OPTIMIZATION OF PRODUCTION CAPACITY IN INTANGIBLE FLOW PRODUCTION SYSTEMS

Przemyslaw Korytkowski

Szczecin University of Technology
Zolnierska 49, 71-210 Szczecin, Poland, pkorytkowski@wi.ps.pl

Abstract: The goal of the paper is to present a model and algorithms for optimizing the production capacity of an intangible flow production system with the given network structure. The production capacity optimization problem is presented as a task of multi-criterion optimization with discrete, non-linear goal functions. The proposed solution enables finding an optimal configuration of the flow production system which is interpreted as a queuing network.

Keywords: modelling, optimization, capacity, queuing theory, simulation.

1. INTRODUCTION

In this paper we address the problem of determining capacity levels in an intangible production. Insufficient capacity can cause elongation of production cycle that leads to late deliveries and high levels of work-in-process. An excess capacity is a waste of expensive resources due to low utilization of equipment. The problem is complicated further by the stochastic behaviour of the production system, like unpredictable time of orders arrival. To take into account the random behaviour, manufacturing systems are often modelled and analyzed using the queuing network theory.

To provide a decision making tool that takes into account capacity costs and the trade-off between cycle times, performance and cooperation, we formulate a capacity optimization model that involves minimizing capacity costs while satisfying a set of systems constrains.

2. INTANGIBLE PRODUCTION

Development in means of communication and digital data exchange leads to establishment of a new kind of goods – intangible products. The increasing computerization of manufacturing processes caused the appearance of a new type of manufacturing – intangible production – in various areas, for example: banking, printing and publishing, entertainment, advertisement and many others. For this kind of production particularly the problem of resource capacity optimization arises because of the highly competitive market, close relationships with customers, high cost of equipment, etc.

An intangible product is a good that possesses a digital form (usually it’s a file), which is a direct result of an intellectual work of a man. An intangible product could be materialized through its recording or printing. The fact of materialization does not change its content, but the form of distribution.

Intangible production is an advanced manufacturing process, where input materials, semi products and final products are in a digital form. Manufacturing and distributing the work articles is realized with a broad use of telecommunication and computer networks. It enables establishing a production process with technological operations that are distributed geographically and a dynamic assignment of works between working posts. One of the main peculiarities of the intangible production is the fact that there are no mechanical operations, excluding final stages of the technological process, when for example the product is printed or burned on CDs. Intangible production is usually a multi-product one. Within one manufacturing system a wide nomenclature of final products is produced basing on the same communication infrastructure and working posts. The intangible production is a customer-oriented one. It gives a unique possibility to customize all final products according to the customer’s requirements if the company is capable of introducing it.

3. INTANGIBLE PRODUCTION NETWORK

The described properties of intangible production enable a manufacturing process to be characterized as a network of sets of working posts, later called production nodes. The network consists of production nodes and telecommunication channels that connect them and allows an exchange of semi-finished articles. The traffic inside the production network due to its multi-product character is a mixture of the traffic generated by the production process of each kind of a final product. The network of an intangible production may be represented as a digraph (fig. 1). The vertices of the digraph are the production nodes and the arcs are the workflows. The listed features of the production network give a possibility to circumscribe it as a network:

- with stochastic workflows of semi-finished articles,
- with unified work stations that could service several types of products,

Under the term of a production node we understand a set of uniform work stations performing the same technical operation with an input buffer that stores semi-articles before sending them to one of the working posts. The output capacity of a production node is the sum of productivities of each workstations being its part. The productivity of a workstation is the number of semi-finished articles that could be processed during a period of time. A production node can be interpreted as a queuing system with parallel servers.

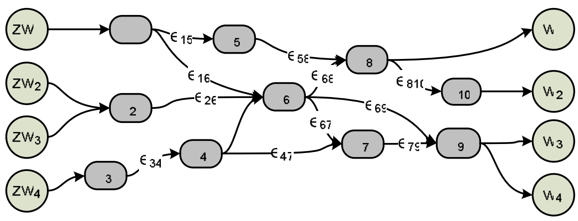

Fig. 1 Production network

In an intangible production buffers that are placed at each production node should be rather considered in terms of time than space or volume. In this case, articles and semi-finished products are in a digital form; it means they don't have dimensions and weight. Nowadays data carriers are cheap, so it's not a problem to store semi-finished products. The challenge is to produce faster and faster. In order to do so, we are considering a buffer as a source of a time loss that is wasted on waiting for servicing and we ought to avoid them. Waiting times in manufacturing are very common and probably they are inevitable. One of the methods to deal with this problem is to pass some work to another cooperating company that is capable of performing that particular technological operation. It is obvious that we cannot pass a lot of work to another company. The trade-off in this case is between the cycle time and the resource utilization level. The increase of cooperation will lower the cycle time; the decrease will hoist the utilization. The level of cooperation can be modelled by a limitation of the buffer's capacity. The buffer overflow indicates the need for cooperation.

That leads us to laying down the objectives of capacity planning in an intangible production:

1. minimize cycle time
2. maximize production system's utilization
3. minimize level of cooperation

The objectives depend on the number of servers and buffers' capacity in an opposite way: while the number of servers increases the cycle time is falling and the idle time of servers rises. In turn, while the buffer capacity increases the cycle time is rising and the costs of cooperation are reduced.

In addition a constrain on the network's construction costs which cannot exceed the given budget is introduced.

Before discussing the capacity planning problem further, we need to define some queuing network terminology. A manufacturing system can be considered as a queuing network that consists of production nodes where works are routed from one node to another to obtain the necessary processing. So the production processes is an open queuing network with deterministic routing.

The formulated task, if considered as Marcovian processes of arrival and servicing, can be easily solved using the Burke's Theorem and decomposition of the network into separate production nodes. However, the problem becomes more complex when the arrival and servicing processes are given in a general form. In this case, Burke's theory is not working any longer and the optimization task requires a combined usage of analytical modelling and simulation.

4. OPTIMIZATION TASK

The given analysis of intangible flow production system allows us to state that the production system, referring to queuing theory, is an open point-to-point connected network with deterministic links, multi-class customers with infinite customer population and with different size of input buffers. For this kind of production network there is no analytical approach that enables analyzing it. For other network configurations, for example for star configuration, some approaches have already been developed, which enable optimizing production capacity solely on the basis of analytical methods, like queuing theory (Dolgui *et al.* 2003; Zaikin *et al.* 2002a; Zaikin *et al.* 2002b).

Analyses in case of the described above production network are usually done using discrete-event simulation, which allows exploring properties of the system being modelled without complete knowledge about all its complicated rules. With regard to the stochastic character of factors influencing the planned production capacity it is necessary to take into account the time factor during analysis. The intangible production is a discrete one, that means it is possible to identify discrete components.

It means that the simulation model has to be dynamic and discrete. Those models could be build with usage of a wide range of simulation packages like: Arena, AutoMod, eM-Plant, Extend, ProModel (Swain 2003).

Conducting experiments allows gaining data about the state of the modelled production system in predetermined conditions. Those data can be a basis for searching for an optimal production system configuration referring to the criterion function.

To do so it is possible to use optimization software, like OptQuest from OptTek. This, in an automated manner, is scanning an area of eligible solutions. The main disadvantage of this approach is the fact that software like OptQuest is using general purpose optimization algorithms, like: simulated annealing, tabu search etc.

An alternative is to take advantage of knowledge about nature of intangible production systems. Additional information about production systems comes from the fact that in every production system it is possible to find a bottleneck, a part of the system that is the slowest, and to examine the properties of a production node as an independent one using analytical or simulation methods.

It is well known that in order to speed up a production system it is necessary to extend the bottleneck or to reorganize production in such a fashion that the bottleneck is working with the highest possible utilization. The optimization algorithm presented in details below is basing on this property. In the first loop of the algorithm it deals with a current bottleneck, it means that we should expect the biggest decrease of the criterion function value after the first bottleneck optimization.

Optimization of one production node is much simpler than optimization of the whole production system. In case when it is possible to employ queuing theory, an optimal solution can be found very quickly, faster than the duration of one simulation experiment. When simulation is needed it is anyway much easier to optimize parameters of one production node knowing which one is the bottleneck.

5. OPTIMIZATION MODEL

The formalization of the optimization model is given as follows:

5.1 Input parameters

1. Production system structure may be represented as a digraph $G(R,E)$. Vertexes of the digraph $R=\{r_1,r_2,r_3,...,r_k\}$ are the production nodes and arcs of the digraph $E=\{e_1,e_2,e_3,...,e_l\}$ where $e_l=(r_i,r_j)$, $l=\overline{1,L}$, are the telecommunication links (channels) connecting them.
2. Customers' orders stream for intangible product is given as a vector $ZW=[zw_1,zw_2,zw_3,...,zw_n]^T$. Those orders concern production of $W=\{w_1,w_2,w_3,...,w_n\}$ final products.
3. Each production node r_i consists of a set of one-type production posts $s_i, i=\overline{1,l}$, performing some kind of technological operation $o_i, i=\overline{1,k}$. The following notation is used: $\lambda_i, i=\overline{1,l}$ – incoming stream intensity, $\mu_i, i=\overline{1,l}$ – productivity of production post, $\rho_i, i=\overline{1,l}$ – production posts' utilization.
4. Each production process $p_i, i=\overline{1,n}$ is carried through a route of production nodes, representing a chain of technological operations $p_i=\{o_{i1},o_{i2},o_{i3},...,o_{ik}\}$. For each technological operation its duration $\tau_{ij}, i=\overline{1,k}, j=\overline{1,n}$ is given.
5. Each production node is also described by the following parameters: $\alpha_i, i=\overline{1,k}$ – production post's labour cost coefficient, $d_i^u, i=\overline{1,k}$ – production post cancellation cost, $d_i^z, i=\overline{1,k}$ – purchase cost of one production post, $\overline{k}_i, i=\overline{1,k}$ – task execution cost in frames of cooperation, $kw_i^+, i=\overline{1,k}$ – stream of incoming tasks into each production node in frames of cooperation, $kw_i^-, i=\overline{1,k}$ – streams of out coming tasks from each production node in frames of cooperation, $rw_i, i=\overline{1,k}$ – streams of task requiring rework.

5.2 Decision variables

Each production node can be represented as a multi-server queuing system. Such structure permits realising the parallel manufacturing process for several tasks simultaneously. Therefore, we have to define the number of servers, operating in parallel for each servicing node and input buffer capacity. $N=[N_1,N_2,N_3,...,N_k]^T$ is the vector of decision variables, where N_i is the number of uniform parallel servers for the r_i production node. $m=[m_1,m_2,m_3,...,m_k]^T$ is the second vector of decision variables, where m_i is the input buffer capacity of r_i production node.

5.3 Criterion functions

In case of intangible flow production, lead time is the time that is counted from order reception up to the moment when all final products required by the order are ready. Further cycle time will be analyzed, which is lead time minus duration of administrative issues, like order's processing, scheduling etc. Cycle time starts with the beginning of the first technological operation on production node and ends when the last item from the order being processed at the last production node of technological chain is finished.

It is quite easy to calculate the cycle time for a production line, where only one product is manufactured and the technological chain is linear.

The cycle time is just the sum of times spent by tasks in each production node. Far more difficult is to find cycle time in case when a technological chain is nonlinear, i.e. when some technological operations are carried out in parallel at different production nodes or when in production system more than one class of final products is manufactured. This is the case of the studied system.

In general, it could be said that cycle time depends on the range of final products (i.e. their technological chains characteristics) and the structure of production system (i.e. types of production nodes) $T(N,m) = F(W(N,m), G(R,E))$. It is known that reduction of a technological operation duration does not lead to an increase in the value of $T(N,m)$, it leads do shortening the cycle time.

For that reason a rough estimate of cycle time was developed, which assumes that changes in cycle time could be approximated by the following function:

$$T(N,m) \approx \sum_{i}^{n} t_i(N_i, m_i), \tag{1}$$

where $t_i(N_i, m_i)$ is duration of technological operation on r_i production node, which is easy to calculate.

Utilization could be considered at the whole production system and also production node levels. We will not consider production post level because each production node consists of identical production posts and tasks are steadily spread. So the second criterion function in this case will have the following form:

$$U(N,m) = \sum_{i}^{n} u_i(N_i, m_i) = N_i(1-\rho_i)\alpha_i, \tag{2}$$

where ρ_i - is a r_i production node utilization and depends on N_i and m_i.

Cooperation level will be considered as a probability of task execution necessity outside home production system. This could happen in every production node in case of task arrival when the input buffer is full. For each production node the probability P_{bl} of this kind of situation can be calculated and depends on N_i and m_i. This factor will be multiplied by the cost of task execution in production node $\bar{k}_i$, which doesn't belong to home production system and input stream intensity λ_i. Outside task execution happens only in one production node and afterwards the task returns to its home system. The third criterion function therefore is as follows:

$$K(N,m) = \sum_{i=1}^{k} P_{bi}(N,m)\lambda_i \bar{k}_i \ . \tag{3}$$

5.4 Constrain

Change in production capacity is connected with the change in production system configuration, i.e. increasing or decreasing the number of production posts working in parallel within each production node. Those changes involve some kind of costs. Production capacity optimization requires considering the total cost of reconfiguration. The decision-maker never has an unlimited budget for this kind of purpose and because of that the budget has to be treated as constrain in this problem. Production system reconfiguration cost can be calculated using the formula:

$$D = \sum_{i=1}^{k} (| N_i - N_i' | \cdot (\delta \cdot d_i^z + (1-\delta) \cdot d_i^u)), \tag{4}$$

where:

$$\delta = \begin{cases} 1 & for\ N_i \le N_i' \\ 0 & in\ other\ cases \end{cases},$$

N_i' - number of production posts after reconfiguration in r_i production node,

N_i - number of production posts before reconfiguration in r_i production node.

Those three criterion functions are in a conflict relation. An increase of the value of the second criterion function leads to an elongation of cycle time and an increase of task execution in cooperation probability. Shortening the cycle time entails a decrease of production system utilization or an increase of cooperation. Reduction of cooperation will increase the production system utilization and cause elongation of cycle time.

In order to find the optimal production system configuration the weighted sum method will be used (Ehrgott, Wiecek 2005). It is one of the most widely spread approaches. For the above-stated problem the global criterion function will be:

$$G(N,m) = w_1 \cdot U(N,m) + w_2 \cdot T(N,m) + w_3 \cdot K(N,m), \tag{5}$$

where w_1, w_2, w_3 are weighting coefficients.

Those coefficients have to be fixed by the decision-maker, basing on her/his knowledge about the production system and its environment. This is usually a difficult assignment and for this purpose the Analytic Hierarchy Process (AHP) method is applied (Saaty 2005).

The AHP is a theory of relative measurement, with this approach a scale of priorities is derived from pair wise comparison measurements that are made with judgements using numerical values taken from the AHP absolute fundamental scale of 1-9. The AHP method consists of three steps (Korpela *et al.* 2001):

- decomposition of a complex multi-criterion problem into hierarchy, where on each level there are some criterions that can be further decomposed into next hierarchy level,
- usage of goals measuring methodology on every level of hierarchy in order to determine their reciprocal importance,
- priorities synthesis.

The main advantage of the AHP method is the possibility to apply it to problems that are in different units, like in case of intangible production capacity optimization where the criterions are in units of time and money.

6. OPTIMIZATION ALGORITHM

The formulated problem is a multi-objective optimization one, with discrete and non-linear criterion functions. The developed optimization algorithm bases on using discrete-event simulation and queuing theory. The optimization algorithm is shown in fig. 2.

1. Arrangement of the initial configuration. The input data could come from an existing production system or they could be estimated. We can consider the processes incoming into each production node as independent ones. This gives us the opportunity to set up initial values of the decision variables for each servicing node independently with the following assumptions: the number of servers for each servicing node is set using the condition of balance between rates of the incoming workflow process and productivity of the servicing node, input buffer capacity can be set as infinite. At this stage all input parameters have to be fixed and loaded into a previously prepared simulation model.

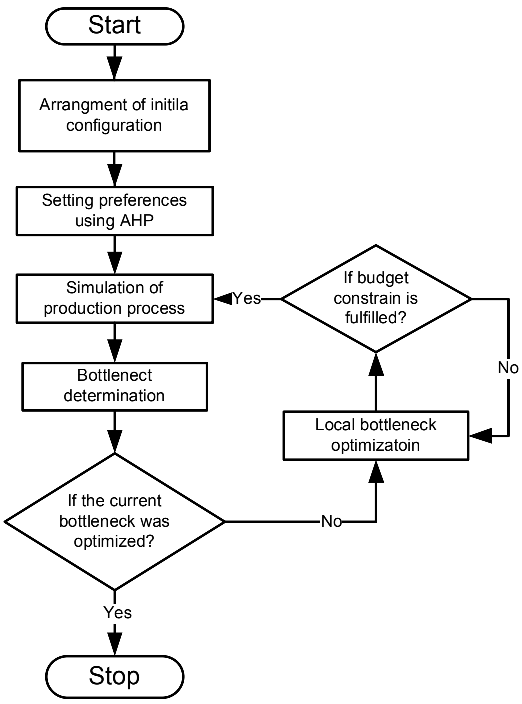

Fig. 2 Optimization algorithm

2. Setting preferences using AHP. At the second stage the AHP method will be applied to take into account the decision-maker's opinion. As a result a set of waging coefficients will be received.

3. Simulation of production process. The aim of this stage is to collect data about the behaviour of the modelled production system. In order to acquire the data it is necessary to conduct a simulation experiment. The warm-up period, numbers of replications and replication length have to be set up, methods of finding those parameters are well known (Law and Kelton, 2000). As a result we receive the following data: average cycle times, utilization and probability of cooperation for every production node.

4. Bottleneck determination. For each production node the criterion function estimation is calculated:

$$g_i(N,m) = w_1 \cdot u_i(N_i,m_i) + w_2 \cdot t_i(N_i,m_i) + + w_3 \cdot k_i(N_i,m_i), \quad (6)$$

where $G(N,M) \approx \sum_{i=1}^{k} g_i(N_i,m_i)$. The estimation is needed because there is no general formula that binds the whole production system cycle time together with its configuration and cycle times of every production node. A production node with the highest value of g_i will be considered as a bottleneck.

5. Local bottleneck optimization. The goal of this stage is to reconfigure the bottleneck production node structure (number of production posts and input buffer capacity) according to the given criterion. This can be done using criterion estimation $g_i(N_i,m_i)$. The reconfiguration problem thus can be stated as an optimization one with two discrete decision parameters. Criterion function components $(t_i(N_i,m_i), u_i(N_i,m_i), k_i(N_i,m_i))$ can be calculated basing on the results obtained from analytical or simulation approaches. In the analytical approach the queuing theory apparatus was used. The queuing theory can be applied in case when the input stream and the technical operation in production node are Markov, Erlang or deterministic (Gross and Harris, 1998). In other cases only using by simulation it is possible to determine the values of criterion function components of a particular production node configuration.

To find an optimal configuration of the bottleneck production node dichotomy algorithm was applied. At first, the number of production posts will be set, because the conducted experiments show that it has a bigger influence on criterion function than the input buffer capacity. Afterwards, the input buffer capacity level will be determined again using dichotomy.

An upper bound for the number of production posts within a production node depends on the budget constrain D. The reconfiguration cost cannot excess the budget. Thus, $N_{\max} \leq N_i + N_n$, where N_n is the number of posts that can be added to a node within the budget, $d_i^z \cdot N_n \leq D_j$. D_j is the remaining part of the budget D after j loops of the algorithm, while each loop decreases the available budget.

6. Stop condition. Decision about continuing the optimization process should be made at each iteration of the optimization algorithm. In order to stop the algorithm it has to be determined if further optimization will minimise the value of the criterion function $G(N,m)$. It can be done by checking whether the currently found bottleneck node r_i was optimized at one of the previous iterations or the available budget has already been used up.

7. ILLUSTRATIVE EXAMPLE

The model and algorithm presented above were tested on a prepress production system. Prepress is the first step in the printing production process, followed by press and post press (Kipphan, 2001). The example was chose not accidentally; prepress is the most complicated stage of the printing process from the management point of view. Production capacity optimization is especially important, due to the cost of production resources and very short lead time of orders execution.

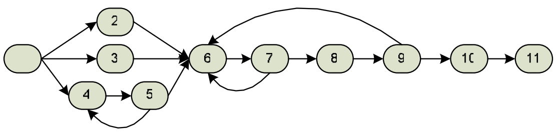

Fig. 3 Prepress production system

In prepress 11 production nodes were distinguished, as it is presented in figure 3. A detailed description of the prepress process can be found in (Kipphan, 2001). The analysed production system is designed to work with different kind of publications, from small leaflets, through books to illustrated albums. During the production process each order is decomposed into three sets of pages: text, grey-scale graphics and colour. For each set of pages within the production system exists a chain of technological operations.

The decision-maker decided that for him the cycle time is a little bit more important than production system utilization, cycle time is strongly more important than the level of cooperation and production system utilization is a little bit more important than the level of cooperation. It enables calculating the waging coefficients w_1, w_2, w_3. In the end, the criterion function had the following form:

$$G(N,m) = 0.26 \cdot U(N,m) + 0.633 \cdot T(N,m) + \\ + 0.106 \cdot K(N,m) \quad (7)$$

In the first loop of the algorithm r_5 production node was found a bottleneck one. Analysis of the input and technological operation showed that local optimization was possible only using simulation. As a result, it was necessary to add to the r_5 node one production post.

In the next iterations r_{10} and r_7 were found bottlenecks. Looking at the value of the criterion function it shows, that it decreased by 14% after optimization. As a result, cycle time dropped from almost 115 hours down to almost 70 hours, what was especially awaited by the decision-maker. A more detailed results are presented in (Korytkowski, 2005).

8. CONCLUSIONS

The problem of production capacity optimization is formulated as a task of optimization of a queuing model parameters. The criterion function depends on stochastic variables, such as distribution law of arriving orders, distribution law of technological operations etc. The developed approach is based on coupling analytical and simulation models. Bottleneck nodes are searched for through simulation and analytical or simulation method is used for local optimization of the current bottleneck node.

During the research it turned out that it's possible to relax assumptions about uniform work stations within a production node. It means that the production node could consist of work stations with different productivity, but performing the same technical operation. The next step would be to consider a model with more flexible production capacity changes.

REFERENCES

Dolgui, A., O. Zaikin, E. Kushtina, P. Korytkowski (2003) Modelling and optimisation of production nodes performance for Automatic Management system in distributed non-material manufacturing. *Automation and Remote Control*, **64**, No. 9, 1501-1506.

Ehrgott, M., M.M. Wiecek (2005) Multiobjective programming. In (J. Figueira et al. (Ed.)) *Multiple criteria decision analysis*, 667-722, Springer, New York.

Gross, D., C.M. Harris (1998) *Fundamentals of Queuing Theory*, 3 ed., Wiley, New York.

Kipphan, H. (Ed.) (2001) *Handbook of Print Media*, Springer, Heidelberg.

Korpela, J., A. Lehmusvaara, M. Tuominen (2001) An analytic approach to supply chain development. *International Journal of Production Economics,* **71**, 145-155.

Korytkowski, P. (2005) *Modeling and optimization of production capacity in intangible flow production systems*, PhD dissertation, Szczecin University of Technology (in Polish).

Law, A.M., W.D. Kelton (2000) *Simulation Modeling and Analysis*, McGraw-Hill, Boston.

Saaty, T.L. (2005) The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making. In (J. Figueira et al. (Ed.)) *Multiple Criteria Decision Analysis*, 345-405, Springer, New York.

Swain, J.J. (2003) Simulation Software Survey, *OR/MS Today*, **30**, No. 4, 46-49.

Zaikin, O., A. Dolgui, P. Korytkowski (2002a) Optimization of Resource Allocation in Distributed Production Networks, *Lecture Notes in Artificial Intelligence*, **2296**, 322-332.

Zaikin, O., Dolgui, A., P. Korytkowski (2002b) Modeling and performance optimization of a distributed production system. In: *Information Control Problems in Manufacturing 2001*, (P. Kopacek et al. (Ed.)), 98-104, Elsevier.