MELIKOV A.Z., FATTAHOVA M.I.

PERFORMANCE ANALYSIS AND OPTIMIZATION OF BUFFER ALLOCATION STRATEGIES: A STATE SPACE MERGING APPROACH

Abstract

Quality-of-Service (QoS) in high speed packet switching networks is largely determined by buffer size and buffer allocation strategies. Performance evaluations of the buffer allocation strategies are computationally difficult problems due to the complexity of the large state space when the number of traffics and/or the buffer size is large. In this paper, we propose the approach based on the state space merging to avoid these difficulties for the systems supporting two traffic flows when buffer size is large enough. The design and optimization problems are discussed for Complete Sharing (CS) strategy more detail and the results of appropriate numerical experiments are carried out. The objective function is to achieve the desirable level of the blocking (loss) probability (PB) under minimal value of the buffer size.

Keywords: packet switching networks, non-push-out strategies, state space merging, analyze and optimization algorithms

1. Introduction.

To evaluate the congestion of traffic at a computer network node in [4], Irland proposed to use the models of multi-stream queuing systems with finite common waiting room and typed channels in which each stream has its own channels. After this work, these models have successfully been used for the analysis of the performance of buffer sharing strategies at a node in a store-and-forward packet switching networks. To obtain optimal system performance of specific sharing strategies, Irland [4] and Latouche [7] developed some heuristic procedures. Performance evaluation of the buffer allocation strategies is computationally a difficult problem due to the complexity of the large state space when the number of traffics and/or the buffer size is large. These problems have intensively been investigated during the recent two decades, especially after the publication of the classical study of Kamoun and Kleinrock [5], where five strategies were proposed. Their showed that for the Poisson arrivals and exponential service times the probability distribution of the buffer occupancy have a well-known product form.

Buffer allocation strategies can be broadly classified into push-out strategies and non-push-out strategies. Strategies, which can accept an arriving packet by dropping another packet from the buffer, are known as push-out strategies. In this paper, we consider the non-push-out type strategies, which do not allow the drop of already accepted packet of any type. Note that the above mentioned five strategies [5] are strategies of non-push-out type. Further references might be found in [3].

There have been a few works in which the finding of the optimal strategy in the class of non-push-out strategies have been addressed. In [1], Foschini and Gopinath proved that in case two output ports for the Markovian systems, the optimal sharing strategy in the sense of minimization of the *PB* (or equivalently of maximization of the throughput) is in the class of

SMXQ strategies. They do not discuss the computational aspects of finding the parameters of the optimal strategy. Yum and Dou [9] proposed algorithms for finding the optimal (minimal) value of parameters of the above mentioned five strategies [5] to support the given level of multiple QoS classes when the latter is defined by indicating the upper bounds of the *PB* for each class of packets. In fact, none of researchers did not deal with the computational complexity the problems due to large value of buffer size and/or the number of types of packets. However, until now the major problem in the performance evaluation of the large scale buffers remains to be the "curse of dimensionality". In this paper, we propose a new approach to avoid this problem for the non-push-out type buffer allocation strategies in case two output ports with large buffer size. Our approach is based on state space merging [6]. Previosly this approach was presented in [8].

2. Model and Algorithms.

The system consists of a buffer shared by packets destined to two output ports. Packets are said to be of type *i*, if they are destined to port $i, i=1,2$. Type *i* packets arrive to the buffer according to a Poisson process with finite rate λ_i , and are transmitted by output port *i* during the random time derived from the exponential distribution with finite rate μ_i. We assume that arriving and transmission processes are mutually independent. The total buffer size is *B*, and a packet releases the buffer when it has completely been transmitted.

Assumption: $\lambda_2 >> \lambda_1$ and $\mu_2 >> \mu_1$. This is a regime that commonly occurs in multimedia networks, in which voice calls arrive and depart more frequently than video calls [2]. Moreover, as it will be shown below, the final results do not depend directly on λ_i and μ_i but depend only on their ratio $v_i = \lambda_i / \mu_i$, $i = 1,2$. However, this assumption is useful to provide rare transition between classes of states underlying Markov chain (MC) for the application of the state space merging algorithms.

As it was mentioned in previous section, for this strategy an arriving packet any type is accepted if any storage space is available. Thus, a two dimensional MC with states $\boldsymbol{n} = (n_1, n_2)$, where n_i is the number of the type *i* packet in the buffer might be used to describe the functioning of the system at equilibrium. The state space E_{CS} of the given MC is defined as follows:

$$E_{CS} := \{\boldsymbol{n} : 0 \le n_i \le B, i = 1,2; 0 \le n_1 + n_2 \le B\}. \tag{1}$$

The elements $q(\boldsymbol{n}, \boldsymbol{n}'), \boldsymbol{n}, \boldsymbol{n}' \in E_{CS}$, of the infinitesimal generator matrix Q_{CS} of the given MC are calculated as follows:

$$q(\boldsymbol{n}, \boldsymbol{n}') = \begin{cases} \lambda_i & \text{if } \boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{e}_i, \\ \mu_i & \text{if } \boldsymbol{n}' = \boldsymbol{n} - \boldsymbol{e}_i, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $i = 1, 2$, $\boldsymbol{e}_1 = (1, 0)$, $\boldsymbol{e}_2 = (0, 1)$.

Let $p(\boldsymbol{n})$ denotes the stationary probability of the state $\boldsymbol{n} \in E_{CS}$. It is well known that the stationary distribution $p(\boldsymbol{n}), \boldsymbol{n} \in E_{CS}$, of this MC possesses a product form [5]:

$$p(\boldsymbol{n}) = G_{CS}^{-1}(B)\prod_{i=1}^{2}\nu_i^{n_i}, \qquad (3)$$

where $G_{CS}(B)$ is called normalizing constant over state space E_{CS},

$$G_{CS}(B) := \sum_{\boldsymbol{n}\in E_{CS}}\prod_{i=1}^{2}\nu_i^{n_i}.$$

Various performance measures can be calculated from the stationary distribution (3). Hence, the stationary blocking probabilities for packets both types are the same:

$$PB_{CS}(B) := \sum_{\boldsymbol{n}\in E_{CS}} p(\boldsymbol{n}) I(n_1 + n_2 = B), \qquad (4)$$

where $I(A)$ is an indicator function of the event A.

But for other performance measures of the system, that is, throughput and utilization are not the same. So, the throughput $TH_{CS}^{i}(B)$ is calculated by

$$TH_{CS}^{i}(B) = \lambda_i[1 - PB_{CS}(B)], \; i = 1, 2 \quad , \qquad (5)$$

where $TH_{CS}^{i}(B)$ denotes the number of completed packets of type i accepted to the buffer per unit time. The utilization, $PU_{CS}^{i}(B)$, of the output port i is a measure fraction of the time that the port is busy and is calculated as follows:

$$PU_{CS}^{i}(B) := 1 - \sum_{\boldsymbol{n}\in E_{CS}} p(\boldsymbol{n}) I(n_i = 0), \quad i = 1, 2. \qquad (6)$$

The buffer utilization by packets of type i, $BU_{CS}^{i}(B)$, is measured by means average number of packets of given type in buffer and is calculated as follows:

$$BU_{CS}^{i}(B) := \sum_{\boldsymbol{n}\in E_{CS}} \sum_{j=1}^{B} j p(\boldsymbol{n}) I(n_i = j) \; . \qquad (7)$$

However, when B is large (as in realistic switches), the computational burden in calculation of the performance measures (4)- (7) is great due to large scale of the state space (1). Now, we propose a new approach to avoid this difficulty.

Consider the following splitting of the E_{CS}:

$$E_{CS} = \bigcup_{m=0}^{B} E_{CS}^{m} \quad , \qquad E_{CS}^{m} \bigcap E_{CS}^{m'} = \varnothing \quad , m \neq m', \qquad (8)$$

where $E_{CS}^{m} := \{ \boldsymbol{n} \in E_{CS} : n_1 = m \}$.

Name the class of states E_{CS}^{m} of the initial MC as the merged state and denote it as $<m>$, $m = 0, 1, \ldots, B$. Starting from the splitting (8), we can form the mergement function $U : E_{CS} \to \hat{E}_{CS}$, $\hat{E}_{CS} := \{ <m> : \; m = 0, 1, \ldots, B \}$ defined as

$$U(\boldsymbol{n}) = <m> \text{ if } \boldsymbol{n} \in E_{CS}^{m} \; . \qquad (9)$$

Thus, mergement function (9) defines the merged model with respect to the initial model which is also MC with state space $\hat{E}_{CS}$. In view of the above assumption, the classes E_{CS}^{m} , $m = 0, 1, \ldots, B$, are the ergodic classes supported by the initial MC. By using (2), their stationary distribution $\rho^{m} = \{ \rho^{m}(\boldsymbol{n}) : \boldsymbol{n} \in E_{CS}^{m} \}$ can easily be determined as the stationary distribution of the one-dimensional birth and death processes describing the classical $M/M/1/B-m$:

$$\rho^m(m,k)=\nu_2^k\frac{1-\nu_2}{1-\nu_2^{B+1-m}},\quad k=0,1,\dots,B-m,\quad m=0,1,\dots,B. \tag{10}$$

Taking into account (2) and (10), we conclude that the elements $\hat{q}(x,y),\ x,y\in\hat{E}_{CS}$ of the generator matrix $\hat{Q}_{CS}$ for the merging model, which also is one-dimensional birth and death process, are :

$$\hat{q}(x,y)=\begin{cases}\lambda_1[1-L(\nu_2,m)] & \text{if} \quad x=<m>,\ y=<m+1>\\ \mu_1 & \text{if} \quad x=<m+1>,\ y=<m>\\ 0 & \text{otherwise}\end{cases} \tag{11}$$

where $L(\nu_2,m)$ denotes the stationary loss probability in $M/M/1/m+1$ system with the offered load ν_2, that is $L(\nu_2,m)=\nu_2^m(1-\nu_2)/(1-\nu_2^{m+1})$.

Therefore, by using (11) the stationary distribution $\pi_c(<m>),\ <m>\in\hat{E}_{CS}$, of the merging model can be written as follows:

$$\pi_c(<m>)=\nu_1^m\prod_{j=B-m+1}^{B}[1-L(\nu_2,j)]\,\pi(<0>),\quad m=1,\dots,B, \tag{12}$$

where $\pi_c(<0>)=[1+\sum_{k=1}^{B}\nu_1^k\prod_{j=B-k+1}^{B}(1-L(\nu_2,j)]^{-1}$.

Then, the stationary distribution of the initial model is

$$p(\boldsymbol{n})\approx\rho^{n_1}(n_1,n_2)\pi_c(<n_1>),\quad \boldsymbol{n}\in E_{CS}^{n_1}. \tag{13}$$

Finally, taking into account (5)-(13), we obtain the following formulae to calculate the performance measures (4)- (7):

$$PB_{CS}(B)=\sum_{m=0}^{B}L(\nu_2,m)\pi_c(<B-m>), \tag{14}$$

$$U_{CS}^1(B)=1-\pi_c(<0>),\ PU_{CS}^2(B)=1-\sum_{m=0}^{B}L_0(\nu_2,m)\pi_c(<B-m>), \tag{15}$$

$$BU_{CS}^1(B)=\sum_{i=1}^{B}i\pi(<i>),\qquad BU_{CS}^2(B)=(1-\nu_2)\sum_{i=1}^{B}i\nu_2^i\sum_{j=i+1}^{B+1}\frac{1}{1-\nu_2^j}\pi(<B+1-j>), \tag{16}$$

where $L_0(\nu_2,m)$ is the stationary probability that the system is empty in model $M/M/1/m+1$ that is, $L_0(\nu_2,m)=(1-\nu_2)/(1-\nu_2^{m+1})$.

Using (14) $T_{CS}^i(B)$ might be calculated from (5).

In summary, we can propose the following algorithm to calculate the performance measures of the given system.

Step 1. Input ν_1, ν_2, B.

Step 2. For $m=0,1,\dots,B$ calculate $\pi_c(<m>)$ from (12).

Step 3. Calculate $PB_{CS}(B)$, $PU_{CS}^i(B)$, $BU_{CS}^i(B)$ and $TH_{CS}^i(B)$ from (14), (15), (16) and (4), respectively.

It should be noted that the complexity of the given algorithm is too low, that is, it can be estimate as $O(B^2)$. Moreover, it is very convenient to calculate because it applies well-known parameters $L(v, m)$ and $L_0(v, m)$ for which there exists a perfect software product (also commonly tabulated).

3. Numerical Results.

In order to show the numerical tractability and some results for calculate the system performance measures in CS strategy , we have solved the set of equations (12)-(16), assuming both the values of v_1 and v_2 are change in the range $0<v_i<1$, $i=1,2$ (i.e. in the normal regime). It is worthwhile to note that the proposed here formulae enable us to realize appropriate calculations in heavy traffic regime (i.e. when loading parameters exceed 1).

Fig. 1-3 show the some results for the symmetric ($v_1=v_2$) and asymmetric ($v_1\neq v_2$) cases. It is noteworthy that difference between approximate values of performance measures calculated by proposed here formulae and their exact values calculated by product form stationary distribution (i.e. by formulae (3)-(7)) is negligible. So, for instance, the values of $PB_{CS}(15)$ and $PB_{CS}(90)$ for $v_1=v_2=0.5$ in exact approach are *1.22087E-04* and *1.83773E-26* , respectively. By comparision these values of $PB_{CS}(B)$ with their appropriate ones in Fig.1 we conclude that with increasing of B this difference is close to zero. Analogous situations there are in other values of buffer size (B) and loading parameters (v_1, v_2). Latter indicate that proposed formulae are asymptotic exact ones for calculating the systems performance measures for large value of B.

Now we consider the optimization results for *CS* strategy. It is clear that for this strategy, the problems of minimization of blocking probability and maximization of throughput are equivalents. That's why, for the sake concretize we assume that QoS is given by indicating the upper limit of blocking probability, and our problem is to find the minimal value of the buffer size that provides the desirable level of QoS. That is, formally this problem may be written as follows:

$$B^* := arg\min_B \{ PB_{CS}(B) \le \varepsilon \}, \tag{17}$$

where $\varepsilon > 0$ is a given upper limit of blocking probability. To solve the problem (17) as in [9], we may use the dichotomous search method since the function $PB_{CS}(B)$ is monotonic decreasing with respect to B. For this purpose, we first should determine the upper and the lower limits for B^* in (17). From (14) it is clear that $PB_{CS}(B) \ge L(v_2, B)$. It means that as a lower limit of B^* in (17) we may use $\underline{B}_2^*$ which is determined by the following condition:

$$\underline{B}_2^* := arg\min_B \{L(v_2, B) \le \varepsilon\}. \tag{18}$$

The solution of the problem (18) is $\underline{B}_2^* = int(x)$ or $\underline{B}_2^* = int(x)+1$ (here $int(x)$ denotes the greatest integer less than or equal to x) if x is not integer, where x is obtained as root of the equation $L(v_2, x) = \varepsilon$, that is,

$$x = \frac{1}{\ln v_2}\left\{ \ln\frac{\varepsilon}{1-v_2} - \ln(1+v_2\frac{\varepsilon}{1-v_2}) \right\}.$$

It is clear that as an upper limit of B^* in (17) we may use $\overline{B}^*$ which is given by $\overline{B}^* = \underline{B}_1^* + \underline{B}_2^*$, where $\underline{B}_1^*$ is the solution of (18) where ν_1 is substituted by ν_2. The results of numerical experiments for the problem (18) carried out by using the indicated above method is shown in Table 1.

Table 1. Solution of the problem (17)

ν_1	0.1	0.9	0.1	0.9	0.1	0.9	0.5	0.5	0.5	0.5
ν_2	0.9	0.1	0.9	0.1	0.9	0.1	0.5	0.5	0.5	0.5
ε	10^{-3}	10^{-3}	10^{-5}	10^{-5}	10^{-6}	10^{-6}	10^{-5}	10^{-9}	10^{-15}	10^{-20}
B^*	44	44	88	88	105	105	19	33	54	70

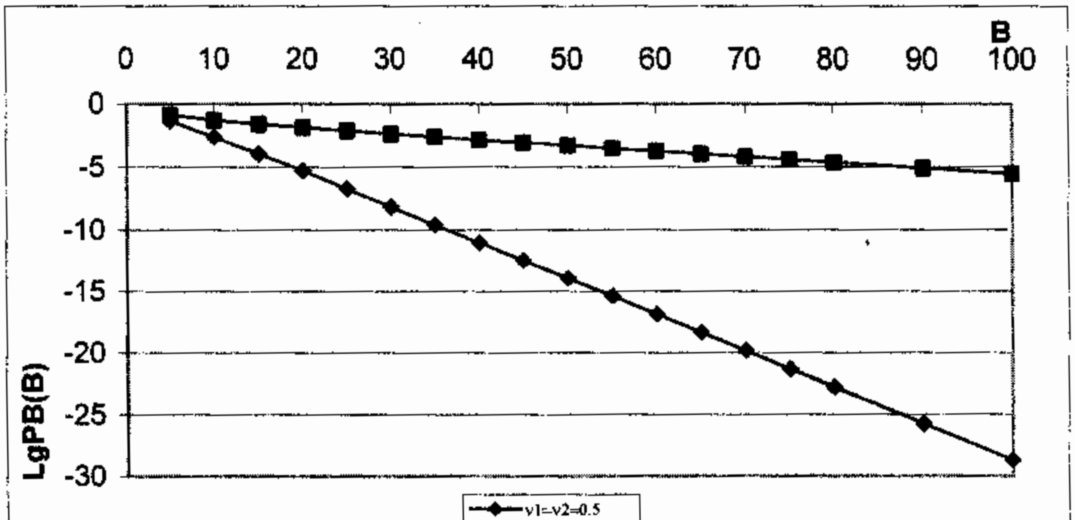

Fig.1. Plot for PB(B) versus B

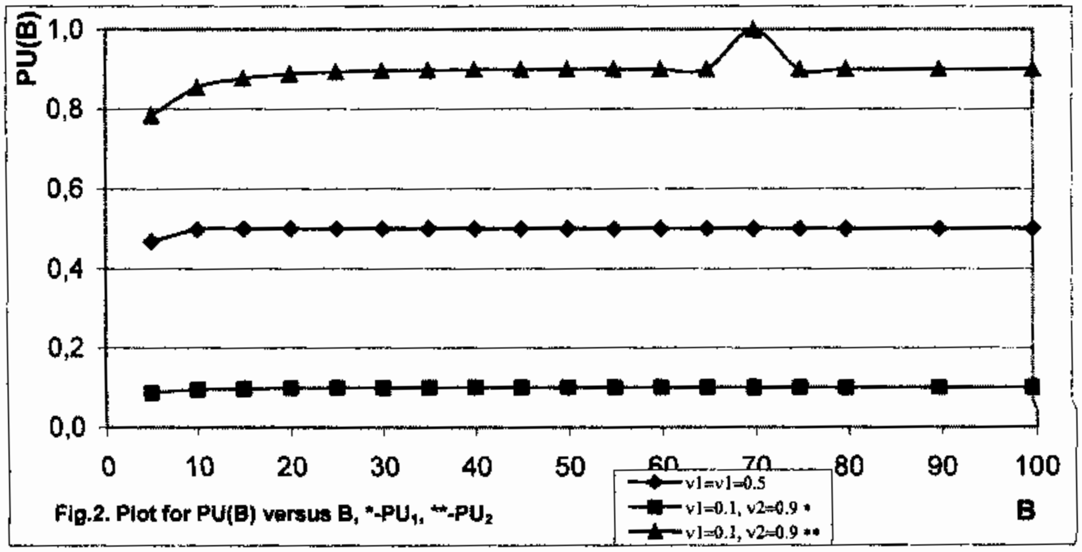

Fig.2. Plot for PU(B) versus B, *-PU_1, **-PU_2

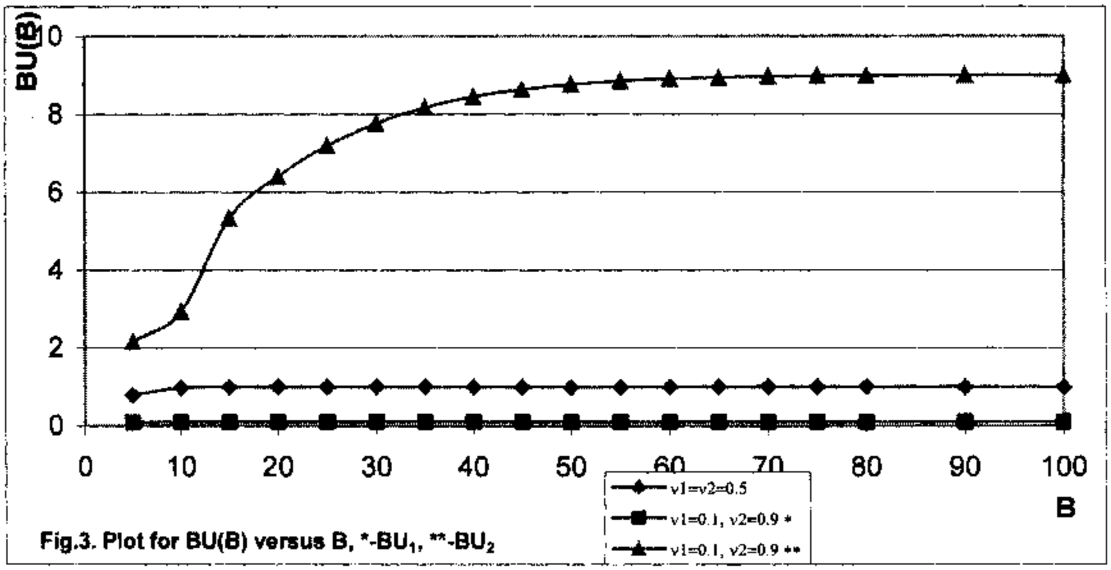

Fig.3. Plot for BU(B) versus B, *-BU_1, **-BU_2

4. Conclusion.

In this paper, we propose a novel algorithmic approach to determine the approximate performance measures of non-push-out type allocation strategies widely used in packet switching networks in case of large-scale buffers. Our approach is based on the state space merging principle and we are develcp efficient algorithms in the sense of low complexity to solve of the problem indicated above for case two output ports. Calculations that are carried out by using the proposed formulae show that they are refined approximations for the system performance measures in engineering practice.

The results that are obtained by using the proposed algorithms show that by appropriate choice of the parameters of the *CS* strategy required QoS may be achieved (analogous problems for other strategies might be carried out also). The latter is carried out by unconstrained optimization of the sample model performance measures to determine the optimal value of the strategy parameter depending on the loads offered by packets of different types. Hence, the proposed approach also permits us to develop the situational algorithm for optimal control of the buffers under a given strategy provided that accurate measurement or prediction of offered loads could be carried out.

Meanwhile, this approach enables us to obtain the accurately appropriate solutions for push-out type strategies as well, if we let only the first type packets (slow time scale packets) may push out the second type packets (fast time scale packets) when the buffer is full since under this assumption, we may provide the rarely transmission between classes of states that is a necessary condition for applying the state space merging theory. Thus, we may study the interesting problems of analysis, optimization and optimal control of the push-out type strategies too and this issue will touch upon in our further studies. Another interesting research area is that the given approach may successfully be applied for models with arbitrary number of output ports.

References

[1]. Foschini G.J., Gopinath B. *Sharing memory optimally*, IEEE Trans.Commun. 31, 1983, p.352-360.
[2]. Greenberg A.G., Srikant R., Whitt W. *Resource sharing for book-ahead and instantaneous-request calls,* IEEE /ACM Trans.Networking, 7, 1999, p.10-22.
[3]. Guerin R., Peris V. *Quality-of-service in packet networks: Basic mechanism and directions*, Comp.Networks, 31, 1999, p.169-189.
[4]. Irland M.I. *Buffer management in a packet switch*, IEEE Trans.Commun. 26, 1978, p.328-337.
[5]. Kamoun F., Kleinrock L. *Analysis of shared storage in a computer network node environment under general traffic conditions*, IEEE Trans.Commun. 28, 1980, p.992-1003.
[6]. Korolyuk V.S., Korolyuk V.V. *Stochastic models of systems*, Dordrecht; Boston: Kluwer Academic Publishers, 1999.
[7]. Latouche G. *Exponential server sharing a finite storage: Comparison of space allocation policies*, IEEE Trans.Commun. 28, 1980, p.910-915.
[8]. Melikov A. Z.., Fattahova M.I. *Performance of allocation strategies for large scale buffers in a packet switching networks*, Proc. of Int. Electrical, Electronics and Comp. Eng. Symp., Cyprus, 23-25 May, 2001, p.237-239.
[9]. Yum T.P., Dou C. *Buffer allocation strategies with blocking requirements*, Perform.Eval. 4, 1984, p.285-295.

Agassi Z. Melikov, Mehriban I. Fattahova
Institute of Cybernetics of National Academy of Science of Azerbaijan.
9, F. Agaev str., 370141, Baku, Azerbaijan.
Tel.: 39-01-51.
E-mail: agassi@dcacs.ab.az

Received February 5, 2001; Revised August 27, 2001.
Translated by authors.

TAGIYEV R.K.

ON THE OPTIMAL CONTROL PROBLEM BY COEFFICIENTS OF A HYPERBOLIC EQUATION

Abstract

In the paper the optimal control problem by coefficients of a hyperbolic equation is investigated. The correctness questions of statement of the problem are investigated, the differentiability of an aim functional is proved and the expression for its gradient is obtained, the necessary optimality condition of a control is established.

The optimal control problems for the processes described by hyperbolic equations with controls in their coefficients are studied in [1-4] and etc. By investigation of these problems a series of difficulties associated with their incorrectness [1,5] arises.

In the present paper one optimal control problem by coefficients at the first derivatives and by solving a linear hyperbolic equation is investigated. For the considered optimal control problem the correctness questions of its statement are investigated, the differentiability of an aim functional is proved and the expression for its gradient is obtained, the necessary optimality condition of a control is established.

1. The statement of the problem and its correctness. Let it be required to minimize the functional

$$I(v)=\alpha_0\|u-y_0\|^2_{L_2(Q_T)}+\alpha_1\|u|_{t=T}-y_1\|^2_{L_2(\Omega)} \tag{1}$$

in the solutions $u=u(x,t)=u(x,t,v)$ of the boundary value problem

$$\frac{\partial^2 u}{\partial t^2}-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}(x,t)\frac{\partial u}{\partial x_j}\right)+\sum_{i=1}^{n}a_i(t)\frac{\partial u}{\partial x_i}+ \tag{2a}$$

$$+a(t)u=f(x,t),\ (x,t)\in Q_T,$$

$$u|_{t=0}=\varphi_0(x),\ \left.\frac{\partial u}{\partial t}\right|_{t=0}=\varphi_1(x),\ x\in\Omega,\ u|_{S_T}=0, \tag{2b}$$

corresponding to the admissible controls

$$v=v(t)\in V=\Big\{v(t)=(a_1(t),...,a_n(t),a(t))\in H\equiv\left(W_2^1[0,T]\right)^{n+1}:|a_i(t),a(t)|\le\mu_1,|a_i'(t),$$

$$a'(t)|\le\mu_2\ \left(i=\overline{1,n}\right)\ a.e.\ in\ [0,T]\Big\}, \tag{3}$$

where Ω is a bounded domain in R_n with a piecewise smooth boundary $S;Q_T=\Omega\times(0,T]$, $S_T=S\times(0,T]$; $T>0$, $\alpha_0,\alpha_1,\mu_2\ge 0$, $\mu_1>0$, $\alpha_0+\alpha_1>0$ are given numbers, $y_0=y_0(x,t)$, $f=f(x,t)\in L_2(Q_T)$, $y_1=y_1(x)$, $\varphi_1=\varphi_1(x)\in L_2(\Omega)$, $\varphi_0=\varphi_0(x)\in$ $\in\overset{\circ}{W}{}_2^1(\Omega)$ are given functions, $a_{ij}(x,t)$ are the known functions having the properties

$$a_{ij}(x,t)=a_{ji}(x,t),\nu\|\xi\|^2\le\sum_{i,j=1}^{n}a_{ij}(x,t)\xi_i\xi_j\le$$

$$\le\mu\|\xi\|^2,\ \forall\xi\in R_n,\ \left|\frac{\partial a_{ij}}{\partial t}\right|\le\mu_3\ \left(i,j=\overline{1,n}\right)$$

a.e. in Q_T, $\nu,\mu=const>0$, $\mu_3=const\ge 0$.