A profit optimizing strategy for congestion control in signaling networks

Stefan Pettersson and Åke Arvidsson

Dept. of Telecommunications and Mathematics,
University of Karlskrona/Ronneby,
S-371 79 Karlskrona, Sweden

Abstract

Congestion control in the signaling system number 7 (SS7) is a necessity to fulfil the requirements of a telecommunication network that satisfy customers' requirements on quality of service. Heavy network load is an important source of customer dissatisfaction as congested networks result in deteriorated quality of service.

With the introduction of a Congestion Control Mechanism (CCM), that annihilates service sessions with a predicted completion time greater than the maximum allowed completion time for the session, network performance improve dramatically. Annihilation of already delayed sessions let other sessions benefit and increase the overall network throughput.

This paper investigates the possibilities of using a decision theoretic approach that base the decision of annihilation on the average loss attached to each of the two possible actions, i.e. annihilate or not. Attributes are attached to each session describing the outcome of any performed CCM action, e.g. the economic loss connected with the annihilation of a session. The attributes are also used to calculate the network loss for a given network load.

The results in this paper indicate that the decision theoretic approach can decrease the network loss up to 40% for the improved CCM (ICCM) compared to an ordinary CCM.

1. Introduction.

1.1 The signaling network, introduction.

The operability of the signaling network is of prime importance in securing telecommunication network performance. A signaling network is engineered in such a fashion that normal load represents about 25-35% of maximum load, suggesting congestion to be very unlikely at normal working conditions. The introduction of mobile communication systems and intelligent networks (IN) has stimulated changes in the demands for signaling services [3]. For example, the hand over procedure in mobile communications must by definition be extremely fast. A mobile station, crossing cell boundaries at normal highway speed, has very little time to exchange essential information with the cellular network and thus perform the switch of base stations. Congestions or other disturbances in the network therefore can give rise to unwanted effects like dropped calls.

The traditional role of a CCM in the signaling network is to resolve an immediate overload situation in a link or a node by throttling the traffic with destination to the congested area without any regards to the impact on the surrounding network. This can sometimes cause congestion in other parts of the network.

A good CCM must be able to resolve the overload situation in such a manner that the entire network benefits. Further more, it must be able to foresee an emerging congestion, and to take adequate prophylactic steps in order to normalize the situation [4].

1.2 The signaling network, definitions.

The signaling network consists of a number of Signaling Points (nodes) and signaling Transfer Points (transit nodes) connected via Signaling Links (links) in a mesh structure [1, 2].

The control information is contained in Message Signal Units (signals), which may be regarded as packets in a packet switched network guided by a routing algorithm.

A service session comprise a sequence of signals, where the length depends on the character and outcome of the service requested.

A service's maximum time for completion is set by timers and system parameters in the signaling network. A service session can not be considered concluded until all signals have successfully reached their destination within the service's maximum allowed time for completion. A service session that exceeds its permitted completion time displeases the customer, increases the network load and deteriorates network performance.

1.3 A congestion control mechanism based on network delays.

The completion time of the session is a metric revealing the state of the network. The completion time of recently completed sessions contains valuable information for the node in order to detect any signs of congestion.

If knowledge of the completion time of a session could be obtained prior to the launch, it would be possible to prevent sessions, which are expected to be delayed, from getting started at all. The load in the congested part of the network is thereby reduced, increasing the probability for other sessions to become successful. The customer would not be aware of the difference between a delayed or an annihilated session, both will result in e.g. a lost call, but there is always a possibility to avoid customer dissatisfaction by sending a message to the customers mobile telephone indicating that the call set up was annihilated.

The estimation of the completion time T_{est} of a signaling session comprising k signals is calculated as

$$T_{est} = \sum_{d=1}^{k} P(i_d, j_d) \quad (1)$$

where $P(i_d, j_d)$ is the predicted completion time for the next signal d sent between the originating node i_d and the destination node j_d. The method is fully described in [6].

An investigation reveals strong correlation between T_{est} and the actual completion time of the session T_{act}, which means that it is possible to make a good prediction of the total completion time of a session.

A simple CCM that annihilates signaling sessions for which the prediction T_{est} is greater than the services maximum allowed completion time T_{max} is also described in [6].

1.4 Network loss as a simple metric.

One way of studying the yield of the CCM is to measure the network profit when applying the CCM to a congested network. A successful service session is profitable and increases operator income. Annihilated or delayed sessions are useless to the operator as they give rise to customer dissatisfaction and thereby decrease the profit. The network profit decrease rapidly in a congested network, which is shown in [7] where also the usage of network profit as a performance metric is described. A class-concept where a signaling session possess attributes like reward for a successful session and penalty for an unsuccessful session is also introduced in [7]. To harmonize with the theoretical background introduced in chapter 2, the loss concept is used in the sequel of this paper. An income, as described in [7], is now defined as a small loss-value and a cost is defined as a larger loss-value.

2. A decision theoretical approach.

2.1 Introduction.

The CCM described in 1.3 is based on the following comparison:

$T_{est} > T_{max}$	Annihilate the session.
$T_{est} < T_{max}$	Do not annihilate the session.

Where T_{max} is a service specific value describing the maximum allowed completion time for a session. The prediction T_{est} is estimated using a state machine and a memory function for each signaling link described in [6]. The prediction is associated with a small degree of uncertainty characterized by the standard deviation σ.

The objective of this paper is to investigate the possibility of refining the CCM, giving an improved CCM (ICCM), by using the uncertainty in the estimation of T_{est}, the loss attached to each action, and finally treating it as a statistical decision theoretic problem. The network loss described in 3.4 is used as a metric.

2.2 Preliminaries in decision theory.

Statistical decision theory describes problems where a decision has to be made on what kind of action a_j, selected from k possible actions, that should be taken. The decision depends on the value of a random variable Θ which is used calculating the *Bayes loss:*

$$B(a_j) \equiv E\,[w(\Theta, a_j)] = \sum_{\forall i} w(\theta_i, a_j) \cdot g_\Theta(\theta_i) \qquad (2)$$

which is the expectation of the loss function $w(\Theta, a_j)$ defined for the finite state space $\Omega = \{\theta_1, \theta_2, ..., \theta_l\}$ and for the finite action space $\Psi = \{a_1, a_2, \ldots, a_k\}$. The loss function describes the loss associated with any action a_j. The discrete probability distribution $g_\Theta(\theta_i) = P(\Theta = \theta_i)$ describes the probability associated with each state θ_i of the system. A *Bayes action* is the action a_B that minimizes *Bayes loss*, which means that a_B *is* the optimal action to be taken.

2.3 An ICCM using a decision theoretical approach.

Considering the problem described in 2.1, two actions are defined:

a_1 = annihilate a session.
a_2 = do not annihilate a session.

and two states:

θ_1 = The session will be completed within the stipulated time, $T_{act} \leq T_{max}$
θ_2 = The session will not be completed within the stipulated time, $T_{act} > T_{max}$

Defining $w(\Theta, a_j)$ for the four combinations of states and actions gives the following loss table:

	θ_1	θ_2
a_1	$w(\theta_1, a_1)$	$w(\theta_2, a_1)$
a_2	$w(\theta_1, a_2)$	$w(\theta_2, a_2)$

Table 1: loss table

Evaluating the *Bayes loss* gives:

$$B(a_1) = w(\theta_1, a_1) \cdot p(\Theta = \theta_1) + w(\theta_2, a_1) \cdot p(\Theta = \theta_2) \qquad (3)$$

and

$$B(a_2) = w(\theta_1, a_2) \cdot p(\Theta = \theta_1) + w(\theta_2, a_2) \cdot p(\Theta = \theta_2) \quad (4)$$

where

$$p(\Theta = \theta_2) = 1 - p(\Theta = \theta_1). \quad (5)$$

2.4 Defining the probability distribution.

The investigations in [6] assume a straight-line relationship and a strong correlation between T_{act}, the actual completion time for a session with length k, and the estimated value, T_{est}. Assume also that the observed estimation error $T_{act} - T_{est}$ for a given network load is Gaussian with mean $\mu = 0$ and variance σ^2.

It is desirable that the probability distribution $p(\Theta = \theta_2)$ is an increasing function of the ratio β where

$$\beta = \frac{T_{est}}{T_{max}} \quad (6)$$

and with the condition $\beta = 1 \Rightarrow p(\Theta = \theta_2) = 0.5$, denoting that if the estimated value is equal to the maximum allowed completion time, there is an equal probability of the actual value being greater then or less then the maximum allowed completion time. As a first approximation, $p(\Theta = \theta_2)$ in the form of a first order polynomial is proposed:

$$p(\Theta = \theta_2) = \begin{cases} 0 & \text{if} \quad T_{est} < T_{max} - \sigma \\ \dfrac{T_{max} \cdot \beta}{2\sigma} - \dfrac{T_{max}}{2\sigma} + \dfrac{1}{2} & \text{if} \quad T_{max} - \sigma \le T_{est} \le T_{max} + \sigma \\ 1 & \text{if} \quad T_{max} + \sigma < T_{est} \end{cases} \quad (7)$$

The approximation is illustrated in figure 1.

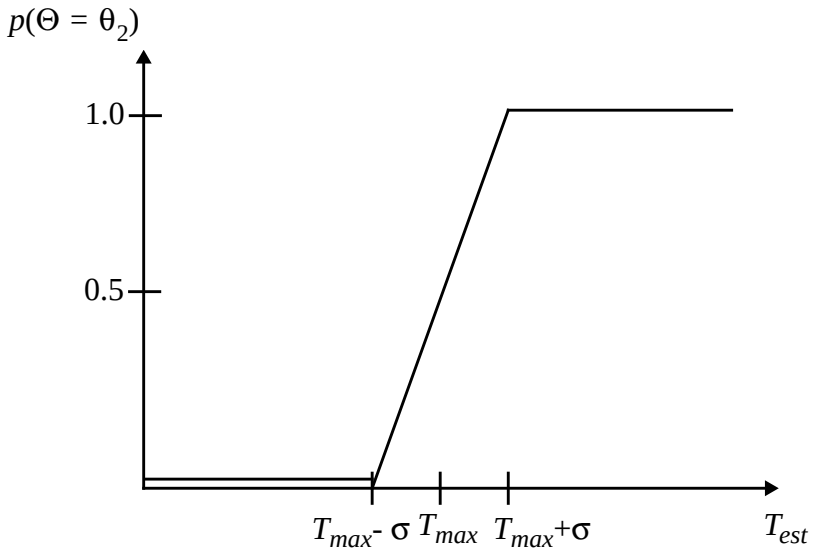

Fig.1. The probability distribution $p(\Theta = \theta_2)$ *proposed in formulae 7, as a function of the estimated time.*

2.5 Defining the loss function.

The loss function should give a definition of the loss attached to each action a_j. The simplest form of the function is

$$w(\theta_i, a_j) = b_{ij} \tag{8}$$

Where b_{ij} is a constant related to the state θ_i and to the action a_j. The loss functions from table 1 are now defined as

$$w(\theta_1, a_1) = b_{11} \tag{9}$$
$$w(\theta_1, a_2) = b_{12} \tag{10}$$
$$w(\theta_2, a_1) = b_{21} \tag{11}$$
$$w(\theta_2, a_2) = b_{22} \tag{12}$$

Combining equations 3, 5, 9 and 11 gives

$$B(a_1) = b_{11}(1 - p(\Theta = \theta_2)) + b_{21}p(\Theta = \theta_2) \tag{13}$$

and combining equations 4, 5, 10 and 12 gives

$$B(a_2) = b_{12}(1 - p(\Theta = \theta_2)) + b_{22}p(\Theta = \theta_2) \tag{14}$$

We may now use any numeric value on b_{ij} and our approximation (7) to find the optimal action (a_1 or a_2).

3. Analysis.

3.1 The signaling network model.

The nodes in the network model comprise both Signaling Point and Signaling Transfer Point functions in the sense that all nodes may initiate or terminate service sessions and they can all transfer incoming signals towards their final destinations.

The signaling network on which this analysis was performed is a symmetrical 20 node mesh network with four links per node. Fixed routing has been employed in such a manner that all signals traversing the network from node A to node B use the same route, while signals from node B to A may use another route. Signals may pass up to three nodes in order to reach their destination, and thereby interact with a total of five nodes. The network model is also described in [5 - 7].

A service session is considered to comprise *k* signals, of which *k/2* are sent from the originating node to *k/2* randomly selected destination nodes. This models the behaviour of a service that will invoke several nodes before completion. All analyses have been performed with the network in a steady state.

3.2 The service class attribute model.

3.2.1 Definition.

The signaling traffic in a signaling network can be divided into several classes with respect to the kind of service each session is initiating, typical examples of a service class is call setup and handover in a mobile communication network. We allow for *M* service classes where each class is defined by the following attributes

$$(k_m, \Lambda_m, b_{11_m}, b_{12_m}, b_{21_m}, b_{22_m}, T_{max_m}),$$

where *k* is the number of signals needed to complete a session and Λ is the node vector containing the nodes the session will pass from the originating to the destination node.

3.2.2 Numeric values.

The attributes described in 3.2.1 are used both by the ICCM and when evaluating the gain of the ICCM as described in 3.3. When observing an annihilated session, no information can be obtained whether it would have been completed within the stipulated time or not. It is then necessary to treat the two states θ_1 and θ_2 equally with respect to action a_1, by defining a constant $b \equiv b_{11} \equiv b_{21}$ giving that (13) is written:

$$B(a_1) = b \qquad (15)$$

The study is performed with the variables b, b_{12} and b_{22} satisfying the inequalities:

$$b_{12} < b < b_{22}$$

otherwise any action will dominate the other, e.g. if $b \geq b_{22}$ then

$$B(a_2) < B(a_1), \text{ for all values of } p(\Theta = \theta_2)$$

The analysis is performed with $m = 3$ classes and three sets of attributes as described in table 2

	Set 1 $\varpi = 10$			*Set 2* $\varpi = 5$			*Set 3* $\varpi = 10$		
attributes	Cl 1	Cl 2	Cl 3	Cl 1	Cl 2	Cl 3	Cl 1	Cl 2	Cl 3
k	40	10	20	40	10	20	40	10	20
b	10	10	10	5	10	15	10	10	10
b_{12}	1	1	1	1	2	3	1	1	1
b_{22}	100	100	100	25	50	75	100	100	100
T_{max}	$3T_1$	$3T_2$	$3T_3$	$3T_1$	$3T_2$	$3T_3$	$5T_1$	$5T_2$	$5T_3$

Table 2. Class attributes used in the study.

The sets are chosen to study the difference in network profit when the magnitude defined as

$$\varpi = \frac{b_{22}}{b} = \frac{b}{b_{12}} \tag{16}$$

between the parameters b, b_{12} and b_{22} is changed and also when the maximum completion time is changed.

The maximum completion time values, given in table 2, are normalized by the transition time for a session traversing the network under minimum load, T_m. The total amount of service sessions started in the network will be in proportion to that 50% belongs to class 1, 30% belongs to class 2 and 20% belongs to class 3. The service sessions are started with the same intensity in all node.

3.3 Definition of the metric used.

The network loss

$$P = \sum_{m=1}^{M} (b_{12_m} \times n_{suc_m} + b_{22_m} \times n_{del_m} + b_m \times n_{ann_m}) \tag{17}$$

where n_{suc_m} is the number of sessions completed in time, n_{ann_m} is the number of sessions annihilated and n_{del_m} is the number of sessions not completed in time, is the metric used to study the

behaviour of the network. The metric use the same parameters as the decision theoretic algorithms described in chapter 2.

3.4 Numerical results.

The impact of the ICCM is negligible at normal network load, and increases dramatically with network load. In other words, it does not interfere with the network under normal working conditions, i.e. a normalized network load below 0.5, but is activated when congestion arises. Applying the ICCM reveal a significant decrease in network loss especially compared to a network without ICCM but also compared with an ordinary CCM as described in [6], i.e. a CCM with no regards to network loss (fig. 2). The difference between the CCM and the ICCM is more significant at higher network loads. Up to 40% lower network loss is possible using the ICCM.

The ICCM yields better performance as the magnitude between the attributes b, b_{12} and b_{22} increases as shown in fig. 3. That means when the magnitude is large, the two actions differ more and hence an intelligent decision whether to annihilate or not, is more important. The gain of the ICCM is similar compared to the ordinary CCM if the maximum allowed completion time alters as shown in fig 4.

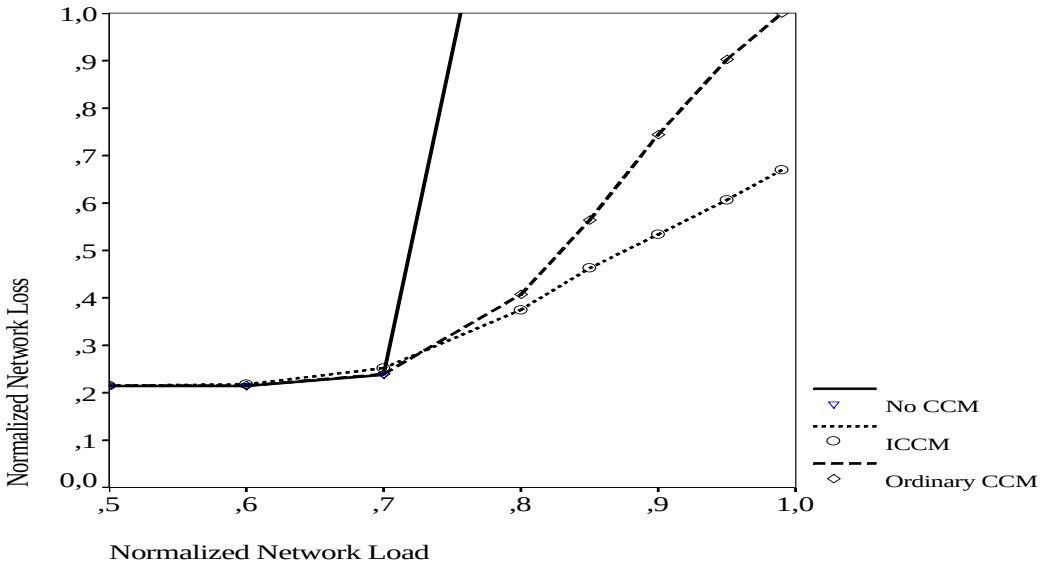

Fig. 2. Comparing network loss for a network with an ICCM, an ordinary CCM and without a CCM. The network loss is normalized with the network loss for an ordinary CCM at high load. Attribute set 1 is used.

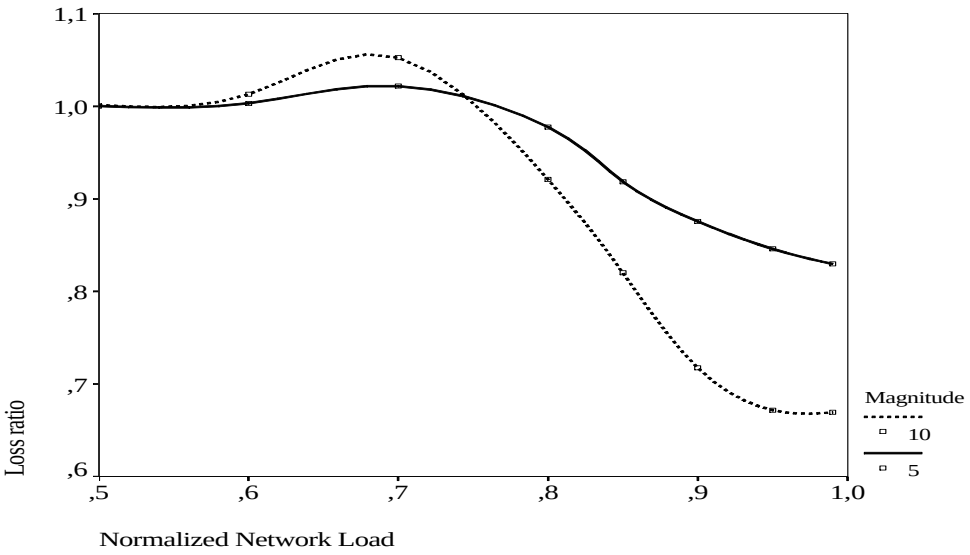

Fig. 3. The loss ratio between the ordinary CCM and the ICCM comparing set 1 and set 2.

All results are produced using the approximation for $p(\Theta = \theta_2)$ proposed in (7). The ratio $\frac{\sigma}{T_{max}} = 0.85$ is used, where σ is derived from measurements on a network using the decision CCM at a network load of 0.99. In the network load interval from 0.5 to 0.75 all three figures indicate that the CCM yields somewhat better performance than the ICCM. This is explained by that the σ used in (7) is not load dependent although a smaller σ should be used for lower loads, where predictions are more accurate, to avoid unnecessary annihilation.

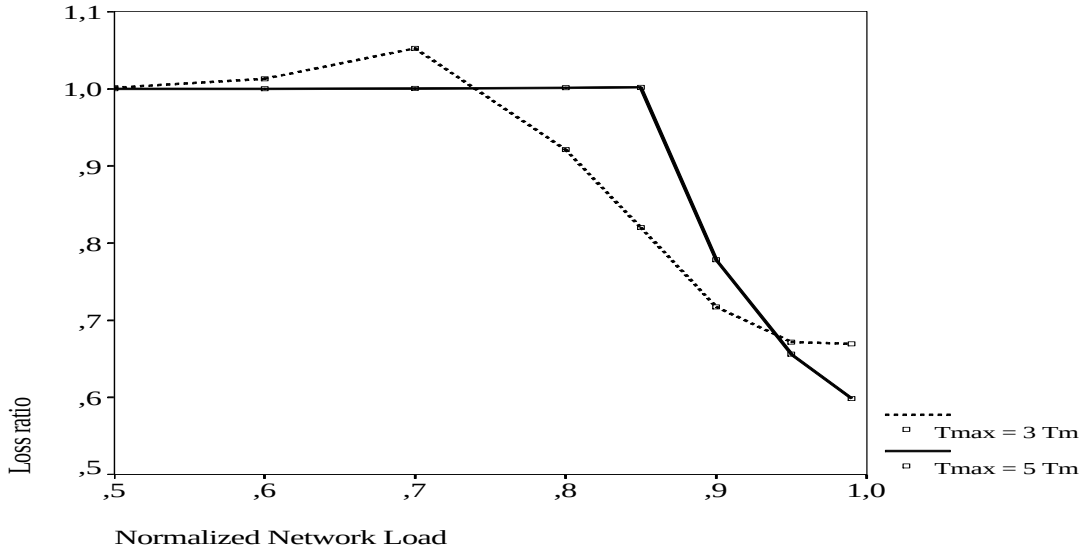

Fig. 4. The loss ratio between the ordinary CCM and the ICCM comparing set 1 and set 3.

4. Conclusions.

The work demonstrates the possibility of using a decision theoretic approach to refine a CCM based on network delays. The ICCM performs the action having the lowest *Bayes loss,* i.e. annihilate or not annihilate a signaling session that is about to start. The decision depends upon the value of the session attributes, and upon the probability of not successful completion $p(\Theta = \theta_2)$. The network loss is used as a metric to study the gain of the ICCM. Applying the ICCM to a network shows significant improvements compared to previously studied CCM, decreasing the network loss up to 40% during periods of congestion.

5. Future work.

The studied CCM can be refined in a number of ways. A more complex and accurate probability distribution, $p(\Theta = \theta_2)$, where σ is load dependent, can be used. The loss functions can also be modified to emulate more realistic network conditions.

The present study can not be expected to reveal all possible flaws or benefits of the CCM unless studied under more realistic circumstances. The assumptions in this paper of a symmetrical mesh network in a steady state, and with uniform service call intensity distribution over the origin-destination pairs, constitute only a small fraction of possible working conditions for a signaling network. A thorough investigation of the CCM performance must include unsymmetric signaling mesh networks exposed to transient loads and non-uniform service call intensities. Focused overload must also be investigated since most congestions are restricted to a small part of a network.

6. References.

[1] P.J. Kühn, C.D. Pack, and R. Skoog, “Common Channel Signaling Networks: Past, Present, Future”, IEEE Journal on Selected Areas in Communications, Vol. 12, No. 3, pp. 383-394, 1994.

[2] A.R. Modarressi and R.A. Skoog, “Signaling System No. 7: A Tutorial”, IEEE Communications Mag., vol. 28, No. 7, pp. 19-35, 1990.

[3] B.A.J. Banh and G. Anido, “Signaling Network Design Aspects For Mobile Services”, Australian Telecommunication Networks & Applications Conference, Melbourne, pp. 695-700, 1994.

[4] J. Zepf and G. Rufa, “Congestion and Flow Control in Signaling System No. 7 - Impacts of Intelligent Networks and New Services”, IEEE Journal on Selected Areas in Communi cations, Vol. 12, No. 3, pp. 501-509, 1994.

[5] L. Angelin, S. Pettersson, and Å. Arvidsson, “A network approach to signaling network congestion control”, St. Petersburg International Teletraffic Seminar, pp. 10-21, 1995.

[6] L. Angelin and Å. Arvidsson, “A congestion control mechanism for signaling networks

based on network delays", Proc. 12th Nordic Teletraffic Seminar, Helsinki, pp. 367-377, 1995.

[7] S. Pettersson and Å. Arvidsson, "Economical aspects of a congestion control mechanism in a signaling network", Proc. 12th Nordic Teletraffic Seminar, Helsinki, pp. 59-70, 1995.

[8] B. W. Lindgren, "Statistical Theory, fourth edition", Chapman & Hall, pp 539 - 574, 1993.

[9] D.G. Kleinbaum, L.L. Kupper, K.E Muller, "Applied Regression Analysis and Other Multivariable Methods", Duxbury Press, pp 41- 63, 1988.

7. Acknowledgements.

We would like to thank Dr. Claes Jogréus, Department of Telecommunications and mathematics, University of Karlskrona/Ronneby, for helpful advice with the decision theoretical problems.