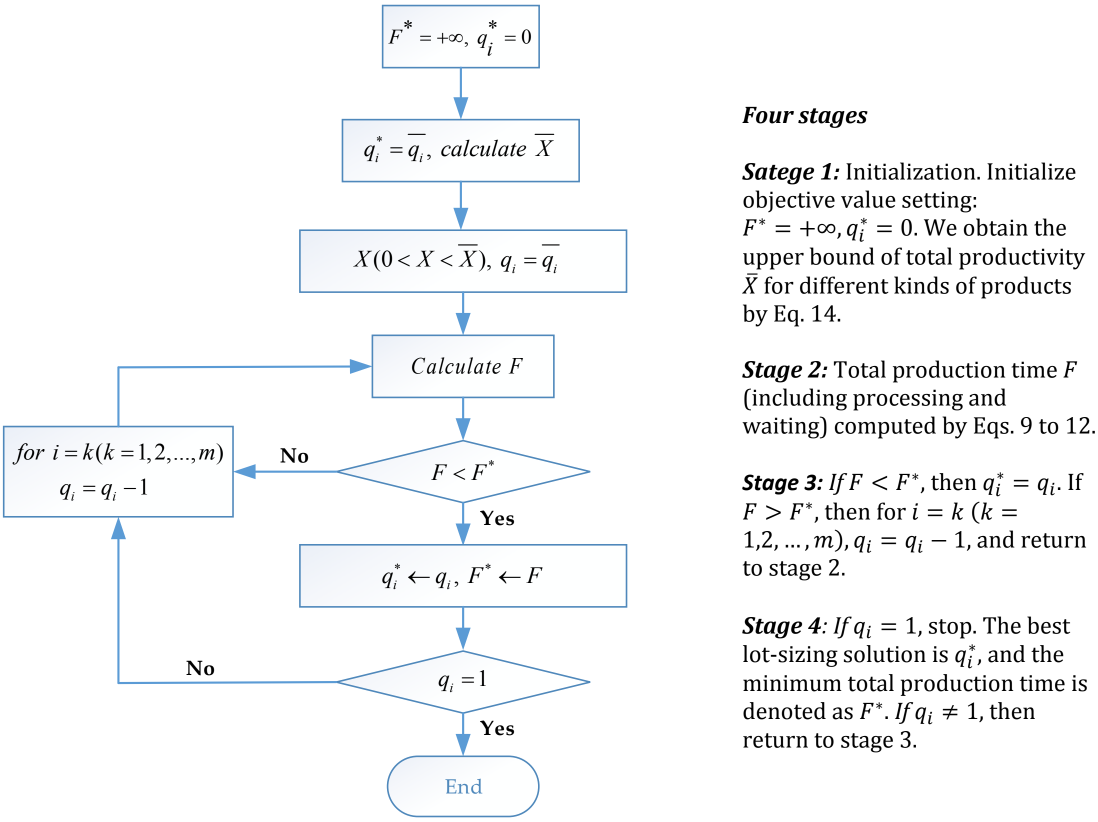

Fig. 2 Two-step optimization algorithm for multi-stage production system

4. Case study and simulation comparison

4.1 Case solutions

We consider a production system with five main procedures, and four classes of products need to be processed (Fig. 3). The information on the different parameters are presented in Table 1 such as processing time, upper bound of each class of products and product ratio. Suppose the arrival times of all products are regarded as Poisson, then $ca_1 = 1$, $cs_j^2 = 0.5$. The best solution of the lot-sizing for each class of products is calculated using the proposed algorithm procedure.

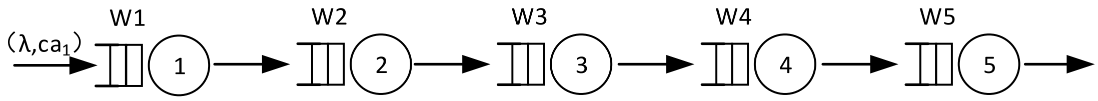

Fig. 3 Serial production system with five procedures

Table 1 Parameter values of four products in serial manufacturing system

Parameters		Product 1	Product 2	Product 3	Product 4
Waiting time	τ_{i1}	0.018	0.012	0.015	0.01
	τ_{i2}	0.01	0.02	0.02	0.01
	τ_{i3}	0.02	0.01	0.02	0.01
	τ_{i4}	0.015	0.005	0.015	0.01
	τ_{i5}	0.012	0.012	0.015	0.01
Processing time	p_{i1}	0.008	0.01	0.006	0.006
	p_{i2}	0.008	0.01	0.006	0.006
	p_{i3}	0.01	0.008	0.005	0.005
	p_{i4}	0.01	0.01	0.006	0.005
	p_{i5}	0.01	0.009	0.005	0.006
Product ratio	r_i	0.3	0.2	0.1	0.4
Upper bound	$\bar{q}_i$	30	30	50	50

From the algorithm, we obtain the best lot sizes under different productivity conditions. For instance, with a productivity of 90 %, the best solution of the four products is 7, 6, 10 and 7 units respectively, and with minimum total production time of 1.230 units. The fifth procedure is the bottleneck with the highest utilization rates of 84.02 %, as shown by Table 2. The results are illustrated in Table 2 for the various productivity situations.

From Table 2, the best lot-sizing solution for each class of products obtained from our algorithm suggests the following findings:

- As the productivity rate declines, the best lot-size for each class of products and the total production time decrease.
- In a relatively balanced manufacturing system, with a variation in lot-sizing, the bottleneck of the manufacturing system drifts as we imagined.
- As the productivity rate declines, the utilization rate of the machine fluctuates slightly, rather than declining. The reason for the utilization rate increases is that the waiting time has been increased with the much more preparation operations.

Table 2 Results for different productivity situations

Productivity (%)	Utilization rate of the machine (%)					Total time (days)	Lot size (piece)			
	Machine 1	Machine 2	Machine 3	Machine 4	Machine 5		Prod 1	Prod 2	Prod 3	Prod 4
90	83.64	83.40	81.56	82.18	84.02	1.230	7	6	10	7
89	82.85	82.67	80.85	81.41	83.24	1.173	7	6	9	7
88	81.92	81.74	79.94	80.50	82.30	1.121	7	6	9	7
87	82.11	81.43	80.27	80.52	82.11	1.073	6	6	9	7
86	81.35	80.74	79.59	79.77	81.35	1.028	6	6	8	7
85	81.89	81.74	80.04	79.94	81.89	0.983	6	5	8	6
84	80.93	80.78	79.10	79.00	80.93	0.941	6	5	8	6
83	79.96	79.82	78.16	78.05	79.96	0.903	6	5	8	6
82	79.22	79.15	77.51	77.33	79.22	0.869	6	5	7	6
81	79.71	78.99	78.18	77.61	79.23	0.835	5	5	7	6
80	79.79	79.09	78.29	77.71	79.31	0.803	5	5	7	5
79	78.80	78.10	77.31	76.74	78.32	0.772	5	5	7	5
78	78.74	78.67	77.11	76.16	78.27	0.743	5	4	7	5

4.2 Simulation verification

A simulation test is implemented to validate the effectiveness and validity of the proposed model. When the variation coefficient of the processing time is $cs_j^2 > 1$, we supposed the processing time complies with the H_2 distribution. When $cs_j^2 = 1$, we obtain the exponential distribution. When $cs_j^2 > 1$, the processing time can be treated as an Erlang distribution [18]. The manufacturing process is simulated by the Flexsim software with the original data information in Table 1, and the simulation is run for ten thousand days (10 times). Fig. 4 compares the simulated total production time and the proposed model.

From Fig. 4, the deviation result between the simulation model and programming model is about 12%. From Wu's research [19], when $cs^2 < 1$, the estimation variation of waiting time for multi-stage queuing system by the QNA method is 6.3%. As there are five procedures in the manufacturing system, the variation of the serial system is much more than the single machine (6.3%) due to the deviation accumulation. As the waiting time estimation method focuses on a single machine, the accuracy of the estimation of waiting time directly influences the performance of the programming model. As for the multi-stage serial manufacturing system with limited procedures, the best solution accuracy of the proposed model is satisfactory and tolerable based on the above comparison analysis.

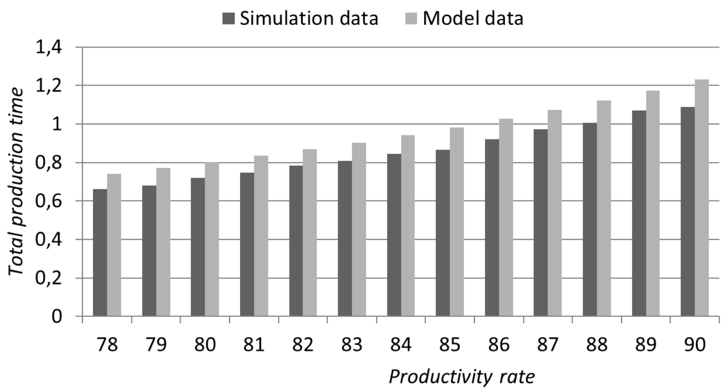

Fig. 4 Comparison of total time for simulation model and programming model

5. Conclusion

In this paper, the non-linear MIP model by the QNA method is proposed to deal with the lot-sizing problem for a multi-stage manufacturing system with multiple classes of products. Not only can the model be applied into a lot-sizing optimization problem of a manufacturing system with fixed bottlenecks, but also this model is applicable to a sensitive production system whose bottleneck fluctuates with the variation of the product combination and lot-sizing. The model is validated against a simulation model run by the *Flexsim* software, and it demonstrates excellent performance. However, from the comparison results, the model shows tolerable variation, which stimulates us to improve the accuracy of the model by revising the waiting time parameters. In future, the lot-sizing decision making problem which jointly considers bottleneck drift and unpredictable events (machine malfunction or failure, and stochastic occurrences) can be studied.

Acknowledgements

The authors are grateful to the anonymous referees for their valuable comments and constructive suggestions on our manuscript. The research is sponsored by the following scientific programmes: Research Fund for Doctoral Program of Higher Education of China (RFDP, 20130191110045), National Key Technology Research and Development Program of China (2015BAH46F01 and 2015BAF05B03). We also wish to thank the editorial team and the reviewers for the fast review process, and the Automotive Collaborative Innovation Centre (ACIC) in Chongqing University (CQU) and CA automotive Co. Ltd. for their assistance and support.

References

[1] Zhou, F.-L., Wang, X., Lin, Y. (2016). Production effectiveness-based system reliability calculation of serial manufacturing with checking machine, *Journal of Computers*, Vol. 27, No. 3, 201-211, doi: 10.3966/199115592016102703017.

[2] Potts, C.N., Van Wassenhove, L.N. (1992). Integrating scheduling with batching and lot-sizing: A review of algorithms and complexity, *Journal of the Operational Research Society*, Vol. 43, No. 5, 395-406, doi: 10.1057/jors.1992.66.

[3] Karimi, B., Fatemi Ghomi, S.M.T., Wilson, J.M. (2003). The capacitated lot sizing problem: A review of models and algorithms, *Omega*, Vol. 31, No. 5, 365-378, doi: 10.1016/s0305-0483(03)00059-8.

[4] Adacher, L., Cassandras, C.G. (2014). Lot size optimization in manufacturing systems: The surrogate method, *International Journal of Production Economics*, Vol. 155, 418-426, doi: 10.1016/j.ijpe.2013.07.026.

[5] Karmarkar, U.S. (1987). Lot sizes, lead times and in-process inventories, *Management Science*, Vol. 33, No. 3, 409-418, doi: 10.1287/mnsc.33.3.409.

[6] Kuik, R., Tielemans, P.F.J. (1999). Lead-time variability in a homogeneous queueing model of batching, *International Journal of Production Economics*, Vol. 59, No. 1-3, 435-441, doi: 10.1016/s0925-5273(98)00029-2.

[7] Vaughan, T.S. (2006). Lot size effects on process lead time, lead time demand, and safety stock, *International Journal of Production Economics*, Vol. 100, No. 1, 1-9, doi: 10.1016/j.ijpe.2004.09.012.

[8] Koo, P.-H., Bulfin, R., Koh, S.-G. (2007). Determination of batch size at a bottleneck machine in manufacturing systems, *International Journal of Production Research*, Vol. 45, No. 5, 1215-1231, doi: 10.1080/00207540600675793.

[9] Koo, P.-H., Koh, S.-G., Lee, W.-S. (2011). Simultaneous determination of lot size and production rate at capacity-constrained multiple-product systems, *Flexible Services and Manufacturing Journal*, Vol. 23, No. 3, 346-359, doi: 10.1007/s10696-011-9091-6.

[10] Kuik, R., Tielemans, P.F.J. (2004). Expected time in system analysis of a single-machine multi-item processing center, *European Journal of Operational Research*, Vol. 156, No. 2, 287-304, doi: 10.1016/s0377-2217(03)00022-5.

[11] Glock, C.H. (2010). Batch sizing with controllable production rates, *International Journal of Production Research*, Vol. 48, No. 20, 5925-5942, doi: 10.1080/00207540903170906.

[12] Maddah, B., Moussawi, L., Jaber, M.Y. (2010). Lot sizing with a Markov production process and imperfect items scrapped, *International Journal of Production Economics*, Vol. 124, No. 2, 340-347, doi: 10.1016/j.ijpe.2009.11.029.

[13] Liu, M., Ling, L., Tang, J. (2013). Research on production lot-size and lead time based on logistics shifting bottleneck in manufacturing shop, *Chinese Mechanical Engineering*, Vol. 24, No. 2, 220-225, doi: 10.3969/j.issn.1004-132X.2013.02.015.

[14] Glock, C.H. (2011). Batch sizing with controllable production rates in a multi-stage production system, *International Journal of Production Research*, Vol. 49, No. 20, 6017-6039, doi: 10.1080/00207543.2010.528058.

[15] Lee, A.H.I., Kang, H.-Y., Lai, C.-M., Hong, W.-Y. (2013). An integrated model for lot sizing with supplier selection and quantity discounts, *Applied Mathematical Modelling*, Vol. 37, No. 7, 4733-4746, doi: 10.1016/j.apm.2012.09.056.

[16] Hopp, W.J., Spearman, M.L., Chayet, S., Donohue, K.L., Gel, E.S. (2002). Using an optimized queueing network model to support wafer fab design, *IIE Transactions*, Vol. 34, No. 2, 119-130, doi: 10.1080/07408170208928855.

[17] Whitt, W. (1983). The queueing network analyzer, *The Bell System Technical Journal,* Vol. 62, No. 9, 2779-2815, doi: 10.1002/j.1538-7305.1983.tb03204.x.

[18] Suresh, S., Whitt, W. (1990). Arranging queues in series: A simulation experiment, *Management Science*, Vol. 36, No. 9, 1080-1091, doi: 10.1287/mnsc.36.9.1080.

[19] Wu, K., McGinnis, L. (2013). Interpolation approximations for queues in series, *IIE Transactions*, Vol. 45, No. 3, 273-290, doi: 10.1080/0740817x.2012.682699.