World Applied Sciences Journal 7 (Special Issue of Computer & IT): 129-137, 2009
ISSN 1818.4952

Analysis of Call Set-up Delay for Different Resource Control Schemes in Next Generation Networks

[1]*Mahmoud Pirhadi and* [2]*Seyed Mostafa Safavi Hemami*

[1]Islamic Azad University, Science and Research Branch, Tehran, Iran
[2]Amirkabir University of Technology, Tehran, Iran

Abstract: Quality of Service (QoS) and resource management in Next Generation Networks (NGN) is provided by a particular architecture called RACF (Resource and Admission Control Function) which is introduced by ITU. In this paper we analyze the impact of different resource control schemes in RACF architecture on Call Set-up Delay (CSD) which is a QoS performance parameter that may be used in specifying, measuring and comparing the speed of call set-up processing in telecommunication networks. The approach taken is, initially, to introduce the RACF architecture and its main elements and protocols. Following this, we present two QoS signaling call flows for two resource control schemes and examine them using a queuing model. The simulation results show the different performance of these two schemes under different network traffic. Using these results, we can have some criteria for choosing the better resource control scheme appropriate to traffic changes.

Key words: Next generation network . quality of service . RACF . call set-up delay . queuing

INTRODUCTION

Next Generation Network (NGN) is a wide term that strives to deliver various services existing in separate technological planes today over an all-IP network, i.e., using packet switching. During recent years, great efforts have been initiated in order to converge various telecommunications networks. According to ITU's definition, NGN is a packet-based network able to provide telecommunication services and able to make use of multiple broadband, QoS (Quality of Service)-enabled transport technologies and in which service-related functions are independent from underlying transport-related technologies. It offers unfettered access by users to different service providers [1].

NGN separates Service layer from transport layer in the new network design. Transport layer is composed of access and core IP networks that will be used to provide global connectivity in all-IP networks, both wired and wireless. Service layer, sometimes referred to as control layer, is to be used to connect services and is defined in an abstract way so that services would not depend on underlying transport network technology.

One of the critical issues in Next Generation Networks (NGN) development is QoS provisioning and resource management. End-to-end QoS and network performance is gaining increasing interest in NGN development and standardization activities. The main task of NGN is to move all currently existing non-IP network technologies to packet switching. The move is complicated by various QoS requirements on the part of various existing technologies. It is difficult for traditional packet switching to support fine QoS granularity. Services which have been offered in PSTN so far should be offered now in NGN with the same or even better quality while these two networks are quite different from a technical point of view. In PSTN, QoS is guaranteed due to allocation of a dedicated circuit to each call, while in packet-based networks there is not any dedicated circuit and usually resources of the network are shared between all users. Different applications generate different types of traffic each of which has its own QoS requirements. Therefore, the network resources have to be managed so that each call gets enough resources to guarantee the quality of service [2].

This has made clear-sighted organizations such as ITU and ETSI to propose models and architectures for provision of resource management in NGN networks. The models and architectures include various elements in each layer and specified protocols are used between these elements. Some of these protocols have evolved and have become mature, while the others are still being developed.

An architecture which has been introduced by ITU-T for the sake of resource management is called RACF (Resource and Admission Control Function),

Corresponding Author: Dr. Mahmoud Pirhadi, Islamic Azad University, Science and Research Branch, Tehran, Iran

which is introduced and analyzed in the next sections. ETSI has also recommended a model for resource and admission control in NGN that is envisaged as an instance of ITU RACF for fixed access networks [3].

QoS provisioning in NGN depends on several parameters such as IP Packet Transfer Delay (IPTD), IP packet delay variation (IPDV), IP packet loss ratio (IPLR) etc., [4], which are mainly dependent on IP transport network performance and also parameters such as Call Set-up Delay (CSD), call misrouting probability and call set-up failure probability [5], which are mainly dependent on call processing nodes' performance [6].

In this paper we present some QoS signaling call flows based on the architecture proposed by ITU for different resource control schemes. The CSD is simulated and the resource control schemes are compared based on their performance in various network traffic. Following this, the optimum points for choosing the best scheme can be found.

The rest of the paper is organized as follows. The next section is an overview of RACF architecture and its main elements and their roles in the architecture. In section 3, the resource control schemes are introduced and two QoS signaling call flows for two main scenarios, i.e., single-phase and two-phase are presented. Queuing models and simulation method is discussed in section 4. In section 5 a short discussion on resource availability for signaling traffic is presented. Section 6 presents the simulation results and finally, Section 7 concludes the paper.

RESOURCE AND ADMISSION CONTROL FUNCTIONAL ARCHITECTURE

Figure 1 illustrates a simplified model of resource and admission control architecture recommended by ITU for supporting end-to-end QoS in NGN [7]. In this architecture RACF acts as the mediator between Service Control Functions (SCF) and transport functions for QoS-related transport. One of the basic functionalities of RACF is to make decisions according to defined policies based on resources status in transport layer and also based on utilization information, Service Level Agreements (SLA), network policy rules and service priorities. The RACF presents a view of transport network infrastructure to the SCF so that service providers do not need to know the details of the transport layer such as network topology, connectivity, resource utilization, QoS mechanisms, etc. The RACF interacts with the SCF and transport functions for the applications that require resource control in the transport layer. SIP-based call flows presented in this paper are examples of such applications.

The SCF represents the functional entities of NGN service layer such as call servers and SIP proxies which can request QoS resource and admission control for media flows of a given service via its interface to RACF.

The RACF applies control policies to transport resources, e.g., routers, upon SCF requests, determines whether transport resource is available and makes admission decisions. The RACF interacts with transport functions to control the following tasks in the transport stratum from the QoS point of view: bandwidth reservation and allocation, traffic classification, traffic marking, traffic policing and priority handling.

As illustrated in Fig. 1, functional entities of RACF are PD-FE (Policy Decision Functional Entity) and TRC-FE (Transport Resource Control Functional Entity).

The main functionality of PD-FE and TRC-FE is to make policy decisions and to determine network resources availability, respectively.

Dividing RACF into two distinct functions, i.e., PD-FE and TRC-FE, enables it to support variant networks within a general resource control framework. Also the PE-FE (Policy Enforcement Functional Entity) in the transport layer is a gateway at the boundary of different packet networks, e.g., edge routers and/or between the CPE (Customer Premises Equipment) and access networks. Dynamic QoS is enforced in PE-FE.

The capabilities of transport networks and associated transport profiles of the subscribers are considered in RACF to support the transport resource control function. The interaction between RACF and Network Attachment Control Functions (NACF) includes network access registration, authentication and authorization, parameter configuration, etc., for checking transport subscriber profiles.

NACF encompasses a collection of functional entities that provide a variety of functions for network management and configuration to provide user access based on the user profiles.

RESOURCE CONTROL SCHEMES AND QOS SIGNALING CALL FLOWS

The QoS resource control process consists of three logical states. These states can occur in one or more steps as described below:

Authorization: The QoS resource is authorized based on policy rules. The authorized QoS bounds maximum amount of resources that can be allocated to a specified user.

Reservation: The QoS resource is reserved based on the authorized resource and resource availability. The

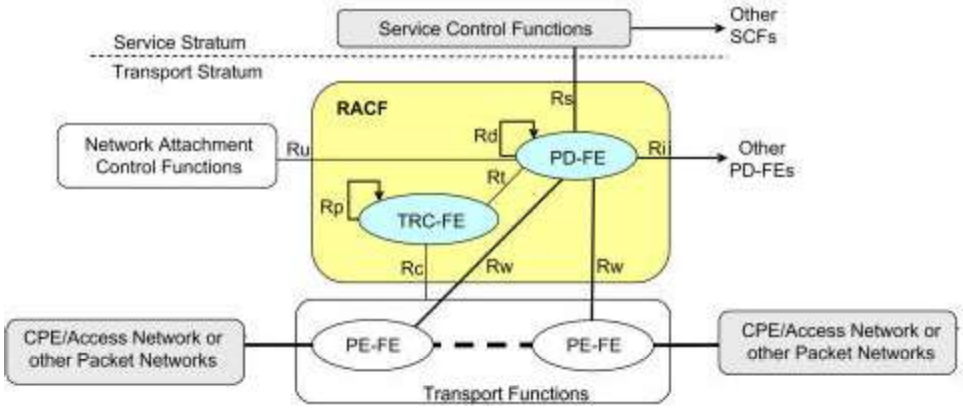

Fig. 1: Resource and admission control functional architecture in NGN

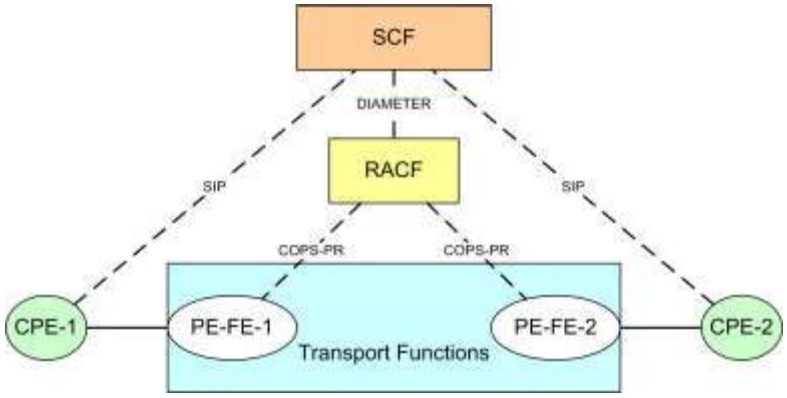

Fig. 2: Generic network architecture and QoS signaling protocols

reserved resource can be used by best effort media flows when the resource has not yet committed in the transport functions.

Commitment: The QoS resource is committed for the requested media flows when the gate is opened and other admission decisions (e.g., bandwidth allocation) are enforced in the transport functions.

According to the diversity of application characteristics and performance requirements, the RACF supports three different schemes of resource control:

Single-phase scheme: Authorization, reservation and commitment are performed in a single step. The requested resource is immediately committed upon successful authorization and reservation.

Two-phase scheme: Authorization and reservation are performed in one step, followed by commitment in another step. Alternatively authorization is performed in one step, followed by reservation and commitment in another step.

Three-phase scheme: Authorization, reservation and commitment are performed in three steps sequentially [7].

This paper focuses on the first two schemes and their impacts on network performance using queuing models for simulating their signaling call flows. The same method can be used also for three-phase scheme.

Figure 2 illustrates the topology, elements and QoS signaling protocols in a simple generic NGN network. In this architecture SCF is assumed to be a SIP proxy server and hence it uses SIP [8] to communicate with the CPEs. Q.3301 [9] (DIAMETER [10]) is used between SCF and RACF and COPS-PR [11] is used between RACF and the transport layer elements according to ITU-T recommendations.

Figure 3 and 4 illustrate two simplified call flows for session establishment and tear down between two end users using SIP signaling. SIP is introduced in RFC 3261 provided by IETF and is used to control the sessions and provide their signaling [8]. These call flows are extracted for two different schemes of resource control, i.e., single-phase and two-phase based on the architecture shown in Fig. 2. For the sake of

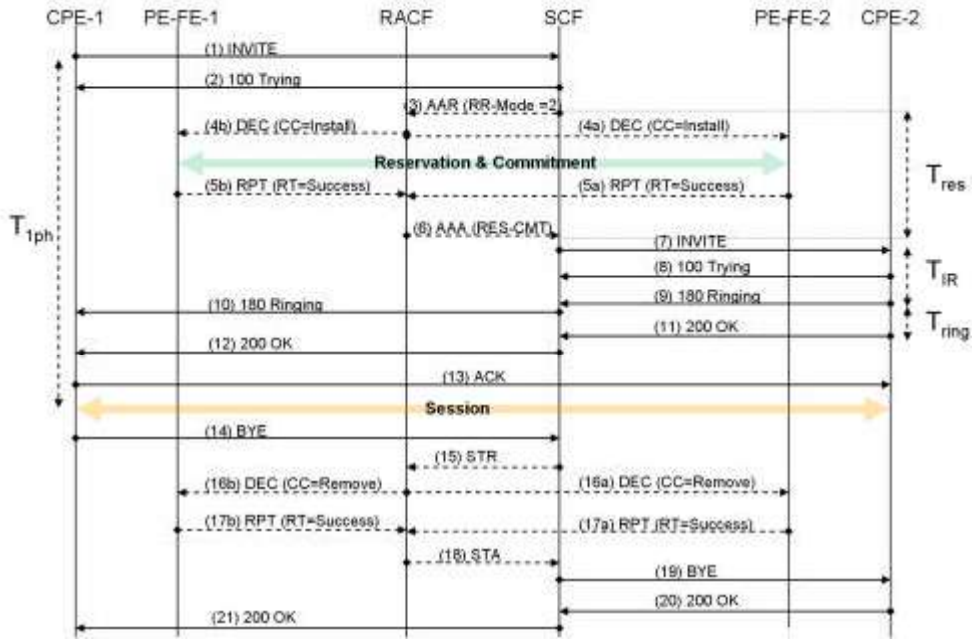

Fig. 3: Signaling call flow of single-phase scheme

simplicity and without loss of generality, the messages exchanged between RACF and NACF are not indicated in the call flows.

Figure 3 depicts the single-phase scheme signaling call flow (set-up and tear down). The call set-up function begins when the calling user issues an *INVITE* request (event 1) and ends (successfully) when the called user receives the corresponding *ACK* to its final *200 OK* (event 13).

- Events 1 & 2: Service is requested by CPE-1 from SCF. When the request is received, reservation and commitment steps are initiated. SCF does not forward the *INVITE* message to CPE-2 before the end of reservation and commitment steps.
- Event 3: Resource reservation and commitment are requested by SCF from RACF. The request is based on the DIAMETER protocol and is sent through the *AAR* command [10]. According to Q.3301.1 this request should have the *Resource-Reservation-Mode=2* option that means authorization, reservation and commitment steps should be performed in a single step [9].
- Event 4 (a & b): A command is issued from RACF to PE-FEs for resource reservation and commitment in transport layer in a bidirectional path. This command is based on COPS-PR protocol and is issued through *DEC* message. This command should have the *Command-Code=1* option that means configuration should be installed [11].
- For the sake of simplicity extra COPS-PR messages and RSVP messages are not indicated.
- Event 5 (a & b): RPT messages are reported from PE-FEs to RACF which means resource reservation and commitment have been performed successfully [11].
- Event 6: Receiving RPT messages from both PE-FEs, RACF answers to SCF through the *AAA* command which means resource reservation and commitment have been performed successfully [9, 10].
- Event 7: *INVITE* request is forwarded to CPE-2.
- Events 8-13: These events relates to the session establishment with respect to the RFC 3261 [8].
- Events 14-21: These events relates to the session termination and releasing the network resources.
- Figure 4 depicts the two-phase scheme call set-up signaling call flow. The differences between the first and second scenarios are as follows:
- Event 3: Resource reservation is requested by SCF from RACF. The request is based on the DIAMETER protocol and is sent through the *AAR* command. According to Q.3301.1 this request should have the *Resource-Reservation-Mode=1* option that means only authorization and reservation should be performed in one step.
- Event 12: Resource commitment is requested by SCF from RACF. The request is based on the DIAMETER protocol and is sent through the *AAR* command. According to Q.3301.1 this request should have the *Resource-Reservation-Mode=3*

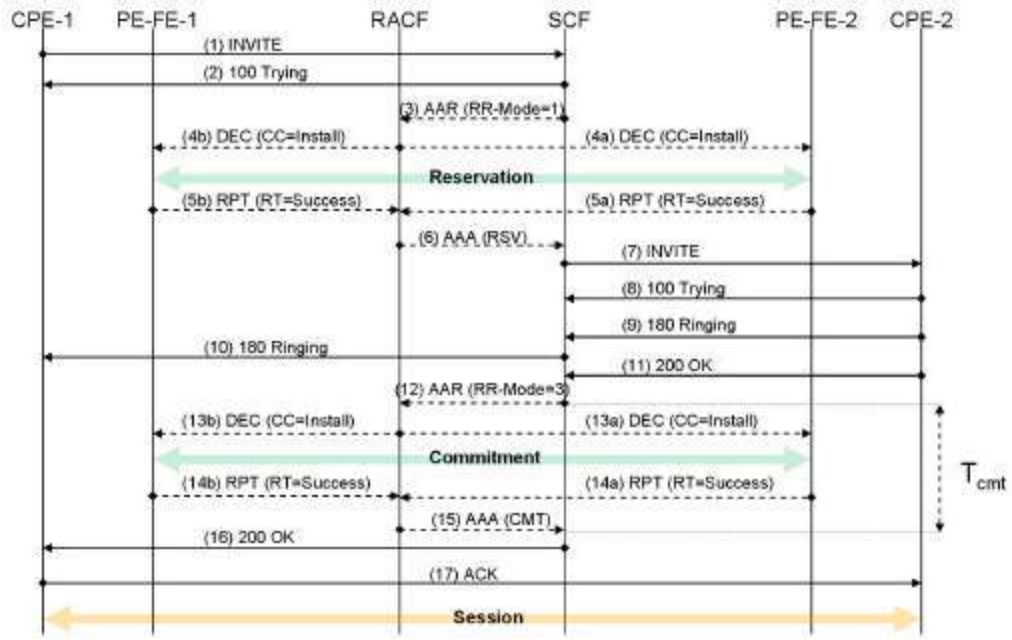

Fig. 4: Signaling call flow of two-phase scheme

option that means only commitment step should be performed. This request will be sent when the message *200 OK* is received from CPE-2 to SCF (off-hook state).

As the figure indicates, signaling flow has more stages in the two-phase scheme than the corresponding single-phase. This causes more amount of queuing and processing delay. It seems that single-phase scheme has a smaller call setup delay than two-phase, however, in networks with limited resources two-phase scheme can outperform single-phase. This is owing to the fact that, in the two-phase scheme, network resources are only reserved but not committed during the time between ringing and going off-hook and hence these resources can be in use by the signaling and best effort traffics. In the next section we are going to examine these issues and show the effects of different parameters on the performance of single and two-phase resource control schemes.

QUEUING MODEL AND SIMULATION OF SIGNALING CALL FLOWS

According to ITU-T recommendations, Call Setup Delay (CSD) is one of the important characteristics of QoS. CSD is the total call establishment time regardless of the delay associated with the called party's answer to the incoming call [5]. It is the elapsed time between the calling user's issuance of an *INVITE* message and the called user's receipt of the corresponding *ACK* message, excluding the called user delay, i.e., the time between the called user's receipt of the *INVITE* message and issuance of the corresponding *200 OK* message [12].

CSD for single-phase scheme call flow can be obtained by means of the following equation according to Fig. 3.

$$CSD_{1ph} = T_{1ph} - (T_{1R} + T_{ring}) \quad (1)$$

In which T_{1ph} is the duration between sending the *INVITE* request and receiving the *200 OK* message by CPE-1, T_{IR} is the duration between receiving the *INVITE* request and sending *180 Ringing* message by CPE-2 and T_{ring} is the average time required for answering to the incoming call (average ringing time).

CSD for two-phase scheme call flow can be obtained by means of the following equation according to Fig. 4:

$$CSD_{2ph} = CSD_{1ph} + T_{cmt} \quad (2)$$

Where T_{cmt} is the required signaling time for committing the reserved resources.

Now we are going to use a queuing model to simulate and calculate the CSD for the two schemes [13-15]. Processing function of each message is assumed as a state and a queuing system is considered for each state. CSD includes the transmission time of the final *ACK* but excludes the response times of the calling and called users. The provisional (*1xx*) SIP

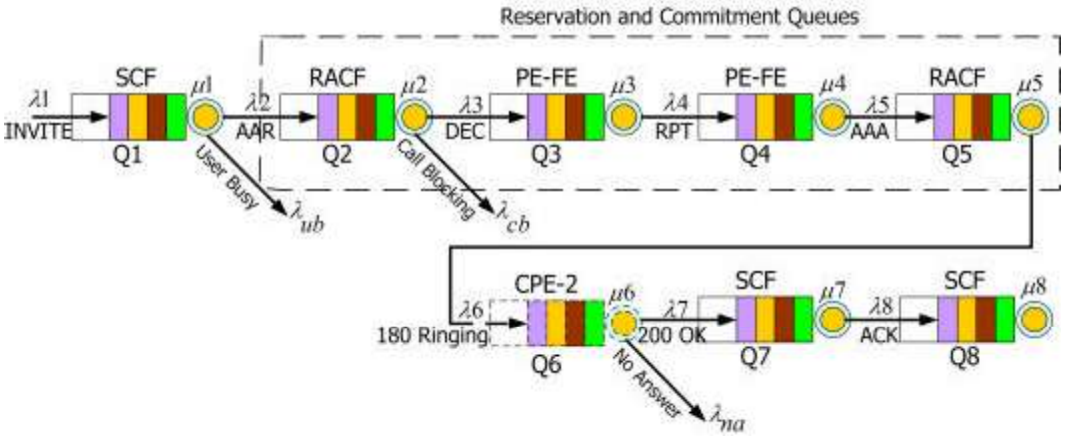

Fig. 5: Single-phase queuing model

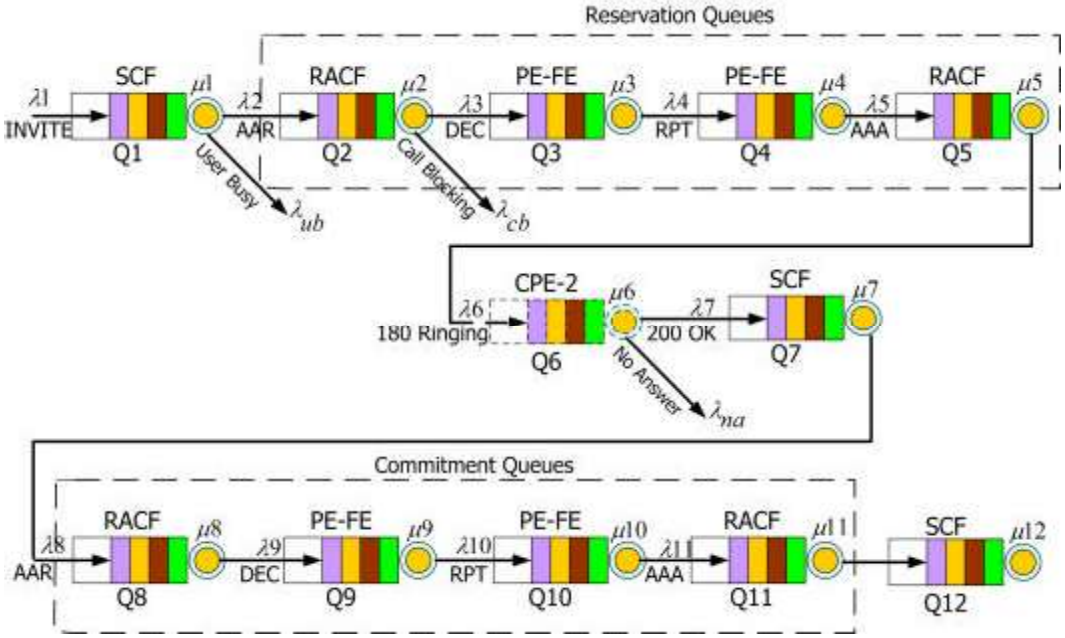

Fig. 6: Two-phase queuing model

responses have no effect on the CSD definition. Therefore the queuing systems of provisioning messages are not considered in the queuing model.

Each queuing station is modeled as an M/M/1 queue. The first M, which represents the arrival distribution, is memoryless. The arrival of one call request is independent of other requests. The second M, representing the serving distribution, also is memoryless. The serving time of a message is independent of all other messages. The 1 means there is only one server in a system. This model is an open, feed forward queuing network, since jobs arrive from an outside source and there is no feedback among queuing stations in the queuing network [16].

The service rate of each queue (μ_i) depends on the processing performance of the related node (e.g., SCF, RACF …) and the message type. The arrival rate (λ_i) depends on the number of network subscribers, offered traffic per subscriber and mean call holding time [14].

The CSD queuing model for single-phase scheme is shown in Fig. 5. It is based on the network architecture illustrated in Fig. 2 and the signaling call flow in Fig. 3.

The model includes eight queues each of which represents the processing and forwarding a specific message in network elements. The incoming message, arrival rate and service rate are indicated for each queue.

There are three other arrival rates called λ_{ub},λ_{cb} and λ_{na} which are user busy (the called party is busy), call blocking (the negative response of RACF due to the lack of network resources) and no answer (the called

party doesn't answer to the call) rates that are subtracted from λ1, λ2 and λ6 respectively. For example the λ_{ub} is the arrival rate of *486 Busy Here* SIP message returned back to CPE-1 by SCF. As mentioned above the provisional (*1xx*) SIP responses have no effect on the CSD definition and according to Equation 1, the queues in CPEs (Q6) should not be considered in CSD calculations.

Figure 6 illustrates the CSD queuing model for two-phase scheme which is derived from signaling flow of Fig. 4.

The main difference between the two models is that the reservation and commitment procedures are separated and hence the number of queues in two-phase model is more than that of single-phase.

The call set-up delay can be found using the above queuing models as the sum of the queuing delay in each node for each message plus the transmission delay (the elapsed time for a message to cross the transport network from one node to another) for all messages.

$$CSD = \sum_{i=1}^{M} \frac{1}{\mu_i - \lambda_i} + nT_{tr} \quad (3)$$

Where *M* is the number of queues, T_{tr} is transmission delay for one message and *n* is the total number of signaling messages crossing the transport network.

The number of queues in two-phase model is more than that of single-phase and this means that the sum of the delays due to these queues will be more. However CSD also depends on some other parameters. The second term in the above equation i.e., the sum of transmission delays depends on some parameters such as the traffic load, the call holding time, etc. Hence the T_{tr} is different for single-phase and two-phase schemes and depends on network resources availability for sending the signaling messages which is investigated in the next section.

Network resources can be utilized for signaling (such as the above call set-up signaling scenarios), best effort and QoS-guaranteed traffics. The resources that have not been committed yet, can be used for signaling and best effort traffics even if they had been reserved before; but committed resources are used only for QoS-guaranteed traffic (e.g., voice, video, etc.) and cannot be used for other traffic flows. Accordingly, commitment would better be postponed as much as possible in order to utilize the resources optimally. This means that using the two-phase scheme can result in better resource utilization.

RESOURCE AVAILABILITY FOR SIGNALING TRAFFIC

In order to compare signaling and best effort resource availability in single-phase and two-phase schemes the below steps can be followed:

Assuming that T_{call} is the average call holding time, T_{res} is the signaling time for resource reservation and $T_{use\text{-}1ph}$ is the time during which the resources are in use in the single-phase scheme, the following equation can be derived from Fig. 3.

$$T_{use-ph} = \frac{T_{res}}{2} + T_{IR} + T_{ring} + T_{call} \quad (4)$$

By the same way the time during which the resources are in use in the two-phase scheme ($T_{use\text{-}2ph}$) equals to:

$$T_{use-2ph} = \frac{T_{cmt}}{2} + T_{call} \quad (5)$$

The ratio of in-use time of two-phase to single-phase shown by α, can be seen as a measure of resource availability for signaling and best effort traffics.

$$\alpha = \frac{T_{cmt}/2 + T_{call}}{T_{res}/2 + T_{IR} + T_{ring} + T_{call}} \quad (6)$$

T_{ring} is constituted of two parts, one part is the mean time to answer (T_{mta}) and the other one is the ringing time limit (T_{rtl}). T_{mta} relates to a situation in which the called party answers to the incoming call while T_{rtl} relates to when the called party does not answer i.e., it is the maximum ringing time. Assuming that p represents the probability of call answering by the called party, the following equation holds:

$$T_{ring} = T_{mta}.p + T_{rtl}.(1-p) \quad (7)$$

Replacing (6) in (5) the following equation is achieved:

$$\alpha = \frac{T_{cmt}/2 + T_{call}}{T_{res}/2 + T_{IR} + T_{mta}.p + T_{rtl}.(1-p) + T_{call}} \quad (8)$$

As mentioned before the T_{tr} in Equation 3 is different for single-phase and two-phase schemes and depends on network resources availability for sending the signaling messages. In other words α can be seen as the proportional coefficient between T_{tr} for single-phase ($T_{tr\text{-}1ph}$) and T_{tr} for two-phase ($T_{tr\text{-}2ph}$) schemes i.e. $T_{tr\text{-}2ph} = T_{tr\text{-}1ph}.\ \alpha$. As shown in [17], α is always less than one, which means that network resources are more available for signaling and best effort traffic in two-phase scheme. Also increasing call holding time will increase α and the difference between the two schemes in terms of their effect on resource availability for signaling and best effort traffic will decrease.

NUMERICAL RESULTS

Assuming that the mean arrival rate of new calls (call intensity) is λ per second and the mean call holding time is T_{call}, then the offered traffic in erlangs is [18]

$$A = \lambda T_{call} \quad (9)$$

According to Equation 9, for a constant traffic load, if the T_{call} is decreased the arrival rate λ will be increased and it causes the CSD to be increased in turn. Two-phase scheme will outperform the single-phase (i.e., the CSD will be less) because it has a better resource utilization in this situation. Figure 7 depicts the CSD versus mean call holding time for single and two-phase schemes. The CSD for two-phase is assumed to be constant and the figure actually shows the relative amount of CSDs.

It can be seen that for T_{call} less than 30 seconds, two-phase scheme has a better performance. Increasing T_{call}, the arrival rate will be decreased and CSD_{1ph} will be decreased to less than the CSD_{2ph}. As mentioned before the two-phase scheme has better resource utilization when the arrival rate is increasing. Figure 8 illustrates the CSD versus T_{call} for different amounts of transmission delay. Using this figure we can choose the better scheme in different network traffic situations.

The effect of p on CSD is depicted in Fig. 9. As mentioned before, in the two-phase scheme, network resources are only reserved but not committed during the time between ringing and answering (going off-hook) and hence these resources can be used by the best effort and signaling traffic. It can be seen that for example for T_{call}=30sec, CSD_{1ph} is more than CSD_{2ph} until p = 0.8. When p is low the wasting time for ringing is more. During this time the resources are not committed in two-phase scheme and hence it has a better performance and lower CSD.

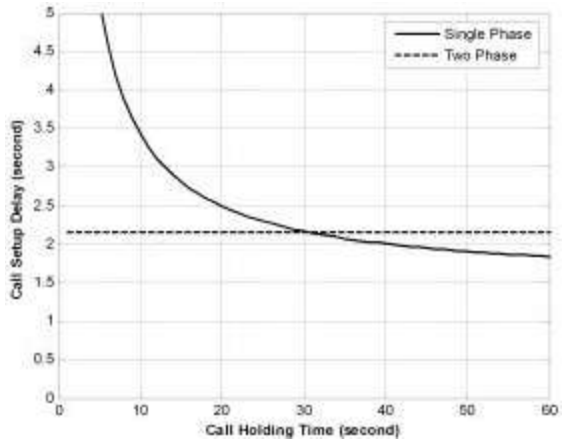

Fig. 7: Call set-up delay versus mean call holding time

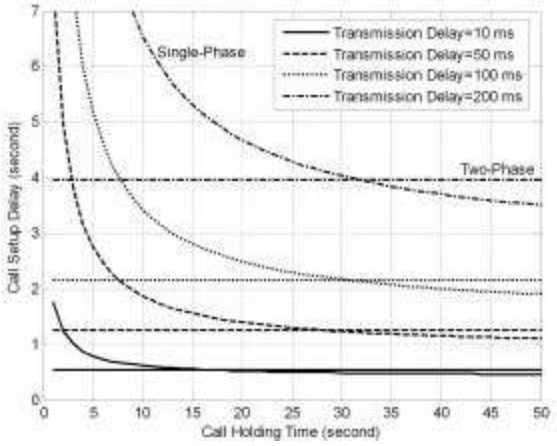

Fig. 8: Call set-up delay versus mean call holding time for different T_{tr}

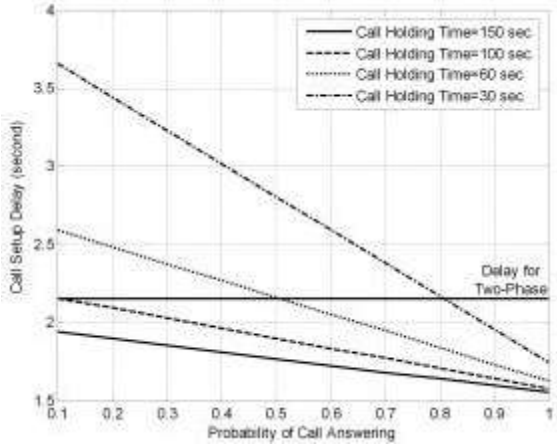

Fig. 9: Call set-up delay versus probability of call answering

CONCLUSION

This paper investigated two resource control schemes in NGN, i.e. single-phase and two-phase and their effect on call set-up delay parameter. The QoS signaling call flows extracted for the two schemes and then the call set-up delay simulated using a queuing model. The simulations showed that the two-phase scheme can have a better performance than single-phase in some situations. The two-phase scheme has better resource utilization for signaling and best effort traffics. Using these results, we can have some criteria for choosing the better resource control scheme appropriate to traffic changes. The numerical results can be used to find the optimal points for dynamical

change of the schemes. Future work is intended to study the multi domain scenarios for resource control schemes in RACF architecture.

REFERENCES

1. ITU-T Rec. Y., 2001. General overview of NGN. 2004
2. Jongtae Song *et al*., 2007. Overview of ITU-T NGN QoS Control. IEEE Communications Magazine.
3. ETSI ES 282 003 v1.1.1, 2006. Resource and Admission Control Sub-system (RACS); Functional Architecture.
4. ITU-T Rec. Y.1541, 2006. Network Performance Objectives for IP-based Services.
5. ITU-T Rec. Y.1531, 2007. SIP-based Call Processing Performance.
6. Taesang Choi *et al*., 2007. NGN Performance Monitoring and Management: Standard Perspective and its Applications. CEC-EEE Conference.
7. ITU-T Rec. Y.2111, 2007. Resource and admission control functions in Next Generation Networks (Release 2). NGN-GSI/DOC-301.
8. Rosenberg J. *et al*., 2002. SIP: Session Initiation Protocol. RFC 3261.
9. ITU-T Q.3301.1, 2006. Resource control protocol no. 1 (rcp1) Protocol at the interface between service control entities and the Policy Decision Physical Entity (PD-PE).
10. Calhoun P. *et al*., 2003. Diameter Base Protocol. RFC 3588.
11. Chan K. *et al*., 2001. COPS Usage for Policy Provisioning (COPS-PR). RFC 3084.
12. ITU-T Rec. Y. 2021, 2006. IMS for Next Generation Networks.
13. Gurbani V.K. *et al*., 2005. Characterizing Session Initiation Protocol (SIP) Network Performance and Reliability. ISAS 2005, LNCS 3694, pp: 196-211.
14. Pirhadi M. *et al*., 2000. Call Set-up Time Modeling for SIP-based Calls in Next Generation Networks. ICACT, Korea.
15. Ibrahiem El Emary *et al*., 2008. New IP QoS Architecture for Voice and Data Convergence Over DSL Lines. World Applied Sciences Journal, 4 (5): 634-645.
16. Gross, D. and C. M. Harris, 1998. Fundamentals of Queuing Theory. John Wiley.
17. Hemami, M.S., M. Pirhadi and A. Iravani Tabrizipoor, 2008. Analysis and Optimization of Resource Control Schemes in Next Generation Networks. ITU/IEEE K-INGN08, Geneva, pp: 63-67.
18. Cooper, R.B. and D.P. Heyman, 1998. Teletraffic Theory and Engineering, Encyclopaedia of Telecommunications. 16: 453-483.