On the properties of a congestion control mechanism for signaling networks based on a state machine

Lars Angelin[1]

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

Abstract

Congestion control in signaling systems is a necessity to fulfil the requirements of a telecommunication network that aims satisfy the customers' requirements on service quality. Heavy network load is an important source of customer dissatisfaction as congested networks result in deteriorated service quality. Sessions of a signaling 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 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. By measuring the network delay on individual signals of a service session, it is possible to perform signaling network congestion control that considers the state in the entire signaling network. Under the assumption that a session comprises a sequence of signals between one originating node and an arbitrary number of destination nodes, it is possible to predict the total duration of a session. The prediction is calculated from previously completed signals using a state machine, which is defined per signaling link. The annihilation of sessions, for which the prediction exceeds a predefined time limit, is an embryo of a simple signaling network congestion control mechanism (CCM). This simple CCM increases the number of successfully completed services with several hundred percent under favorable circumstances. The state machine approach is proven to perform well in all types of environments. The robustness and stability of the proposed CCM is demonstrated in a wide range of environments. The fairness in the admission of signaling services into the network at very high loads are also shown to be good.

1. Introduction

In the face of growing competition in the telecommunication industry caused by the removals

[1] e-mail: larsa@itm.hk-r.se, phone: + 46 455 78042, facsimile: + 46 455 78057

of monopolies and the rapid deployment of technological advances, network operators can not spare any effort to keep their network and its services at the leading edge of technology. Customers are attracted to networks with a wide range of sophisticated services. Services that must perform at its best all times, not to cause custumer dissatisfaction or disappointment *i.e.* it is obvious that long delays and lost calls are unacceptable and displeasing to custumers [1]. Solid business reasons now exist for operators and service providers to tune their networks to peak performance, previously a technical and a planning issue. Today, the financial survival of the network depends upon it. This calls for new weapons in the ongoing war to enhance network performance. The signaling network clearly is one of many battlefields in this war.

The evolution of IN and mobile communications requires services of high complexity, and alters signaling traffic patterns, compared to ordinary PSTN services [2]. 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 services in the PSTN [3,4]. 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.

The subject of this work is to study a new mechanism for robust load control. A mechanism that uses a network oriented approach to the problem of congestion control in overloaded networks.

2. Signaling in telecommunication networks

2.1 Introduction to the Common Channel Signaling System No. 7

The present signaling system, CCITT Signaling System No. 7 (SS7), was introduced in 1980 [5,6]. SS7 is a common channel signaling system and such systems are characterized by the use of separate channels for the signaling information and for the transfer of user data. An other feature of SS7 is datagram mode of operation. An SS7 network can be viewed as a packet switched network with a predetermined packet size and a parallel network to the trunk network. The signaling network is needed for the exchange of messages regarding *e.g.* call set-up, supervision and call release. SS7 enables the capability of messages to act as queries and responses and this is essential to the operability and flexibility in modern telecommunication networks.

Some of the benefits of SS7 compared to older signaling systems are:

- - reduced call set-up times
- - enhanced signaling capabilities, security and flexibility
- - IN capabilities are enabled
- - a worldwide standard is provided

2.2 The SS7 protocol model

The protocol model used in SS7 has a structure similar to the OSI-model, a simplified structure is shown i fig. 2.1.

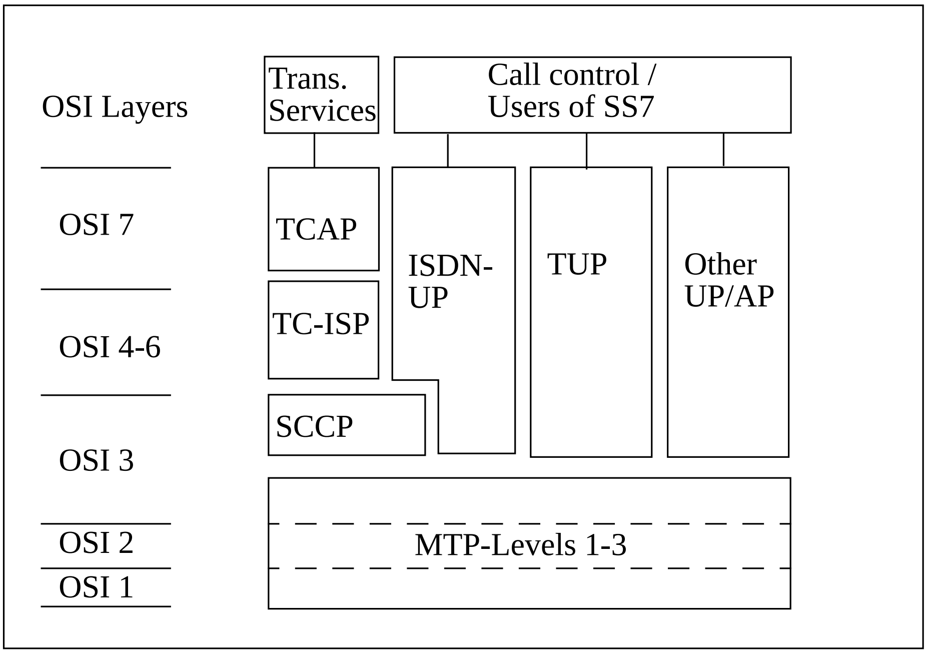

Figure 2.1: *SS7 protocol architecture*

The message transfer part (MTP) is the basic building block for reliable connectionless transport of signaling information within a signaling network [5,6,7]. The Signaling Connection Control Part (SCCP) enhances the addressing function of the MTP and also enables the SS7 to offer connection oriented services to applications of higher layers. The MTP in conjunction with the SCCP, referred to as the Network Service Part (NSP), then provides both connectionless and connection oriented services. The upper SS7 protocol layers use the signaling functions offered by the NSP as a network independent transport vehicle.

The ISDN User Part (ISUP) provides the signaling functionality necessary for switched voice and data. The ISUP may use the services offered by the SSCP as a method for end-to-end signaling. It also provides the signalling functionality needed to support advanced IN-services. Prior to the ISUP, the Telephone User Part (TUP) was commonly used to support control of telephone calls. The Transaction Capabilities (TC) protocol caters for control and transfer of noncircuit related information such as the location of mobile roamers, operation and maintenance or any kind of service that may require a query/response type of interaction. The TC consists of the Transaction Capability Application Part (TCAP) and the Intermediate Services Part (ISP). The IN application protocol (INAP), is a protocol within the component sublayer of TCAP, and is of prime importance to the operability of IN services.

Several other User Parts/Application Parts (UP/AP) exist, *e.g.* the Mobile User Part for analogue mobile networks and the Mobile Application Part for digital mobile networks.

The signalling requirements of new service are easily met by the flexibility of SS7. The NSP acts as the independent bearer of all signalling traffic and only an interface between the NTS and the SS7 user is necessary, *i.e*. a new UP.

2.3 Messages and signaling schemes in SS7

The signaling information is contained in Message Signal Units (MSU). An MSU may contain up to 272 octets of data and its payload is the UP-messages. The MSU also contains information for addressing and protection against delivery of corrupt UP-messages. There is a large set of predefined message types where each message type performs a specific function. Typical examples of ISUP-messages are Initial Address Message (IAM) which initiates a call set-up and Address Complete Message (ACM) which indicates the complete receipt of the called party address at the final destination exchange. In figure 2.2 an ISUP call set-up is shown. The Answer Message (ANM) indicates that the call has been answered

.

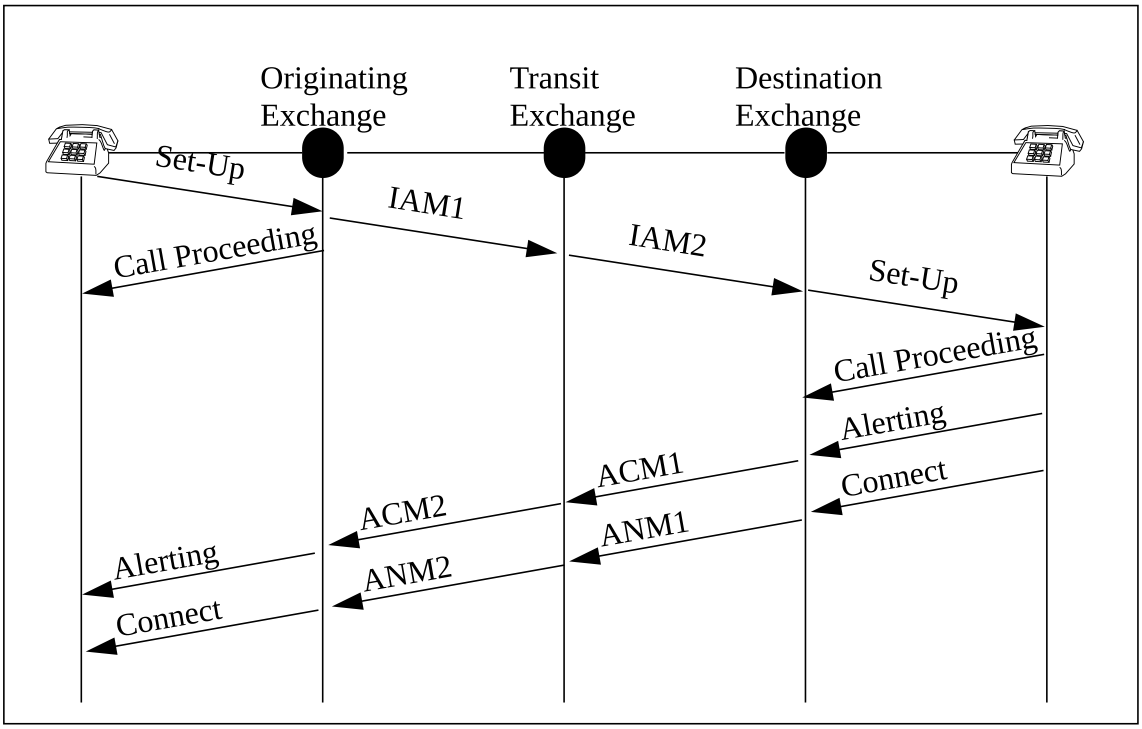

Figure 2.2: *ISUP call setup example*

2.4 The signaling network, a generic model

In the sequel of this report, a generic model of a signaling network with *N* nodes and *M* links is used. All nodes in the network may serve as Signaling Points (nodes) or Signaling Transfer Points (transit nodes) connected via Signaling Links (links) in a mesh structure where Signaling Points originate and terminate signaling.

A service in the network is generalized to comprise a distinct number of MSUs (from now on: signals), and each service has its own, specific sequence of signals which is executed each time the service is requested. Each such 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 depend on the outcome of the request (e.g. answer, no answer, and busy).
A service's maximum time for completion, T_S, is set by timers and system parameters in the signaling network, or by the customers' patience, *i.e.* the call is abandoned if it is delayed. 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.

3. An introduction to Intelligent Networks

3.1 What is Intelligent Networks?

The term Intelligent Network (IN) was first used in 1984 by Bellcore in order to help the Regional Bell Operating Companies to become more competitive in the deregulated telecommunications environment [8,9]. The original goal was to provide network operators with the ability to introduce, control and manage services more efficentlly by using a centralized database in a Service Control Point (SCP).

The basic idea of IN today is to facilitate provisioning of new services quickly and independently from the telecommunications network and equipment vendors. The IN will act as a distributing and centralizing framework for rapid and cost effective deployment of new telecommunication services.

IN is applicable to a wide variety of networks *e.g.* public switched telephone networks (PSTN), mobile networks (GSM etc.), or Broadband ISDN (B-ISDN).

3.2 Functional architecture of the IN

The main elements of the IN architecture are:

- Service Switching Point (SSP)
- Signaling Transfer Point (STP)
- Service Control Point (SCP)
- Service Data Point (SDP)
- Service Management System (SMS)
- Service Creation Environment (SCE)
- Intelligent Peripheral (IP)

The SCP form the core of the IN and contains the service logic that controls the IN-based services, assisted by the SDP which provides data about the customers and the network. The IP contains special functions like announcement machines and is used to communicate with the users of a service. The SSP provides the end-user access to the IN-services. It contains functions for detecting service access codes and sending service requests to the SCP. The SCP and SSPs may be interconnected via STPs. The SMS is a platform for operation and maintenance tasks, such as provision of services or collection of statistics. The SCE contains functions for defining, developing and testing IN-based services. An overview of the IN is shown in fig 3.1.

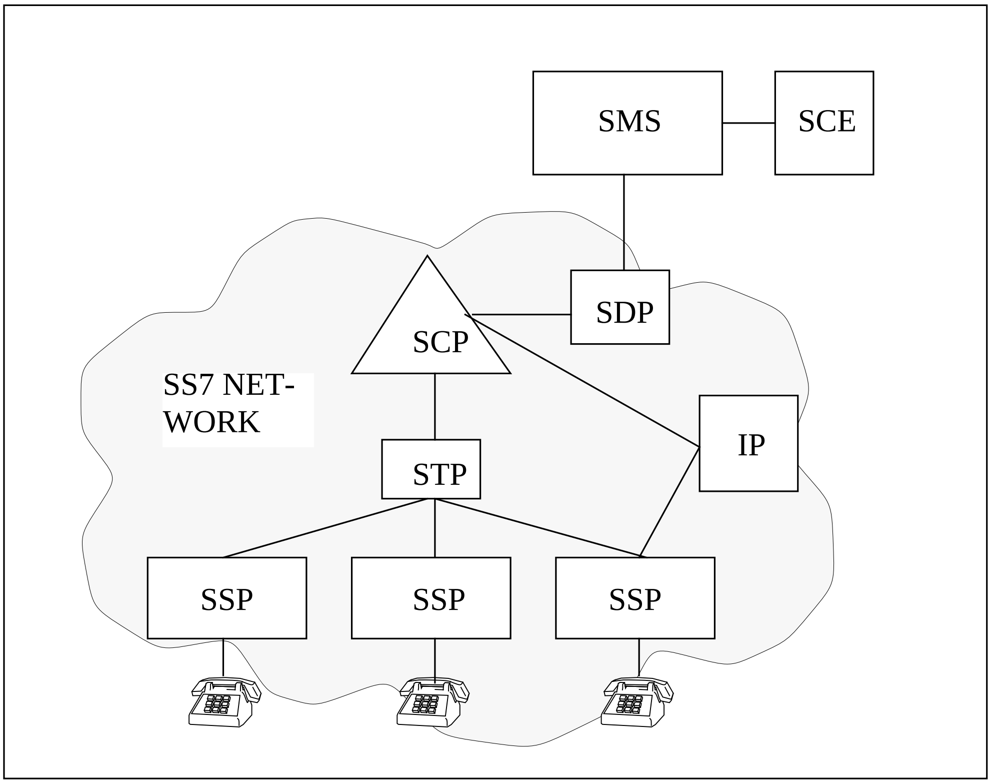

Figure 3.1*: Functional Architecture of the IN*

3.3 Services in the IN

One of the driving forces behind the introduction of IN is the possibilities to invent and introduce new, advanced services in the network. This evolution will meet the customer demands for more advanced services, but will also contribute to higher revenues for the operator. Typical examples of IN services are:

- **Freephone**, the called party is charged for the call, whatever the origin or length of the call. The calling party only need one phone number, regardless of its location, to become connected to the called party. With additional intelligence in the service the calling party may even be connected to the local/nearest office of *e.g.* a nation wide insurance company. The IN will perform the call distribution to determine the nearest office.

- **Premium rate**, allow the sharing of revenues between the network operator and an external service provider.

- **Personal number**, the subscription is associated to an individual and not to a subscriber line. The user may log in to an arbitrary telephone set, thus linking the telephone set/subscriber line to the personal number. Outgoing calls will be charged to the personal number.

- **Televoting**, voting via the telephone network. Votes are counted by the local exchange and subsequently passed to the management/IN database where they are collected and formatted in a suitable form for the customer. This service can be used for elections or opinion polls in *e.g.* TV or radio shows.

3.4 Signaling in the IN

The use of SS7 is essential to facilitate proper IN functionality. IN has brought changes to the demands on signaling services both with respect to the number of signals in a session and the timing requirements for a session [4,10]. 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 mobile network and thus perform the switch of base stations. A call set-up in a GSM system may consist of up to 40 signals invoking large number of nodes in the network [11].

4. Congestion control in signaling networks

4.1 Introduction

The operability of the signaling network is of prime importance in securing telecommunication network performance. Signaling networks may suffer congestion like all other packet switched networks and therefore need some scheme of protection against congestion. The effects of congestions or other disturbances in the signaling network encompass unwanted effects like *e.g.* long delays and lost calls which will lead to customer dissatisfaction. A typical example of actions that can cause congestion in the network is a mass call-in scenario within an IN application [12,13]. Excessive traffic to one node will cause a local or focused overload in the network. The re-routing of the signaling through other parts of the network or repeated call attempts at lost calls are likely to make local congestion spread to the rest of the network.

The ever ongoing introduction of new advanced services cause an increase in both communication and processing requirements. The new services may also change the traffic patterns in the network. These are all potential threats to signaling network stability. Other reasons for congestion are network failures as breakdowns of signaling links or signalling processors.

The main bottlenecks, where congestion is most likely to arise, in a signaling network are the link capacity and the protocol processing capacity. The level of link and processor load mainly depends on the message intensity and types of messages to be processed. The present congestion control system in SS7 is designed to handle congestion in signalling networks where telephony is the predominant service. History has, on a few occasions, proven present congestion control mechanisms not to be flawless [14].

Today telecommunications encompass a wide variety of services such as mobile communications, data communications, and IN. All with new, and different demands *vis-a-vis* the signalling system. Present Congestion Control Mechanisms (CCMs) operate in a node or a link perspective, and are thus not always able to give good response in the perspective of the signaling network. All in all, there is a need for efficient solutions for signaling network congestion control which use information from the entire network [15].

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, which keeps the load of a particular piece of signaling equipment 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. Most of these have been inherited from general packet switched networks, such mechanisms as throttling and flow control, and call gapping and random admittance from the circuit switched networks. These methods may work properly in a network architecture with only one central processing entity to protect [12]. The major drawback with these traditional methods is that they, in more distributed architectures, may give rise to congestion in other, not congested, parts of the network. Congestion situations may arise within each individual SS7 level, *i.e.* in level 2 (link congestion), in level 3 (route set congestion), and in level 4 (User Part congestion).

4.2 Requirements on congestion control mechanisms

The main objective of a useful congestion control mechanism is the ability to resolve an overload situation in such a manner that the entire network benefits. The mechanism must avoid resource utilization by unproductive work (like repeated attempts) and maximize call completion in an overload situation. Further more, it must be able to foresee an emerging congestion, and to take adequate prophylactic steps in order to normalize the situation [4].

Most existing overload control schemes in public switches are derived from pure telephone switch applications. These schemes are normally not well suited to cope with the signaling traffic. The requirements to be put on load control in signaling networks are defined as [16,12]:

- **Efficiency**, The message flow should be unaffected during normal conditions and adjusted to the right level during periods of overload.
- **Robustness**, The load control should be able to handle various grades of overload as well as rapid changes in the network load.
- **Fairness**, *S*ervices with identical priority and properties are treated equally.
- **Stability**, The probability of successful service completion should not fluctuate randomly or too much.
- **Simplicity,** The mechanism should be easy to implement.

4.3 Congestion control mechanisms

To fulfil the requirements on congestion control, a spectrum of congestion control mechanisms have been proposed in the literature. The most important ones are presented below [17]:

- **Message dropping**, Signals may be dropped, *e.g.* at buffer overflow, without any notification to the originator. It is only efficient during short periods of overload, while repeated requests will worsen the situation should the condition remain.
- **Message throttling**, New call requests are rejected if certain requisites are fulfilled. Examples of this method are the *window method*, *i.e.* the number of outstanding requests between the SSP and the SCP is limited, or the *random method*, only a certain percentage of all calls between the SSP and the SCP is accepted [16], or *call gapping*, a limited number of calls may be accepted per time unit.
- **Prioritizing**, Messages of certain types may be prioritized to make sure they are processed.
- **Routing**, New calls may be routed such that the congested area is bypassed.
- **Redundancy**, At least two different alternative paths are available between two nodes, messages may be directed to the alternative path if the ordinary path is congested or damaged.
- **Reactive control**, Examples of this method is the explicit notification of the sender by the receiver with the messages like Receiver Ready and Receiver Not Ready.

These methods may also be combined to achieve optimal control, depending on thecause of the overload.

In the literature, only few examples are given of congestion control mechanisms that operate in a network oriented way. A typical example is described in [18] where end-to-end congestion control using the window method is described for *e.g.* the ARPANET.

4.4 Congestion control mechanisms in SS7

Several congestion control or overload protection mechanisms are incorporated into SS7. Each mechanism is dealing with the congestion aspects of a certain layer in the SS7 protocol stack [17]. A short review of them:

- **MTP layer 2**, Protection by the change of signaling links to uncongested ones.
- **MTP layer 3**, Protection by explicit notification of load status by the receiver to the sender. Many versions of the congestion control mechanism exist for this layer.
- **SCCP layer**, The SCCP offers several connection modes and the congestion control must abide by the connection mode. Among the congestion control mechanisms we find flow control methods for connection oriented packet communication, particularly the notification of the source to reduce traffic.
- **UP layer**, A number of different suggestions have been made to solve the impact of congestion for the various applications of the UPs. Some solutions are close to those of the SCCP layer.

- **Application layer**, Here we find the rare opportunity to reduce the signaling traffic at the source. Automatic Congestion Control (ACC) is an application part function and thus lies outside the signaling network domain [19]. It supports the sending of explicit congestion notification between application parts located in different nodes.

4.5 Current research on congestion control in Intelligent Networks

Current research on congestion control in the IN are focused on protecting the SCP from overload and several papers have been published on this issue. Kihl and Nyberg studied fair and efficient throttling in [12]. The importance of load control communication between SCP and SSP is studied by Nyberg and Olin in [16]. The efficiency of the congestion control mechanisms in SS7, during severe overload *e.g.* mass call-in situations, is studied by Rumsewicz in [13] and he addresses the problems of multi-operator signaling networks in [20]. The use of congestion control in distributed multi-processor systems is studied by Ahlfors in [21]. In both [12] and [21], the efficiency of the PID-controller as a control mechanism for message throttling is investigated. Network architecture is a key element in the overall performance of the network, so also for congestion aspects. This view is discussed by Mostrel in [22] and by Leung and Wainberg in [23]. The sustained overload is investigated by Manfield, Millsted and Zukerman in [24], while the transient case has been studied by Zepf and Rufa in [25]. The effect of the delay in the notification of congestion to other areas in the network is studied by Smith in [26].

The ease with which new services are made and introduced in a network is reflected in the signaling network. New signaling traffic scenarios are expected and so are an overall increase of signaling network load. Studies in this field are presented by Zepf and Rufa in [4], and by Fujioka and Wakahara in [10].

S. Pettersson and Å. Arvidsson take a network- and profit oriented approach to congestion control in signaling networks [27,28,29,30]. This means that the control mechanisms focus on performance as a measure of operator profit and hence indirectly of customer satisfaction. The approach differs from traditional ones in that:

- Single signals has little or no value on their own since customers neither accept nor pay for anything but actually delivered services *i.e.* successfully terminated, complete signaling sessions.
- Signaling sessions may differ in value depending on the service they support, for example handovers are more important than new connections in cellular systems.

This concept focuses on service requests, or rather the associated sequence of signals. The goal is to make sure that resources are spent wisely *i.e.* that processing- and transmission resources are spent so that all accepted service requests are completed and that important ones are given priority.

5. Foundations to the proposed congestion control mechanism

5.1 General remarks

The information communicated between the nodes to conclude a signaling service session is transported in MSUs 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, i.e. the robustness of the routing algorithm is not negotiable [31]. This suggests that the properties of the routing algorithm are inseparable from flow and congestion control in setting the boundaries for signaling 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 signaling network is engineered in such a fashion that normal load represents about 25-40% of maximum load, suggesting congestion to be very unlikely at normal working conditions. Congestion is more likely to arise from traffic redirections at network component failure, or by an extremely high call intensity to one specific node [19]. 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. Furthermore, it must be able to foresee an emerging congestion, and to take adequate prophylactic steps in order to normalize the situation [32,33].

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 overall performance 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 (CCM) [34]. The introduction of a state machine and a memory function for each signaling link makes it possible, even before any signal of the session has left the originating node, to predict the completion time of a service session with good accuracy and to detect an emerging congestion [35,36].

5.2 Network delays as a foundation to a CCM

An increase in offered signaling network load will increase the signaling session completion time, and a further increase in offered load will eventually cause congestion with session completion times approaching infinity (fig. 5.1) [7]. Two conclusions may be derived instantaneously:

i) The signaling session completion time contains information about network load, and thus indirectly information concerning the congestion state in the network. This information may be used both as a parameter in a routing algorithm or in a CCM.

ii) Signaling sessions with real time demands will not easily be able to meet these demands during high network load.

The carried load, *i.e.* the number of sessions completed within their allowed service completion time divided by the number of a generated sessions at an offered network load of 1.0, increases in proportion to the increase in offered load (fig. 5.2). When the offered load increases beyond a certain value, the load threshold, the carried load reaches its maximum and then falls dramatically. The load threshold is determined by the real time demands of the signaling sessions, the physical topology and the service architecture of the network. A reduction of the real time demands moves the load threshold to a higher offered network load, and *vice versa*. The effect may be interpreted as a virtual congestion, more severely experienced by services with high real time demands, long before an actual congestion arises. This implies that congestion control is a necessity at all conceivable network loads when real time demands are present.

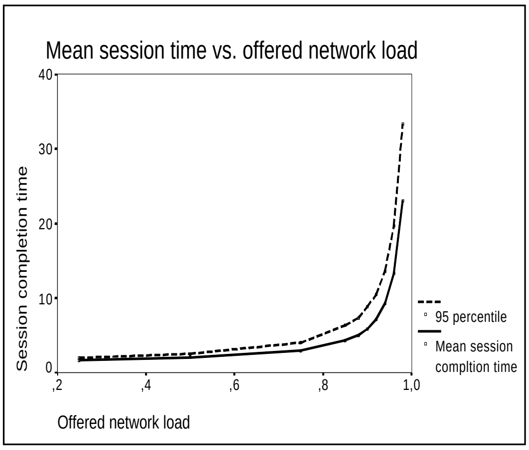

Figure 5.1: *The relationship between work load and session times.*

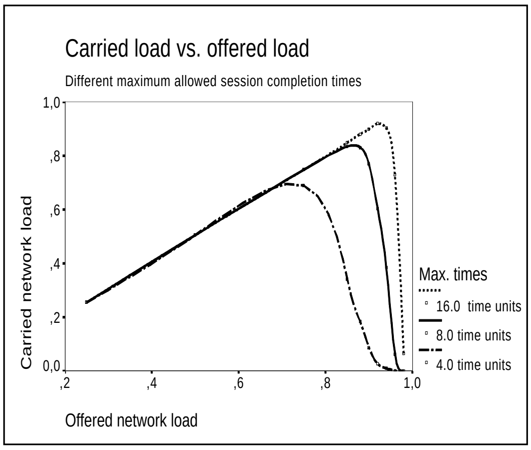

Figure 5..2: *The relationship between of- offered net-*
fered network and carried network load. completion
The 95% confidence intervals are within +/-
0.05 of the curves.

A signaling service session that exceeds its allowed completion time displeases the customer and deteriorates network performance by occupying buffer space and processor capacity without contributing to the carried network load. In a normally engineered network, signals of such sessions have with high probability encountered a congested part of the network. Signaling sessions encountering congestion fuel the congestion, and consume much time in penetrating the congested part of the network. The annihilation of such sessions would serve the dual purpose of reducing the load of the congested part as well as freeing communication facilities, and thus enhancing the possibility for other sessions to meet their real time demands. 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.

6. A CCM state machine

6.1 An estimate of the network load

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. *E.g.* consider a 20 node symmetrical network with a uniform load. To achieve a correlation above 0.8 in this network between two events, “too long” means more than one average signal completion time at an offered network load of 0.25, and about 7 at an offered network load of 0.95. This in spite of the average signal completion time being roughly 10 times greater at the 0.95 offered network load as 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, *j* where *j* = 1, 2,.. .,*N* and $j \neq i$, and event *n* is to take place, is given by

$$L(i, n) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{N} P(i, j, n-1)}{\sum_{\substack{j=1 \\ j \neq i}}^{N} M(i, j, n)} \qquad 6.1$$

where

P(i,j, n–1) = the present prediction of the signal completion time between the originating node *i* and the destination node *j*, calculated with *L(i,n-1)*

and

M(i,j,n) = the smallest measured signal completion time between the originating node *i* and the destination node *j* at signalling event *n*.

L(i,n) is a load measure in the interval $1 \leq L(i, n) \leq \infty$ and it increases with load increase.

6.2 The state machine

The two ways of using the completion time information may be molded together onto a state machine to produce predictions of signaling session completion time (fig. 6.1). The prediction of a service session completion time is made when a signal is about to leave the originating node, i.e. even before the first signal of the session has entered any link.

There is one state machine per origination-destination pair, and it consists of three states, and of three transitions. A brief explanation of the states and transitions is presented below.

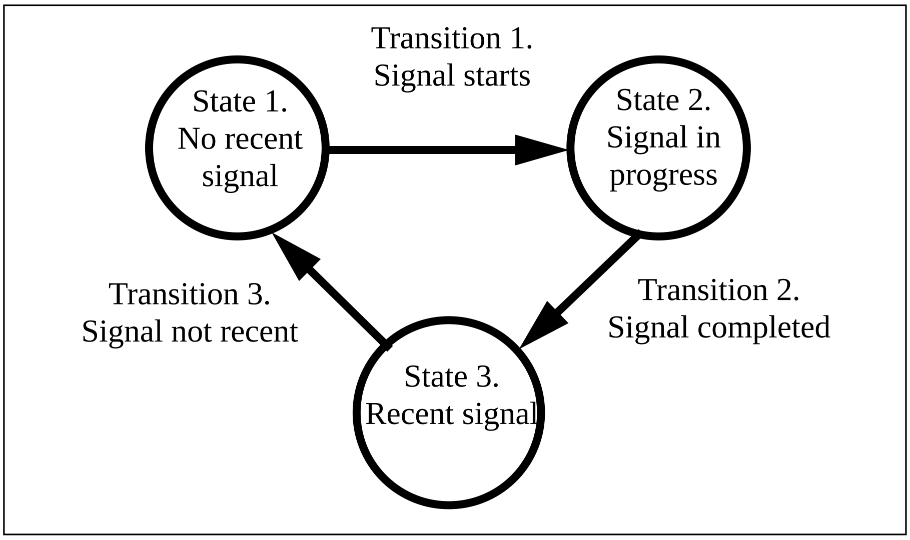

Figure 6.1: *The state machine with its states and transitions.*

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. We then set *P(i,j,n) = aM(i,j,n)L(i,n),* where *a* is a constant scaling factor.

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

State 2. The signal causing Transition 1 has not yet returned to node *i*. We set *P(i,j,n) = t+bM(i,j,n)L(i,n),* where *t* is the time so far consumed by the signal and *b* is a constant scaling factor.

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

State 3. There exists a recent signal completion time, *R(i,j,k),* that can be used as a prediction for the next signal. Here we set *P(i,j,n) = cR(i,j,k)* and *c* is also a constant scaling factor. *k* denotes the *k*th signal to node *j* from node *i*

Transition 3. Too long time has elapsed since the last signaling event. This happens when $T_e(i,j,n)$= *dR(i,j,k)L(i,n),* where $T_e(i,j,n)$ is the time elapsed since signalling event *n* between OD pair *i,j* and *d* is a constant scaling factor.

6.3 Prediction of session completion time and annihilation criteria

A service session in our model has one originating node, *i,* and *k* randomly selected destination nodes, j_m: $j_m \in \{1, \ldots, N\} \wedge j_m \neq i$, and m: $m \in \{1, \ldots, k\}$. A service session then comprises *k* signals. The signal *m* to node j_m is divided into two parts. The first part traverses the network from the node *i* to node j_m and then the second part of signal *m* completes the round trip back to node *i*. The prediction of the completion time of a signaling session originating in node *i* and comprising *k* signals of which *l* signals are already completed is calculated as

$$D(i, n, k, l) = \sum_{m=1}^{l} t_m + \sum_{m=l+1}^{k} P(i, j_m, n) \qquad 6.2$$

where t_m is the actual time consumed for signal *m*. *D(i,n,k,0)* is the initial prediction for a session of *k* signals which is made before the first signal of the session has left the originating node.

A simple CCM is to annihilate signaling sessions for which the prediction *D(i,n,k,0)* is greater than a set time limit depending on the maximum allowed session completion time. The time limit may be derived from time critical services in the network, such as the hand over procedure in cellular networks, or simply be set in such a fashion that it protects the network from congestion.

The annihilation procedure in this study is as follows:

i) Determine the shortest possible completion time for session *s*, originating in node *i* and comprising *k* signals

$$min(s) = \sum_{m=1}^{k} min(j_m) \qquad 6.3$$

ii) If *D(i,n,k,0) > A min(s),* session *s* is annihilated.

and the real time demands of a session are, in this study, treated in the following manner:

If *D(i,n,k,k) > B min(s),* session *s* has not met its real time demands and is considered not successfully completed.

To make the annihilation criteria and the session real time demands work together it is obvious that the relation between *A* and *B* is determined by the actual network and it's services. The annihilation procedure can be refined by estimating variance of the predictions and including this in the decision weather to annihilate or not [28].

Since several services are supported in a real signaling network, each with its unique service characteristic, one single time limit is not satisfactory as an annihilation criteria for service sessions. The annihilation criteria may by added as a service characteristic for each specific service, this is possible without corrupting either the proposed CCM or the service [29]. In a wider perspective it is not only the network itself that must benefit from a CCM. The network operators' objective is to optimize the financial profit from the signaling network. The CCM can be one tool in achieving this task, using the CCM to annihilate the least profitable sessions first [28].

An investigation reveals a correlation in the order of 0.8-0.95 between *D(i,n,k,0)* and the actual completion time of the session through a wide range of networks, network characteristics and offered network loads [36].

In other words, it is possible to make a good prediction of the completion time of a session, and to predict how it will meet its real time demands. 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.

7. Stability and optimization of the state machine

7.1 Introduction

A CCM state machine may be implemented in networks with great variety in both node-to-node connectivity and routing algorithms. The state machine must be stable and give predictions with good accuracy while functioning well in arbitrary networks. If the predictions do not approach infinity the CCM is considered to be stable. The stability and the accuracy is to some extent governed by the parameters *a*, *b*, *c* and *d*. The accuracy of the prediction can be defined by considering the r^2 factor in a linear regression analysis. Every network will need its own set of parameter values and it is therefore desired that the parameters can assume a wide range of values without disturbing the stability or the accuracy of the state machine.

A study of how the parameter values affect the state machine is an essential task in determining the generic functionality of the state machine. To find analytical results for arbitrary networks is certainly a non-trivial task. Even to obtain these results when the network is given is a paramount affair.

The networks refered to in figures 7.1 to 7.5 are described in section 8.1.

7.2 General assumptions

A signalling service session comprises a number of signals traversing the network in what seems to be a random pattern. Each signal in the signalling session involve an *OD* pair and the state machine makes the predictions for each and every signal individually, see section 6.3. If the state machine can produce a prediction with good accuracy for a signal without jeopardizing its own stability, then it can also produce a prediction with good accuracy for an entire signalling session without causing instability to itself. Hence, the need to study the behavior of the state machine over an entire signllling session is obsolete.

A few general assumptions are presented below, and they are not all needed in every subsequent discussion. The assumptions are:

i) It is enough to study the signalling between one *OD* pair since all *OD* pairs have identical behavior. This assumption gives

$$\begin{aligned} P(i,j,n) &\to P(n) \\ R(i,j,r) &\to R(r) \\ L(i,n) &\to L(n) \end{aligned} \qquad 7.1$$

and

$$M(i, j, n) \rightarrow M(n) \qquad 7.2$$

ii) The shortest possible signal completion time is constant and equal to *M*.

$$M(n) \rightarrow M \qquad 7.3$$

iii) Constant network load implies a network in eqelibrium. Let *T(n*) be the actual signal completion time for signal *n*. This assumption then yeilds

$$\begin{aligned} &\lim_{n \rightarrow \infty} P(n) \rightarrow \tilde{P} \\ &\lim_{n \rightarrow \infty} R(r) \rightarrow \tilde{R} \\ &\lim_{n \rightarrow \infty} T(n) \rightarrow \tilde{T} \end{aligned} \qquad 7.4$$

where the ~ denotes a stocastic variable and *L(n)* can be considered to be

$$\lim_{n \rightarrow \infty} L(n) \rightarrow \tilde{L} \qquad 7.5$$

The assumption states that the delays of signalling events of identical character share statistical distribution.

7.3 State 1

7.3.1 Stability in State 1

a is the only parameter in State 1 and its impact on the state machine stability may be investigated without to many simplifications. The stability criteria can determined under thr assumptions *i*) and *ii*) in the previous section. State 1 dominates the predictions at low offered network loads and *L(n)* varies little.

The prediction of the signal completion time is given by *P(i,j,n) = aM(i,j,n)L(i,n)*. With the assumptions above, the prediction can be written

$$P(n) = aML(n) \qquad 7.6$$

Now, assume that State 1 determines the prediction for *m* consecutive signals. These signals may belong to a number of signalling sequences. This gives

$$P(n) = a^m ML(n-m) \qquad 7.7$$

since *L(n)* is derived as

$$L(n) = \frac{P(n-1)}{M} \qquad 7.8$$

This gives the following stability properties

$a > 1$, the state machine is not stable and $P(n) \to \infty$ as $n \to \infty$
$a = 1$, the state machine is stabile and $P(n) \to ML(n-m)$ as $n \to \infty$
$a < 1$, the state machine is stabile and $P(n) \to 0$ as $n \to \infty$

The only possible value for stability and accuracy is a=1, and if only stability is desired $a < 1$ may also be used.

7.3.2 Optimal prediction in State 1

The prediction of the signal completion time in State 1 is *P(i,j,n) = aM(i,j,n)L(i,n)*. With assumptions *i*), *ii*) and *iii*) the the prediction of the signal completion time can written

$$\tilde{P} = aM\tilde{L} \qquad 7.9$$

The optimal prediction of the signal completion time is the prediction that minimizes the squared error of the actual and the predicted signal completion time. The expectation of the squared error is

$$E\left[(\tilde{T}-\tilde{P})^2\right] = E\left[\tilde{T}^2\right] + E\left[\tilde{P}^2\right] - 2E[\tilde{T}\tilde{P}] \qquad 7.10$$

The squared error can be reformulated with 7.9

$$E\left[(\tilde{T}-\tilde{P})^2\right] = E\left[\tilde{T}^2\right] + (aM)^2 \cdot E\left[\tilde{L}^2\right] - 2aM \cdot E[\tilde{T}\tilde{L}] \qquad 7.11$$

and minimum for this expression is found when

$$a = E[\tilde{T}\tilde{L}] / ME\left[\tilde{L}^2\right] \qquad 7.12$$

To get an indication to the choice of a: the fact that $\tilde{L}$ varies little and slowly gives the oppertunity to consider it constant and equal to L. We now get

$$a = E[\tilde{T}] / ML \qquad 7.13$$

The optimal prediction is $P = E[\tilde{T}]$, but $E[\tilde{T}]$ is unknown and load dependent, and therefore unreachable. But on the other hand, if ML is a good prediction, i.e. $ML = E[\tilde{T}]$, then a = 1 is the best choice.

7.3.3 Simulated results for State 1

Simulations indicate that a = 1 is the best choice, as shown in fig. 7.1. The minimum of $E\left[(\tilde{T}-\tilde{P})^2\right]$ is located at a = 1 with a fairly steep decent on the lower a side. This indicates that the accuracy of the prediction of the signal completion time is severely reduced if a is set to a too small value. The state machine is unstable and can not produce any results regarding $E\left[(\tilde{T}-\tilde{P})^2\right]$ for $a > 1$. This makes the choice of a fairly simple, just let a be smaller than, but as close to one as possible. The simulations are performed with an offered network load of 0.25, this being a region with a high probability to be in State 1.

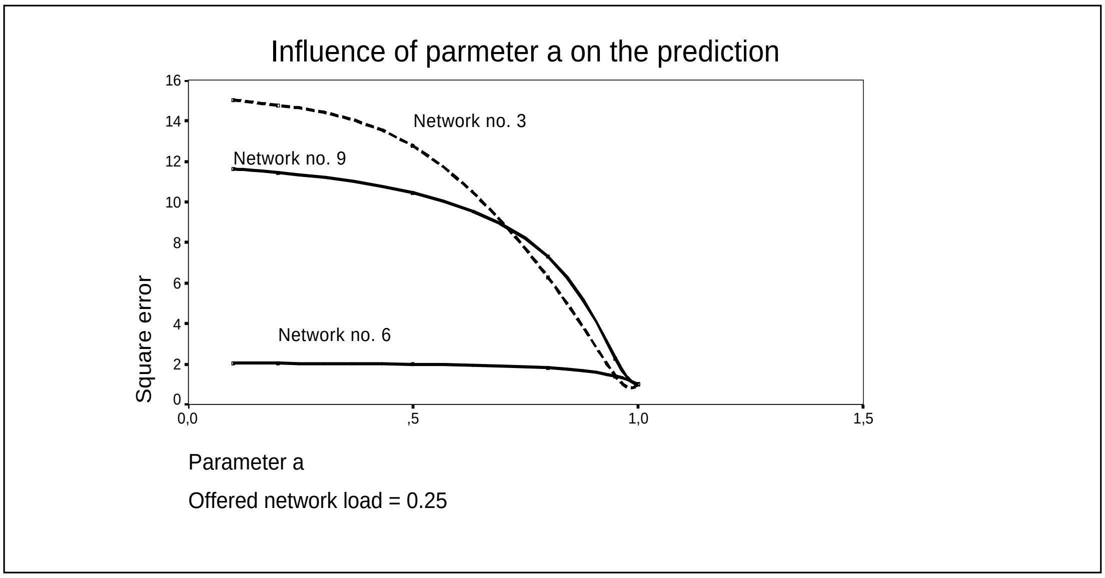

Figure 7.1: *The influence of the parameter a on the prediction of the signal completion time.*

7.3.4 Conclusions regarding optimal predictions in State 1

The best prediction of the signal completion time is found for $a = 1$. An $a > 1$ causes instability in the state machine and is can therefore not be an advisable choice. It is possible to have $a < 1$, but this will always be at the expense of prediction accuracy. The prediction accuracy is especially important at load transients. If the prediction starts at too low values and the transient is positive or vice versa, the state machine must perform the double task of catching up with both the error and the transient.

7.4 State 2

7.4.1 Stability in State 2

Assume that we start studying the system when signal *n* just has put the state machine in State 2. The prediction of the signal completion time of first signal to leave State 2 after signal *n* is

$$P(n+1) = t + bML(n) \qquad 7.14$$

where *t* is the time elapsed since the initiation of signal *n*.

The prediction of the *k*th signal completion time in State 2 after signal *n* is then given by

$$P(n+k) = t_k + bML(n+k) \qquad 7.15$$

Signal *n* started in State 1 and the prediction for signal *n* , with *a*=1, was

$$P(n) = L(n)/M \qquad 7.16$$

which can be rewritten as

$$L(n) = P(n) / M \tag{7.17}$$

Expression 7.15 may now be expressed as a function of *L(n)*

$$P(n+k) = t_k + \sum_{i=1}^{k} b^i t_{k-i} + b^{k+1} ML(n) \tag{7.18}$$

Now, set the completion time of signal *n* to *x* and let exactly *k* signals arrive in State 2 during signal *n*'s completion time. Since $t_i \leq x$, an upper bound of the expectation of the prediction time, P_{max} , can be made

$$E(P_{max} | k \text{ arriving signals in, } (0, x)) = x + \sum_{i=1}^{k} b^i x + b^{k+1} ML(n) \tag{7.19}$$

And it is obvious that

$$E[P_{max} | k \text{ arriving signals in, } (0, x)] \geq E[P(n+k)] \tag{7.20}$$

Remove the conditioning on exactly *k* arrivals

$$\begin{aligned} &E[P_{max}, k \text{ arriving signals in, } (0, x)] = \\ &\quad E[P_{max} | k \text{ arriving signals in, } (0, x)] \cdot P(k \text{ arriving signals in } (0, x)) \end{aligned} \tag{7.21}$$

The arrival streams to the network are Poissonian with arrival intensity λ. The probability of *k* arrivals during the time interval (0, *x*) is

$$P(k \text{ arriving signals in } (0, x)) = \frac{(\lambda x)^k}{k!} e^{-\lambda x} \tag{7.22}$$

An upper bond of the expectation of prediction time is now at hand

$$E[P_{max}, k \text{ arriving signals in, } (0, x)] = \left(x + \sum_{i=1}^{k} b^i x + b^{k+1} ML(n) \right) \frac{(\lambda x)^k}{k!} e^{-\lambda x} \tag{7.23}$$

State 2 is stable since

$$\lim_{k \to \infty} E[P_{max}, k \text{ arriving signals in, } (0, x)] \to 0 \tag{7.24}$$

The conclusion is that the parameter *b* does not influence the stability of the state machine.

7.4.2 Optimal prediction in State 2

The tuning parameter *b* also determines the accuracy of the prediction in State 2. As the prediction in State 2 is given by: $P(n+1) = t + bML(n)$, where *t* is the time elapsed since the initiation of signal *n*. The assumptions in the section 7.2 are also applied in this section. The assumptions give

$$\tilde{P} = t + bM\tilde{L} \qquad 7.25$$

Let T be actual signal completion time. The optimal prediction of the signal completion time is found when the expectation of the squared error of the actual and the predicted signal completion times have a minimum. That is, minimizing

$$\begin{aligned} &E\left[(\tilde{T}-\tilde{P})^2\right] = \\ &= E\left[\tilde{T}^2\right] + E\left[t^2\right] + 2E[tbML] + E\left[(bML)^2\right] - 2E[\tilde{T}t] - 2E[\tilde{T}bML] \end{aligned} \qquad 7.26$$

gives the optimal prediction of the signal completion time.

The minimum of $E\left[(\tilde{T}-\tilde{P})^2\right]$ with respect to *b* is found by setting

$$\frac{d}{db}\left(E\left[(T-P)^2\right]\right) = 0 \qquad 7.27$$

and the minimum is found for

$$b = \frac{E[\tilde{T}] - E[t]}{ML} \qquad 7.28$$

Let *B(x)* be the distribution of the service time and thus

$$E[\tilde{T}] = \int_0^\infty x dB(x) \qquad 7.29$$

The quest is now to find *E*[*t*]. Let *A(t)* be the distribution of the interarrival times and let *N(t)* denote the number of arrivals in the interval *(0,t)*. The situation is depicted in fig. 7.2 below.

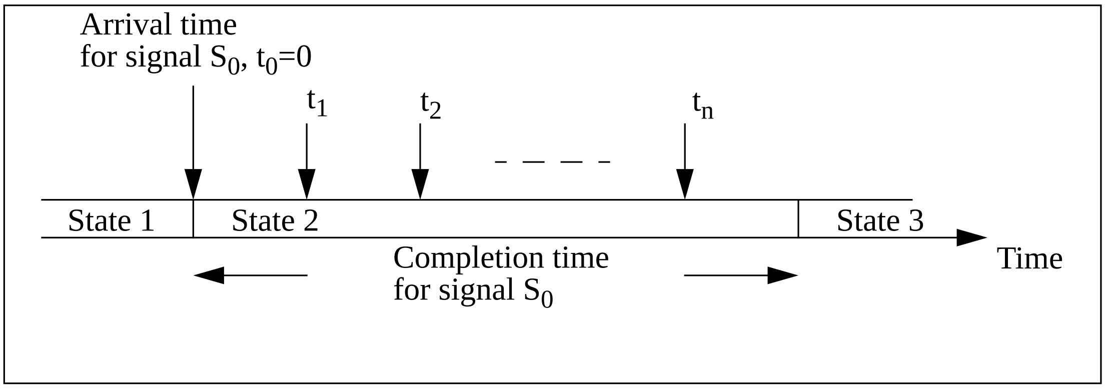

Figure 7.2: *The arrival process in State 2*

A theorem concerning the conditional distribution of arrival times from the theory of stochastic process states [37]

> *Given that N(t)=n, the n arrival times t_1,.....t_n have the same distributions the order of statistics corresponding to n independent random variables uniformly distributed on the interval (0,t).*

This implicates

$$E[t|n \text{ arrivals during a service time, } n > 0] = \frac{E[\tilde{T}]}{2} \qquad 7.30$$

$$\Rightarrow$$

$$E[t] = \int_0^\infty \frac{E[\tilde{T}]}{2} \cdot (1 - P(N(x) = 0)) \cdot dB(x) \qquad 7.31$$

Hence the optimal *b*

$$b = \frac{E[\tilde{T}]}{2ML} \cdot \left(1 + \int_0^\infty P(N(x) = 0) \cdot dBx\right) \qquad 7.32$$

Assuming that *ML* is a good estimate of the signal completion time, that is $ML = E[\tilde{T}]$, yields

$$b = \frac{1}{2} \cdot \left(1 + \int_0^\infty P(N(x) = 0) \cdot dBx\right) \qquad 7.33$$

The conclusion is that the optimal prediction of the signal completion time is found when $0.5 \le b \le 1.0$.

7.4.3 Simulated results for State 2

Simulations indicate that the best choice is $0.5 \le b \le 1.0$ as shown in fig.7.3. The minima of $E\left[(\tilde{T} - \tilde{P})^2\right]$ are located at $b \ge 0.5$ with somewhat steep descents on both sides. This indicates that the accuracy of the prediction of the signal completion time is reduced if *b* is not optimally chosen. The state machine is stable for all investigated values of *b*. This makes the choice of *b* a delicate matter as the optimal *b* is network dependant. But on the other hand, the probability to be in State 2 is small and therefore is the impact of parameter *b* on the overall state machine performance also small. The simulations are performed with an offered network load of 0.75, since this is where the highest probability to be in State 2 is found.

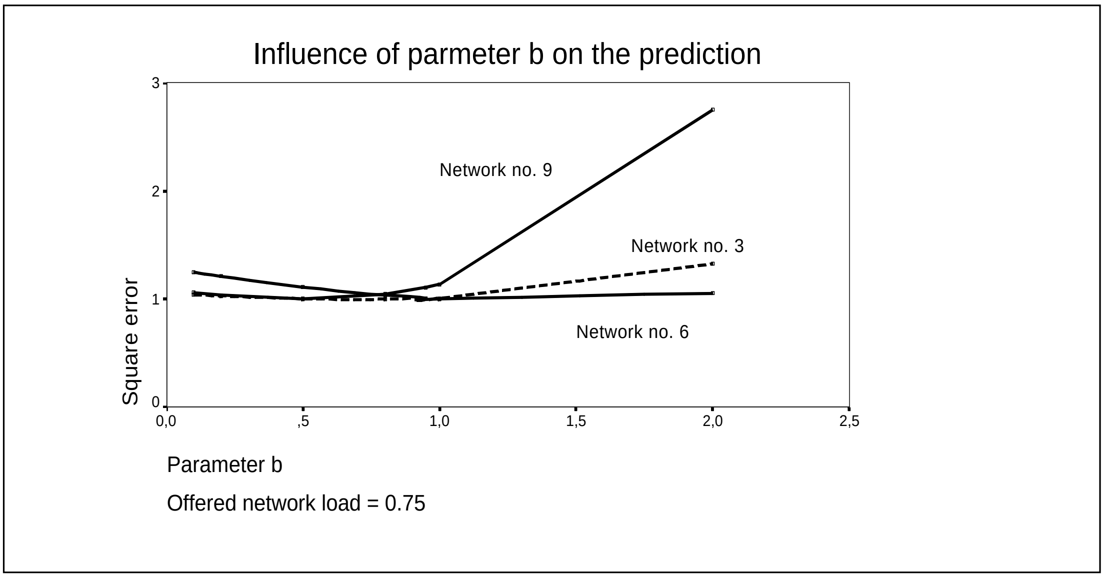

Figure 7.3: *The influence of the parameter b on the signal completion time prediction.*

7.4.4 Conclusions regarding State 2

The optimal state machine performance is sensitive to the choice of *b*, but for normal network loads the optimal prediction is found within 0.5 ≤ *b* ≤ 1.0. Further, the parameter *b* does not influence the stability of the state machine.

7.5 State 3

7.5.1 Stability in State 3

The prediction of the signal completion time in State 3 is *P(i,j,n) = cR(i,j,r)* and the prediction one term in the sum that defines *L(i,n)*. Then, there is not any memory of *P(i,j,n)* involved in the calculation of *L(i,n+k)* or *P(i,j,n+*k*) = cR(i,j,r+l), r+l>n+k*. The lack of memory in State 3 yields the conclusion that *c* can not influence the stability of the state machine.

7.5.2 Optimal prediction in State 3

The statistical behavior of signals is not known and it varies from network to network. Fortunately many analyses can be performed without explicit assumptions of the statistical nature of the signals. The prediction in State 3 is controlled by the parameter *c*. The assumptions in section 7.2 gives the prediction

$$P(n) = cR(n, r) \qquad 7.34$$

where *n-r* is the number of signals initiated since the initiation of the signal lending its completion time to this prediction. *R* is an actual signal completion time that is *n-r* signals old. The assumption of constant offered network load then gives that $\tilde{T}$ and $\tilde{R}$ have identical statistical distributions

$$\tilde{R} = \tilde{T} \tag{7.35}$$

An estimate of the squared difference between the actual signal completion time and the prediction of the signal completion time is given by

$$E\left[(\tilde{T}-\tilde{P})^2\right] = E\left[(\tilde{T}-c\tilde{R})^2\right] \tag{7.36}$$

This gives a minimum for $c = 1$.

7.5.3 Simulated results for parameter *c*

Simululations indicate that the best choice is $c = 1$ as shown in fig. 7.4. The minima of $E\left[(\tilde{T}-\tilde{P})^2\right]$ are located at $c = 1$ with fairly steep descents on both sides, indicating a noticeable loss of accuracy for the prediction of the signal completion time if $c \neq 1$. Within the range simulated, *c* does not reveal any tendencies to cause instability in the state machine. This makes the choice of *c* rather simple, just let $c = 1$. The simulations are performed with an offered network load of 0.95, since this is an area where State 3 prevails.

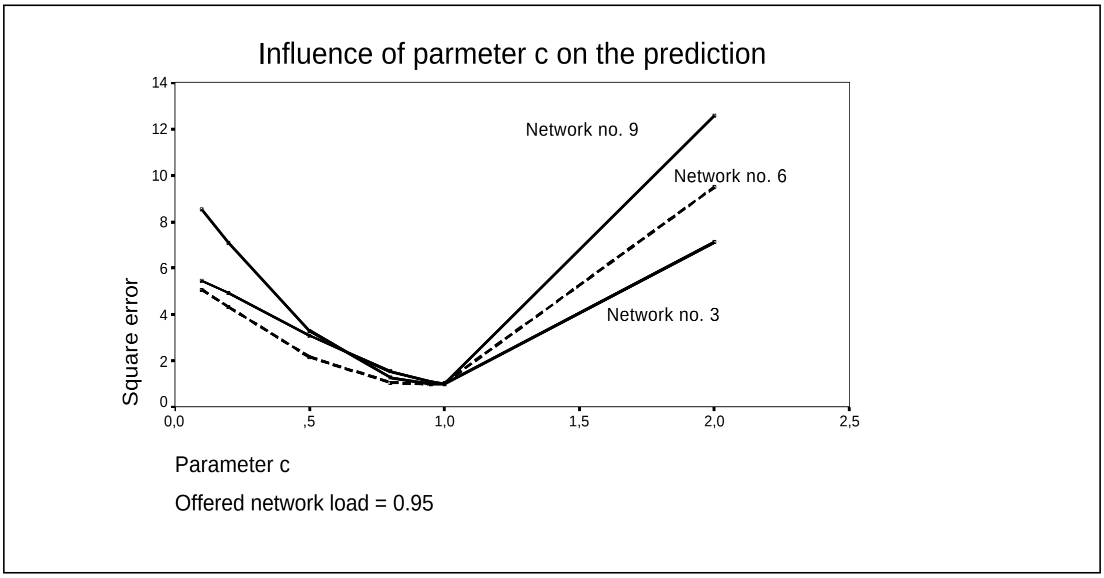

Figure 7.4: *The influence of the parameter c on the prediction of the signal completion time.*

7.5.4 Optimal time in State 3

The investigation of the optimal time to spend in State 3 is a non-trivial task. The squared error of the prediction of and the actual signal completion time is given by

$$E\left[(\tilde{T}-\tilde{P})^2\right] = E\left[(\tilde{T}-c\tilde{R})^2\right] \tag{7.37}$$

In section 7.5.3 *c*=1 proved to be the optimal choice. The time spent in State 3 after the last signalling event is $T_e(i,j,n)= dR(i,j,k)L(i,n)$. To simplify the notation, let $T_e(i,j,n)= t_e$ and to introduce time dependacy into 7.38, let T and R be a function of the time at their instigation

$$\begin{aligned} \tilde{R} &\rightarrow \tilde{T}(t) \\ \tilde{T} &\rightarrow \tilde{T}(t+\tau) \end{aligned} \tag{7.38}$$

τ is thus time difference bewteen the signal's departure times. Expression 7.38 may now be written as

$$E\left[(\tilde{T}(t+\tau) - \tilde{T}(t))^2\right] \tag{7.39}$$

State 3 yields the best prediction if not too long time has passed since the last signalling event, then State 1 can be expected to render the better prediction, this can be formulated as

$$E\left[(\tilde{T}(t+\tau) - \tilde{T}(t))^2\right] \le E\left[(\tilde{T}(t+\tau) - \tilde{P}_1)^2\right] \tag{7.40}$$

where $\tilde{P}_1$ now is the prediction made in State 1. The prediction $\tilde{P}_1$ is be made in process that only depends on the average network load, this makes which possible to consider $\tilde{P}_1$ constant. To make the predictions in State 1 challenge those of State 3, assume that

$$E[\tilde{P}_1] = E[\tilde{T}(t+\tau)] \tag{7.41}$$

The essence of equation 7.40 can now be written as

$$E[\tilde{T}(t) \cdot \tilde{T}(t+\tau)] \ge \frac{E\left[\tilde{T}(t)^2\right] + E[\tilde{T}(t+\tau)]^2}{2} \tag{7.42}$$

To once more worsen our predicament assume the process to be a first order weakly stationary process, i.e. the first moment of the process is time invariant and the covariance only depends on τ. This may be interpreted as the network is subject to a constant network load, something that makes it less obvious why to make a transition to State 1, unless the network load is very low. Using the covariance *r(*τ) we may write 7.42 as

$$r(\tau) \ge \frac{E\left[\tilde{T}(t)^2\right] - E[\tilde{T}(t)]^2}{2} = \frac{r(0)}{2} \tag{7.43}$$

This result may be illustrated by assuming a normal process and to let the network be represented by an M/G/1 queuing model. If we take a covariance function from a simple normal process we get

$$r(\tau) = \sigma^2 e^{-a|\tau|} \qquad 7.44$$

where σ^2 is the variance and α defines the memory of the process. As an example, let the memory be defined as

$$a = 1/E[\text{duration of a busy period}] \qquad 7.45$$

The expectation of a busy period in an M/G/1 queuing model with an arrival intensity λ and an expected service time $\bar{x}$ is

$$E[\text{duration of a busy period}] = \frac{\bar{x}}{1-\lambda\bar{x}} \qquad 7.46$$

The smallest τ can now be calculated

$$\tau = \frac{\bar{x}}{1-\lambda\bar{x}}\log 2 \qquad 7.47$$

From the beginning of this section we have t_e= *dR(i,j,k)L(i,n)* or simpler put: t_e= *dRL*. t_e is emidiatly identified as τ and so is *R* as one average time spent in the system conditioned on a busy period,

$$\frac{\overline{x^2} + 2\bar{x}^2(1-\rho)}{2\bar{x}(1-\rho)} \qquad 7.48$$

The load measure *L* has the same behavior as 1/ (1-$\lambda\bar{x}$). For this specific case we get

$$d \geq \frac{2\bar{x}^2(1-\rho)}{\overline{x^2} + 2\bar{x}^2(1-\rho)}\log 2 \qquad 7.49$$

7.5.5 Simulated results for optimal time in State 3

Simulations indicate the best choice is $d \geq 1.0$ as shown in fig. 7.5. The minimum of $E\left[(\tilde{T}-\tilde{P})^2\right]$ are located at *d* = 1with fairly steep decents on the lower *d* side and with rather flat decents on the other side. This implyes a fair loss of accuracy for the prediction of the signal completion time if *d* is set to be too small. On the other hand, overshooting *d* has only a small influence on the accuracy. Within the simulated range, *d* does not reveal any tendencies to cause instability in the state machine. This makes the choice of *d* rather simple, let *d* be slightly larger than 1. The simulations are performed with an offered network load of 0.75, since this is where the highest number of transitions from State 3 to State 1 can be expected.

.

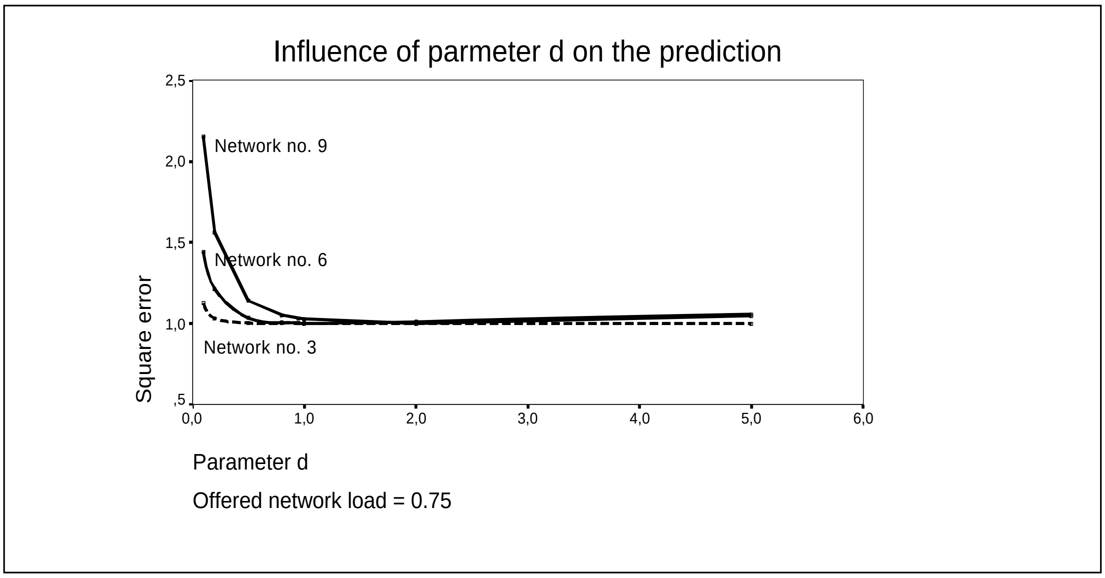

Figure 7.5: *The influence of the parameter d on the prediction of the signal completion time.*

The results above are generated from all states of the state machine and do therefore only depict the overall impact of the parameter *d* on state machine. The poor accuracy of the predictions for $d \leq 1.0$ are the result of too quick transitions to State 1. This will result in to many predictions made in State 1 when State 1 is not the optimal state to make those prediction in. The impact of the parameter $d \geq 1.0$ coincides with the results for the previous section.

7.5.6 Conclusions for State 3

The best choice is to let $c = 1$ and $d \geq 1.0$. This is not in all cases the optimal choice, but it gives a good prediction of the signal completion time and it keeps state machine in a stable state in all situations. Furthermore, it is easy to implement.

8. Numerical results

8.1 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 the final destinations. Each node comprises two parts: the lower layers and the upper layers, representing the OSI layers 1-3 and 4-7 respectively. Or equivalently expressed: One queue represents the Message Transfer Part (MTP) and the other represents the User Parts (UP) and Transaction Capabilities (TC). 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 to an outgoing link for further transport in the network (fig. 8.1).

Each composite layer is represented by a queue with an FCFS queuing discipline and with the service time being the sum of a constant time and a time derived from a negative exponential distribution. The mean service time of the server in the lower layers is fixed, while the mean service time for the upper layers is variable to model the complexity of the processing performed by the upper layers. The processing times in the upper layers are chosen to make congestion highly probable in either the lower or the upper layers. The two cases represent distinctly different congestion scenarios: one where MTP functions and one where UP/TC functions suffer from congestion. In the sequel, the two cases are referred to as lower layer congestion and upper layer congestion respectively.

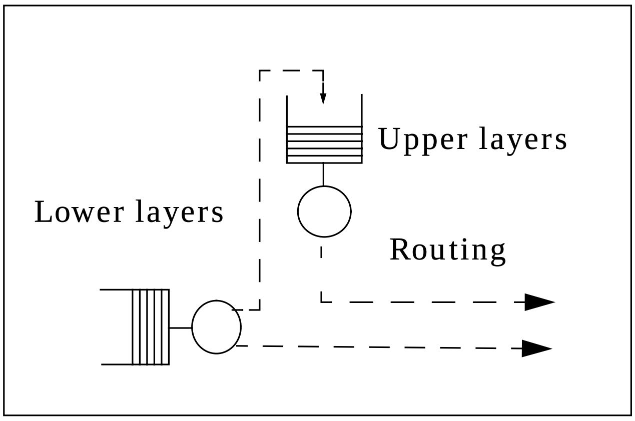

Figure 8.1: *Queuing model of node interior.*

The analysis are performed on 8 asymmetrical networks and on one symmetrical network. In asymmetrical networks all nodes have their own unique vicinity both in terms of connectivity and traffic, while in the symmetrical network all nodes have identical physical and statistical surroundings. The networks differ both in the number of nodes and the number of outgoing/incoming links per node. The network topologies are generated by a random process. Their characteristics are presented in the table 8.1 below.

Fixed routing is 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. The routing tables are calculated by a shortest path algorithm, where shortest is in terms of nodes traversed. The transmission delays are incorporated in the lower layers service times. The selected routing strategy complies with the strategy of SS7 [38].

In this investigation is the number of signals, *k*, per service session derived from a uniform distribution and is in the range of $2 \leq k \leq 19$. The number of signals per service session will then span from 4 to 38. In the PSTN is the number of signals per service session less than 10 for all services, while in GSM it may even exceed 40.

The generated sessions are uniformly distributed between the originating nodes, and is derived from a negative exponential distribution. The destination nodes are selected in the same fashion.

Network no.	Number of nodes	Congestion in	Average no. of links per node	Real time demand: *B*
1	10	lower layers	3.0	10.5
2	10	lower layers	4.0	10.5
3	10	upper layers	3.0	10.5
4	10	upper layers	4.0	10.5
5	40	lower layers	3.0	10.5
6	40	lower layers	4.5	10.5
7	40	upper layers	3.0	10.5
8	40	upper layers	4.5	10.5
9 Symmetrical	20	lower layers	4.0	8

Table 8.1: *Properties of the studied networks.*

8.2 The metric

We use the carried load, i.e. the number of sessions completed within their allowed service completion time divided by the number of a generated sessions at an offered network load of 1.0, as a metric. The metric discloses the network's ability to handle the present offered network load under the constraint of service related real time demands. It also reveals the possibility for a session to fulfill its mission as requested by a customer, and is thereby closely related to the part of customer satisfaction that is derived from network performance.

The real time demands are expressed as $D(i,n,k,k) < Bmin(s)$, where $B = 10.5$. This is a moderate demand for the networks that suffer from lower layer congestion and a reasonably harsh demand from the networks that suffer from upper layer congestion. The real time demand in network no. 9 is in-between the above mentioned cases.

The carried load cannot reach 1.0 if signaling session real time demands are present. See fig. 6.2 for maximum carried load at a given maximum allowed session completion time in an unadulterated network. E.g. the maximum carried load is 0.83 and the offered load threshold is 0.86 for an allowed session completion time of 8.0 t.u, or equivalently $D(i,n,k,k) > 8min(s)$.

8.3 Static and transient performance

The impact of the proposed CCM is negligible at normal offered network load and increases dramatically with offered network load [34,35]. In other words, it does not interfere with the network under normal working conditions, i.e. an offered network load below the load threshold, but steps into action when congestion arises. Simulations reveal significant improvements of throughput, compared to an unadulterated network, at offered network loads above the load threshold. The proposed CCM also performed better than our interpretation of the existing CCM in SS7 in a comparison [36].

The proposed CCM also performs well during transients, even at instantaneous offered load changes with a offered load peak magnitude of 10.0 [33].

8.4 Robustness

The topologies of signaling networks may vary widely and the proposed CCM must be able to function in all network topologies without any change in its behavior. The networks display variation in size, connectivity, traffic demand, real time demands, and the location of where a congestion is likely to arise. Robustness is the ability to maintain peak performance independent of the environment. The robustness of the signaling network is crucial and thus is the robustness of a signaling network congestion control mechanism not negotiable[39].

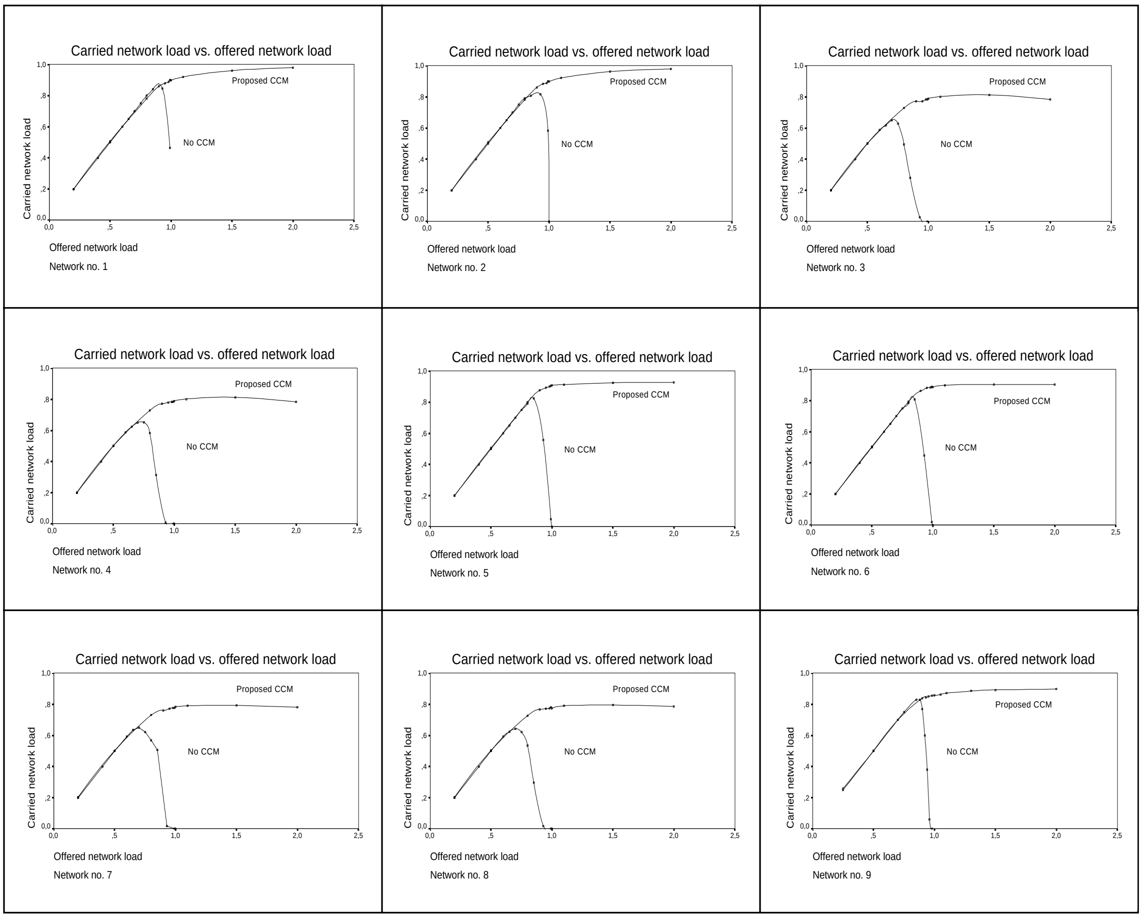

Figure 8.2 a-i: *Carried network load vs. offered network load. The static behavior of the networks no. 1-9 utilizing the proposed CCM are represented by the upper lines and the lower lines represent the carried load without any CCM.*

The proposed CCM handles all nine networks well during steady state as shown in fig. 8.2 a-i. The conclusion must be that the CCM with identical tuning parameters is robust regarding network size, topology, traffic demands, real time demands, and congestion location.

The diagrams presented in the sequel only cover network no 3, 6 and 9 since these network containes the essence of the presented networks. The proposed CCM must also handle transients in offered load without any signs of instability or of delay in reaction time. The networks are subject to a pulse in offered load, with a magnitude of 2.0, to facilitate a study of its behavior during transients (fig. 8.3 a-c). Before initiating the pulse, the networks are kept at low offered load (0.25) until a steady state is reached. The duration of the pulse is long enough to let the networks enter a new steady state. The offered load of 0.25 is resumed after the offered load pulse. The proposed CCM follows the pulse well.

The slope of the carried load at the transient in fig. 8.3 a-c is partly due to the time it takes to build up the new load in the network, and partly to the reaction time of the proposed CCM. Nevertheless, the time span of the slope is not longer than a few session completion times. The conclusion must be that the proposed CCM is able to protect the network at rapidly emerging congestions, while maintaining a high carried load.

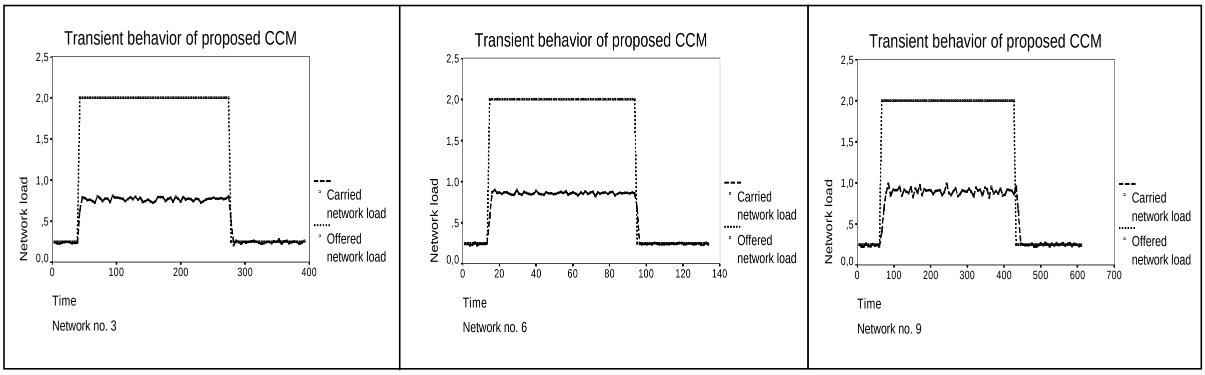

Figure 8.3 a-c: *The transient behavior of the proposed CCM for networks no. 3, 6 and 9 respectively. Dashed lines represent the offered network load and solid lines represent the carried load while utilizing the proposed CCM. Time is defined so that 100 sessions are initiated per time unit in the diagrams.*

The proposed CCM is able to produce a carried load higher than obtainable from an unadulterated network at offered loads in excess of the load threshold. This is due to the annihilation of sessions with a predicted completion time exceeding *Amin(s)* gives “extra space” in the network for sessions with a predicted completion time below *Amin(s)*, and that extra space increases the probability for the surviving sessions to meet their real time demands, thus the increase of carried load. This effect is more evident at higher real time demands as proportionally more sessions are annihilated in this case. The CCM can not entirely free the network of too long sessions, and therefore is the carried network load less than 1. The phenomenon is obvious in fig. 8.2 a-i and visible in fig.8.3. The maximum carried network load for a network without a CCM can be obtained from fig. 8.2 a-i.

8.5 Admission

The interarrival times for the generated sessions follow a negative exponential distribution, and our analysis shows that the distribution of the interarrival times of admitted sessions are close to the negative exponential distribution, but with a slight increase in the probability for the longer interarrival times of admitted sessions.

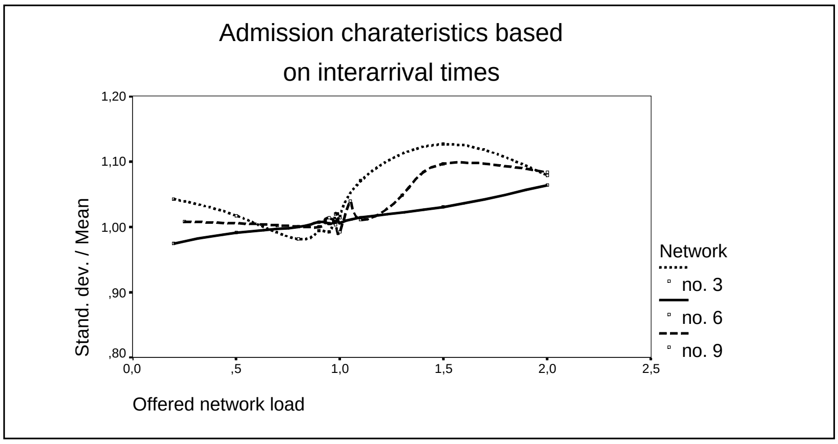

Figure 8.4: S*tandard deviation of the interarrival times divided by the mean of the interarrival times for networks no. 3, 6 and 9.*

This indicates that the proposed CCM is both fair in its session admittance policy, and is free from exaggerated oscillating between time periods of admittance and non-admittance (fig. 8.4).

9. Focused overload

9.1 Introduction

In the previous section the attention has been on entire networks and congestion is assumed to prevail throughout the network. This scenario is unlikely as some of the major causes of congestion are component failures and events like mass call-ins. These reasons point to the conclusion that only one or a few nodes of the network may suffer from overload, at least in an initial stage of the overload situation and then may the whole network gradually become congested if the situation is left unattended. The situation where a single node is congested is referred to as focused overload. The behavior of the proposed CCM at focused overload can not be overlooked in this investigation.

The SS7 comprises functionality from the MTP level to the UP/TC level and congestion may occur in all of the protocol levels. The signaling network node model in this study contains to layers, representing the MTP and UP/TC functionality, and the issues of congestion in each of them need to be addressed to thoroughly investigate the properties of the proposed CCM.

The networks studied in this section are identical to those in the previous section, the only exception is the omittance of network no. 9.

The node with the highest arrival intensity of signals is selected to become the congested node in both the studied cases. The arrival intensity is a concern of both the physical network architecture and the routing algorithm, but in the model is the node with most links also most likely to have the highest arrival intensity. In the lower layer case this corresponds to a worst case scenario as the congestion will have maximum impact on network performance. In the upper layer case can a high number of links imply a important facility being located in the node.

The overload is induced by increasing the service time in the concerned server. This does not correspond to a realistic cause of overload, but it had the advantage of “clean” implementation in the model. The interpretation of “clean” is the comparatively small impact of the behavior of the surrounding network so that the focused overload can be investigated without interference from the surrounding network. This implementation reduces the effectiveness of the CCM since the load increase does have the corresponding increase in completion time information as would be the case if the load increase was caused by an increase in the signalling volume.

9.2 Results on focused overload

The ability of the proposed CCM to function well in all network topologies without any change in its behavior is equally important in the focused overload case. As mentioned in section 8.1 do the networks display variation in size, connectivity, traffic demand, and real time demands.

The loads referred to on the axis are the loads in the overloaded node.

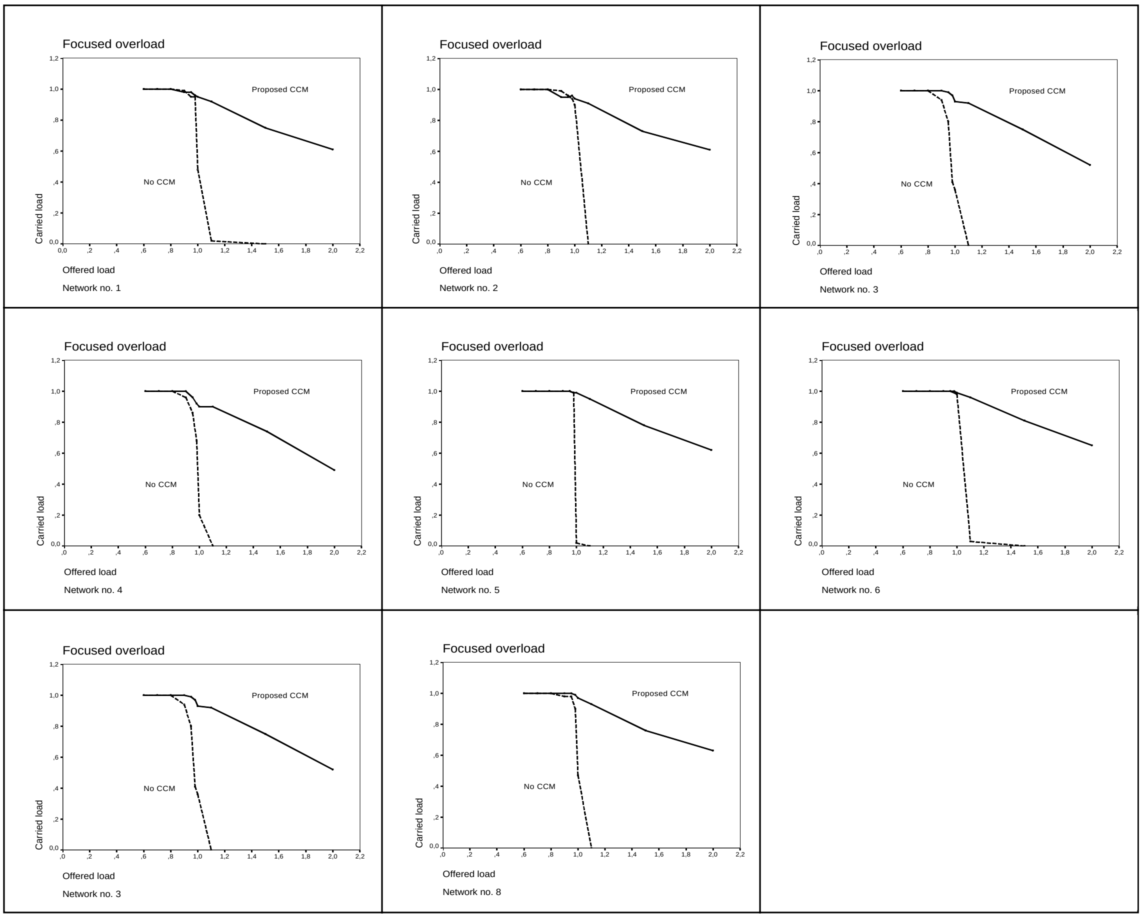

Figure 9.1 a-h: *Carried network load vs. offered network load. The static behavior of the networks no. 1-8 utilizing the proposed CCM are represented by the upper lines and the lower lines represent the carried load without any CCM.*

The proposed CCM handles all eight networks well during focused overload at steady state as shown in fig. 8.2 a-h. The conclusion must be that the CCM with identical tuning parameters is robust regarding network size, topology, traffic demands, real time demands, and congestion location.

10. Conclusion

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

A simple CCM that predicts the session completion time from the most reliable signaling events of a node, and annihilates the session if the predicted completion time is found to be too long, improves network performance at congestion significantly. This applies to a signaling network both during steady state and transients loads. The proposed CCM has proven to handle very high overloads. It is even able to increase the carried load beyond the maximumum of carried load for a network without a CCM.

The proposed CCM is both stable and robust. It handles a wide range of signaling network topologies, connection densities, traffic demands, real time demands, and congestion location. Further, its performance is close to peak performance in the studied networks without any tuning.

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 [27], to all call attempts, and to services traversing any number of nodes.

11. Future work

The studied CCM can be refined in a number of ways. One way is to calculate $L_n(i,j)$ for each individual outgoing signaling link of a node and not as present, calculated per on information from all outgoing links.

The proposed CCM cannot be expected to reveal all possible flaws or benefits unless studied under more realistic circumstances. The assumptions in this paper of uniform service call intensity distribution over the nodes of the network is not entirely realistic.

Congestion control is not separable from routing. The proposed CCM must also be able to work in conjunction with routing algorithms. This aspect is not investigated in this report.

12. References

[1] B. Cotton, T.A. Pappas, “Effect of Service Completion Time and Network Interactions on the Acceptability of Advanced Intelligent Network Services”, Proceedings of the XIV international switching symposium, Yokohama, 1992.

[2] G. Pollini, K. Meier-Hellstern and D. Goodman, “Signaling Traffic Volume Generated by Mobile and Personal Communications”, IEEE Communications Mag., no. 6, pp. 60-65, 1995.

[3] B.A.J. Banh and G. Anido, “Signaling Network Design Aspects For Mobile Services”, Proceedings of the Australian Telecommunication Networks & Applications Conference, pp. 695-700, Melbourne, 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 Communications, vol. 12, no. 3, pp. 501-509, 1994.

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

[6] B. Jabbari, “Common Channel Signaling System No. 7 for ISDN and Intelligent Networks”, Proceedings of the IEEE, vol. 79, no. 2, pp. 155-169, 1991.

[7] G. Willmann and P. Kühn, “Performance Modeling of Signaling System No. 7”, IEEE Communications Mag., no. 7,pp. 44-56, 1990.

[8] J. Thörner, Intelligent Networks, Artech House, 1994.

[9] J. Harju, T. Karttunen, O. Martikainen, “Intelligent Networks”, Chapman & Hall, 1995.

[10] M. Fujioka and Y. Wakahara, “Consideration on Common Channel Signaling Evolution for Global Intelligent Networking”, IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, pp. 510-516, 1994.

[11] “CME 20 System Training Document”, Ericsson, EN/LZT 120 226 R2A, 1991

[12] M. Kihl, “On Overload Control in Intelligent Networks”, Technical Report 118, Department of Communication Systems, Lund University of Technology, 1996.

[13] M.P. Rumsewics, “A comparison of SS7 Congestion Control Options During Mass Call-In Situations”, IEEE/ACM Transactions on Networking, vol. 3, No. 3, pp. 1-9, 1995.

[14] D.J. Houck, K. S. Meier-Hellstren, F. Saheban and R. A. Skoog, “Failure and Congestion Propagation Through Signaling Controls”, Proceedings of the ITC 14, Juan-Les-Pins, pp. 367-376, 1994.

[15] U. Körner, C. Nyberg and B. Wallström, “The impact of New Services and New Control Architectures on Overload Control”, Proceedings of ITC 14, pp. 275 - 283, 1994.

[16] H. Nyberg, B. Olin, “On Load Control of an SCP in The Intelligent Network”, Proceedings of Australian Telecommunication Networks & Application Conference, pp. 751 - 756, 1994.

[17] P.J. Kuhn, “Overload Control For Signalling Networks”, Proceedings of the St. Petersburg regional International Teletraffic Seminar, 1995.

[18] J. Walrand, “Communication Networks: A First Course”, Aksen Associates Incorporated Publishers, 1991.

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

[20] M.P. Rumsewics, “Critical control issues in the evolution of Common Signaling Networks”, Proceedings of the ITC 14, Juan-Les-Pins, pp. 115-123, 1994.

[21] U. Ahlfors, “Overload Control In Distributed-Memory Systems”, Technical Report 119, Department of Communication Systems, Lund University of Technology, 1996.

[22] M.M. Mostrel, “Issues on the design of Survivable Common Channel Signaling Networks”, IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, pp. 526-532, 1994.

[23] A.T. Leung and S. Wainberg, “Deployment issues of CC Links on Self-Healing Rings”, IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, pp. 539-543, 1994.

[24] D.R. Manfield, G. Millsteed and M. Zukerman, “Performance Analysis of Congestion Controls Under Sustained Overload”, IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, pp. 405-414, 1994.

[25] J. Zepf, G. Willman, “Transient Analysis of Congestion and Flow Control Mechanisms in Common Channel Signalling Networks”, Proceedings of ITC-13, pp. 413-419, 1991.

[26] D.E. Smith, “Effects of Feedback Delay on the Performance of the Transfer-Controlled Procedure in Controlling CCS Network Overloads”, IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, pp. 424-432, 1994.

[27] S. Pettersson and Å. Arvidsson, “A decision theoretic approach to congestion control in signaling networks”, Proceedings of NTS-13 in Trondheim, 1996.

[28] S. Pettersson and Å. Arvidsson, “A profit optimizing strategy for Congestion Control in Signaling Networks”, Proceedings of the ITC-Seminar in Bangkok, 1995.

[29] S. Pettersson and Å. Arvidsson, “Economical aspects of a Congestion Control Mechanism in a Signaling Network”, Proceedings of NTS-12 in Helsinki, 1995.

[30] S. Pettersson, “Some results on optimal decisions in network oriented load control in signaling networks”, Technical Report, University of Karlskrona/ Ronneby, 1996.

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

[32] 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.

[33] M.P. Rumsewicz, “Analysis of the Effects of SS7 Message Discard Schemes on Call Completion Rates During Overload”, IEEE/ACM Transactions on Networking, vol. 1, no. 4, 1993

[34] L. Angelin, S. Pettersson and Å. Arvidsson, “A network approach to signaling network congestion control”, Proceedings of the ITC Seminar in St. Petersburg, 1995.

[35] L. Angelin and Å. Arvidsson, “A Congestion Control Mechanism for Signaling Networks based on Network Delays”, Proceedings of NTS-12 in Helsinki, 1995.

[36] L. Angelin and Å. Arvidsson, “A Congestion Control Algorithm for signaling Networks Based on Network Delays”, Proceedings of the ITC-Seminar in Bangkok, 1995.

[37] S.M. Ross, “Stocastic Processes”, Wiley Series in Probability and Mathematical Statistics, ISBN 0-471-09942-2, 1983.

[38] W. Klein and R. Kleinewillinghöfer-Kopp, “Performance Analysis of a Large-Scale Common Channel signaling Network”, Teletraffic and Datatraffic in a Period of Change, ITC-13, Elsevier Scientific Publishers, 1991.

[39] V. Karmarkar, “Assuring SS7 Dependability: A Robustness Characterization of Signaling Network Elements”, IEEE Journal on Selected Areas in Communications, vol. 12, no. 3, 1994.