A CONGESTION CONTROL MECHANISM FOR SIGNALING NETWORKS BASED ON NETWORK DELAYS

Lars Angelin, and Åke Arvidsson

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

Abstract

Congestion control in Signaling System #7 faces new challenges as mobile communication systems and Intelligent Networks grow rapidly. New services change traffic patterns, add to signalling network load, and raise demands on shorter service completion times. To handle new demands, the congestion control mechanisms must foresee an overload situation, and respond to it so that the network can maintain high probability for successful service completion. With the introduction of a state machine and a memory function for each signaling link it is possible to predict the completion time of a service session and to detect an emerging congestion. If the predicted completion time of a service session is too long, the session is annihilated. This is the foundation of a congestion control mechanism that reacts fast on information supplied by the congested part of the network. The congestion control mechanism increases the ratio of successfully completed service sessions during congestion by several hundred percent.

1. Introduction

The evolution of IN and mobile communications requires services of high complexity, and alters signaling traffic patterns. A complex service, such as the hand over procedure in mobile communications, needs more signals before completion, has higher demands on real time efficiency, and involves more nodes, than any service in the PSTN [1] [2]. Moreover, as new services are introduced, the number of simultaneous sessions to be handled by the signaling network increases, thereby increasing network load. Present congestion control mechanisms in Signaling System #7 (SS7) are primarily designed to cope with traditional call set-up and call release in the PSTN. All in all, this necessitates a new approach to efficient network solutions for signaling network congestion control.

Sessions of a service with high real time demands which are subject to unacceptable delays may be obsolete, or prematurely terminated by the customer; in either way, they are just a burden to the signaling network. It would ease the load of the network and improve the performance of all sessions in progress if such delayed sessions could be aborted as quickly as possible. The annihilation of sessions for which the first two signals consume more time than an allowed fraction of the allowed service completion time, has proven to be a well functioning congestion control mechanism (CCM) [3].

2. Congestion control in signalling networks

2.1 Congestion control functions in signalling networks

An SS7 network is a packet switched network with the sole mission to support telephone networks. The signalling network is a number of Signalling Points (nodes) and Signalling Transfer Points (transit nodes) connected via Signalling Links (links) in a mesh structure [4] [5]. The information communicated between the nodes to conclude a signalling service session is transported in signals guided by a routing algorithm. In case of link outage, or congestion, the routing algorithm must redirect the signals through the network in such a fashion that healthy parts of the network are not overloaded and thus causes other parts of the network to become congested, i.e. the robustness of the routing algorithm is not negotiable [6]. This suggests that the properties of the routing algorithm are inseparable from flow and congestion control in setting the boundaries for signalling network performance. A large number of routing algorithms have been thoroughly investigated, and their properties are well known, all ranging from fixed routing to very sophisticated adaptive routing algorithms [5].

A signalling 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. Congestions are more likely to arise from traffic redirections at network component failure, or by an extremely high call intensity to one specific node [7]. The traditional role of a CCM in SS7 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.

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].

2.2 Network delays as foundation to a CCM

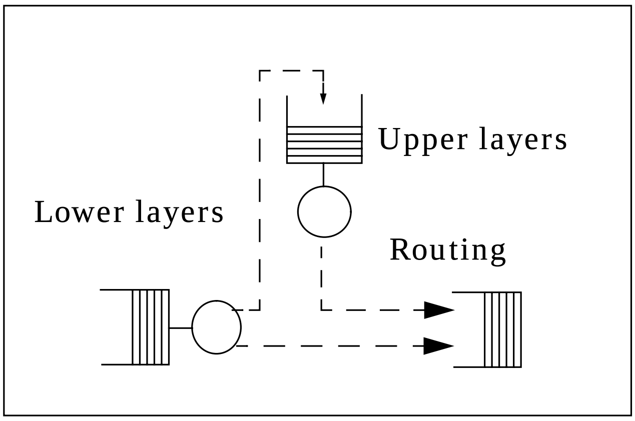
A signalling service session that exceeds its allowed completion time displeases the customer and deteriorates network performance. In a normally engineered network, signals of such sessions have with high probability encountered a congested part of the network. Signalling sessions encountering a congestion fuel the congestion, and consume much time in penetrating the congested part of the network. Information about the completion time of recently completed sessions may be used both as a parameter in a routing algorithm or in a CCM. The annihilation of such sessions would serve the dual purpose of reducing the load of the congested part as well as freeing communication facilities which may be used by other sessions. If knowledge of the duration of sessions could be obtained prior to their completion, it would be possible to annihilate sessions with too long completion time or to prevent them from getting started. This is the foundation of a benign CCM, one that detects a congestion at an early state and acts to reduce the flow in the congested direction.

3. Analysis of a CCM state machine

3.1 The signalling network model

The nodes in the network model comprise both Signalling Point and Signalling Transfer Point functions in the sense that all nodes may initiate or terminate service sessions and they can all transfer incoming signals towards the proper destination. Each node is divided into two parts: the lower layers and the upper layers, representing the OSI layers 1-3 and 4-7 respectively. In the lower layers there is also a signal discrimination function for routing an incoming signal to either the upper layers of the node or for further transport in the network (fig.1).

Each composite layer is represented by a queue with a FCFS queueing strategy and with the service time being the sum of a constant time and a time derived from a negative exponential distribution. This enables the processing time of a signal in a node to be modeled as an exponential service time which is longer than a shortest possible service time. The mean service time of the server for the lower layers is fixed. The mean service time for the upper layers is variable to model the complexity of the processing performed by the upper layers.



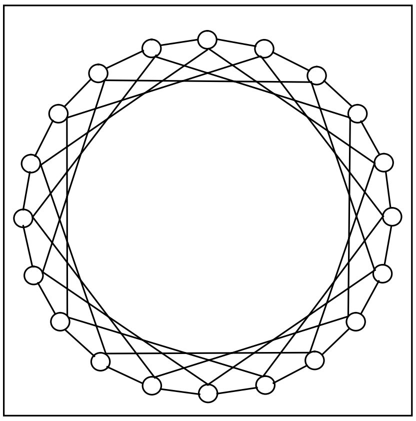


Figure 1. Queueing model of node interior

Figure 2. Network based on nodes and signalling links

The network, on which this analysis was performed, is a symmetrical 20 node mesh network (fig.2) 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 traffic load between the originating-destination pairs is uniform, and is derived from a negative exponential distribution. All analyses have been performed with the network in a steady state.

A service session is considered to comprise 20 signals, i.e. 10 “round trips” between the originating node and 10 randomly selected destination nodes. Such a service session is longer compared to the service call set-up in PSTN but shorter compared to the same service in a mobile network.

3.2 The congestion control mechanism

The signal completion time contains information about the network load. This information may be used in two ways, one regarding the link between the originating and the destination nodes and one regarding the overall load situation in the network.

The completion time of the most recent signal on a link is a good estimate of the completion time for the next signal to traverse that link if not too long time has elapsed between the two events. A study of the relationship between the completion time of the most recent signal as a prediction of the next signals completion time and the actual completion time of the next signal shows that recent, at a correlation of 0.8, is 1 average signal completion time at a network load of 0.25, and about 7 at a network load of 0.95. This in spite of the average signal completion time being roughly 10 times greater at the 0.95 network load compared to the 0.25 load.

An estimate of the overall network load from an originating node *i*'s perspective in a network with N possible destination nodes, and event *n* is to take place is given by $L_n(i)$, where

$$L_n(i) = \frac{\sum\limits_{\substack{d=1 \\ d \neq i}}^{N} P(i, d, n-1)}{\sum\limits_{\substack{d=1 \\ d \neq i}}^{N} M(i, d)}$$

where

$P(i,d,n-1)$ = the present prediction of the signal completion time between the originating node *i* and the destination node *d*, calculated with $L_{n-1}(i)$

and

$M(i,d)$ = the smallest measured signal completion time between the originating node *i* and the destination node *d*.

The two ways of using the completion time information may be molded together onto a state machine to produce a prediction of the signaling session completion time. The prediction of a service session completion time is updated when a signal is about to leave the originating node, i.e. even before the first signal of the session has entered the link.

There is one state machine per originating-destination pair, and it consists of three states and of three transitions. A brief explanation of the states and transitions is presented below. The constant scaling factors *a*, *c*, and *d* are set to 1 and *b* is set to 0.5. Their values are primarily chosen to make the model simple.

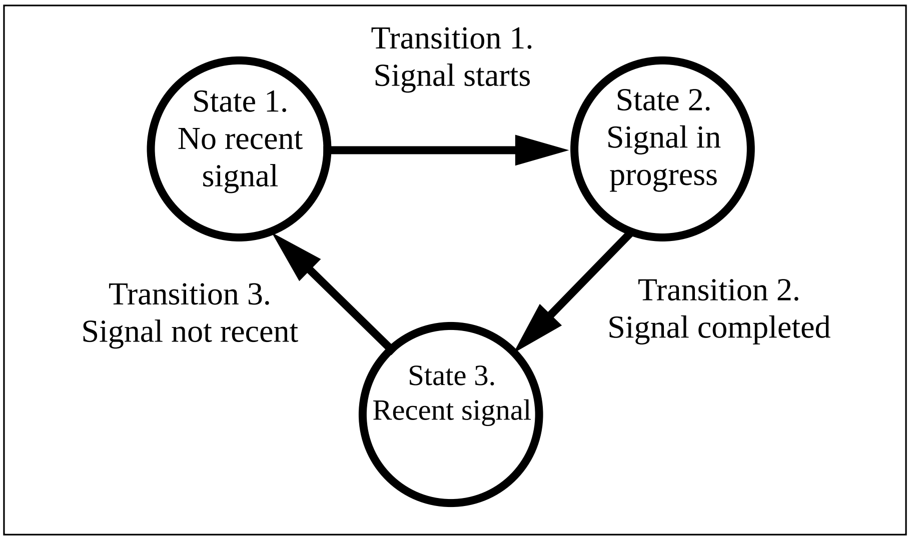


Figure 3. The state machine with it's states and transitions.

State 1. The link has been idle for such a long time that the most recent signal completion time is no longer valid as a prediction for the completion time of the next signal.
$P(i,d,n) = a\ M(i,d) \cdot L_n(i)$

Transition 1. A signal is sent from node *i* to node *d*.

State 2. The signal causing Transition 1 has not yet returned to node *i*.
$P(i,d,n) = \textit{time of signal in progress} + b\ M(i,d) \cdot L_n(i)$

Transition 2. The signal in State 2 has returned to node *i*.

State 3. There exists a resent signal completion time, *RSCT(i,d)*, that can be used as a prediction for the next signal.
$P(i,d,n) = c\ RSCT(i,d)$

Transition 3. Too long time has elapsed since the last signaling event.
$\textit{Elapsed time} = d\ RSCT(i,d) \cdot L_n(i)$

The prediction of the completion time of a signaling session comprising k signals of which l signals are already completed is calculated as

$$D(k, l) = \sum_{j = l + 1}^{k} P(i, d_j, n) + \sum_{m = 1}^{l} time(signal_m)$$

The case where $l = 0$ is the initial prediction for the session and it is made before the first signal of the session has left the originating node.

An investigation reveals high correlation between the *D(k,0)* and the total completion time of the session. Linear regression analysis gives an r^2 in the order of 0.9 (fig. 4). In other words, it is possible to make a good prediction of the total completion time of a session, and to predict how it will meet its real time demands.

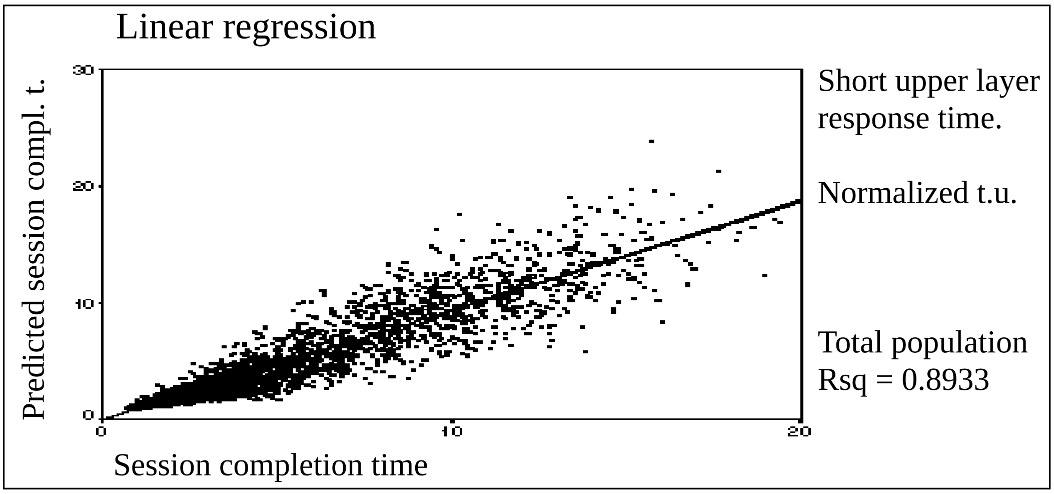


Figure 4. Linear regression analysis of the completion time of a session and the predicted completion time of the session.

A good prediction of the completion time of sessions also makes it possible to detect an actual or an emerging congestion with good accuracy since session completion time is closely related to network induced signal delays and thereby related to network load.

A simple CCM is to annihilate signaling sessions for which the prediction $D_n(k,0)$ is greater than a set time limit depending on the allowed session completion time. The time limit may be derived from time critical events in the network, such as the hand over procedure in mobile networks.

3.3 The metric

We have used the session completion ratio, i.e. the number of sessions completed within their allowed service completion time divided by the number of a generated sessions, as a metric. The metric discloses the probability for a session to fulfill its mission as requested by the customer, and is thereby closely related to that part of customer satisfaction that is derived from network performance.

3.4 Numerical results

The impact of the proposed CCM 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 steps into action when congestion arises. The simulations reveal improvements of up to 500-600% in the session completion ratio at normalized network load of 0.95 (fig.5). Varying the mean upper layer response time, the time scales alters, but the general behavior of the algorithm remains. This suggests the proposed CCM to be robust in terms of the processing complexity required in the upper layers. Longer upper layer response time gives a comparatively smaller fraction of the session completion time to be caused by the lower layers and thus yields less to network load. The upper layer response time is 10 times greater in the right diagram below compared to the left diagram. Normalized network load is 0.95 in the figures, and times are normalized with respect to shortest possible session completion time.

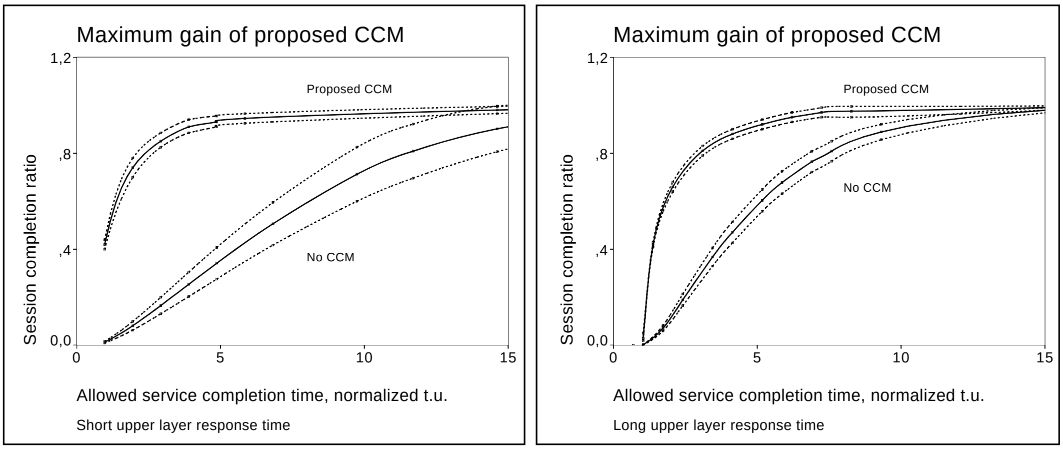


Fig. 5. Maximum gain in the session completion ratio of the proposed CCM (upper lines) compared to a network without CCM (lower lines). The central line in each group represents the mean value and the others represent the 95% confidence intervals.

Congestion control already exists in SS7 and it has some impact on the signalling network. The proposed CCM challenges the existing SS7 CCM and a comparison between the two is inevitable. The SS7 CCM is based on traffic information sent between nodes in Link Status Signalling Units, which are internal SS7 administrative signals [5] [7]. The information can originate from the signalling link layer levels or the User Part levels and concerns the state of the input buffers of these levels. The receivers of the information have then to take proper action either, i.e. to cease or throttle the signalling towards the congested node. Affected signals and sessions are discarded or terminated as quickly as possible. A model of the SS7 CCM functions above mentioned was incorporated into the studied model. After finding the best buffer sizes for the SS7 CCM model it was possible to conduct a comparison between the two CCMs. The result is given in figure 6 and the session completion ratio for a network without CCM is also shown in the diagram.

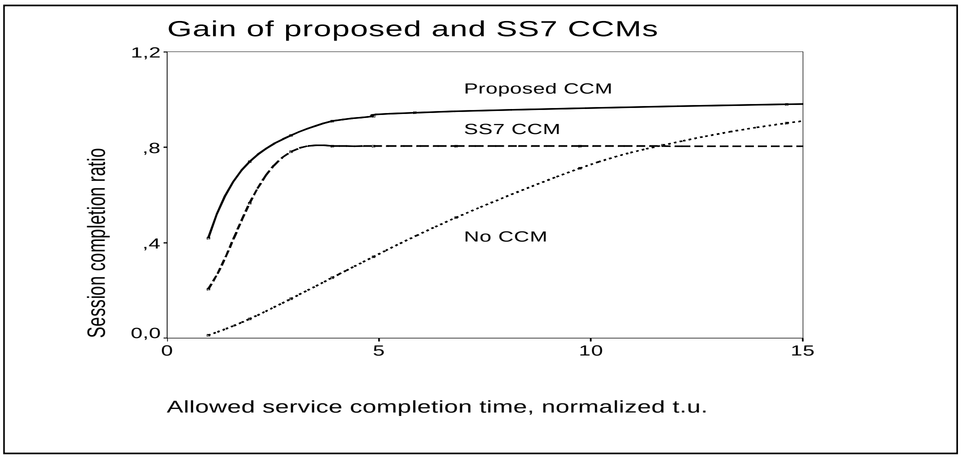


Fig. 6. Maximum gain of the proposed CCM (upper line) and existing congestion control in SS7 compared to network behavior without congestion control (lower line). The 95% confidence intervals are within +/- 0.04 of the CCM curves.

The proposed CCM favors sessions with low time consumption. The delay experienced by a particular session depends on the number of nodes traversed and on the present congestion in each of these nodes. All sessions benefit from the proposed CCM, regardless of the number of nodes traversed to reach the destination node [3].

The proposed CCM may easily be converted to take into account the diverse traversing distances of a service session. Instead of using one single time limit as the annihilation criteria for sessions, the annihilation criteria may by an individual time limit set for each specific service [8].

4. Conclusion

The work demonstrates the possibility to use information derived from the completion time of the signals in an SS7 network to gain knowledge of network performance, and thereby detect congestion. This information may also be used to design a signalling network CCM that operates independently of applications, and independently of nodes.

A simple CCM that predicts the session completion time from the most reliable signalling events of a node, and if the predicted completion time is found to be too long the CCM annihilates the session, shows great improvement on network performance at congestion. The proposed CCM steps into action when a congestion is detected and reduces the load in the proper directions while being fair to all service classes and to services traversing any number of nodes.

The proposed CCM has shown to be superior to the existing CCM in SS7 with our interpretation of congestion control.

5. Future work

The studied CCM can be refined in a number of ways. One way is to choose the constants a,b,c, and d with more care. An other way is to consider $L_n(i)$ for each outgoing signalling link and not as present, per node, calculated on information from all outgoing links in that node. Yet an other way is to let Transition 1 take place after the signal has been out $P(i,d,n) = a\ M(i,d) \cdot L_n(i)$.

The proposed CCM can not be expected to reveal all possible flaws or benefits 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 nodes, constitute only a small fraction of possible working conditions for a signalling network. A thorough investigation of the CCM performance must include unsymmetrical signalling mesh networks exposed to transient loads and non-uniform service call intensities. Focused overloads must also be investigated since most congestions are located to a small part of a node or a network.

Finally, in order to create good routing algorithms and CCMs, a good metric for signalling network performance is a necessity.

6. References

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

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

[3] L. Angelin, S. Pettersson, and Å. Arvidsson, 1995, “A network approach to signalling network congestion control”, ITC Seminar

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

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

[6] B. Jabbari, 1992, “Routing and Congestion Control in Common Channel Signaling System No. 7”, Proceedings of the IEEE, Vol. 80, No. 4, pp. 607-617

[7] D.R. Manfield, G. Millsteed, and M. Zukerman, 1993, “Congestion Controls in SS7 Signaling Networks”, IEEE Communications Mag., No. 6, pp. 50-57

[8] S. Pettersson, and Å. Arvidsson, 1995, “Economical aspects of a congestion control mechanism in a signaling network”, NTS-12