NUMERICAL METHODS OF STRUCTURE OPTIMIZATION OF HOMOGENEOUS QUEUING NETWORKS

Boris V. Sokolov (Dr. Sci., Prof.), Sergey Kokorin

Saint-Petersburg Institute for Informatics and Automation of RAS, Information Technologies for Systems Analysis and Modeling laboratory, 39, 14 line, Saint-Petersburg, Russia

sokol@iias.spb.su (Boris V. Sokolov)

Abstract

An approach to an implementation of the numerical optimization for a structure of homogeneous queuing networks is considered. The restrictions on optimizing parameters of different nature is embedded into the algorithm. Examples for using the method for efficiency and fail-safety of networks are presented. The usage of optimization for network degradation is included.

Keywords: Queuing networks, numerical optimization, restrictions, efficiency, fail-safety.

Presenting Author's biography

BORIS V. SOKOLOV was born in Leningrad (now Saint-Petersburg), Russia in 1951. He obtained his main degrees in Mozhaisky Space Engineering Academy, Leningrad. MS in Automation Control Systems of Space Vehicles in 1974. Candidate of Technical Sciences subject the area of planning automation and decision making in 1982. Doctor of Technical Sciences subjects the area of military cybernetics, mathematical modeling and methods in military research. Professional Interests: Basic and applied research in mathematical modeling and mathematical methods in scientific research, optimal control theory, mathematical models and methods of support and decision making in complex organization-technical systems under uncertainties and multi-criteria. At present he is a Deputy Director for research of St.-Petersburg Institute for Informatics and Automation. His e-mail address is: sokol@iias.spb.su and his Web-page can be found at http://www.spiiras-grom.ru.

1 Introduction

Calculation and analysis of the characteristics of queuing networks is one of the main directions of research in the field of analytical and simulation modeling. The optimum choice of those or other characteristics is often beyond the scope of formulating and solving the classical problems of analysis for systems and networks, or considered in relation to the simplest models: an analytic solutions exist for the considering problems (an exponential distribution of service time and interarrival time of orders [1]) . In this paper we consider a broader class of models of the queuing theory, allowing an arbitrary service time distribution at the nodes in the task of searching for optimal parameters of the network.

2 Queuing network properties calculation

The calculation of open homogeneous networks, including the conversion of flows, in the general case includes the following steps [2]:

The input data are approximated by the method of moments with exponential (M), or hyper-exponential (phase, H) distributions, depending on the features of the model itself.

Initial approximations for flows of orders emerging from each network nodes is set λ_i. These rates are taken from the solution of the system flows balance, defined below:

$$\lambda_i = \Lambda r_{0i} + \sum_{j=1}^{M} \lambda_j r_{ij}, \quad i = 1, \dots, M. \qquad (1)$$

where Λ – total intensity of the flow from external sources, r_{ij} – elements of the matrix R (stationary transition probabilities between nodes), M – number of network nodes. The source of orders will be referenced as 0 node and sink as the node number M+1. A similar but more complex system of equations could be written for a situation, when input flow is approximated with hyper-exponential distribution (H).

The limitation of the numerical methods for the calculation of stationary characteristics of the network, namely the lack of the overload for each of the network nodes should be taken into account:

$$\lambda_i b_{i1} / n(i) < 1 \qquad (2)$$

Non-Markovian coefficients of interarrival intervals is set to $\xi_i = 0, \quad i = 1, \dots, M.$

For nodes $i = 1, \dots, M$:

- calculation of the thinned flows on the output of an each node to the input of i-th node;
- summation of all this thinned flows on the input of an each node;
- Recalculation of non-Markovian coefficients ξ'_i of the i-th node and defining of the preciseness coefficient $\Delta_i = |\xi_i' - \xi_i|$;
- calculation of the node as an isolated queuing system of the given type;
- recalculation of the flow emerging from the i-th node;
- if $\max_i \Delta_i > \epsilon, \quad i = \{1, \dots, M\}$ return to the start of the algorithm.

The calculation of moments of the sojourn time distribution of orders in the node at each visit.

Calculation of moments of the the sojourn time distribution of orders in the network as a whole.

Using this algorithm, we can calculate the following network features: moments of the sojourn time distribution of orders in the network $(\nu_1, \dots, \nu_m)$, moments of waiting time for an each node $(\mu_{11}, \dots, \mu_{Mm})$ and coefficient of loading for an each node of the network $(\rho_1, \dots, \rho_M)$.

3 Optimization methods

We are going to investigate the influence of the structure of a homogeneous network at its capacity. In this case the controlled parameters are the elements of the matrix of transition probabilities R, as well as uncontrolled parameters are all other characteristics of the network such as the number of channels and distributions of the service time of orders in each node. We consider two functions, that determine the effectiveness and the network resiliency, respectively, as objective functions:

$$\nu_1(R) \to \min_{R \in \Omega}, \qquad (3)$$

where Ω – set of admissible values of the matrix of stationary transition probabilities and $\nu_1(R)$ is the mean sojourn time for the network with the given transition matrix.

$$1/M \sum_{i=1}^{M} \rho_i^2 - 1/M^2 \left(\sum_{i=1}^{M} \rho_i\right)^2 \to \min_{R \in \Omega}. \qquad (4)$$

The formula (3) can be interpreted as minimizing the average transit time of the order through the queuing network, and formula (4), as the uniformity of loading between the available network nodes, we denote this objective function as $\rho(R)$.

For optimization of the objective functions described earlier in the paper are used two methods: global search method – the method of psi-transform [3], and a method of numerical optimization without

calculating derivatives - the method of principal axes of Brent [4].

The method of psi-conversion is a method for searching the global extremum of the objective function. It is not critical to the choice of an initial approximation, but requires of significant computational resources in the case when the dimension of parameters to be optimized is increasing. We chose a probability measure on the set of modifiable parameters which the value of a given objective function above a predetermined levels as a psi-function at. Thus, the problem of optimization reduces to finding a solution to the equation with many variables (parameters to be optimized). Using this method as an independent method of optimization often yields results of the very low accuracy.

The algorithm of the method is stated below:

1. The estimation of the spread of values of the objective function by the random test.
2. Choose of the value levels $\nu_1(R) \geqslant \zeta_l, \quad l \in \{1, \ldots, L\}$
3. Calculation of mean values of the objective function for each level by the means of the random test.

$$\Psi_l = 1/n \sum_{\{R^{(k)}:\nu_1(R^{(k)}) \geqslant \zeta_l\}} (\nu_1(R^{(k)}) - \zeta_l), \quad (5)$$

n – number of random generations, $R^{(k)}$ – parameters of the k-th generation, $l \in \{1, \ldots, L\}$.

4. Calculation of the mean values for optimizing parameters for each level.

$$x_{ijl} = n/\Psi_l \sum_{R^{(k)}:\nu_1(R^{(k)}) \geqslant \zeta_l} r_{ij}^{(k)} (\nu_1(R^{(k)}) - \zeta_l)^{\alpha}. \quad (6)$$

5. The parabolic approximation and the extrapolation to the level 0.

Because of the high computational complexity of the method of psi-transformation and its lack of the precision while solving problems of large dimensions is proposed the method of principal axes of Brent. This method focuses on local optimization of functions of several variables without calculating derivatives. In practice, the method proved effective in the solving problems of optimization of network structure, including, for the case when the parameters' space has a large dimension and fairly tight restrictions, the introduction of possible restrictions will be described below. The main drawback of the algorithm, it implements, is the need to specify the initial approximations, which should be calculated for an each task separately. In this case, the algorithm is characterized by two main parameters: the index of accuracy of the target function and the magnitude of step changes in parameters to be optimized. The first of these determines the moment to stop the iterative process, the second determines the rate of convergence of the algorithm.

The algorithm of the method is stated below:

1. Calculation of initial approximation $R^{(0)}$.
2. The initial directions is defining $U^0 = \{(u_i)^{(0)}\}_{i=1}^{N(N+2)} = I$, where I – the identity matrix of the given dimension.
3. Alternately, the optimal value is searched along each direction.
4. The direction vector with minimal index is dropped and new vector is substituted in the end of direction matrix $R^{(N(N+2))} - R^{(0)}$.
5. After a complete change of a set of directional vectors $U^{N(N+2)}$, a set of directional vectors is replacing with the orthogonal matrix that approximates the Hessian of the objective function in the point of current value.
6. Steps 3 – 5 is repeating till the achieving the given factor of preciseness.

For choosing and a determined initial approximation for the implementation of the method of principal axes is proposed the using of Jackson network model [4]. There is an analytic solution of the problem for it.

Multi-channel nodes are replaced with single-channel ones with proportionally increased intensity $\mu'_i = 1/b_{il} \cdot n(i)$, where b_{il} – first moment of service time distribution in the node, $n(i)$ – the number of channels in the i-th node. The exponential service time distribution at each node is approximated by taking into account the first moment of a given distribution. Sequential bypassing of the network in width allowed to choose the initial estimates for each node, according to the solving of the simple maximization problem:

$$\sum_{j=1}^{M+1} \frac{r_{ij}}{\mu'_j - r_{ij}\lambda_j} \to \min_i, \quad i = \{1, \ldots, M\}. \quad (7)$$

The usage of this method of choosing the initial estimates is effective in the case of a small number of channels at the nodes and a low value of the coefficient of variance of service time in the nodes. As an alternative to this method is the usage of the psi-transformation method: its is preferable for networks where the service time distribution differs significantly from the exponential. (For example, the uniform distribution on an interval or a gamma distribution with high coefficient of variance.)

Tab. 1: Initial estimations comparison

	Det. method		Method Ψ		
Node num	value	It. num	value	It. num	Opt. value
3	2.91	319	2.91	651	2.18
5	3.35	895	3.57	1293	3.19
8	4.9	6067	4.83	5419	4.53
13	7.02	21939	7.92	41345	6.57
21	9.05	212939	8.84	201323	8.64

Table 1 shows the comparative initial values estimates for the mean of the distribution of the sojourn time of orders on the network for different numbers of nodes, the objective function with initial estimates obtained using the deterministic method and the method of psi-transformation, the number of calls of the objective function for achieving the optimal value by means of the method of principal axes and the value of the objective function after optimization.

4 Constraints

Consideration of the set of allowable values Ω includes technical constraints $\sum_{j=0}^{j=M+1}(r_{ij})=1, \ \forall i\in\{0,\dots,M+1\}$, thus, $(r_{i0},\dots,r_{i,M+1})$ – discrete distribution. The same reasons implies that not all nodes are connected with each other, so many elements have the restriction of the form $r_{ij}=0$. These limitations are taken into account while the calculation of the optimization method. It is assumed that we have the right to control not the whole matrix R, but only a certain subset of its elements. In addition, we can impose additional restrictions on the parameters of the matrix R, as follows:

$$\mu'(x,R)=I(x\in\Omega')\cdot\mu(R), \quad (8)$$

where $I(\cdot)$ is an indicator of a set membership Ω', defined by the user in an arbitrary way.

Constraints allow to simulate natural restrictions associated to the technical features of the adequate use of the simulated system, as well as considering additional claims regarding the behavior of the system.

We consider the constraints of following types:

- $a\leqslant r_{ij}\leqslant A$, where a, A – arbitrary constants. We do not check the consistency of constraints for different r_{ij}. In the case when constraints are not consistent, method won't find any suitable solution.
- $a\leqslant\rho_i\leqslant A$, where a, A – arbitrary constants, ρ_i – loading coefficient for node i.

The imposition of constraints on the loading coefficients of nodes can decrease the chance of deterioration of equipment while working at the limit of allowable capacity and reservation additional "durability" for execution in emergency situations.

The imposition of restrictions on the coefficients of the matrix of stationary transition probabilities can reflect the differences in routes and features of transport links between nodes.

In some cases, the use of more complex kind of penalty function, can significantly increase the speed of convergence, but this issue requires additional studies.

5 Degradation

An example of the applicability of the approach described for optimization of the network structure can serve as a rapid redistribution of the flows at the moment of the failure of any network nodes [5], when you need to analyze the current situation and propose an optimal structure of the new network to report on is full refusal of service.

Full refusal of service depends on the network structure (even failure of one node can cause that): the network may break up into independent subnetworks, it may cause nodes become overloading or cause a situation when orders can not access the network or leave it. Implemented algorithms can signal the occurrence of such situations. If such a situation did not happen, we construct a new network, corresponding to the original network without the failed node and try to optimize the matrix of stationary transition probabilities.

Construction of the subnetwork involves conversion of initial estimates, the transfer of existing constraints, the analysis of the matrix R by the following algorithm:

1. Remove from matrix R the row and column corresponding to failed node.
2. If the new matrix contains nodes, such that all incoming flows for them are 0, then the node is marked as failed and the algorithm goes back to 1.
3. Checking that there is at least one element for which $r_{i,M+1}>0$, also searching for the element for which $r_{0,j}>0$.
4. For all nodes we are checking the condition of no overloading (2).

The optimization is carried out for the resulting network as for a new one with a given structure.

6 Simulations

As an example we introduce the network of fairly simple structure: Optimization of the network can be done only on elements of the matrix R, highlighted in Fig. 1 by black arrows.

The structure of this network is symmetrical, so the degradation is sufficient to consider only the cases where the nodes in the network is failed: 1, 2, 4, 5, 7, as the failure of the node 3 is equivalent to the failure of the node 1, 4 – 6, 7 – 8.

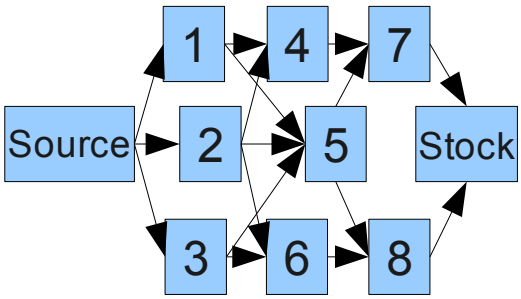

Fig. 1: Network example

As an example (fig. 2) shows a graph of the mean sojourn time of orders in the network for the case when the loading coefficient of the node 5 is limited to 0.2. On the *x-axis* plotted numbers of failed nodes. The blue line indicated the mean sojourn time for orders in the original (with all nodes in service) network.

Orange line shows the deterioration of the network capacity while one node is failed and the optimal transition matrix, calculated with methods described above, is used. As expected, the nodes 7 and 8 are the most critical in terms of sustainability of the system, while the failure of nodes 1, 2, 3, has almost no effect on the efficiency of the system.

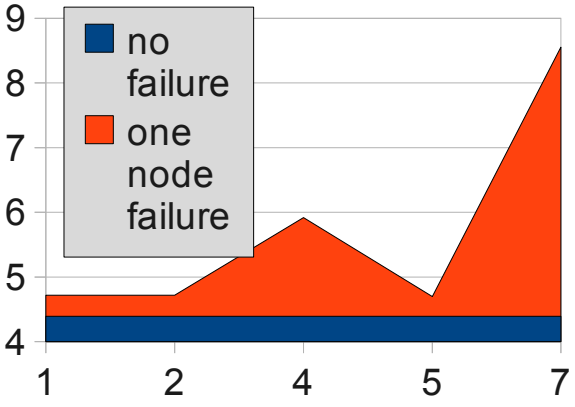

Fig. 2: Optimization of network with failures

7 Conclusions

We represent a numerical approach for a problem of optimization of characteristics of homogeneous queuing networks. It's fast enough to deal with networks of the hundreds of nodes. It's general to include different types of service time and interarrival distributions. This paper includes the certain example of implementation of the approach to the optimization of transition matrix in cases of nodes failure for redistributing the loading among other nodes in service. The approach allows minimize the losses while repairing period. The main benefits of the proposed combined techniques are interrelated with superposition of advantages of psi-transformation and Brent principal axis methods which compensate disadvantages of corresponding approaches.

This research was supported by the grants of Russian Foundation for Basic Research (RFBR) 10-07-00311, 09-07-00066a, 09-07-11004, 10-08-90027-Bel-a, 08-08-00346-a, 10-08-0906-a.

8 References

[1] Bramson M. Stability of queuing networks. *USA: Probability surveys*, 2008.

[2] Рыжиков Ю.И. Машинные методы расчёта систем массового обслуживания. *СПб: ВИКИ им А. Ф. Можайского*, 1979.

[3] Чичинадзе В.К. Решение невыпуклых нелинейных задач оптимизации. Метод пси-преобразования. *Москва: Наук*а, 1983.

[4] Brent R.P. Algorithms for minimization without derivatives. *Englewood Cliffs, NJ: Prentice-Hall, Inc.*, 1973.

[5] Охтилев М. Ю., Соколов Б. В., Юсупов Р. М. Интеллектуальные информационные технологии мониторинга и управления структурной динамикой сложных технических объектов. *Москва: Наука,* 2006.