PERIODICA POLYTECHNICA SER. SOC. and MANAG. SCI. VOL. 4, NO. 1, PP. 5-28 (1996)

THE ILLUSTRATION OF THE ROUTING SENSITIVITY CALCULATION OF FLEXIBLE MANUFACTURING SYSTEMS WITH PERTURBATION ANALYSIS

Tamás KOLTAI* and Sebastian LOZANO**

*Department of Industrial Management
Technical University of Budapest
H-1521 Budapest, Hungary
Fax: + 36 1 463-1606, email: koltai@vvg.bme.hu
**Department of Industrial Organization
School of Engineering, University of Seville
Avda. Reina Mercedes, 41012 Sevilla, Spain
Fax: + 34 5 462 92 05, email: sebastian@esi.us.es

Received: Oct. 30, 1995

Abstract

The effect of the change of the routing mix on the throughput is an important question to be answered by operations management both at the planning and control phases of an FMS. This paper presents a method for the calculation of the gradient of the throughput with respect to the routing mix. The algorithm is based on perturbation analysis (PA) and on the product form properties of certain type of queuing networks. The provided gradient estimate is exact when conditions on the product form equilibrium distribution of the network are met, but it can also give good approximative results in other cases. Numerical examples are presented to illustrate the proposed calculation.

Keywords: flexible manufacturing systems, perturbation analysis, queuing networks, routing flexibility, simulation.

1. Introduction

One of the many aspects of flexibility of a flexible manufacturing system (FMS) is the possibility of alternative routing (BROWNE, DUBOIS, RATHMIL, SETHI and STECKE, 1984). The objective of the routing problem is to find the appropriate ratio of the routes followed by the various part types so as to maximize certain system performance measures. The practical relevance of this problem is twofold:

- In production planning, a routing mix providing the best possible system performance should be found;
- In production control, when the planned routing mix has to be modified, it should be done with the smallest possible deterioration of the system performance measure.

Although the routing of part types in an FMS can not be separated from other planning and control problems, from practical reasons hierarchical approaches are recommended (VAN LOOVEREN, GELDERS and WASSENHOVE, 1986; STECKE, 1986a).

The objective of this paper is to determine the sensitivity of the system throughput with respect to the routing mix using perturbation analysis. This sensitivity is important information for operations management when both planning and control decisions are to be made. In both cases these are the variables an operation manager can directly be interested in. He controls the routing mix by directing parts to different routes and observes the result in terms of the throughput.

Due to the complexity of the throughput function most of the work in this field tackles this question indirectly. The mathematical properties of the throughput as a function of the number of entities in the system and as a function of the relative workload in closed queuing networks were studied by YAO (1985). STECKE (1985; 1986) and STECKE and SOLBERG (1985) examined the characteristics of the workload belonging to the optimum throughput. KOBAYASHI and GERLA (1983) worked out an algorithm to optimize the throughput as a function of routing, for which they used the gradient of the system delay with respect to the relative visiting ratio. Their analytical method of computing this gradient works only for the single server case.

The various mathematical tools for modelling FMS can not provide efficient methods for the examination of routing sensitivity either. Mathematical programming models using the throughput as objective function are generally nonlinear mixed integer problems (STECKE, 1983), and sensitivity results cannot be gained without considerable computational burden if at all. Heuristic methods are frequently recommended to cope with the complexity of the loading problem but the sensitivity information they can provide is very limited (KIM and YANO, 1993; SHANKER and TZEN, 1985).

Representing FMS as closed queuing networks SURI, (1984) used mean value analysis (MVA) to study the effect of various control parameters on the throughput, but sensitivity data was gained by recalculating the throughput for the changed value of the parameter in question.

Simulation is one of the most frequently used methods for studying the performance of FMS. In this case sensitivity analysis means complicated and computation intensive experimental design methods (LAW and KELTON, 1991; CARRIE, 1988; STEUDEL and BERG, 1986). Perturbation analysis was recommended by HO and CAO (1985) to reduce the computational burden when discrete event systems are studied by simulation, but they examined the effect of the change of routing *probability* and not the change of the routing *mix*. Making a distinction between these two con-

cepts is important. While routing probability expresses the ratio of parts going from one station to another, routing mix expresses the ratio of the parts following the various routes. The latter one is a decision variable for operations management, while the first one is a consequence of that decision.

It can be concluded that:

- Neither most mathematical programming approaches that use the system throughput as objective function, nor conventional simulation can provide efficient methods for obtaining routing sensitivity information.
- Both simulation based PA and queuing network models seem to be appropriate approaches but they have so far been applied indirectly (e.g. via the relative workload, routing probability) instead of using the routing mix as decision variable.

The main contribution of the paper is to examine the direct relationship between the routing mix and the throughput. The application of the simulation based PA helps to examine realistic cases while the use of queuing theory principles ensures the robustness of some approximations required throughout the calculation.

In the following the quantitative formulation of the basic problem is provided in section 2. Then the calculation of the throughput gradient with respect to the routing mix is presented in section 3 and the validity of the calculated gradient is examined in section 4. Finally numerical examples illustrate the proposed method in section 5.

2. Problem Formulation

Assume an FMS consisting of a load/unload station and a number of already pooled machines with unlimited buffers. Parts can follow alternative routes, in all of which the load/unload station (denoted by subindex 0) is visited once. The number of parts produced is measured by the parts leaving the system through the load/unload station. The total production time is measured by the time necessary to produce a required amount of parts. The part types using the same type of pallets belong to a pallet class. The number of pallets available for each class is limited.

Two groups of data will be defined to formally describe the problem:

1.) System configuration and part type data:

P – number of part types
M – number of workstations
W – number of pallet class
C_f – number of circulating pallets of class f

$L(f)$ – set of part types using pallet class f
n_t – number of alternative routes of part type $t, t = 1, ..., P$
d_{ts} – number of operations of part type t on route $s, t = 1, ...P, s = 1, ..., n_t$
z_{tsk} – k-th station visited by part type t
on route $s, k = 1, ..., d_{ts}, t = 1, ..., P, s = 1, ..., n_t$
h_{tsk} – k-th operation time of part type t
on route $s, k = 1, ..., dts, t = 1, ..., P, s = 1, ..., n_t$
m_t – part mix ratio for part type $t, t = 1, ..., P$,
defined such that $\sum_{t \in L(f)} m_t = 1, f = 1, ..., W$
Θ_{ts} – ratio of route s of part type $t, t = 1, ..., P, s = 1, ..., n_t$,
defined such that $\sum_s \Theta_{ts} = m_t, t = 1, ..., P$

2.) Observational data, which can be gained by collecting information about the real or simulated performance of the system:

T – total production time
N_{fj} – total number of visits of parts belonging to class f at station j

Based on the previously defined parameters some further data can be derived:

y_{tsi} – number of visits at station i by part type t on route s

$$y_{tsi} = \sum_{k=1}^{d_{ts}} \delta(z_{tsk} = i), \tag{1}$$

where $\delta(z_{tsk} = i) = 1$ if part type t on route s goes to station i in the k-th step, and $\delta(z_{tsk} = i) = 0$ otherwise.

b_{tsi} – sum of processing requirements on station i along route s of part type t

$$b_{tsi} \sum_{k=1}^{d_{ts}} h_{tsk} \delta(z_{tsk} = i). \tag{2}$$

h_i – mean operation time at station i. It is assumed that the average operation time at any station is the same for all part types. Since at the aggregation of the processing times at station i only the routes that visit station i need to be considered, let $S(t, i) = \{s : y_{tsi} > 0\}$. Then h_i is calculated as follows,

$$h_i = \frac{b_{tsi}}{y_{tsi}} \qquad \forall t \forall s \in S(t, i). \tag{3}$$

v_{fi} – relative visiting ratio of class f at station i, which is defined as

$$\frac{N_{fj}}{N_{fi}} = \frac{v_{fj}}{v_{fi}}. \tag{4}$$

TP_{fj} – throughput of class f at station j,

$$TP_{fj} = \frac{N_{fj}}{T}. \tag{5}$$

TP_j – throughput of station j

$$TP_j = \frac{\sum_{f=1}^{W} N_{fj}}{T} = \sum_{f=1}^{W} TP_{fj}. \tag{6}$$

Considering these data three remarks must be made:

- It will be assumed that the operation time at a station is exponentially distributed and independent from the pallet class. This assumption permits the application of product form equilibrium distributions for multi class closed queuing networks (BASKETT, CHANDY, MUNTZ and PALACIOS, 1975).
- Since v_{fi} are relative values one of them must be fixed arbitrarily for every class. As every route goes through the load/unload station once, the visiting ratio of that station will be set to one for every class ($v_{f0} = 1$).
- The throughput of class f is defined as TP_{fj}/v_{fj} for any $j = 1, ..., M$. Since $v_{f0} = 1\,(f = 1, ..., W)$ are chosen, the system throughput can be determined based on the observed data, that is,

$$TP_0 = \sum_{f=1}^{W} TP_{f0} = \sum_{f=1}^{W} \frac{N_{f0}}{T}. \tag{7}$$

The objective is to determine the gradient of the system throughput with respect to the routing mix, that is,

$$\frac{\partial TP_0}{\partial \Theta_{ts}} = \frac{\sum_{f=1}^{W} TP_{f0}}{\partial \Theta_{ts}} = ? \qquad \begin{matrix} t = 1, \ldots, P, \\ s = 1, \ldots, n_t. \end{matrix} \tag{8}$$

To facilitate further discussion (8) will be examined with the help of the relative workload of class $r(v_{ri}h_i)$. Transforming (8) we get

$$\frac{\partial TP_0}{\partial \Theta_{ts}} = \sum_{f=1}^{W} \sum_{r=1}^{W} \sum_{i=1}^{M} \frac{\partial TP_{f0}}{\partial (v_{ri}h_i)} \frac{\partial (v_{ri}h_i)}{\partial \Theta_{ts}}, \qquad \begin{matrix} t = 1, \ldots, P, \\ s = 1, \ldots, n_t. \end{matrix} \tag{9}$$

3. The Steps of Gradient Calculation

The calculation is based on the combination of analytical and simulation results. According to the system configuration and part type data the performance of the system is simulated and the observational data are recorded. Two sets of random variables are used in the simulation.

- The operation times at the workstations are considered random variables with a mean equal to the mean operation time calculated by (3). This is an approximation but used frequently at the planning phase of FMS (STECKE, 1986a).
- The routing of the part types is considered as a random process with expected values equal to Θ_{ts}. There exist many alternative system performances at the same Θ_{ts} depending on the detailed scheduling considerations. The long term performance, however, can be approximated by this random process.

The calculation of (9) is performed in three steps. First $\partial TP_{f0}/\partial(v_{ri}h_i)$ is transformed and shown to be a function of the gradient of the total production time with respect to an operation time. Then this gradient is calculated using perturbation analysis. Finally $\partial(v_{ri}h_i)/\partial\Theta_{ts}$ is computed.

3.1. Calculation of the Gradient of the Throughput with Respect to the Relative Workload

Due to the product form of the equilibrium distribution the throughput of class f is a function of the relative workloads of all the classes, that is, $TP_{f0} = f(v_{ri}h_i)$. Using the differentiation rules of compounded functions we get,

$$\frac{\partial TP_0}{\partial h_i} = \sum_{r=1}^{W} \frac{\partial TP_{f0}}{\partial(v_{ri}h_i)} \frac{\partial(v_{ri}h_i)}{\partial h_i} = \sum_{r=1}^{W} v_{ri} \frac{\partial TP_{f0}}{\partial(v_{ri}h_i)} \tag{10}$$

(10) is an implicit expression for $\partial TP_{f0}/\partial(v_{ri}h_i)$. To avoid the difficulties of solving a linear equation system for its calculation a more efficient method is suggested. Let us assume that the operation time of the various classes at workstation i are distinguished (h_{ri}). Note that for every class $h_{ri} = h_i$ at the point where the throughput sensitivity is calculated. Differentiating TP_{f0} according to h_{ri} we get

$$\frac{\partial TP_{f0}}{\partial h_{ri}} = \sum_{g=1}^{W} \frac{\partial TP_{f0}}{\partial(v_{gi}h_i)} \frac{\partial(v_{gi}h_i)}{\partial h_{ri}}. \tag{11}$$

Since

$$\frac{\partial(v_{gi}h_i)}{\partial h_{ri}} = \begin{cases} v_{ri} & \text{if} \quad g = r, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

the explicit form of $\partial TP_{f0}/\partial(v_{ri}h_i)$ is the following,

$$\frac{\partial TP_{f0}}{\partial(v_{ri}h_i)} = \frac{1}{v_{ri}} \frac{\partial TP_{f0}}{\partial h_{ri}}. \tag{13}$$

Using (5) the throughput gradient of the workstations with respect to the mean operation times is the following,

$$\frac{\partial TP_{fj}}{\partial h_{ri}} = -\frac{N_{fj}}{T^2} \frac{\partial T}{\partial h_{ri}}. \tag{14}$$

Applying (4) when $j = 0$ and $v_{ro} = 1$, $r = 1, \ldots, M$ the gradient of the system throughput with respect to the relative machine workload is the following,

$$\frac{\partial TP_{f0}}{\partial(v_{ri}h_i)} = -\frac{N_{f0}N_{r0}}{T^2 N_{ri}} \frac{\partial T}{\partial h_{ri}}. \tag{15}$$

(15) allows the computation of the gradient of the throughput with respect to the workload as a function of $\partial T/\partial h_{ri}$.

3.2. The Gradient of the Total Production Time with Respect to an Operation Time

To obtain the value of (15) $\partial T/\partial h_{ri}$ should be known. This term can be gained using perturbation analysis. PA was developed for the gradient estimation of performance measures with respect to certain system control variables in discrete event systems (DES), when the performance measure is obtained by discrete event simulation. The basic idea is that a sample path of the simulation contains information about certain system characteristics and therefore it is not necessary to rerun the simulation when the performance measure sensitivity is estimated (HO and CASSANDRAS, 1983; HO and CAO, 1985). This approach may efficiently provide sensitivity information concerning the performance measures of discrete type manufacturing systems (HO, 1987; KOLTAI, LARRAÑETA and ONIEVA, 1993; KOLTAI, LARRAÑETA, ONIEVA and LOZANO, 1993).

In this case the gradient of the expected total production time with respect to a mean operation time should be determined, that is,

$$\frac{\partial E[T(h_{ri}, \xi]}{\partial h_{ri}} = ? \tag{16}$$

where ξ is representing a particular realization of all the random variables in the system. Writing (16) in differential form

$$\lim_{\Delta h_{ri}\to 0} E[\frac{T(h_{ri}+\Delta h_{ri},\xi)-T(h_{ri},\xi)}{\Delta h_{ri}}] =? \tag{17}$$

PA states that in certain conditions inferences can be made about the evolution of the sample path when the control variable changes, that is, without repeating the experiment for $h_{ri}+\Delta h_{ri}, T(h_{ri}+\Delta h_{ri})$ can be known (HO and CAO, 1991). To apply this approach two problems should be discussed.

a.) **Generation of perturbations** provides values of Δh_{tsk} if Δh_{ri} is known, where i is the station at which the corresponding operation is performed (that is, $i = z_{tsk}$). Then Δh_{tsk} should be generated so, that

$$E[h_{tsk}+\Delta h_{tsk}] = h_{ri}+\Delta h_{ri}, \qquad t \in L(r). \tag{18}$$

Applying the perturbation generation rules introduced by SURI (1983a) if h_{ri} is scale parameter of the operation time distribution the change of the operation times can be calculated as follows,

$$\Delta h_{tsk} = h_{tsk}\frac{\Delta h_{ri}}{h_{ri}}, \qquad t \in L(r). \tag{19}$$

and if h_{ri} is location parameter of the operation time distribution then

$$\Delta h_{tsk} = \Delta h_{ri}, \qquad t \in L(r). \tag{20}$$

Although the real reason of the change of h_{ri} is the change of the routing mix, we can substitute this by a generated operation time perturbation.

b.) **Propagation of perturbations** provides the information for examining the effect of all the generated Δh_{tsk} on T. Using the perturbation propagation rules of infinitesimal perturbation analysis it can be seen how the generated perturbation affects the operation of other stations and how they accumulate or die throughout the sample path (HO and CAO, 1991). To keep record on this changes the final result in case of scale parameter is as follows,

$$\Delta T = \frac{\Delta h_{ri}}{h_{ri}}\sum_{A} h_{tsk}, \tag{21}$$

where A denotes the set of all the accumulated perturbations affecting T. The sum in (21) is an observational data and is determined by keeping

record throughout the simulation about all the perturbations influencing T. In (21) if Δh_{ri} is small enough, then T is a linear function of Δh_{ri} (HO and CAO, 1991) and the gradient can be calculated as follows,

$$\frac{\Delta T}{\Delta h_{ri}} = \frac{1}{h_{ri}} \sum_{A} h_{tsk}. \tag{22}$$

Analogously, for location parameters the total operation time gradient is as follows,

$$\frac{\Delta T}{\Delta h_{ri}} = \frac{1}{\Delta h_{ri}} \sum_{A} \Delta h_{ri}. \tag{23}$$

3.3. The Gradient of the Relative Workload with Respect to the Routing Mix

The expected value of the relative workload of class r at station i can be calculated by summing the mean processing times of all operations performed at that station by parts belonging to that class and it is, therefore, a function of the routing mix variables ts. If the experiment is long enough this approximation is correct and the relative workload can be calculated as follows,

$$v_{ri}h_i = \sum_{t \in L(r)} \sum_{s=1}^{n_t} \Theta_{ts} \sum_{k=1}^{d_{ts}} h_{tsk}\delta(z_{tsk} = i). \tag{24}$$

Derivating (24) with respect to the routing mix,

$$\frac{\partial(v_{ri}h_i)}{\partial\Theta_{ts}} = \begin{cases} \sum_{k=1}^{d_{ts}} h_{tsk}\delta(z_{tsk} = i) & \text{if} \quad t \in L(r), \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

(25) shows that this gradient can be determined from the system configuration and part type data. Using (2), (25) can shortly be written as

$$\frac{\partial(v_{ri}h_i)}{\partial\Theta_{ts}} = b_{tsi}\delta(t \in L(r)), \tag{26}$$

where $\delta(t \in L(r)) = 1$ if part type t belongs to class r and $\delta(t \in L(r)) = 0$ otherwise.

4. The Routing Mix Sensitivity and its Validity

Finally writing (15) and (26) in (9) the throughput gradient with respect to the routing mix is obtained,

$$\frac{\partial TP_0}{\partial \Theta_{ts}} = -\frac{1}{T^2}\sum_{f=1}^{W} N_{f0} \sum_{r=1}^{W} N_{r0}\delta(t \in L(r)) \sum_{i=1}^{M} \frac{b_{tsi}}{N_{ri}} \frac{\partial T}{\partial h_{ri}} \quad t = 1, \ldots, P, \quad s = 1, \ldots, n. \tag{27}$$

The validity of (27) depends on three factors:

- the validity of the product form assumption;
- the validity of the $\partial T/\partial h_{ri}$ estimate;
- the validity of the application of steady state formulae.

a.) **The validity of the product form assumption.** Although this condition is rather restrictive, there are many cases when it is satisfied. One of the most frequently mentioned arguments against product form approximation is the existence of blocking. YAO and BUZACOTT (1987) showed that in the majority of real FMS blocking is avoided by the special organization of buffers (central and local buffers) and in these systems product form equilibrium distribution holds in case of exponential service times. General service time distribution networks can be approximated by an equivalent exponential service time distribution network (YAO and BUZACOTT, 1986), although, in case of complicated networks the calculation burden is considerable.

When the product form assumption is not met the results can still be good approximations for practical applications. SURI (1983) showed that queuing formulas can be very robust in cases when conditions on the product form equilibrium distribution are not met. HO and CAO (1985) presented several examples in which this condition was not valid but perturbation based sensitivity calculation concerning the sensitivity of the throughput with respect to the routing probability provided good results. SURI and DILLE (1985) also presented several perturbation based sensitivity calculation results for non product form queuing networks.

b.) **The validity of $\partial T/\partial h_{ri}$ estimate.** One of the main issues in PA is the problem of interchangeability of differentiation and expectation. If this interchangeability holds, then the sample path gradients of the different experiments can be used for calculating the expected value of the gradient (HO and CAO, 1991). The intensive research activity on this field had two main directions. First, unbiasedness of gradient estimation was proved for a considerable amount of cases (HEIDELBERGER et al., 1988, HO and CAO, 1991). Secondly, when the estimate is biased, methods

were worked out to modify the original technique of infinitesimal perturbation analysis to ensure the required statistical properties of the estimated gradient. Smoothed (GONG and HO, 1987) and extended (HO and LI, 1988) perturbation analysis are examples for overcoming the difficulties of the discontinuities of the sample path, however, generally at the cost of increased computational requirement. It has to be mentioned that even when the estimate is biased the calculated gradient gives better results than the 'brute-force' finite difference approximation.

Flexible manufacturing systems are generally modelled as queuing networks with multi-class customers. It is proven that in this type of networks the infinitesimal perturbation analysis gives unbiased estimates just in very limited cases (CAO, 1988, HO and CAO, 1991). Interestingly in practical FMS configuration it is not rare to meet these special conditions. If, for example, every workstation visited by more than one class of customers is fed by only a single source, then PA gives unbiased estimate. In practice, workstations which are visited by more than one type of products are frequently fed by one load/unload or measuring station, therefore the above condition can be met. In most of the cases, however, either special methods of PA must be used or the results have to be considered as approximations. SURI and DILLE (1985) showed how these approximate gradients can be used for the efficiency improvement analysis of an FMS.

c.) **The validity of the application of steady state formulae.** The application of steady state formulae for the rough cut modelling of queuing network type systems is an accepted method due to its rapidness and relative robustness (SURI, 1983). For control purposes, however, it can be used only if the real performance of the system approaches the steady state conditions. Although in FMS generally relatively small batches are used, there are cases (e.g. common due dates and short operation times) when steady state approximations are realistic (KIM and YANO, 1993).

5. Numerical Examples

The following two examples are to illustrate the proposed method. For both problems the optimal routing mix was determined by KOBAYASHI and GERLA (1983) using MVA. Since at the optima the gradient should satisfy the optimality conditions this can be a reference to check the results. Using Lagrangian multipliers it can be shown that the components of the gradient

should satisfy

$$\frac{\partial TP_0}{\partial \Theta_{ts}} = \lambda_r, \qquad \begin{matrix} t \in L(r), \\ s = 1, \ldots, n_t, \\ r = 1, \ldots, W, \end{matrix} \tag{28}$$

that is, the partial derivative of the throughput with respect to every route of parts belonging to the same class should be equal.

The simulation of the examined systems was carried out using the SIMAN IV simulation language (PEDGEN, SHANNON and SADOWSKI, 1990). Perturbation analysis was performed by user written subroutines.

5.1. Single Class Queuing Network Example

Suppose we have a FMS consisting of four single server workstations $(W_j, j = 0, \ldots, 3)$, where the load/unload station (W_0) is the central server (*Fig. 1*). One type of product is produced (PR) and five pallets are used $(K = 1, C_1 = 5)$. The parts may follow three routes. The sequence of operations with their respective mean operation times is given in *Table 1*. The operation times are exponentially distributed. 10000 parts should be produced. The optimum routing mix for this system was determined by KOBAYASHI and GERLA (1983) and its result is also presented in *Table 1*. Since in this example there is just one part type the subindices expressing the part type and pallet class can be suppressed $(v_{ri} = v_i, TP_{fj} = TP_j, \Theta_{ts} = \Theta_s)$.

Table 1
Basic data of example 1

Part type	Route number	Operation sequence (station no./mean oper.time)	$\Theta_s^{(OPT)}$
PR	1	(0/0.25)→ (1/0.5)	0.693
	2	(0/0.25)→ (2/1)	0.239
	3	(0/0.25)→ (3/2)	0.068

The system throughput as a function of Θ_1 and Θ_2 is depicted in *Fig. 2*. These values were calculated using the exact MVA algorithm described in REISER and LAVENBERG (1980). The results of the simulation, using the optimal routing mix as the expected value of random routing, are summarized in *Table 2*. The second row of the table shows the relative visiting

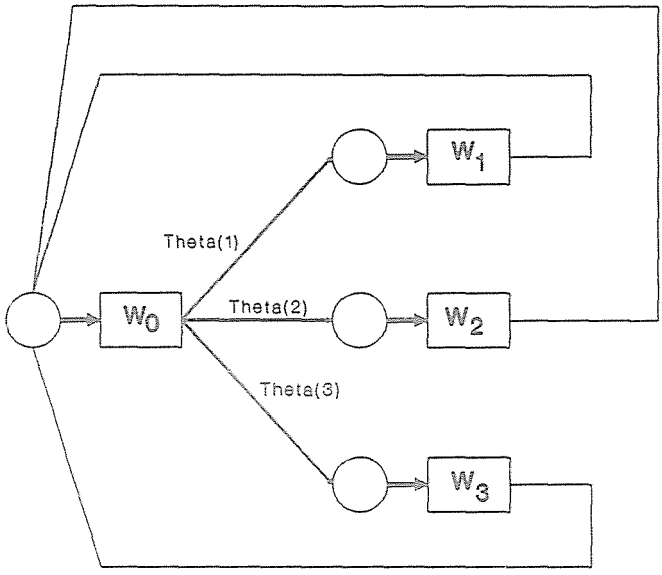

Fig. 1. Single-class queuing network model

frequencies (v_i) and the third row provides the throughput at the workstations (TP_j). The value of the throughput obtained by simulation deviated from the analytical value found by KOBAYASHI and GERLA (1983) using MVA, by 0.67%. The gradient of the total operation time calculated by perturbation analysis can be seen in the last row ($\partial T/\partial h_i$).

Table 2
Simulation results of example 1

Work stations	W_0	W_1	W_2	W_3
Rel. visits (v_i)	1	0.696	0.235	0.068
Throughput (TP_j)	2.3877	1.6611	0.5618	0.1638
$\partial T/\partial h_i$	3514.93	4699.06	739.41	110.21

The results of the calculation of the throughput gradient with respect to the routing mix can be seen in *Table 3*. The gradients of the throughput of all the stations are provided, although we are basically interested in the system throughput gradient ($\partial TP_0/\partial\Theta_s$). Change of the routing mix means that at least two Θ_s change with opposite signs, since $\Theta_1+\Theta_2+\Theta_3 = 1$. It can be seen that the difference of any two systems throughput gradient

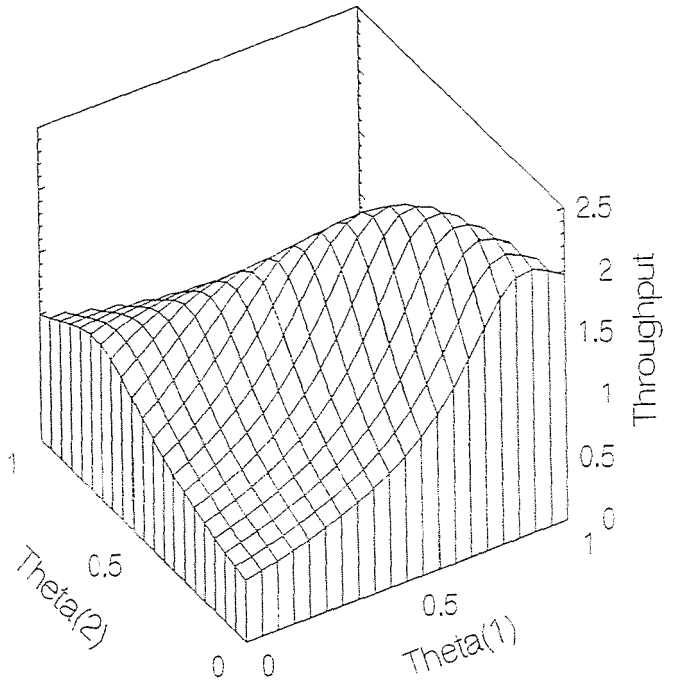

Fig. 2. The throughput function of the single class example

Table 3
Routing sensitivity information of example 1

Route	$\partial TP_0/\partial\Theta_s$	$\partial TP_1/\partial\Theta_s$	$\partial TP_2/\partial\Theta_s$	$\partial TP_3/\partial\Theta_s$
1	-2.426	-1.688	-0.571	-0.166
2	-2.292	-1.595	-0.539	-0.157
3	-2.333	-1.623	-0.549	-0.160

component is relatively small. It coincides with the fact that these gradients belong to the maximum value of the throughput function.

Figs. 3 and *4* show the throughput as a function of the routing mix. In *Fig. 3* Θ_3 is held constant and equal to its optimal value (0.068). The system throughput and its corresponding directional derivative are drawn in the same chart. The change of the routing mix means that Θ_1 increases from 0 to 0.932, while at the same time Θ_2 decreases from 0.932 to 0. The directional derivative $(\partial TP_0/\partial\Theta_1 - \partial TP_0/\partial\Theta_2)$ contains the effect of both changes. This derivative indicates inflection points around $\Theta_1 = 0.4$ and

around $\Theta_1 = 0.9$ and a maximum at $\Theta_1 = 0.69$ which is the optimum according to the solution of KOBAYASHI and GERLA (1983). *Fig. 4* shows the same function when Θ_2 is held constant and equal to its optimum value (0.234). The change of routing mix in this case means that Θ_1 increases from 0 to 0.766, while in the same time Θ_3 decreases from 0.766 to 0. The optimum is found around $\Theta_1 = 0.69$ again.

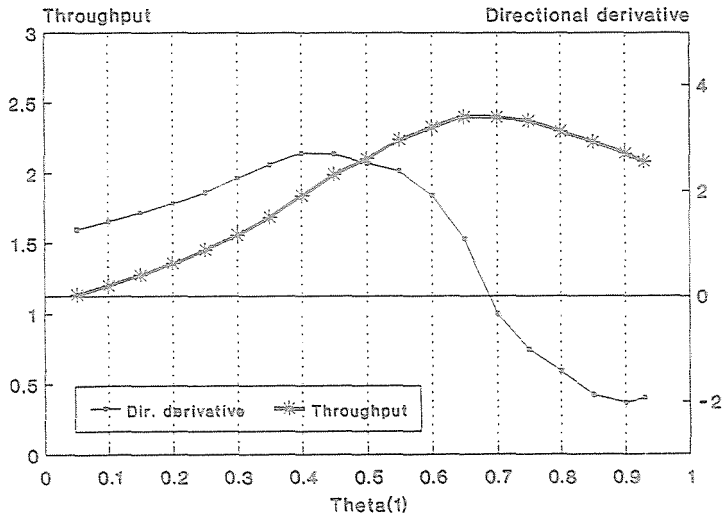

Fig. 3. Change of the throughput ($\Theta_3 = 0.068$)

It can be observed that the throughput function depicted in *Fig. 3* corresponds to a cut of the surface of *Fig. 2* along $\Theta_1 + \Theta_2 = 0.932$. Analogously, *Fig. 4* corresponds to a cut of the same surface along $\Theta_2 = 0.234$.

5.2. Multiple Class Queuing Network Example

Suppose we have an FMS consisting of four single server workstations ($W_j, j = 1, \ldots, 4$) (*Fig. 5*). Two types of products are produced (PR1 and PR2) and two types of pallets are used ($K = 2$). Three pallets are available in each class ($C_1 = 3, C_2 = 3$). Each part type uses a different pallet type ($L(1) = 1, L(2) = 2$). Every part type may follow two different routes. The sequence of operations with the respective mean operation times is given in *Table 4*. The operation times are exponentially distributed. In total, 10000 parts should be produced. The two local maxima found by KOBAYASHI and GERLA (1983) for this problem are also presented in *Table 4* ($\Theta_{ts}^{(M1)}, \Theta_{ts}^{(M2)}$).

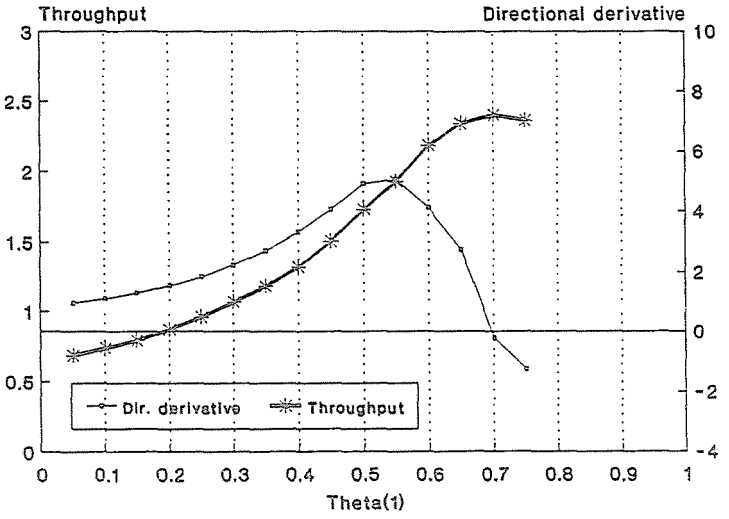

Fig. 4. Change of the throughput ($\Theta_2 = 0.234$)

Table 4
Basic data of example 2

Part type	Route number	Operation sequence (station no./mean oper.time)	$\Theta_{ts}^{(M1)}$	$\Theta_{ts}^{(M2)}$
PR1	1	(1/0.5)— (3/0.5)	0.544	1.000
	2	(1/0.5)— (4/1.0)	0.456	0.000
PR2	1	(2/1.0)— (3/0.5)	1.000	0.025
	2	(2/1.0)— (4/1.0)	0.000	0.975

Fig. 5 shows that there is not a unique load/unload station, however. W_1 is visited by every PR1 once and not visited at all by PR2. At the same time W_2 is visited by every PR2 once and not visited at all by PR1. This can be interpreted as if each class had a different load/unload station. In this case the relative visiting ratios of each class at its own load/unload station should be set to one. Alternatively one could consider a virtual unique load/unload station (W_0) – with zero operation time – which is visited by all part types prior to starting their true sequence of operations.

The throughput as a function of $\Theta_{1,1}$ and $\Theta_{2,1}$ are given in *Fig. 6*. These values were calculated using, again, the MVA algorithm. The results of the simulation, using the first local maximum as the expected value of

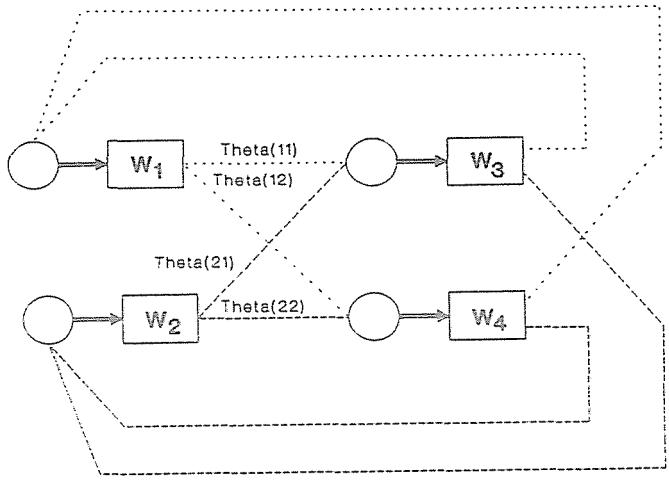

Fig. 5. Multiple class queuing network model

random routing, are summarized in *Table 5*. The second and third row of the table shows the relative visiting frequencies of class 1 and class 2 (v_{ri}). The fourth and fifth rows show the gradients of the total operation time with respect to the operation times of class 1 and class 2 ($\partial T/\partial h_{ri}$). The value of the system throughput ($TP_0 = TP_{1,1} + TP_{2,2}$) is given in the last row. This value deviates from the theoretical value found by KOBAYASHI and GERLA (1983) using MVA, by 0.1%.

Table 5
Simulation results of example 2 at $\Theta_{ts}^{(M1)}$

Work stations		W_1	W_2	W_3	W_4
Rel. visits (v_{ri})	Class 1	1	0	0.544	0.456
	Class 2	0	1	0.999	0.001
$\partial T/\partial h_{ri}$	Class 1	1394.3	0	1693.3	668.62
	Class 2	0	1493.6	1938.0	1.7036
Throughput (TP_0)			2.1382		

The results of the calculation of the throughput gradient with respect to the routing mix can be seen in *Table 6*. The gradients of the throughput of all the classes with respect to the four routing variables are provided in

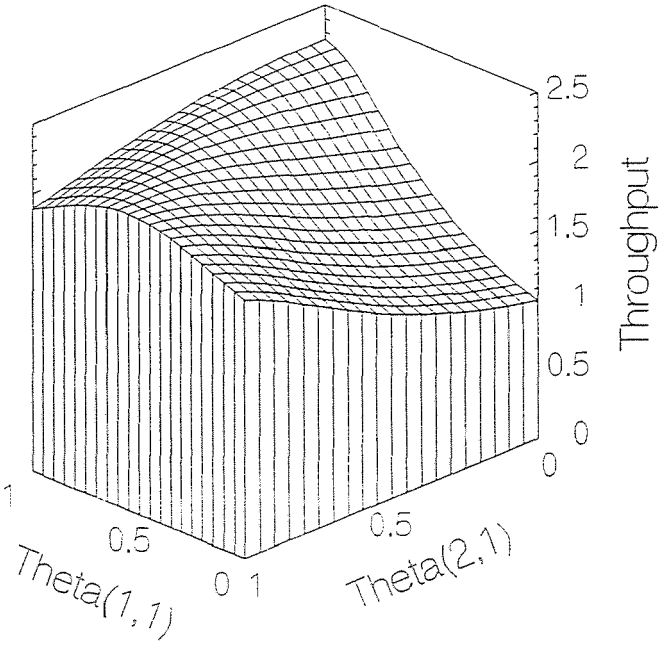

Fig. 6. The throughput function of the multiple class example

the second and third row. The last row contains the gradient of the system throughput ($\partial TP_0/\partial\Theta_{ts}$). Change of the routing of any part type means that at least two Θ_{ts} change with opposite signs. It can be seen that the components of the gradient corresponding to class 1 satisfy the optimality conditions. However, since $\Theta_{2,1}^{(M1)}$ takes its maximum feasible value the optimality condition for the gradient components corresponding to class 2 is $\partial TP_0/\partial\Theta_{2,1} \geq \partial TP_0/\partial\Theta_{2,2}$. This condition is also met according to *Table 6*.

Table 6
Routing sensitivity information of example 2 at $\Theta_{ts}^{(M1)}$

$\partial TP_0/\partial\Theta_{ts}$	Θ_{11}	Θ_{12}	Θ_{21}	Θ_{22}
Class 1	−0.6170	−0.5924	−0.6746	−0.0326
Class 2	−0.4132	−0.3967	−0.4518	−0.6916
$\partial TP_{f0}/\partial\Theta_{ts}$	−1.0303	−0.9891	−1.1263	−1.7242

Figs. 7 and *8* show the throughput as a function of the routing mix. In *Fig. 7* Θ_{21} is held constant and equal to its value at the first local maximum (1.00). The system throughput and the corresponding directional derivative along $\Theta_{1,1}$ are drawn in the same chart. The change of routing mix means that $\Theta_{1,1}$ increases from 0.00 to 1.00, while at the same time $\Theta_{1,2}$ decreases from 1.00 to 0.00. This directional derivative $(\partial TP_0/\partial\Theta_{1,1} - \partial TP_0/\partial\Theta_{1,2})$ contains the effect of both changes. This derivative indicates a maximum around $\Theta_{1,1} = 0.54$ which is the local maximum according to KOBAYASHI and GERLA (1983). *Fig. 8* shows the same function when $\Theta_{1,1}$ is held constant and equal to its value at the first local maximum (0.544). The change of routing mix in this case means that $\Theta_{2,1}$ increases from 0.00 to 1.00 and $\Theta_{2,2}$ decreases from 1.00 to 0.00. The maximum is found at $\Theta_{2,1} = 1.00$.

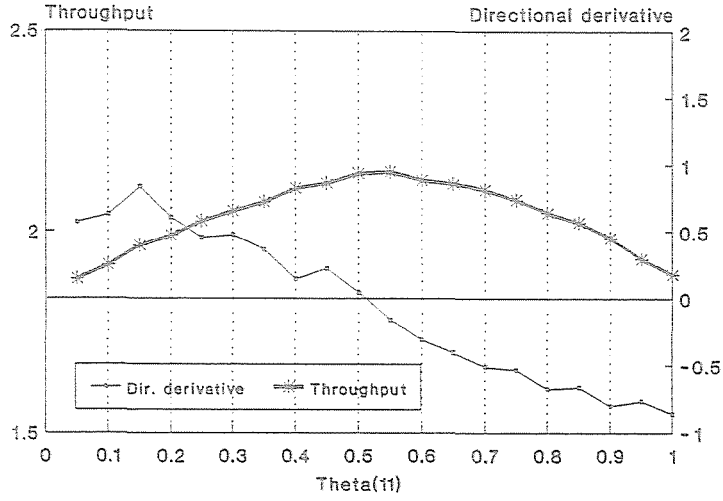

Fig. 7. Change of the throughput ($\Theta_{2,1} = 1$)

Tables 7 and *8* contain the simulation and sensitivity calculation data for the other local maximum found by KOBAYASHI and GERLA (1983) while *Figs. 9* and *10* contain the throughput functions and their respective directional derivatives. Since $\Theta_{1,1}^{(M2)}$ takes its maximum feasible value the optimality condition for the gradient components of class 1 is $\partial TP_0/\partial\Theta_{1,1} \geq \partial TP_0/\partial\Theta_{1,2}$. This condition is met according to *Table 8*. The gradient components corresponding to class 2 should be equal according to (28). This is not reflected in *Table 8*. The reason for this is that, due to the random nature of simulation, the observed value of $\Theta_{2,1}$ is not exactly equal

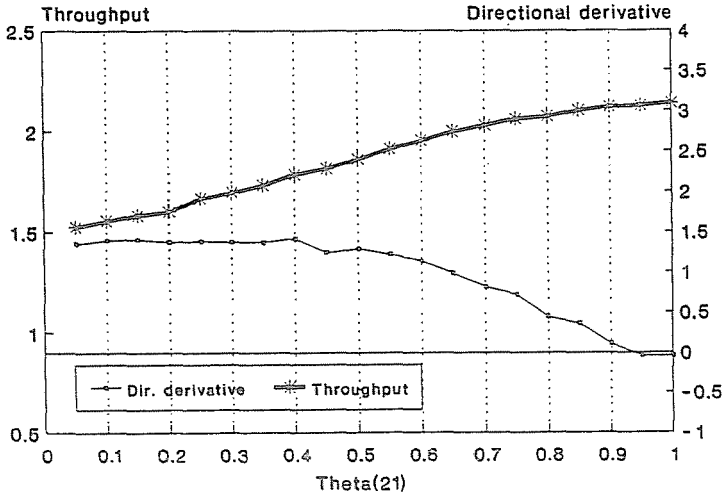

Fig. 8. Change of the throughput ($\Theta_{1,1} = 0.544$)

to the expected $\Theta_{2,1}^{(M2)}$ (0.028 instead of 0.025). *Fig. 10*, however, shows that the optimum is really at $\Theta_{1,1} = 1$ and $\Theta_{2,1} = 0.025$ as it was expected. This figure also shows that the directional derivative along $\Theta_{2,1}$ reaches zero again around 0.22. This is, however, not a local maximum because the directional derivative along $\Theta_{2,1}$ at this point is equal to 0.2.

Table 7
Simulation results of example 2 at $\Theta_{ts}^{(M2)}$

Work stations		W_1	W_2	W_3	W_4
Rel. visits (v_{ri})	Class 1	1	0	0.999	0.001
	Class 2	0	1	0.028	0.972
$\partial T/\partial h_{ri}$	Class 1	2479.7	0	2371.1	3.1504
	Class 2	0	933.14	23.979	1062.9
Throughput (TP_0)			2.254		

It can be observed that the throughput function depicted in *Fig. 7* corresponds to a cut of the surface of *Fig. 6* along $\Theta_{2,1} = 0.025$. Analogously, *Fig. 8* corresponds to a cut of the same surface along $\Theta_{1,1} = 1$.

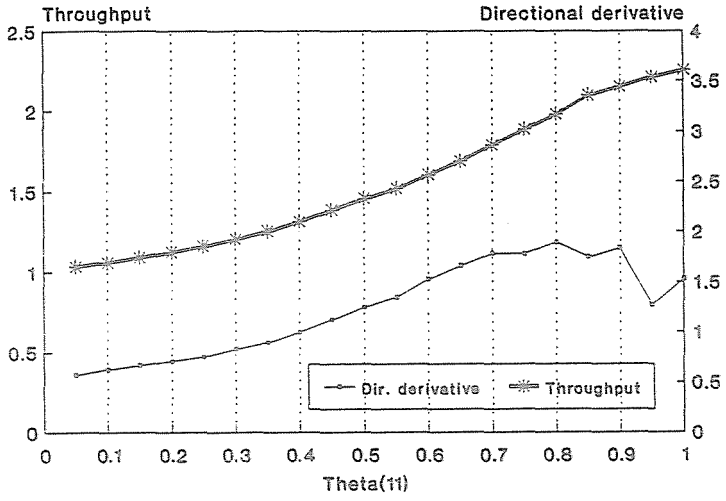

Fig. 9. Change of the throughput ($\Theta_{2,1} = 0.025$)

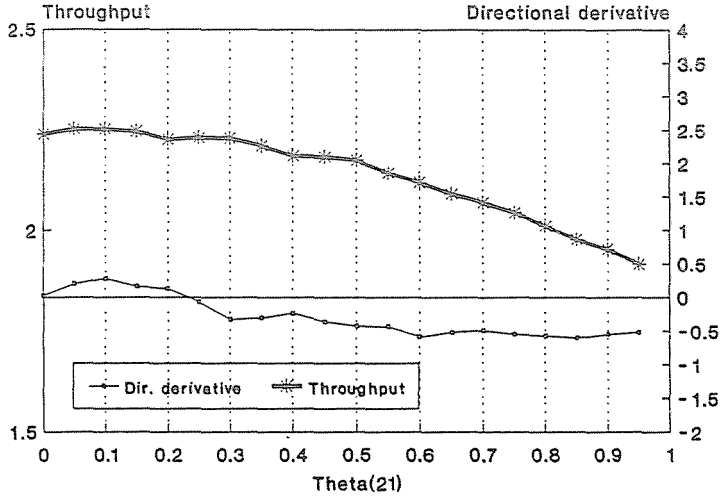

Fig. 10. Change of the throughput ($\Theta_{1,1} = 1$)

Table 8
Routing sensitivity information of example 2 at $\Theta_{ts}^{(M2)}$

$\partial TP_{f0}/\partial\Theta_{ts}$	Θ_{11}	Θ_{12}	Θ_{21}	Θ_{22}
Class 1	-0.8201	-1.8359	-0.4597	-0.6850
Class 2	-0.4126	-0.9236	-0.2313	-0.3446
$\partial TP_0/\partial\Theta_{ts}$	-1.2327	-2.7596	-0.6910	-1.0297

5.3. Discussion

In both examples the conditions on the product form distribution requirements were met. Some noise, however, can be found in the results. The directional derivatives at the local optima are not exactly equal to zero. One reason of this noise is that, although the simulation run length was rather long, the effect of the transient period still could have been experienced. Another reason is that a random process was used for selecting the various routes, with probabilities equal to the required routing mix. Due to the random nature of the process, the real visiting ratios were slightly different from the expected ones. The noise is especially strong at the extreme points ($\Theta_{ts} \approx 1$ or ≈ 0) where small deviations from the expected routing mix have a strong effect on the results. Therefore a careful experimental design is required when the throughput gradient is computed near these values.

6. Summary

In this paper, the problem of estimating the sensitivity of the throughput of a FMS with respect to the routing mix has been studied. The complexity of the throughput function precludes an analytical approach for the general case (multiple class, multiple server, etc.). The proposed method is based on using perturbation analysis of discrete event dynamic systems for gradient estimation. Some theoretical results from queuing network theory are also taken advantage of. Numerical results confirmed the feasibility of the proposed approach. The robustness of the calculation when the required conditions are not met, the utilization of the results in optimization procedures, or to design efficient routing control, are topics for further research.

References

BASKETT, F. – CHANDY, K. M. – MUNTZ, R. R. – PALACIOS, F. G. (1975): Open, Closed and Mixed Networks of Queues with Different Classes of Customers, *Journal of the ACM*, Vol. 22, No. 2, pp. 248–260.

BROWNE, J. – DUBOIS, D. – RATHMILL, K. – SETHI, S. P. – STECKE, K. E. (1984): Classification of Flexible Manufacturing Systems, *The FMS Magazine*, April, pp. 114–117.

BUZEN, J. P. (1973): Computational Algorithms for Closed Queuing Networks with Exponential Servers, *Communications of the ACM*, Vol. 16, No. 9, pp. 527–531.

CAO, X. (1988): Realization Probability in Multi-class Closed Queuing Networks, *European Journal of Operational Research*, Vol. 36. pp. 393–401.

GONG, W. – HO, Y. (1987): Smoothed (Conditional) Perturbation Analysis of Discrete Event Dynamical Systems, *IEEE Transactions on Automatic Control*, Vol. 32. No. 10. pp. 858–866.

HEIDELBERGER, P. – CAO, X. – ZAZANIS, M. A. – SURI, R. (1988): Convergence Properties of Infinitesimal Perturbation Analysis Estimates, *Management Science*, Vol. 34. No. 11. pp. 1281–1302.

CARRIE, A. (1988): Simulation of Manufacturing Systems, Chichester, John Wiley & Son.

HO, Y. C. – CASSANDRAS, C. (1983): A New Approach to the Analysis of Discrete Event Dynamic Systems, *Automatica*, Vol. 19, No. 2, pp. 149–167.

HO, Y. C. – CAO, X. R. (1985): Performance Sensitivity to Routing Changes in queuing Networks and Flexible Manufacturing Systems using Perturbation Analysis, *IEEE Journal of Robotics and Automation*, Vol. RA–1, No. 4, pp. 165–172.

HO, Y. C. (1987): Performance Evaluation and Perturbation Analysis of Discrete event Dynamic Systems, *IEEE Transactions on Automatic Control*, Vol. AC-32, No. 7, pp. 563–572.

HO, Y. C. – CAO, X. R. (1991): Perturbation Analysis of Discrete Event Dynamic Systems. Boston, Kluwer Academic Publisher.

HO, Y. C. – LI, S. (1988): Extension of Infinitesimal Perturbation Analysis, *IEEE Transactions on Automatic Control*, Vol. 33, No. 5, pp. 427–438.

HO, Y. C. – CAO, X. R. – CASSANDRAS, C. (1983): Infinitesimal and Finite Perturbation Analysis for Queuing Networks, *Automatica*, Vol. 19, No. 4, pp. 439–445.

KIM, Y. D. – YANO, C. A. (1983): Heuristic Approaches for Loading Problems in Flexible Manufacturing Systems, *IIE Transactions*, Vol. 25, No. 1, pp. 26–39.

KOBAYASHI, H. – GERLA, M. (1983): Optimal Routing in Closed Queuing Networks, *ACM Transactions on Computer Systems*, Vol. 1, No. 4, pp. 294–310.

KOLTAI, T. – LARRAÑETA, J. A. – ONIEVA, L. (1993): Examination of the Sensitivity of an Operation Schedule with Perturbation Analysis, *International Journal of Production Research*, Vol. 31, No. 12, pp. 2777–2787.

KOLTAI, T. – LARRAÑETA, J. A. – ONIEVA, L. – LOZANO, S. (1994): An Operations Management Approach to Perturbation Analysis, *Belgian Journal of Operations Research, Statistics and Computer Science*, Vol. 33, No. 4, pp. 17–41.

LAW, A. M. – KELTON, D. W. (1991): Simulation Modelling and Analysis. New York, McGraw-Hill.

PEDGEN, D.C. – SHANNON, R. E. – SADOWSKI, R. P. (1990): Introduction to Simulation Using SIMAN. New York, McGraw-Hill.

REISER, M. – LAVENBERG, S. S. (1980): Mean-value Analysis of Closed Multichain Queuing Networks, *Journal of the AMC*, Vol. 27, No. 2, pp. 313–322.

SHANKER, K. – TZEN, Y. J. (1985): A Loading and Dispatching Problem in a Random Flexible Manufacturing System, *International Journal of Production Research*, Vol. 23, No. 3, pp. 579–595.

STECKE, K. E. (1983): Formulation and Solution of Nonlinear Integer Production Planning Problems for Flexible Manufacturing Systems, *Management Science*, Vol. 29, No. 3, pp. 273–287.

STECKE, K. E. AND MORIN, T. L. (1985): The Optimality of Balancing Workloads in Certain Types of Flexible Manufacturing Systems, *European Journal of Operational Research*, Vol. 20, pp. 68–82.

STECKE, K. E. – SOLBERG, J. J. (1985a): The Optimality of Unbalancing both Workloads and Machine Group Sizes in Closed Queuing Networks of Multiserver Queues, *Operations Research*, Vol. 33, No. 4, pp. 882–910.

STECKE, K. E. (1986): On the Nonconcavity of throughput in Certain Closed queuing Networks, *Performance Evaluation*, Vol. 6, pp. 293–305.

STECKE, K. E. (1986): A Hierarchical Approach to Solving Machine Grouping and Loading Problems of Flexible Manufacturing Systems, *European Journal of Operational Research*, Vol. 24, pp. 369–378.

STEUDEL, H. J. – BERG, L. E. (1986): Evaluating the Impact of Flexible Manufacturing Cells via Computer Simulation, *Large Scale Systems*, Vol. 11, pp. 121–130.

SURI, R. (1983): Robustness of Queuing Network Formulas, *Journal of the ACMachinery*, Vol. 30, No. 3, pp. 565–594.

SURI, R. (1983a): Implementation of Sensitivity Calculations on a Monte Carlo Experiment, *Journal of Optimization Theory and Applications*, Vol. 40, No. 4, pp. 625–630.

SURI, R. – HILDEBRANT, R. R. (1984): Modelling Flexible Manufacturing Systems using Mean-value Analysis, *Journal of Manufacturing Systems*, Vol. 3, No. 1, pp. 27–38.

SURI, R. – DILLE, W. L. (1985): A Technique for On-line Sensitivity Analysis of Flexible Manufacturing Systems, *Annals of Operations Research*, Vol. 3, pp. 381–391.

YAO, D. D. (1985): Some Properties of the Throughput Function of Closed Networks of Queues, *Operations Research Letters*, Vol. 3, No. 6, pp. 313–317.

YAO, D. D. – BUZACOTT, J. A. (1986): The Exponentialization Approach to Flexible Manufacturing System Models with General Processing Times, *European Journal of Operational Research*, Vol. 24, pp. 410–416.

YAO, D. D. – BUZACOTT, J. A. (1987): Modeling a Class of Flexible Manufacturing Systems with Reversible Routing, *Operations Research*, Vol. 35, No. 1, pp. 87–93.

VAN LOOVEREN, A. J. – GELDERS, L. F. – VAN WASSENHOVE, L. N. (1986): A Review of FMS Planning Models, in: A. Kusiak (Ed.) Modelling and Design of FMS., pp. 3–31, Amsterdam, Elsevier.