Network oriented load control in Intelligent Networks

Stefan Pettersson
Research
Ericsson Software Technology
Box 518
SE-371 23 Karlskrona, Sweden

Abstract
Heavy network load in the signaling system number 7 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 rejects service sessions with a predicted completion time greater than a maximum allowed completion time for the session, network performance improves dramatically, and thus customer satisfaction. Rejection of already delayed sessions let other sessions benefit and increase the useful overall network throughput. The decision of rejection is based upon Bayesian decision theory that takes into account the cost or revenue attached to each action, i.e. whether to reject the session or not. More valuable sessions are then given priority through the network, at the expense of less valuable sessions. To clearly display the benefit from this approach we propose to use network profit as a performance metric.

This paper summarises the ongoing research and discusses the future direction of this project. Of special interest is deployment of new services in the IN and the implications this has to network load and the profit made by the operator.

1 Introduction
1.1 Setting the scene
The telecommunications scene is presently undergoing a period of fundamental changes such as increasing transmission rates, increasing user mobility, and higher service sophistication. This development both leads to and is driven by a growing usage of telecommunications, not only in terms of number of users and applications but also in terms of information volumes.

Customers are very concerned about the level of service provided. This means that, in a situation with rapid changes in telecommunications usage and requirements, there are solid business reasons for operators and providers to take a profit oriented look at performance issues and related matters, which previously might have been regarded as purely technical issues. It is also clear that there should be a connection between performance and service degradation in the network and customer dissatisfaction caused by e.g. delayed or lost calls.

1.2 Intelligent Networks
The basic idea of Intelligent Networks (INs) is to facilitate provisioning of new services quickly and independently from the telecommunications network and equipment vendors. The IN acts as a distributing and centralising framework of telecommunication services whereby new services may be introduced rapidly and cost efficiently.

The mechanism of communication in the IN is by exchanging signals over the signaling network. Typically, ITU-T Common Channel Signaling System No. 7 (SS7) is deployed. In SS7, signaling information is transmitted as Message Signal Units (MSUs), which by the signaling network essentially are treated as packets in a packet switched network.

The major impact the introduction of IN has to signaling networks is the changes in signaling patterns and real time requirements with respect to complexity. The new requirements for signaling networks also affect the requirements on overload control, which is a necessary tool to maintain the operability of a signaling network.

Processing and transmission resources of INs are dimensioned in order to meet relevant real time demands. Congestions or other disturbances in the network can give rise to unwarranted effects like e.g. long delays and lost calls as response times exceed time outs of signaling system protocols. Clearly, such delayed or failed service requests will lead to displeased customers. The situation is worsened by the fact that incidents of this kind caused by a temporary, light overload may cause repeated attempts which will increase loads further, and in the worst case a circle of increasing loads is started which eventually may lead to complete service disruption.

1.3 Congestion Control
Traditionally, congestion in signaling networks is considered as an overload problem in a local node or link. The overload problem is also resolved locally by using some kind of congestion control algorithm. The algorithm keeps the load of individual IN-elements on a level where normal service can be provided for longer or shorter periods, regardless of the impact on other parts of the network.
A whole range of mechanisms have been proposed to cope with situations of overload in nodes or links. The major drawback with these traditional methods is that they, in more distributed architectures, may give rise to congestion in other, initially not congested, parts of the network.

The main objective of a useful congestion control mechanism is that it must be able to resolve the overload situation in such a manner that the entire network benefits. The mechanism must prevent wasting resources on unproductive work (like repeated attempts) and maximise call completion in an overload situation.

The objective of this paper is to study a new mechanism for robust load control that use a profit-and-network-oriented approach to the problem of congested networks.

2 Network oriented load control

2.1 Preliminaries

We assume that when an IN service is requested, it has its own, specific signaling sequence which is executed. Such an execution is referred to as a signaling session. The number of signals in each session and the list of nodes invoked may vary from one service to another (e.g., calling a fixed or mobile subscriber), and also from one session to another depending on the outcome (e.g., answer, no answer, or busy).

To each service \( s = 1, \ldots, S \) we then associate a maximum signaling session completion time \( T_s \), which is set with respect to protocol standards and users’ expectations or, for pure load control purposes without stringent real-time requirements, simply engineered to match completion times for similarly loaded networks.

We distinguish three distinct outcomes of a signaling session, viz., successful, unsuccessful, and rejected. We say that a request for an \( s \)-service is successful only if the completion time of the associated session is less than or equal to \( T_s \), otherwise the request is unsuccessful. A rejected service is terminated before any signals are sent to the destination node and thus before consuming any network resources but the ones required to make the decision of rejection.

Our approach is based on the recognition that single signals have little or no value on their own since customers neither accept nor pay for anything but actually delivered services, and that signaling sessions may differ in value or urgency depending on the service they support. The overall goal is then to ensure that all accepted service requests can be successfully completed and that important ones are given priority.

We also make explicit use of some features which are exclusive to signaling traffic as compared to general data traffic, viz., that the traffic essentially consists of a large number of more or less predictable sequences each of which spans over a relatively short time. In our model we say that a session of class \( s \) consists of \( L_s \) signals.

2.2 Network costs and revenues

Each outcome of a session may be given a value representing the revenue or cost the operator gain from it.

Successful sessions generate revenues to operators in terms of call or service charges. The income associated with class \( s \) sessions is denoted by \( I(s) \). The costs for successful sessions is defined as a fixed processing cost \( C_p(s) \) plus a variable transport cost \( (C_t + C_v) L_s \). \( C_t \) is the cost for using transmission equipment, etc., \( C_v \) is the marginal cost the session represents in terms of increased completion times and possible failure for other sessions, and \( t_L \) is the completion time of the session.

Unsuccessful sessions do not generate any revenues. The costs are the same as for successful sessions, but an extra bad will cost \( C_{ba}(s) \) is added. The cost is the measure of customer dissatisfaction of e.g., losing a call, and could be defined as e.g., marketing costs, discounts, or costs for redesigning the network to increase performance.

Rejected sessions are similar to unsuccessful ones in that they do not generate any revenues. The costs, however, are smaller since there are no transport costs. Moreover, the bad will cost \( C_{ba}(s) \) for rejected sessions might be lower than the one for unsuccessful sessions if, e.g., the user is immediately informed by some kind of announcement or message that the service request can not be accepted due to temporary overload.

2.3 Network Profit as performance metric

We define network profit as revenues minus costs per time unit (t.u.), using the definitions of costs and revenues in 2.2. The main advantage of using network profit as a network performance metric is that other aspects than throughput, i.e., the number of signal sessions passing through the network within a specific time, are considered when measuring performance. We believe the proposed metric is more directly relevant to operators. It is also an important metric to achieve improved customer satisfaction as certain sessions are more valuable to the user than others. An example of this is that it may be more annoying to loose an ongoing call during a hand over than having a call set up failure.

Maximising network profit obviously means maximising the number of successful signaling sessions, and in particular the number of valuable sessions. The latter notation refer to the fact that some signaling sessions are more valuable and profitable than others, both for the operator and for the user, and should definitely be allowed to pass through the network unaffected by network delays. This can be achieved if less valuable and profitable sessions are rejected, thus releasing network capacity for the more valuable sessions. In fact, even if all sessions are considered equally important may the overall throughput in terms of successful signaling sessions be increased by a load control that rejecting a few requests, at an early stage, viz., the ones which are least likely to become successful.
The network profit principles are used by the algorithm proposed in this paper both at the decision procedure in section 2.6, and when evaluating the performance of the algorithm.

2.4 The algorithm
Assuming that only successful sessions generate revenues, it is immediately observed that unsuccessful ones, occurring e.g. during equipment failures or unexpected demand peaks, represent a double nuisance: Not only do they fail to generate an income, but the fact that they consume processing- and transmission resources while in the network, means that they actively contribute to delaying other sessions and thus further increase the number of unsuccessful ones.

Clearly, the only way to completely eliminate unsuccessful sessions, is to expand network resources or in some other way remove the overload. This is also the natural option for repeated or permanent overloads.

However, for unexpected and exceptional occurrences of overload, current equipment and load conditions set the limits, and the best one can do is to maximise the number of successful sessions. To do so, network resources should be spent on useful work, i.e. successful sessions, while unsuccessful ones should be removed from the network. This is also the base of our congestion control mechanism [CCM]: To predict the outcome of a session and prematurely terminate the ones which eventually will become unsuccessful.

2.5 Estimating completion times
We can view signal round trip times \( \tau(o,d) \) between two nodes \( o \) and \( d \) in a signaling network as providers of two kinds of information, one regarding the link between the two nodes \( o \) and \( d \) and another regarding the overall load situation in the network. Summing up over all links, a measure of the present overall network load \( \rho = \rho(o,k) \) from the perspective of an originating node \( o \) from which the \( k \)th signal is about to be sent is given by

\[
\rho(o,k) = \sum_{d=1}^{N} \bar{t}(o,d,k-1) / \sum_{d=1}^{N} t_{\text{min}}(o,d,k-1) \tag{1}
\]

where \( \bar{t}(o,d,k-1) \) is the present prediction of the round trip time between \( o \) and \( d \) calculated from \( \rho(o,k-1) \) as shown below, and \( t_{\text{min}}(o,d,k-1) \) is the smallest round trip time between \( o \) and \( d \) measured up to event \( k-1 \). Note that \( \rho(o,k) : 1 \leq \rho(o,k) \leq \infty \) and not identical to the conventional notation of load as system utilisation.

The two ways of using the information of round trip times may be melded together into a state machine to produce predictions of signal session completion times. The idea is to deliver a prediction before the first signal of the session has left the originating node. There is one such state machine per origination-destination (OD-) pair, and it consists of three states and three transitions as described in [3, 4].

Using this estimate, we may apply load control by accepting a signaling session if \( t_{L}(o,k,0) \leq T_{s} \) and rejecting it if \( t_{L}(o,k,0) > T_{s} \).

2.6 Decision using Bayesian theory
We now apply Bayesian decision theory to combine satisfaction/profits with uncertain predictions in order to determine the most profitable action for any service request. The most profitable action is the action \( a_{j} \) that maximises the expected value of the gain function

\[
E[w(\Theta,a_{j})] = \sum_{\forall \theta_{i}} w(\theta_{i},a_{j})g(\theta_{i}) \tag{2}
\]

where \( A = \{a_{j}\} \) is the set of possible actions to be taken, \( \Theta = \{\theta_{i}\} \) is the set of possible states or outcomes, \( g(\theta_{i}) \) is the probability of the system being in state \( \theta_{i} \), and \( w(\theta_{i},a_{j}) \) is the gain obtained in state \( \theta_{i} \) for action \( a_{j} \).

System state \( \theta_{1} \) means that the session will be successful and \( \theta_{2} \) that it will not. The probabilities can formally be written as the probabilities that the completion time is less than or equal to the maximum completion time and greater than the maximum completion time respectively. Assuming that the estimation error \( t_{L} - t_{L} \) is Gaussian with mean zero and standard deviation \( \sigma \), it follows that

\[
g(\theta_{1}) = \phi(\frac{t_{L} - T_{s}}{\sigma}) \tag{3}
\]

\[
g(\theta_{2}) = 1 - \phi(\frac{t_{L} - T_{s}}{\sigma}) \tag{4}
\]

where \( \phi \) is the probability function of a standard (0,1) normal distribution. The derivation of \( \sigma \) involves two models, one of the prediction state machine, and another one of the network. The details may be found in other works by the authors [15].

Action \( a_{1} \) means to annihilate a session and \( a_{2} \) to accept it. The gains for the two actions, some of which take negative values, are written as

\[
w(\theta_{1},a_{1}) = b_{1,1} \tag{5}
\]

\[
w(\theta_{1},a_{2}) = b_{1,2} + Kt_{L} \tag{6}
\]

\[
w(\theta_{2},a_{1}) = b_{2,1} \tag{7}
\]

\[
w(\theta_{2},a_{2}) = b_{2,2} + Kt_{L} \tag{8}
\]

where \( b_{1,1} \) and \( b_{2,1} \) is the (negative) gain for an annihilated session \(- (C_{p}(s) + C_{ba}(s)) \), \( b_{1,2} \) is the revenue minus processing costs \( I(s) - C_{p}(s) \), \( b_{2,2} \) is the (negative) gain for an unsuccessful session \(- (C_{p}(s) + C_{ba}(s)) \), and \( K \) is a (negative) variable transportation gain \(- (C_{t} + C_{o}) \).

2.7 Results
In [5, 6, 15] we study the yield of the algorithm using a simulated network model. We study the robustness of the algorithm for several load conditions (i.e. transient and focused overload) and for several network topologies. The results show significant performance improvements, both compared to a network without any load control, and compared to a network
using the node based load control proposed by ITU-T. Figure 1, which shows network throughput in terms of successfully completed signaling sessions under various conditions, gives an example of these results. The diagram to the left refer to the case of no control while the diagram to the right refer to a case of control with all signaling sessions are assigned equal profit. It is immediately noted that the algorithm significantly improves throughput and thus network profit.

3 Further research

This chapter will discuss some open issues with respect to research in network oriented load control of signaling networks. Possible solutions and enhancements of present results are also discussed.

3.1 Industrialisation

The overall goal for research in the field of telecommunications is (or should be) to industrialise the ideas and results for possible implementation in commercial products.

With respect to the present project, it is realised that prediction of completion times must be fast, accurate, and simple in order to implement a set of predictors in each node without significantly adding to the delay of a session or increasing the load of the processors. With regards to the present approach, other methods than the state machine are feasible for prediction of completion time estimates and their variances. Various forms of simplified state machines might be considered, possibly extended to cover variances by on line measurements. Another approach is to deploy Kalman predictors for completion times and use the variances provided by these.

3.2 Interaction with other controls

Despite the obvious benefits of network oriented load control, it does not seem likely that traditional, node based solutions will be removed. An obvious reason for this is the fact that node based solutions already are standardised and implemented world wide. Another, equally important reason is that no vendor wants to see its equipment go down under overload, no matter who or what caused the conditions. Hence all signal network equipment will come with whatever control facilities required to make it stay up. This indicates that the node and network based methods will coexist for a foreseeable future. An essential question is then how our method will behave when interacting with autonomic node based control algorithms, like the ones proposed by the ITU-T.

3.3 Implications on routing

Clearly there is a strong connection between our method of predicting completion times and how an optimal routing method is implemented [4]. An open issue is whether it is possible to use predicted completion times when selecting the best/fastest path in a routing algorithm. Another issue is to investigate the behaviour of the algorithm when using other routing methods then static.

3.4 Deployment of new services

A problem of special interest is the trade off between increased revenues at deployment of new services and the costs for increased signaling traffic associated with the service, in particular when the customer demand for the services increase. The main goal for the operator is of course to maximise the profit, something which we have seen also may include to keep the increase in signaling load at a reasonable level.

The scenario can be described as follows: Consider a network with a certain level of background traffic. A new service, which generates excessive signaling to one node $i$ is deployed in the network. The service generates $N$ signaling sessions to node $i$ per t.u. Then consider that only a fraction $M$ of the sessions are successful per t.u. If the revenue for each successfully completed service session is $I$ we have a total revenue of $MI$ per t.u.

At the same time costs will go up as the total traffic is increased in the network and causes local overload at node $i$. The costs relate to badwill costs and at least to some extent affect sessions from all over the network and amount to $C$ per t.u. The total profit for this service is then $P = MI - C$ per t.u.

The following experiments are defined to identify the characteristics of the profit function:

- Study the behaviour both with and without the CCM.
- Study the behaviour if the signaling load is both constant and variable during the test period.
- Study if the location of the service to an arbitrary node affects the behaviour.

Another interesting issue is how the service should be charged. This is not only a question of optimal pricing according to the laws of supply and demand, but it must be taken into account that the effective marginal costs of supplying the service are affected by the demand through the overall load on the network and its impact on the quality of service offered. We can thus foresee the development of a complete cost and performance model for new services.
3.5 Considerations for the B-ISDN

With the introduction of broadband telecommunications services and systems we introduce a range of changes to the signaling system. New services imply new signaling sessions with completely different characteristics and technologies like ATM extends the task of advanced routing since not only the traffic can be controlled but there is also a possibility to dynamically reshape the signaling network.

4 Conclusions

Current developments in telecommunication technologies and markets point at increasing demands on INs and signaling networks. These demands can partly be translated to requirements on system engineers to design robust control mechanisms that maximise a possibly weighted throughput of service requests under real time considerations. We have developed a congestion control mechanism which aims at maximising an arbitrarily weighted sum of successfully completed service requests under all load conditions. It takes a network wide approach to overload control, operates on complete requests rather than individual signals, and consists of a load predictor described as a state machine, and a decision procedure formulated in Bayesian terms. Numerical results indicate that the mechanism offers considerable improvements in throughput, in particular under severe overloads. Some remaining open issues include interaction with node based load controls and with routing mechanisms, profit optimal provisioning of new services, and a variety of implications with respect to the B-ISDN.

References


