Automation and Remote Control, Vol. 64, No. 9, 2003, pp. 1501–1506. Translated from Avtomatika i Telemekhanika, No. 9, 2003, pp. 155–161.

AUTOMATIC CONTROL SYSTEMS

Modeling and Optimization of the Throughput of the Processing Nodes in Computer-aided Control Systems of Distributed Production of Printer Matter

A. Dolgii*, O. Zaikin, E. Kushtina**, and P. Koritkovskii****

** University of Technology, Troyes, France*
*** Technical University, Szczecin, Poland*

Received February 13, 2003

Abstract—A computer-aided control system of distributed production of printed matter which is interpreted as an open queuing network was considered. Optimal resource allocation to the processing network nodes is one of the key problems in this class of systems. An algorithm based on a combination of the branch-and-bound method and simulation was proposed for determining the optimal solution. The branch-and-bound method enables one to establish the direction of search, and simulation is used to verify each step of search. High convergence of the algorithm was corroborated by numerical experiments.

1. INTRODUCTION

Allocation of resources is the most frequently encountered problem in the technical support of the distributed manufacture, corporate, and telecommunication networks. Despite different physical nature of their processes, these systems can be studied using a unique approach based on the models of queuing networks and methods of discrete optimization [1].

Broad experience has been gained in using queuing models of the distributed manufacturing systems. Various aspects of their analysis were discussed in the fundamental publications [2–4]. Methods of simulation-based discrete optimization were considered in [5–8].

Development of the communication technologies in the field of corporate networks, on the one hand, and wide introduction of the information technologies into the modern printing systems, on the other hand, gave rise to a new kind of production, the so-called distributed publishing and printing complexes (PPC) that are characterized by the following:

(1) investments in the PPC's are comparable with those in other high-tech industries;

(2) global scale of the (interregional, interstate, transnational) network infrastructure;

(3) beginning from the 1990's, the number of workplaces in the field of publishing and printing services grows at a higher rate than in other industries;

(4) wide product mix, the raw materials, half-finished and finished products being usually represented in electronic form;

(5) high-performance expensive equipment at the processing nodes (both at the pre-printing and printing stages); and

(6) multiple variants of relations between the processing nodes manufacturing one kind of product.

These characteristics allow one to specify the PPC's as a new kind of production which features (1) stochasticity of the repeated manufacturing processes, (2) merging of alternative processing

0005-1179/03/6409-1501$25.00

nodes with different throughputs into a unique network structure, and (3) the possibility of servicing more than one input flow in a single node.

The aforementioned gives ground to modeling this manufacturing process as an open queuing network with probabilistic transitions between the nodes of different types.

2. FORMULATION OF THE PROBLEM

The network consisting of a central node and several remote (local) printing nodes is a standard structural element defining the PPC architecture. The local nodes are equipped with relatively inexpensive offset machines, whereas the central node has expensive universal digital machines. In the ordinary mode, the local nodes execute the user requests. However, if the local node is completely loaded, the orders can be transmitted to the central node through the telecommunication facilities (network). It is clear that the expensive universal high-performance equipment makes the speed of request processing and its cost higher than in the local nodes.

Intensity of request arrival to each local node and the mean times of their servicing by each unit of equipment are known from experience. It is required to determine the number of offset machines at each local node and the number of digital machines at the central node that provide the minimum total cost of processing all requests with regard for different cost and productivity of both kinds of printing machines.

This structure is representable in terms of the queuing theory as an open queuing system with independent flows and limited (or unlimited) queue. This network can be described by the oriented graph $G = \{S, F\}$ whose set of vertices $\{S\}$ represents the processing nodes and that of arcs $\{F\}$, the stochastic flows of requests between the local and central nodes. For the above organization of request flows, the graph G has star-shaped configuration. Each processing node has a set of parallel machines (servers).

Let us formulate the conditions for request servicing in this queuing system configuration:

(1) in the ordinary mode, all arriving requests are processed by dedicated servers of the corresponding local node;

(2) if all dedicated servers are occupied, then the arriving request is routed to the input buffer of the corresponding local node which has a limited number of waiting places;

(3) if the input buffer of the local node is full, then the arriving request is sent for servicing to the central node;

(4) if all universal servers of the central node are occupied, then the request is sent to the common infinite-capacity buffer of the entire queuing system; lack of constraints on the buffer size at the central node is due to the fact that in the PPC its role can be performed by the mail server;

(5) as some fixed repeating time passes, the request is again sent for servicing;

(6) the central and local nodes process all requests according to the FIFO discipline.

Therefore, one has to determine the number of dedicated local servers and universal central servers for the given parameters of request arrival and servicing in the queuing system with regard for the differences in throughputs, costs, and penalties for idling of the servers of both types. The total number of servers in the queuing system will be referred to below as the volume resource. The aforementioned suggests the following model of resource allocation in the open queuing system.

Source data:

$S = \{s\}$ is the set of processing nodes,

$F = \{f_s\}$ is the set of request (job) flows arriving to the local nodes,

$\Pi(f) = \{\Psi_f(n,t), \lambda_f\}$ are the parameters of the flow f, where $\Psi_f(n,t)$ is the law of distribution of the arriving requests, λ_f is the intensity of the arriving requests,

τ_1 and τ_2 are the times of processing one request, respectively, at the local and central nodes.

Controlled parameters:

m_s is the number of dedicated servers assigned to each local node $s \in S$,

m_c is the number of universal servers assigned to the central node.

Optimization criterion.

The objective function comprises the following three components:

(1) the general time of processing the entire request flow

$$T_\Sigma = \sum_f \lambda_f \widetilde{\tau}_f, \tag{1}$$

(2) costs due to server idling at the local and central nodes,

$$U_\Sigma = \left(\sum_S \frac{\lambda'_s}{m_s \mu_1} \delta_1 + \frac{\lambda'_c}{m_c \mu_2} \delta_2 \right), \tag{2}$$

(3) total cost of equipment at the local and central nodes

$$C_\Sigma = \left(m_c \gamma_1 + \sum_S m_s \gamma_2 \right), \tag{3}$$

where $\widetilde{\tau}_f$ is the mean efficient time of processing one request (job) from the flow f (with regard for the waiting time and denied servicing); λ'_s and λ'_c are the efficient intensities of request arrivals to the local and central processing nodes, respectively; μ_1 and μ_2, $\mu_1 < \mu_2$, are the intensities of request processing by a dedicated server and the universal server, respectively; δ_1 and δ_2, $\delta_1 < \delta_2$, are the costs caused by idling of a dedicated server and the universal server, respectively; and γ_1 and γ_2, $\gamma_1 < \gamma_2$, are the costs of one dedicated server and the universal server, respectively. The coefficients γ and δ in (2) and (3) are reduced to the time unit.

By introducing the weight coefficients α and β, the components of the objective function can be united into the unique criterion $CR = \alpha T_\Sigma + \beta U_\Sigma + C_\Sigma$. The ratio of α and β is established from the expert estimate of the manager. The components of the objective function depend differently on the chosen control parameters of m_s and m_c. As the number m_c of universal servers at the central node grows, the total time of processing decreases, whereas the idling costs increase. On the contrary, as the number m_s, $s = \overline{1, S}$, of the dedicated servers at the local nodes increases, the total time of executing all jobs increases, whereas the idling costs decrease. Therefore, it is possible to determine the control parameters that minimize the objective function

$$CR = \alpha T_\Sigma + \beta U_\Sigma + C_\Sigma = \min. \tag{4}$$

3. METHOD OF SOLUTION

3.1. Analytical Model

The above problem does not yield to analytical solution in the general form. However, for systems with Markovian arrivals and processing of jobs, such a solution is feasible [9]. Let us consider the simplest queuing system for the case under study that consists of one central node (CN) and one local node (LN). The above conditions for request servicing allow one to apply the reasoning below concerning an arbitrary "LN–CN" pair to the entire network of the form under consideration. To simplify the model in point, we assume that the local node has a zero-capacity input buffer.

We introduce the following notation:

m_s and m_c are the numbers of dedicated and universal servers, respectively;

τ_1 and τ_2 are the mean times of request processing by the dedicated and universal servers, respectively;

τ_W is the established repeating time, that is, the time after which the request is again sent from the buffer for processing; obviously, $\tau_W \gg \tau_1 > \tau_2$,

P_1^S is the probability of processing the request arriving to the local node (at the instant of arrival, at least one dedicated server of the local node is free),

P_2^S is the probability of processing the request arriving to the central node (at the instant of arrival, all dedicated servers of the local node are occupied and at least one universal server of the central node is free),

P_3^S is the probability of request arrival to the waiting queue (all dedicated and universal servers are occupied).

Obviously, it follows from the normalization condition that

$$P_1^S + P_2^S + P_3^S = 1. \tag{5}$$

With the conventional notation, the mean time $\widetilde{\tau}_f$ of processing one request arriving to the local N_s is as follows:

$$\widetilde{\tau}_f = \left(\tau_1 P_1^S + \tau_2 P_2^S + \tau_3 P_3^S\right) \frac{1}{1 - P_3^S}. \tag{6}$$

The probabilities P_1^S, P_2^S, and P_3^S can be established from the following assumptions.

According to Kendall, a queuing system with m parallel servers and no input buffer can be considered as a multiple-line Erlangian system with losses $\{M/M/m/m\}$. For this kind of queuing systems, the probability of blocking P_b follows

$$P_b = P_m = \frac{\dfrac{\rho^m}{m!}}{\sum\limits_{k=0}^{m} \dfrac{\rho^k}{k!}} = P_0 \frac{\rho^m}{m!}, \tag{7}$$

where P_0 is the probability of free (nonoccupied) state of the system

$$P_0 = \left[\sum_{k=0}^{m} \frac{\rho^k}{k!}\right]^{-1}.$$

Expressions (5) and (7) allow one to obtain the following formulas for the probabilities P_1^S, P_2^S, and P_3^S:

$$P_1^S = 1 - P_m = 1 - \frac{\dfrac{\rho_s^{m_s}}{m_s!}}{\sum\limits_{k=0}^{m_s} \dfrac{\rho_s^k}{k!}}, \tag{8}$$

$$P_2^S = (1 - P_1^S)\left(1 - \frac{\dfrac{\rho_c^{m_c}}{m_c!}}{\sum\limits_{k=0}^{m_c} \dfrac{\rho_c^k}{k!}}\right), \tag{9}$$

$$P_3^S = 1 - P_2^S - P_1^S, \tag{10}$$

where $\rho_c = \lambda'_c/\mu_2$ and $\rho_s = \lambda'_s/\mu_1$, $s = \overline{1,S}$, are the traffic intensities for the central and local nodes, respectively.

By analogy, the efficient intensities λ'_s and λ'_c of request arrivals to the local and central nodes are, respectively, as follows:

$$\lambda'_s = P_1^S \lambda_s, \quad \lambda'_c = \sum_S P_2^S \lambda_s. \tag{11}$$

The values $\widetilde{\tau}_f$, λ'_s, and λ'_c from (6)–(11) can be used in (1)–(4) to calculate the objective function.

3.2. Optimization Algorithm

The algorithm of optimal solution is based on the branch-and-bound method and lies in performing successive analytical calculations and simulation at each step. The proposed analytical model is used to determine the primary—that is, lower—bound of the objective function and choose the direction of search. Simulation is used to verify the analytical calculations and complete the algorithm.

The algorithm comprises three components: construction of the variant tree, determination of the lower bound, and branching procedure. Search of solution begins with determining the initial value of the distributed volume resource defined by the given number of hypothetical universal servers installed at the central node. This number is determined from the relation between the total intensity of the input flows and throughput of the universal servers. The algorithm reallocates the volume resource from the central node to the local ones by moving a server from the central node to a local one and changing simultaneously its status from universal to dedicated, which corresponds to its lower cost and throughput. The algorithm steps are as follows:

(1) determination of the volume resource;

(2) calculation of the value of the criterion function for each "LN–CN" pair;

(3) calculation of the lower bound for the current step;

(4) organization of branching according to the lower bound. If the lower bound is smaller than the best value (record) obtained at the preceding steps, then branching goes on; otherwise, return to the preceding vertex of the variant tree;

(5) simulation to specify the criterion in the solution obtained;

(6) the algorithm is completed if $CR_i > CR_{i-1}$, where CR is the minimum value of the criterion function, respectively, at the ith and $(i-1)$st steps.

The minimum value of the criterion function for all "LN–CN" pairs that is calculated from (6)–(11) is the lower bound for the step $i \in I$. At that, consideration is given at each step of the algorithm to the structure of the queuing system with $(m_c^{k+1} - 1)$ universal servers and m_s^{k+1} or $(m_s^{k+1} + 1)$ dedicated servers. The objective function is calculated from the assumption that all processing local nodes are independent. Some examples of efficient use of the above methods in production of printed matter were realized.

4. CONCLUSIONS

(1) The problem of resource allocation in the computer-aided control system of production of printed matter can be formulated as that of optimizing the parameters of an open queuing system.

(2) The proposed analytical model of the star-shaped queuing system can be used for preliminary estimation of the possible solutions under certain assumptions about the form of the input flow and the time of processing.

(3) To optimize the parameters of the queuing system model, an algorithm based on the branch-and-bound method can be used. At that, simulation is used to determine the value of the criterion function, and the analytical method, to calculate the lower bound.

(4) Simulations bear out the fact that for the open queuing system (allocated resources) the number of servers at a node of each kind is critical for minimization of the objective function.

(5) Simulation and solution of actual examples corroborate precision and adequacy of the proposed analytical model, and numerical examples testify to high convergence of the proposed algorithm.

REFERENCES

1. Zaikin, O.A., A Mathematical Model of Resource Allocation in the Computer-aided Control System of Nonmaterial Production, *Avtom. Telemekh.*, 2002, no. 8, pp. 160–168.
2. Buzacott, J. and Shanthikumar, J., *Modeling and Analysis of Manufacturing Systems*, New York: Wiley, 1993.
3. Chee Hock, Ng., *Queuing Modelling Fundamentals*, New York: Wiley, 1997.
4. Hall, R.W., *Queuing Methods for Service and Manufacturing*, New York: Prentice Hall, 1991.
5. Azadivar, F. and Lee, Y., Optimization of Discrete Variable Stochastic Systems by Computer Simulation, *Math. Comput. Simul.*, 1988, vol. 2, pp. 331–345.
6. Biethan, J. and Nissen, V., Combinations of Simulation and Evolutionary Algorithms in Management Science and Economics, *Ann. Oper. Res.*, 1994, vol. 32, pp. 183–208.
7. Dolgui, A. and Zaikin, O., Queuing Models for a Capacity Allocation Problem, in *Proc. 14th Eur. Simul. Multiconf. "Simulation and Modeling: Enables for Better Quality of Life"*, Ghent, Belgium, May 23–26, 2000, pp. 315–317.
8. Guariso, G., Hitz, M., and Werthner, H., An Integrated Simulation and Optimization Modeling Environment for Decision Support, *Decision Support Syst.*, 1996, vol. 1, pp. 103–117.
9. Zaikin, O., Dolgui, A., and Kraszewski, P., Queuing Modeling of the Resource Assignment in the High Tec Assembly Manufacturing, in *New Frontiers in Computational Intelligence and Its Applications*, Amsterdam: IOS Press, 2000, pp. 340–346.

This paper was recommended for publication by V.V. Kul'ba, a member of the Editorial Board