www.czasopisma.pan.pl PAN www.journals.pan.pl
jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2013, VOL. 59, NO. 2, PP. 113–123
Manuscript received August 23, 2012; revised April, 2013. DOI: 10.2478/eletel-2013-0014

Traffic Model of IMS/NGN Architecture with Transport Stratum Based on MPLS Technology

Sylwester Kaczmarek, Magdalena Kosek, and Maciej Sac

***Abstract*—Growing expectations for a fast access to information create strong demands for a universal telecommunication network architecture, which provides various services with strictly determined quality. Currently it is assumed that these requirements will be satisfied by Next Generation Network (NGN), which consists of two stratums and includes IP Multimedia Subsystem (IMS) elements. To guarantee Quality of Service (QoS) all NGN stratums have to be correctly designed and dimensioned. For this reason appropriate traffic models must be developed and applied, which should be efficient and simple enough for practical applications. In the paper such a traffic model of a single domain of NGN with transport stratum based on Multiprotocol Label Switching (MPLS) technology is presented. The model allows evaluation of mean transport stratum response time and can be useful for calculating time of processing requests in the entire NGN architecture. Results obtained using the presented model are described and discussed. As a result of the discussion, elementary relationships between network parameters and transport stratum response time are indicated.**

***Keywords*—IMS, mean transport stratum response time, MPLS, NGN, traffic model, transport stratum**

I. Introduction

CURRENTLY we can observe a significant growth in the amount of distributed information. To standardize this process International Telecommunication Union Telecommunication Standardization Sector (ITU-T) proposed the Global Information Infrastructure (GII) concept [1] and appropriate telecommunication network architecture dedicated to its realization, called the Next Generation Network (NGN) [2]. NGN is a packet-based network, which consists of service stratum as well as transport stratum and provides various services with precisely defined quality.

Nowadays it is assumed that NGN service stratum includes elements of IP Multimedia Subsystem (IMS) [3], a platform initially designed for delivering multimedia services in 3G mobile networks, utilizing mainly SIP [4] and Diameter [5] communication protocols.

Transport stratum in Next Generation Network is, comparing to service stratum, dependent on the used transport network technology, which must support carrying IP packets. Any transport technology satisfying this condition can be used in NGN. From existing technologies one of the most promising for core of Next Generation Network is Multiprotocol Label Switching (MPLS) [6].

In order to fulfill specified quality requirements, both service stratum and transport stratum of NGN have to be correctly designed and implemented. For this reason proper traffic models should be proposed and utilized, which should be possibly uncomplicated and appropriately describe operation of network elements. Performed review of current work regarding traffic engineering in IMS-based NGN (abbreviated in the next part of the paper as IMS/NGN) [7], [8] demonstrated that this area is out of the scope of standardization bodies. Moreover, existing traffic models [9]–[15] are not fully compatible with IMS/NGN architecture as they do not take into consideration standardized resource and admission control elements [16]. Furthermore, there is a small number of models explicitly applicable for NGN service stratum since many of them focus only on VoIP networks with SIP protocol or IMS architecture [12]–[15].

Taking these facts into account, we decided to propose our own traffic model of a single domain of IMS/NGN with transport stratum utilizing MPLS technology, which allows evaluation of mean transport stratum response time. The aim of this paper is to present the proposed analytical model and indicate the parameters, which have the largest impact on mean MPLS-based transport stratum response time. The paper is organized as follows. Architecture of IMS-based ITU-T NGN is presented in section II. The proposed traffic model is described in section III. Section IV is devoted to the results of performed transport stratum response time investigations. Conclusions and future work regarding the proposed model are presented in section V.

II. IMS/NGN Architecture

The IP Multimedia Subsystem solution [3], [17] was proposed by 3GPP in 2002 as a key component of the 3G mobile network architecture. The IMS was designed as a universal set of service control servers independent of the used transport network technology. Taking this fact into consideration, IMS elements were incorporated into service stratum of ITU-T and European Telecommunications Standards Institute Telecommunications and Internet converged Services and Protocols for Advanced Networking (ETSI TISPAN) NGN architectures [18], [19]. ITU-T Next Generation Network proposition [18] is more advanced, for example in application of existing technologies (Ethernet [20], [21], Flow State Aware – FSA [22], MPLS [23], [24]) to transport stratum as well as user

This research work was partially supported by the system project "InnoDoktorant – Scholarships for PhD students, Vth edition" co-financed by the European Union in the frame of the European Social Fund.

S. Kaczmarek, M. Kosek, and M. Sac are with the Department of Teleinformation Networks, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, 11/12 Gabriela Narutowicza Street, 80-233 Gdańsk, Poland (e-mails: kasyl@eti.pg.gda.pl; kosekm@gmail.com; Maciej.Sac@eti.pg.gda.pl).

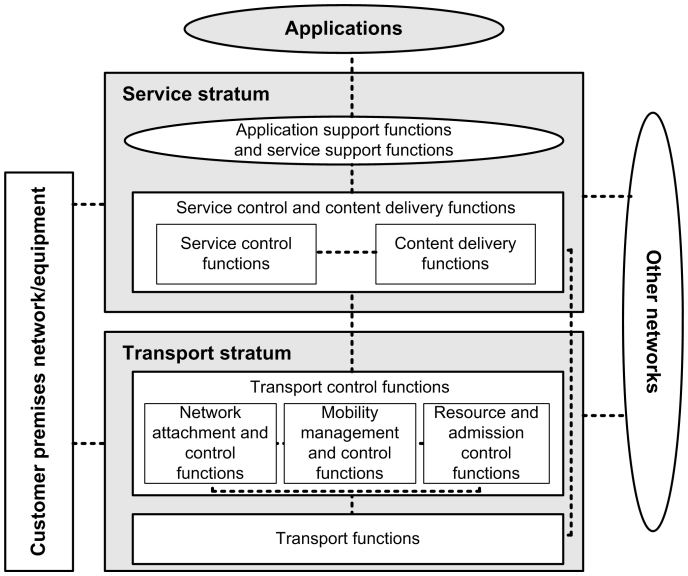

Fig. 1. ITU-T NGN Release 2 functional architecture [18], [27].

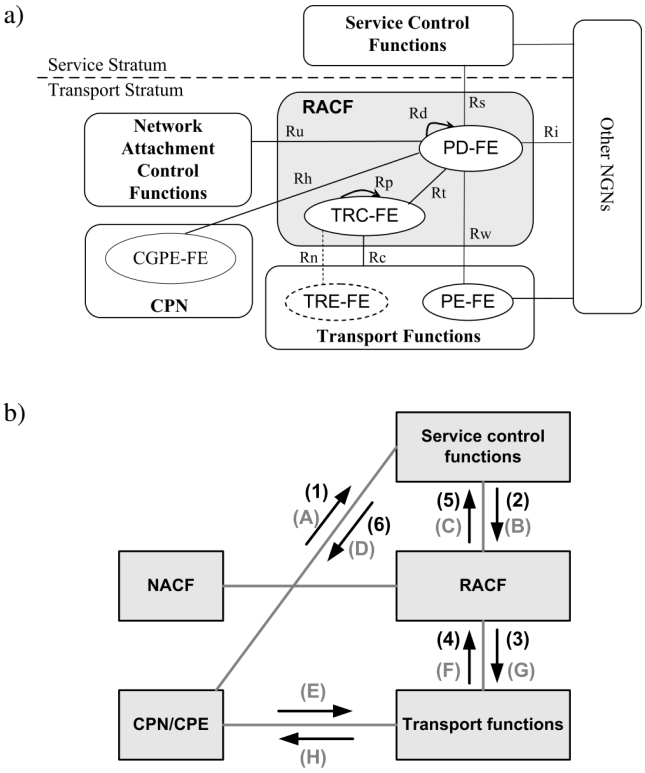

Fig. 2. ITU-T NGN RACF functional architecture (a) and resource control modes (b): push mode (black numbers) and pull mode (gray letters) [16], [27].

mobility [25], [26]. Therefore, ETSI TISPAN solution is not considered in the next part of the paper. A comparison of ITU-T and ETSI TISPAN NGN architectures is available in [8].

Functional architecture of ITU-T Next Generation Network is depicted in Fig. 1 and includes transport stratum, service stratum and applications. ITU-T NGN delivers services to various Customer Premises Equipments (CPEs – NGN terminals, legacy PSTN/ISDN terminals) often forming Customer Premises Networks (CPNs) and interworks with NGN networks, IP non-NGN networks and PSTN/ISDN networks.

Service stratum of ITU-T NGN works together with applications in order to deliver services to users. Application Support Functions and Service Support Functions (ASF&SSF) are responsible for the functionality of gateway, registration, authentication and authorization at the application level. They cooperate with Service Control and Content Delivery Functions (SC&CDF) to handle service requests of CPNs/CPEs and applications. SC&CDF units contain Service User Profile Functions (databases which store information concerning service users) as well as service components. For providing NGN terminals with multimedia and traditional PSTN/ISDN services the most vital is IP Multimedia Service Component, which includes IMS functional elements [28]. For this reason, the Next Generation Network architecture is called IMS/NGN.

Transport stratum responsible for providing IP connectivity services in ITU-T solution is controlled by Transport Control Functions: Network Attachment Control Functions (NACF), Resource and Admission Control Functions (RACF) as well as Mobility Management and Control Functions (MMCF). NACF unit is mainly used during connecting CPN/CPE to access transport network. Its functions include dynamic provisioning of IP addresses and other parameters, authentication, authorization as well as location management. MMCF element delivers IP-based mobility services to NGN terminals with support for handover across access networks of different technologies.

For traffic engineering the most important entity of NGN transport stratum is RACF (Fig. 2a) [16], which performs admission control and resource allocation functions. RACF can be regarded as an arbitrator in terms of Quality of Service (QoS) between Service Control Functions (SCF) and Transport Functions, which makes the final decision concerning the demanded resources. In order to formulate the final decision, RACF analyses among others transport subscription information, Service Level Agreements (SLAs), network policy rules, service priority as well as transport resource status and utilization.

As can be observed in Fig. 2a, detailed RACF architecture includes Policy Decision Functional Entity (PD-FE) and Transport Resource Control Functional Entity (TRC-FE), which manage the following transport units:

1) CGPE-FE (CPN Gateway Policy Enforcement Functional Entity) – element typically residing in a gateway to which a Customer Premises Network is connected; responsible for traffic filtering, classifying and marking as well as resource allocation, traffic control, shaping and maintaining resource utilization status; affects only up-stream traffic,
2) PE-FE (Policy Enforcement Functional Entity) – functional element located typically in a gateway between two IP networks; additionally to CGPE-FE functionality provides operations connected with Network Address and Port Translation (NAPT) and firewall; affects up-stream and downstream traffic,

3) TRE-FE (Transport Resource Enforcement Functional Entity) – functional element responsible for traffic management in a network which technology supports traffic aggregation (eg. VLAN, VPN, MPLS).

PD-FE entity is a decision element independent of the controlled transport network technology. Its main tasks are processing and coordination of resource demands from service stratum (SCF elements) and transport stratum (PE-FE element). It acts as a final decision point which accepts or rejects a resource demand with respect to the network policy rules, service information, transport subscription information (from NACF) and decision on resource availability made by TRC-FE. Transport Resource Control Functional Entity (TRC-FE) is a transport technology dependent decision point which hides the aspects concerning the transport technology from PD-FE and collects information about available resources as well as network topology.

Division of RACF architecture into two decision elements (transport technology independent PD-FE and dependent TRC-FE) provides an abstract view of transport network infrastructure to SCF and make service stratum functions agnostic to the details of transport facilities.

In the ITU-T Next Generation Network solution RACF can control resources in two modes: push mode and pull mode (Fig. 2b) [16]. Push mode is a target mode for the NGN architecture, which is used for CPEs capable of negotiating QoS at service stratum level (using e.g. SIP as well as SDP protocols and their appropriate extensions) or without such a capability. Service demand including or not requested resource amount is transmitted to SCF (1). Service Control Functions extract or determine the amount of resources necessary for a demanded service and forward the request to RACF (2), where the final decision is made and required resources are assigned (3).

Pull mode is supported for interworking of NGN with existing transport technologies and utilized for CPEs capable of negotiating QoS at transport stratum level (using e.g. RSVP protocol). Before requesting transport resources for a service, CPE may optionally send a message including or not service level description of QoS requirements to SCF (A). In Service Control Functions the information about QoS is extracted from the message or determined and transmitted to RACF for authorization (B). RACF responses with an authorization token, utilized to bind service request at service and transport stratums, which is sent to SCF (C) and CPE (D). Subsequently, a resource request to transport stratum elements is generated by CPE (E) and forwarded to RACF (F), which makes the final policy decision and allocates the requested resources (G).

III. Traffic Model of IMS/NGN with MPLS-Based Transport Stratum

In the paper a single domain of IMS/NGN core based on MPLS technology is considered (Fig. 3). A very general idea how quality parameters can be analyzed in this architecture was first presented in [8] without many details regarding the implementation of the traffic model and without results of any investigations. This paper is a thoroughly extended version of that work and includes extensive details about calculations of

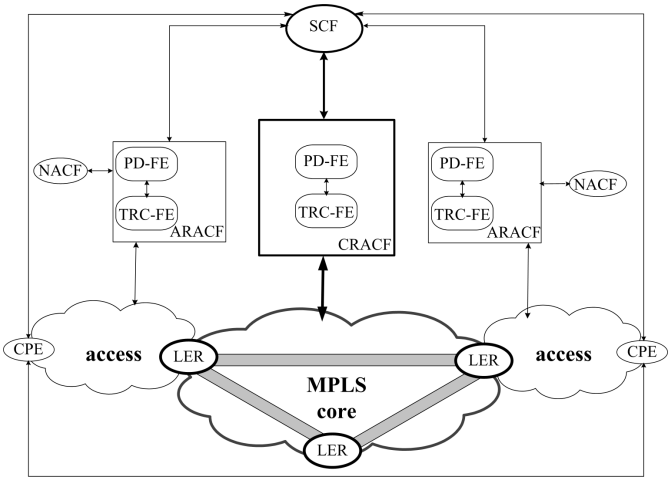

Fig. 3. Model of IMS/NGN core with transport stratum based on MPLS technology [8], [24].

mean transport stratum response time, $E(T)$. This involves among others the aspects of computing communication times, which were not covered in [8]. Apart from that, contrary to [8], this paper includes the results of investigations performed for several representative data sets (section IV). The presented results are provided with comments on the influence of network parameters on $E(T)$.

Physical resources in the IMS/NGN core (Fig. 3) are centrally controlled by the CRACF (Core RACF) element as proposed in [24] (another possible solution, a distributed resource control for MPLS-based transport stratum, is described in [23]). We assume that in the network push resource control mode is utilized (Fig. 2b) and resource operations are coordinated by Service Control Functions (SCF). As a result, RACF elements depicted in Fig. 3 do not communicate directly with each other. Therefore, access networks under the control of the ARACF (Access RACF) units are not considered in the next part of the paper.

User requests regarding the demanded services are sent to SCF and result in transport stratum resource reservation, modification or release. These resource operations are performed by MPLS routers under the control of the CRACF element, which is responsible for the following tasks [24]:

1) authorization and handling of resource requests sent by SCF,
2) storing information concerning transport resource utilization in a local database,
3) monitoring controlled MPLS network state,
4) sending resource reservation, modification and release requests to controlled MPLS elements and processing responses from these elements,
5) sending final responses regarding requested resource operations to SCF.

In the model it is assumed that routing as well as changes in bandwidth of Label Switched Paths (LSPs, logical channels carrying aggregated data) are performed using MPLS in-band signaling [6], [24]. CRACF communicates directly only with a Label Edge Router (LER), which begins or ends a particular LSP. It is assumed that all Label Switched Paths are set

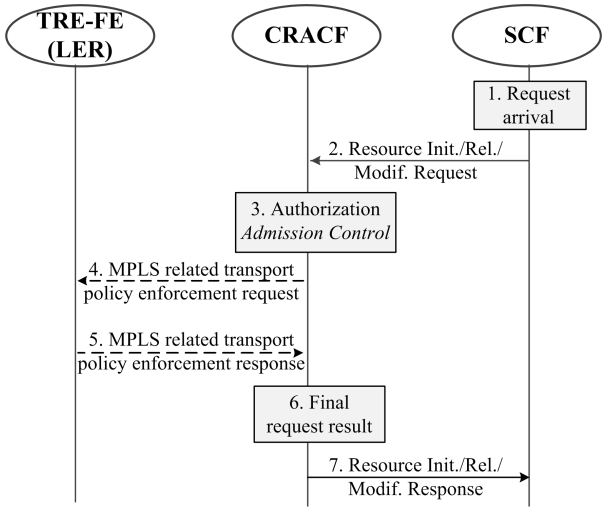

Fig. 4. Communication procedures for resource reservation, release and modification [16], [24].

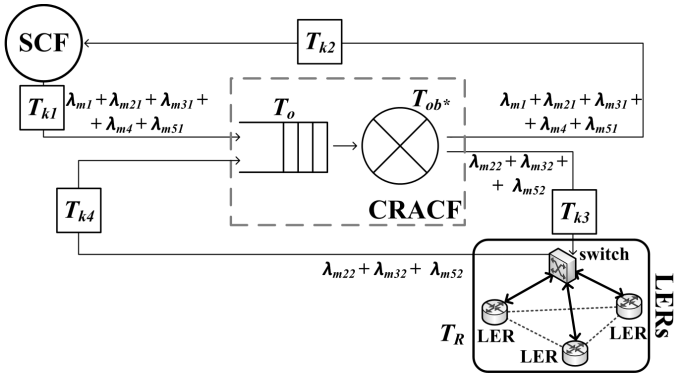

Fig. 5. Structure of the proposed traffic model with marked intensities of messages sent through links (Tab. I, Fig. 4).

up administratively with strictly determined initial bandwidth (static bandwidth reservation [24]).

For reasonable resource management not all resource requests sent by SCF involve changes in bandwidth of LSPs, some requests result only in update of resource state in the CRACF local database [8], [29], [30]. In case of requests concerning bandwidth reservation or increase CRACF queries the local database for the amount of free bandwidth of the particular LSP. If there exists enough free bandwidth, the requested resources are allocated without communication with LERs. Otherwise the LSP bandwidth is increased by Label Edge Router with some reserves, so that another request will most likely not involve operations on LSP and communication of RACF with controlled MPLS elements. For requests regarding bandwidth release or decrease CRACF checks the utilized LSP bandwidth state in the local database. If the LSP bandwidth utilization is low after the demanded resource operation, CRACF sends a request to LER in order to release a part of the free LSP bandwidth.

The CRACF unit in the assumed network model (Fig. 3) is a single physical entity having the functionality of PD-FE and TRC-FE elements (Fig. 2a). The PD-FE subunit is responsible for authorizing and handling resource requests. It operates based on the resource utilization information stored in a local database or on the decision of the TRC-FE subunit, which communicates with LERs and adjusts LSP resources when necessary. Label Edge Routers play the role of the TRE-FE elements defined by ITU-T [16], [24] and are capable of changing bandwidth of LSPs. For this reason they communicate with other routers on Label Switched Paths using MPLS in-band signaling.

In the network depicted in Fig. 3 standardized communication procedures are utilized for resource reservation, release and modification (Fig. 4) [16], [24]. The procedures are performed in the following steps:

1) SCF receives a service stratum request which involves transport resource reservation, release or modification.
2) After processing the service request, SCF determines required transport resource operation and sends a Resource Initiation/Release/Modification Request message to CRACF for resource reservation, release or modification respectively.
3) The Resource Initiation/Release/Modification Request is authorized and processed by CRACF. When the request involves bandwidth reservation or increase, free bandwidth of the LSP is checked in the local database and, if there is enough free bandwidth, the demanded resources are allocated without communication with LERs (the procedure goes to step 6). Otherwise, steps 4 – 5 are performed for increasing the bandwidth of the LSP. When the request concerns bandwidth release or decrease, the utilized resource level is checked in the local database. In case of low LSP bandwidth utilization after the demanded resource operation, steps 4 – 5 are additionally performed to decrease the bandwidth of the LSP. Otherwise, the procedure goes to step 6.
4) CRACF communicates with a Label Edge Router, which begins or ends the LSP in order to increase or decrease the LSP bandwidth.
5) LER performs the requested operations using MPLS in-band signaling and sends a response with their result.
6) CRACF makes final decision regarding the Resource Initiation/Release/Modification Request based on the information stored in the local database or obtained from LER.
7) A Resource Initiation/Release/Modification Response message with the result of resource reservation, release or modification respectively is sent to SCF by CRACF.

It is important that, according to [16], [24], CRACF final decision (6) may be preceded by sending a network policy enforcement request to LER (which in this case acts as a PE-FE unit described in section II) in order to install final admission decisions at the edge of the domain. However, we regard this optional procedure as not increasing the load of LERs significantly and thus it is not considered in the paper.

Taking into account the network model and communication procedures depicted in Figs. 3 and 4, a traffic model of a single domain of IMS/NGN core based on MPLS technology was proposed. The structure of the model is presented in Fig. 5. The SCF, CRACF and LERs elements correspond to the

elements of the network model described in Fig. 3. SCF is a generator of transport resource reservation, release and modification requests, which are handled by CRACF controlling a network of LERs connected through a switch (due to the fact that requests are transported in the network as messages in the next part of this paper we use terms "request" and "message" interchangeably). T_o and T_{ob*} represent message waiting time in the queue and message handling time by the CRACF processor respectively. T_{ob*} values vary for different message types and request processing paths. T_{k1} – T_{k4} blocks correspond to communication times between particular elements of the network, which include message buffering if the link is busy, message transmission times dependent on the message lengths and link bandwidth as well as propagation times dependent on the link length. T_R represents request processing time by LER, which includes communication time concerning sending a request from the switch to appropriate LER, LSP bandwidth adjustment time by the LER and communication time regarding sending a response from the LER to switch (Fig. 5). λ_{mn} (n =1, 21, 22, 31, 32, 4, 51, 52) parameters correspond to intensities of messages sent through particular links. Characteristics of messages sent from SCF and LERs to CRACF are presented in Tab. I. It is worth noting that (according to Fig. 4) CRACF sends the same number of messages to SCF and LERs as it receives from these units. Therefore, total intensity of messages sent from SCF to CRACF is the same as total intensity of messages sent from CRACF to SCF. Similar situation takes place in case of communication between CRACF and LERs.

Intervals between aggregated requests regarding bandwidth reservation, release, increase and decrease are given by exponential distributions with λ_1, λ_2, λ_3 and λ_4 intensities correspondingly. Although message arrivals to CRACF in the model (Fig. 5) do not generally follow a Poisson process, as in the case of resource requests, we assume that due to high intensities and several message sources (network elements in the model) resultant inter-arrival times can be approximately described using an exponential distribution. Thus, the operation of the CRACF processor can be modeled using M/G/1 queuing system [27], [31].

Comparing to this paper an analogical approach to network analysis is used in [27], however, the aim of the work described in [27] is to examine the behavior of IMS/NGN service stratum, while in this paper IMS/NGN MPLS-based transport stratum is investigated. Due to the fact that the model presented in [27] is verified by simulations, which results are convergent with theoretical solutions, we expect that the similar methodology used in this paper will have similarly good accuracy in case of transport stratum based on MPLS technology.

The aim of the proposed model is to evaluate mean MPLS-based transport stratum response time $E(T)$ in a single domain of IMS/NGN, which is defined as the mean time between sending a resource request by SCF to CRACF and receiving a response. For this reason the following input variables are defined:

1) λ_1 – λ_4 – intensities of resource requests regarding bandwidth reservation, release, increase and decrease respectively,
2) T_A – time of message authorization and request type determination by CRACF,
3) T_{proc} – time of performing elementary database operations concerning checking and updating resource state by the CRACF processor,
4) T_{resp} – time of processing a response from LER by CRACF,
5) T_R – time of processing a request by LER,
6) p_{11} – probability of a successful bandwidth reservation or increase without the necessity of increasing LSP bandwidth,
7) p_{12} – probability of a successful bandwidth reservation or increase with the necessity of increasing LSP bandwidth,
8) p_{13} – probability of an unsuccessful bandwidth reservation or increase,
9) p_{21} – probability of a bandwidth release or decrease without the necessity of decreasing LSP bandwidth,
10) p_{22} – probability of a bandwidth release or decrease with the necessity of decreasing LSP bandwidth,
11) l_i – length of optical link i, b_i – bandwidth available on optical link i, $\mathbf{l_{mi}}$ – vector with lengths of messages transmitted over optical link i (values necessary to calculate communication times T_{ki}, i = 1, 2, 3, 4).

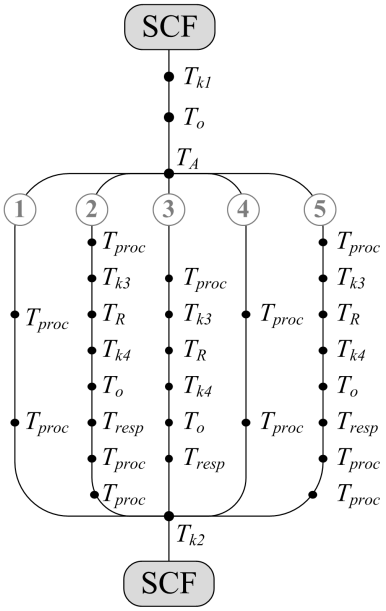

Fig. 6. Graph with request processing paths (1 – 5) in the system [8].

In order to calculate mean IMS/NGN transport stratum response time, a set of request processing paths must be determined. In the considered network (Fig. 3) there are five ways of handling requests (Fig. 6). Paths 1 – 3 regard resource reservation, while paths 4 – 5 concern resource release. It is important that there are no dedicated paths for resource modification as such requests are handled in the same way as resource reservation requests (when allocated bandwidth needs to be increased) and resource release requests (when allocated bandwidth needs to be decreased).

TABLE I
MESSAGES HANDLED BY CRACF AND THEIR CHARACTERISTICS

Message	From	Handling time	Intensity	Participation in total CRACF message intensity λ_{CRACF} (11)
$m1$ – bandwidth reservation or increase request (Resource Initiation/Modification Request message – Fig. 4, request processing path 1 – Fig. 6)	SCF	$T_{ob1} = T_A + 2 \cdot T_{proc}$	$\lambda_{m1} = (\lambda_1 + \lambda_3) \cdot p_{11}$	$p_{m1} = \lambda_{m1}/\lambda_{CRACF}$
$m21$ – bandwidth reservation or increase request (Resource Initiation/Modification Request message – Fig. 4, request processing path 2 – Fig. 6)	SCF	$T_{ob21} = T_A + T_{proc}$	$\lambda_{m21} = (\lambda_1 + \lambda_3) \cdot p_{12}$	$p_{m21} = \lambda_{m21}/\lambda_{CRACF}$
$m22$ – LER response to LSP bandwidth increase request (MPLS related transport policy enforcement response message – Fig. 4, request processing path 2 – Fig. 6)	LERs	$T_{ob22} = T_{resp} + 2 \cdot T_{proc}$	$\lambda_{m22} = (\lambda_1 + \lambda_3) \cdot p_{12}$	$p_{m22} = \lambda_{m22}/\lambda_{CRACF}$
$m31$ – bandwidth reservation or increase request (Resource Initiation/Modification Request message – Fig. 4, request processing path 3 – Fig. 6)	SCF	$T_{ob31} = T_A + T_{proc}$	$\lambda_{m31} = (\lambda_1 + \lambda_3) \cdot p_{13}$	$p_{m31} = \lambda_{m31}/\lambda_{CRACF}$
$m32$ – LER response to LSP bandwidth increase request (MPLS related transport policy enforcement response message – Fig. 4, request processing path 3 – Fig. 6)	LERs	$T_{ob32} = T_{resp}$	$\lambda_{m32} = (\lambda_1 + \lambda_3) \cdot p_{13}$	$p_{m32} = \lambda_{m32}/\lambda_{CRACF}$
$m4$ – bandwidth release or decrease request (Resource Release/Modification Request message – Fig. 4, request processing path 4 – Fig. 6)	SCF	$T_{ob4} = T_A + 2 \cdot T_{proc}$	$\lambda_{m4} = (\lambda_2 + \lambda_4) \cdot p_{21}$	$p_{m4} = \lambda_{m4}/\lambda_{CRACF}$
$m51$ – bandwidth release or decrease request (Resource Release/Modification Request message – Fig. 4, request processing path 5 – Fig. 6)	SCF	$T_{ob51} = T_A + T_{proc}$	$\lambda_{m51} = (\lambda_2 + \lambda_4) \cdot p_{22}$	$p_{m51} = \lambda_{m51}/\lambda_{CRACF}$
$m52$ – LER response to LSP bandwidth decrease request (MPLS related transport policy enforcement response message – Fig. 4, request processing path 5 – Fig. 6)	LERs	$T_{ob52} = T_{resp} + 2 \cdot T_{proc}$	$\lambda_{m52} = (\lambda_2 + \lambda_4) \cdot p_{22}$	$p_{m52} = \lambda_{m52}/\lambda_{CRACF}$

Paths depicted in Fig. 6 illustrate elementary processing times and communication times forming total transport stratum response time for a request in a particular network state. Symbols in Fig. 6 conform to these previously introduced in the paper. In the next part of the section a description of all request processing paths is provided.

1) The first request processing path regards a successful bandwidth reservation or increase without the necessity of increasing LSP bandwidth. SCF sends a new bandwidth reservation or increase request to CRACF (T_{k1} communication time), which waits in the queue for being handled (T_o). After that, the request is authorized by CRACF (T_A), and available LSP bandwidth is checked in the local database (T_{proc}). As there is enough free bandwidth, the request results only in updating LSP resource utilization in the database (T_{proc}). Finally, a positive response is sent to SCF (T_{k2} communication time). As a result, the time of handling the request/message by the CRACF processor is given by the following equation:

$$T_{ob1} = T_A + 2 \cdot T_{proc} \tag{1}$$

Total transport stratum response time for the request processing path can be defined as follows:

$$T_1 = T_{k1} + T_o + T_{ob1} + T_{k2} \tag{2}$$

2) The second request processing path concerns a successful bandwidth reservation or increase with the necessity of increasing LSP bandwidth. Similarly to the first path, SCF sends a new bandwidth reservation or increase request to CRACF (T_{k1} communication time), which waits in the queue for being handled (T_o). After that, the request is authorized by CRACF (T_A), and available LSP bandwidth is checked in the local database (T_{proc}). As there are not enough free resources for the request, LSP bandwidth must be increased. For this reason CRACF sends a proper request to Label Edge Router (T_{k3} communication time), which processes it and configures all MPLS routers in the LSP using in-band signaling (T_R). In the considered case LSP bandwidth is successfully increased and a positive response is sent by LER to CRACF (T_{k4} communication time). The LER response waits in the CRACF queue (T_o) and is processed by CRACF (T_{resp}), which updates total LSP bandwidth (T_{proc}) and LSP resource utilization level (T_{proc}) in the database. Finally, a positive response is sent to SCF

(T_{k2} communication time). As a result, request/message handling times by the CRACF processor are given by the following equations:

$$T_{ob21} = T_A + T_{proc} \qquad T_{ob22} = T_{resp} + 2 \cdot T_{proc} \quad (3)$$

Total transport stratum response time for the request processing path can be defined as follows:

$$T_2 = T_{k1}+T_o+T_{ob21}+T_{k3}+T_R+T_{k4}+T_o+T_{ob22}+T_{k2} \quad (4)$$

3) The third request processing path regards an unsuccessful bandwidth reservation or increase. This scenario is very similar to the second request processing path, however, in this case LER fails to increase LSP bandwidth due to lack of free transport resources and sends a negative response to CRACF (T_{k4} communication time). The LER response waits in the CRACF queue (T_o) and is processed by CRACF (T_{resp}), which sends final negative response concerning the resource operation to SCF (T_{k2} communication time). Consequently, request/message handling times by the CRACF processor are given by the following equations:

$$T_{ob31} = T_A + T_{proc} \qquad T_{ob32} = T_{resp} \quad (5)$$

Total transport stratum response time for the request processing path can be defined as follows:

$$T_3 = T_{k1}+T_o+T_{ob31}+T_{k3}+T_R+T_{k4}+T_o+T_{ob32}+T_{k2} \quad (6)$$

4) The fourth request processing path concerns a bandwidth release or decrease without the necessity of decreasing LSP bandwidth. SCF sends a new bandwidth release or decrease request to CRACF (T_{k1} communication time), which waits in the queue for being handled (T_o). After that, the request is authorized by CRACF (T_A). This step is followed by the check in the local CRACF database if free LSP bandwidth after the requested resource release or decrease operation is acceptable (T_{proc}). In considered case the LSP is satisfactorily utilized and there is no need for decreasing its bandwidth. Therefore, utilized LSP bandwidth level is updated in the CRACF local database (T_{proc}) and the final response regarding the requested resource operation is send to SCF (T_{k2} communication time). As a result, the time of handling the request/message by the CRACF processor is given by the following equation:

$$T_{ob4} = T_A + 2 \cdot T_{proc} \quad (7)$$

Total transport stratum response time for the request processing path can be defined as follows:

$$T_4 = T_{k1} + T_o + T_{ob4} + T_{k2} \quad (8)$$

5) The fifth request processing path regards a bandwidth release or decrease with the necessity of decreasing LSP bandwidth. Similarly to the fourth path, SCF sends a new bandwidth release or decrease request to CRACF (T_{k1} communication time), which waits in the queue for being handled (T_o). Subsequently, the request is authorized by CRACF (T_A). In this case LSP bandwidth is not efficiently utilized after resource release or decrease in the local CRACF database (T_{proc}) and a part of unused LSP bandwidth must be freed. For this reason CRACF sends a proper request to Label Edge Router (T_{k3} communication time), which processes it and configures all MPLS routers in the LSP using in-band signaling (T_R). After performing these operations a response is sent by LER to CRACF (T_{k4} communication time). The LER response waits in the CRACF queue (T_o) and is processed by CRACF (T_{resp}), which updates total LSP bandwidth (T_{proc}) and LSP resource utilization level (T_{proc}) in the database. Finally, a positive response is send to SCF (T_{k2} communication time). Consequently, request/message handling times by the CRACF processor are given by the following equations:

$$T_{ob51} = T_A + T_{proc} \qquad T_{ob52} = T_{resp} + 2 \cdot T_{proc} \quad (9)$$

Total transport stratum response time for the request processing path can be defined as follows:

$$T_5 = T_{k1}+T_o+T_{ob51}+T_{k3}+T_R+T_{k4}+T_o+T_{ob52}+T_{k2} \quad (10)$$

Based on the formulas (2),(4),(6),(8),(10) we can calculate mean transport stratum response times $E(T_1) - E(T_5)$ for the request processing paths depicted in Fig. 6. For this reason the following elements are necessary:

1) Mean message handling times $E(T_{obn})$ (n = 1, 21, 22, 31, 32, 4, 51, 52) by the CRACF processor for particular processing paths, which are determined by mean values of T_A, T_{proc} and T_{resp} input variables. In order to simplify calculations, we assume that the above mentioned variables are replaced by constant values representing the maximum time of message authorization and request type determination by CRACF, maximum time of performing elementary database operations by CRACF and maximum time of processing a response from LER by CRACF respectively. As a result of such an estimation, $E(T_{obn})$ values are equal to T_{obn} (n = 1, 21, 22, 31, 32, 4, 51, 52) (1),(3),(5),(7),(9).
2) Mean time of processing a request by LER $E(T_R)$, which is the mean value of the random variable describing T_R.
3) Mean message waiting time $E(T_o)$ in the CRACF queue.
4) Mean communication times $E(T_{k1}) - E(T_{k4})$.

Mean message waiting time $E(T_o)$ in the queue of the CRACF processor can be estimated using formulas for M/G/1 queuing system [31]. In order to perform calculations we need:

1) Intensity of messages sent to the CRACF processor

$$\lambda_{CRACF} = (\lambda_{m1} + \lambda_{m21} + \lambda_{m31} + \lambda_{m4} + \lambda_{m51}) + \\ + (\lambda_{m22} + \lambda_{m32} + \lambda_{m52}) \quad (11)$$

which can be taken from Fig. 5.
2) Mean value $E(T_{ob*})$ and variance $V(T_{ob*})$ of message handling time T_{ob*} by the CRACF processor. Based on the previously stated assumption that T_A, T_{proc} and T_{resp} times are constant, values of $E(T_{ob*})$ and $V(T_{ob*})$

TABLE II
INPUT DATA SETS

Data set	$\lambda_1+\lambda_3$ [1/s]	$\lambda_2+\lambda_4$ [1/s]	T_A [ms]	T_{proc} [μs]	T_{resp} [ms]	$E(T_R)$ [ms]	p_{11}	p_{12}	p_{13}	p_{21}	p_{22}	Link parameters (length l and bandwidth b)
1	1 – 400	1 – 400	0.05 – 2	50	0.5	5	0.4	0.59	0.01	0.4	0.6	0 km
2	1 – 400	1 – 400	0.5	5 – 500	0.5	5	0.4	0.59	0.01	0.4	0.6	0 km
3	1 – 400	1 – 400	0.5	50	0.05 – 2	5	0.4	0.59	0.01	0.4	0.6	0 km
4	1 – 550	1 – 550	0.5	50	0.5	1 – 1000	0.4	0.59	0.01	0.4	0.6	0 km
5	300	300	0.5	50	0.5	20	0 – 1	0 – 1	0	0 – 1	0 – 1	0 km
6	1 – 550	1 – 550	0.5	50	0.5	10	0.4	0.59	0.01	0.4	0.6	0 – 1000 km, 10 Mb/s
7	1 – 550	1 – 550	0.5	50	0.5	10	0.4	0.59	0.01	0.4	0.6	0 – 1000 km, 100 Mb/s

can be calculated using the information presented in Tab. I and the following formulas

$$E(T_{ob*}) = \sum_{n=1,21,22,31,32,4,51,52} p_{mn} \cdot T_{obn} \quad (12)$$

$$V(T_{ob*}) = \sum_{n=1,21,22,31,32,4,51,52} p_{mn} \cdot (T_{obn} - E(T_{ob*}))^2 \quad (13)$$

The last unknown parts of mean transport stratum response times $E(T_1) - E(T_5)$ for particular request processing paths 1 – 5 are mean communication times $E(T_{k1}) - E(T_{k4})$, which consist of propagation times, message transmission times as well as message buffering delays before sending them through busy links. Propagation time is a constant value dependent only on the distance between network elements and assuming optical links is equal to 5μs/km. Message transmission time is a fixed time necessary to send a particular message, which can be calculated by the message length division by the link bandwidth.

As lengths of messages exchanged in the network are not precisely known, in the paper we assume mean length l_m and approximately estimate mean message buffering delay before sending through a link using M/M/1 queuing model [31]. For calculations of this delay we need message transmission time and message intensity λ_{link} (the average number of the messages sent through the link in a unit time period), which is given by the following formula (Fig. 5)

$$\lambda_{link} = \begin{cases} \lambda_{m1} + \lambda_{m21} + \lambda_{m31} + \lambda_{m4} + \lambda_{m51}, & \text{for } T_{k1}, T_{k2} \\ \lambda_{m22} + \lambda_{m32} + \lambda_{m52}, & \text{for } T_{k3}, T_{k4} \end{cases} \quad (14)$$

After computing mean message waiting time $E(T_o)$ in the CRACF queue and mean communication times $E(T_{k1}) - E(T_{k4})$, we can calculate mean transport stratum response times $E(T_1) - E(T_5)$ for the request processing paths depicted in Fig. 6. Finally, using these values as well as request processing paths probabilities $p(1) - p(5)$ mean MPLS-based transport stratum response time $E(T)$ can be obtained

$$E(T) = \sum_{j=1}^{5} p(j) E(T_j) \quad (15)$$

where

$$p(1) = p_1 \cdot p_{11}, \quad p(2) = p_1 \cdot p_{12}, \quad p(3) = p_1 \cdot p_{13},$$
$$p(4) = p_2 \cdot p_{21}, \quad p(5) = p_2 \cdot p_{22} \quad (16)$$

In formulas (16) p_1 and p_2 are probabilities that a request generated by SCF concerns bandwidth reservation/increase or release/decrease respectively. These probabilities are described as follows based on request intensities $\lambda_1 - \lambda_4$

$$p_1 = \frac{\lambda_1 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}, \quad p_2 = \frac{\lambda_2 + \lambda_4}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \quad (17)$$

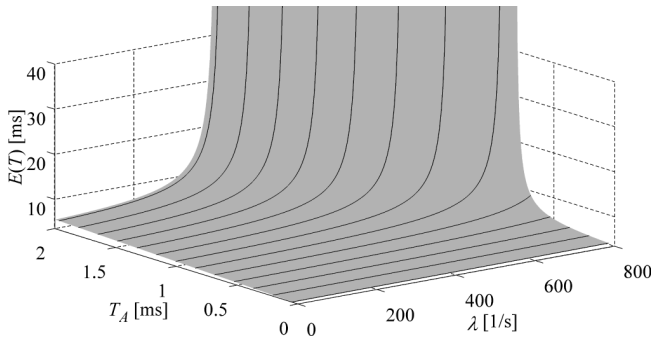

Fig. 7. Mean transport stratum response time $E(T)$ versus message authorization time T_A and total resource request intensity λ (data set 1).

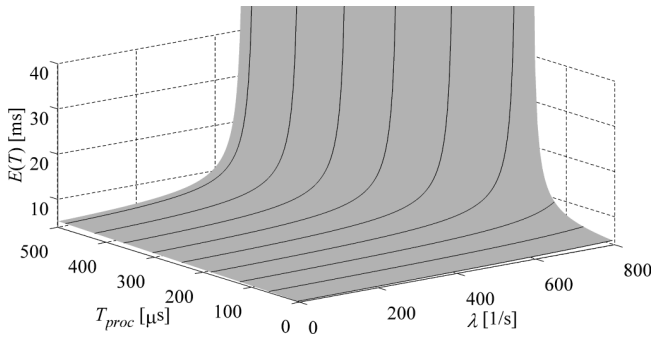

Fig. 8. Mean transport stratum response time $E(T)$ versus elementary database operation time T_{proc} and total resource request intensity λ (data set 2).

IV. ANALYSIS OF IMS/NGN TRANSPORT STRATUM RESPONSE TIME

In this section we present the results of MPLS-based transport stratum response time investigations in a single domain of IMS/NGN architecture obtained using the model described in section III and implemented in the MATLAB environment [32]. The results demonstrated in the next part of the paper were achieved using the data sets presented in Tab. II. Additionally, mean message length l_m of 750 bytes was assumed.

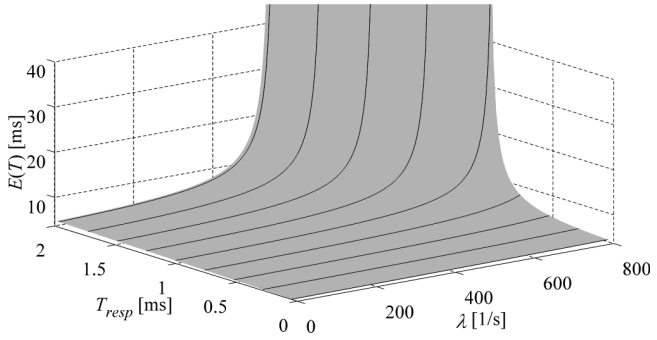

Fig. 9. Mean transport stratum response time $E(T)$ versus time of processing response T_{resp} from LER and total resource request intensity λ (data set 3).

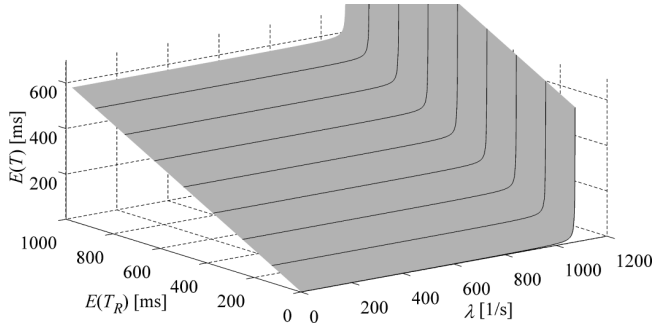

Fig. 10. Mean transport stratum response time $E(T)$ versus mean time of processing request $E(T_R)$ by LER and total resource request intensity λ (data set 4).

Results presented in Figs. 7–9 demonstrate mean MPLS-based transport stratum response time $E(T)$ dependence on total resource request intensity $\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$ as well as T_A, T_{proc} and T_{resp} times. Greater values of T_A, T_{proc} and T_{resp} increase message handling times T_{obn} (n = 1, 21, 22, 31, 32, 4, 51, 52) (1),(3),(5),(7),(9) by CRACF directly and also result in higher message waiting times (T_o) in the CRACF queue due to higher load offered to this unit, which is also affected by total resource request intensity λ. The described influence on $E(T)$ is, however, strong only when the CRACF processor is overloaded, which is avoided in practice. Under normal conditions the modeled system is characterized with high performance and can handle several hundreds of resource requests per second even for quite high λ, T_A, T_{proc} and T_{resp} values.

It is important that T_A, T_{proc} and T_{resp} parameters indicate the performance of the CRACF processor and their influence on $E(T)$ depends on their values as well as the number of occurrences in request processing paths (Fig. 6). Authorization (T_A) is performed for each request incoming to CRACF, while T_{resp} time occurs only for requests forwarded to LERs. As elementary database operations (T_{proc}) are used very often, it is crucial to implement a high performance local database for CRACF. The influence of database systems on telecommunications systems performance was also investigated in [33], where we tested different database solutions and their impact on request handling time in the laboratory testbed for ASON/GMPLS technology. Test results indicated that dedicated database systems (e.g. storing data in the device memory as C/C++ structures) offer better performance than standard open source database solutions (e.g. PostgreSQL database), often reading from hard disks.

In Fig. 10 $E(T)$ dependence on mean time of processing request $E(T_R)$ by LER is illustrated. $E(T_R)$ values result from the architecture and complexity of the MPLS domain as well as the performance of MPLS routers. As can be observed in Fig. 10, assuring proper processing power of the MPLS routers in the domain is very important since $E(T)$ is proportional to $E(T_R)$. The proportionality is defined by p_{11}, p_{12}, p_{13}, p_{21} and p_{22} probabilities. Higher values of p_{12}, p_{13} and p_{22} (equivalently lower values of p_{11} and p_{21}) indicate that more requests are sent to LERs, which results in increased mean transport stratum response time $E(T)$. These properties

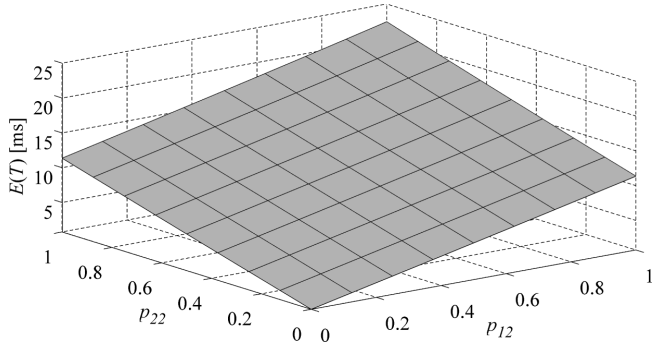

Fig. 11. Mean transport stratum response time $E(T)$ versus probabilities p_{12} and p_{22} describing request processing path in the system (data set 5).

can be observed in Fig. 11, in which for simplification it is assumed that there are no unsuccessfully handled requests ($p_{13} = 0$). In order to decrease the values of p_{12}, p_{13} and p_{22}, LSP bandwidth should be allocated with more reserves so that more resource requests will result only in update of the CRACF local database. This, however, leads to worse LSP bandwidth utilization. Therefore, a network designer should strike a balance between the above mentioned criteria.

Results presented in Figs. 7–11 are obtained based on the assumption that communication times $T_{k1} - T_{k4}$ are equal to zero, which means that all elements illustrated in Fig. 5 are in the same place. The influence of non-zero distances between elements on mean transport stratum response time $E(T)$ is demonstrated in Figs. 12–13. For simplification of calculations it is assumed that all links have the same length $l_i = l$ and bandwidth $b_i = b$. As can be noticed in Figs. 12–13, non-zero distances between network elements may increase $E(T)$ quite significantly, especially for larger link lengths l. Mean transport stratum response time $E(T)$ increases linearly with l values, which results from distance-dependent propagation times. It is important that 10 Mb/s links are sufficient for carrying signaling traffic regarding MPLS resource control (Fig. 12) even for high request intensities $\lambda_1 - \lambda_4$. It is not worth utilizing higher throughputs as increasing link bandwidth b even 10 times (Fig. 13) only slightly improves $E(T)$ values.

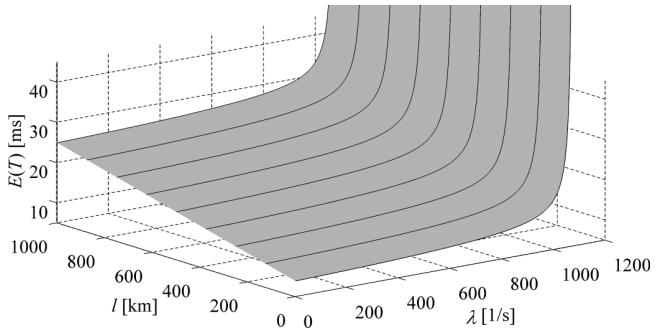

Fig. 12. Mean transport stratum response time $E(T)$ versus length l of optical links and total resource request intensity λ (data set 6).

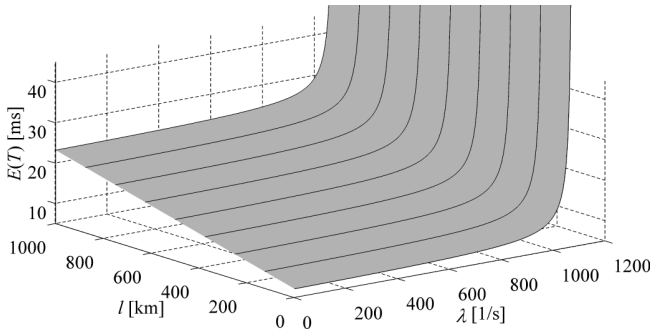

Fig. 13. Mean transport stratum response time $E(T)$ versus length l of optical links and total resource request intensity λ (data set 7).

V. Conclusions

In the paper a traffic model of a single domain of NGN architecture with transport stratum based on MPLS technology is proposed, which allows evaluation of mean transport stratum response time $E(T)$. The model conforms to the latest ITU-T standards and utilizes central resource control with push control mode. Available MPLS resources are reasonably managed so that only a part of requests sent by SCF involve changes in bandwidth of LSPs (using MPLS in-band signaling), some requests result only in update of resource state in the CRACF local database.

The paper also contains results of the investigations, which demonstrate elementary relationships between network parameters and mean MPLS-based transport stratum response time $E(T)$. For typical parameters the modeled system offers satisfactory performance and can handle several hundreds of requests per second. The most influential factors on $E(T)$ are distances between network elements (link lengths l) as well as mean time of processing request $E(T_R)$ by LER dependent on the structure of the MPLS domain and performance of MPLS routers. The impact of request processing in LERs can be decreased when LSP bandwidth is allocated with more reserves so that more resource requests will result only in update of the CRACF local database. This way, however, MPLS resources are not efficiently utilized, which creates the necessity to balance between lower transport stratum response time and better resource utilization.

Our future work will in the first step concern more thorough research regarding mean MPLS-based transport stratum response time in a single domain of IMS/NGN architecture. We will start our investigations with examining other than M/G/1 queuing models, which have complexity acceptable for engineering applications and can possibly more properly describe the operation of the CRACF unit. At the beginning known approximations of G/G/1 queuing systems will be applied and investigated. We are also going to extend the presented model by introducing an algorithm with threshold bandwidth utilization [8], [29], [30], which decides whether to communicate with LERs for LSP bandwidth adjustment or not. After that, different thresholds and their influence on $E(T)$ will be investigated. Moreover, one of our goals is to verify the described model using a proper simulator of a single domain of IMS/NGN with transport stratum based on MPLS technology, which will be implemented in the near future. The simulation model will be also helpful in determination of the best queuing model describing the operation of CRACF in the analytical model. Apart from that, we are simultaneously working on a traffic model of a multi-domain IMS/NGN focused on the behavior of the service stratum. After finishing this task, we are planning to extend the service stratum model by adding transport stratum with elements specific for MPLS technology, which are described in this paper. This will allow us to perform investigations in the two-layer multi-domain NGN architecture consisting of service stratum and transport stratum.

References

[1] ITU-T Rec. Y.100, "General overview of the Global Information Infrastructure standards development," June 1998.
[2] ITU-T Rec. Y.2001, "General overview of NGN," December 2004.
[3] 3GPP TS 23.228, "IP Multimedia Subsystem (IMS); Stage 2 (Release 11)," March 2011, v11.0.0.
[4] J. Rosenberg, et al., "SIP: Session Initiation Protocol," *IETF RFC 3261*, June 2002.
[5] P. Calhoun, J. Loughney, E. Guttman, G. Zorn, and J. Arkko, "Diameter Base Protocol," *IETF RFC 3588*, September 2003.
[6] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture," *IETF RFC 3031*, January 2001.
[7] S. Kaczmarek and M. Sac, "Traffic modeling in IMS-based NGN networks," *Gdańsk University of Technology Faculty of Electronics, Telecommunications and Informatics Annals*, vol. 1, no. 9, pp. 457–464, 2011.
[8] ——, "Traffic engineering aspects in IMS-based NGN networks," in *Teleinformatics library, vol. 6. Internet 2011*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2012, pp. 63–115 (in Polish), ISBN 978-83-7493-685-9.
[9] N. Lin and H. Qi, "A QoS model of Next Generation Network based on MPLS," in *Proceedings of IFIP International Conference on Network and Parallel Computing NPC 2007*, Dalian, China, 2007.
[10] I. K. Cho and K. Okamura, "A centralized resource and admission control scheme for NGN core networks," in *Proceedings of 23th International Conference on Information Networking ICOIN 2009*, Chiang Mai, Thailand, 2009.
[11] J. Joung, J. Song, and S. Lee, "Flow-based QoS management architectures for the Next Generation Network," *ETRI Journal*, vol. 30, no. 2, pp. 238–248, April 2008.
[12] P. S. Gutkowski and S. Kaczmarek, "Service time distribution influence on end-to-end call setup delay calculation in networks with Session Initiation Protocol," in *Proceedings of First European Teletraffic Seminar,*, Poznań, Poland, 2011, pp. 37–42.
[13] ——, "The model of end-to-end call setup time calculation for Session Initiation Protocol," *Bulletin of the Polish Academy of Sciences. Technical Sciences*, vol. 60, no. 1, pp. 95–101, January 2012.
[14] A. Hernandez, M. Álvarez-Campana, and E. V. Haddadzadeh, "Quality of Service in the IP Multimedia Subsystem," in *Proceedings of 5th COST 290 Management Committee Meeting*, Delft, The Netherlands, 2006.
[15] V. S. Abhayawardhana and R. Babbage, "A traffic model for the IP Multimedia Subsystem (IMS)," in *Proceedings of IEEE 65th Vehicular Technology Conference VTC2007-Spring*, Dublin, Ireland, 2007.

[16] ITU-T Rec. Y.2111, "Resource and admission control functions in next generation networks," November 2008.
[17] 3GPP TS 23.002, "Network architecture (Release 10)," March 2011, v10.2.0.
[18] ITU-T Rec. Y.2012, "Functional requirements and architecture of next generation networks," April 2010.
[19] ETSI Standard ES 282 001, "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN functional architecture," September 2009, v3.4.1.
[20] ITU-T Rec. Y.2112, "A QoS control architecture for Ethernet-based IP access networks," June 2007.
[21] ITU-T Rec. Y.2113, "Ethernet QoS control for next generation networks," January 2009.
[22] ITU-T Rec. Y.2121, "Requirements for the support of flow state aware transport technology in an NGN," January 2008.
[23] ITU-T Rec. Y.2174, "Distributed RACF architecture for MPLS networks," June 2008.
[24] ITU-T Rec. Y.2175, "Centralized RACF architecture for MPLS core networks," November 2008.
[25] ITU-T Rec. Y.2018, "Mobility management and control framework and architecture within the NGN transport stratum," September 2009.
[26] ITU-T Rec. Y.2807, "MPLS-based mobility capabilities in NGN," January 2009.
[27] S. Kaczmarek and M. Sac, "Traffic Model for Evaluation of Call Processing Performance Parameters in IMS-based NGN," in *Information Systems Architecture and Technology*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2012, pp. 85–100, ISBN 978-83-7493-699-6.
[28] ITU-T Rec. Y.2021, "IMS for next generation networks," September 2006.
[29] S. Kaczmarek and P. Żmudziński, "Bandwidth Broker as the element of the dynamic controlling DiffServ Domain," in *Proceedings of Internet – Wrocław 2005*, Wrocław, Poland, 2005, (in Polish).
[30] Z. Zhang, Z. Duan, and Y. Hou, "On Scalable Design of Bandwidth Brokers," *IEICE Transactions on Communications*, vol. E84–B, no. 8, pp. 2011–2025, August 2001.
[31] R. B. Cooper, *Introduction to queueing theory*, 2nd ed. New York: Elsevier, 1981.
[32] "Mathworks – MATLAB and Simulink for Technical Computing," www.mathworks.com/products/matlab/.
[33] S. Kaczmarek, M. Młynarczuk, M. Narloch, and M. Sac, "Evaluation of ASON/GMPLS Connection Control Servers Performance," in *Information Systems Architecture and Technology. Service Oriented Networked Systems*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2011, pp. 267–278, ISBN 978-83-7493-625-5.