Scientific
Research
Publishing

Real-Time Resource Availability Signaling in IP Multimedia Subsystem Networks

Tanır Özçelebi[1], Igor Radovanović[1] & Debraj Sengupta[2]

[1]*Eindhoven University of Technology, PO Box 213, 5600 MB, Eindhoven, The Netherlands*
[2]*ICT Embedded BV. PO Box 6420, 5600 HK, Eindhoven, The Netherlands*
Email: [1]{t.ozcelebi, i.radovanovic}@tue.nl, [2]sengupta.debraj@gmail.com

ABSTRACT

IP Multimedia Subsystem (IMS) allows the use of unlicensed, non-dedicated and nondeterministic access networks for delivering IP multimedia services. Providing end-to-end Quality-of-Service (QoS) for resource demanding real-time services (e.g. real-time multimedia) over such networks is a challenging task due to varying resource availability of the network and the end-device[i] during a single session[ii]. A common solution to this problem is adapting service quality and type according to availability of system resources, which requires end-to-end monitoring and signaling of resource availability during a single session. This paper presents an extension of the IMS architecture for real-time resource availability monitoring and signaling. The novelty of the proposed extension is twofold. It takes into account uncontrolled access networks with no resource reservation on the one hand, and end devices on the other. Two system architecture components are introduced for monitoring and signaling of the real-time resource availability in both networks (e.g. bandwidth, buffer space), and end-devices (e.g. battery, CPU, memory, storage), namely a Resource Manager (RM) and a Resource Availability Server (RAS). Our solution does not require any modifications of the IMS architecture and can be implemented as a plug-in.

Keywords: *end-to-end quality of service, IP multimedia subsystem, resource management, session initiation protocol*

1. Introduction

Today's smart phones and personal digital assistants (PDA) with wireless local area network (WLAN) connectivity are capable of accessing voice, video and data services outdoors over the Internet. Unlike the general practice in traditional mobile telecom world, the quality of these services is not always guaranteed and users can only get best-effort service. Nevertheless, such services have recently attracted a lot of interest from the mobile telecom community and upgrading from plain voice services to data and real-time multimedia services was necessary. Such services are currently being enabled by the emerging 3rd Generation (3G) mobile phone technologies.

Mobile telecom operators used to own not only the usage license of their wireless frequency band but also the network infrastructure in which their mobile services are available, including access points. This has several advantages in terms of QoS. Licensing of the wireless frequency band assures that the wireless link is free of external signal pollution, i.e. non-telecom signals can not use the specific frequency band. Ownership of the network infrastructure allows the operator to monitor availability of network resources and enforce QoS mechanisms such as admission control (AC) [1], resource reservation (RSVP) [2] and traffic engineering [3] for their

services. Using this approach, it is possible achieve the best service quality that meets the real-time system resource constraints.

Traditionally, all service types offered for certain devices were assumed to be available disregarding the availability of end-device and network resources. Moreover, mobile communications domain is currently migrating towards an All-IP network, which is accessible not only via mobile telecom access points but also via any IP access point. In the converged Next Generation Networking (NGN) framework [4] introduced recently, the functionality of services is independent from the underlying network, where managed and unmanaged IP networks are used together.

Note that QoS enforcement mechanisms used in mobile networks are no longer valid for NGN. For instance, applying AC and RSVP on NGN sessions in an unlicensed access network (e.g. a computer network) would not prevent QoS degradation at peek hours due to capacity overload and interference from non-NGN flows. The IMS [5] is an NGN architecture that was standardized by the 3rd Generation Partnership Project (3GPP) group [6], where the Session Initiation Protocol (SIP) [7] is employed at the application-layer as a control safeguard integrating

IMS with the Internet. According to IMS specifications, the users should be able to access services anytime, from anyplace and using any SIP-enabled device with IP version 6 (IPv6) [8].

Therefore, it is currently envisioned that the operators must provide services to users even if they are in access networks that are completely out of their control, hence the need for new end-to-end QoS mechanisms. This new approach is shown in Figrue. 1. Generally, the term "end-to-end" refers to the connection from the user-agent or the proxy to the server in telecom networks. According to this definition, the path from one end to the other lies within the core network owned and controlled by the operator. On the other hand, in this paper, the term "end-to-end" is used to refer to the connection from one user agent to the other, passing through the access networks and the core network. Thus, the access networks on the path constitute the variable-resource bottlenecks of the end-to-end network.

Guaranteeing network resource availability for resource demanding services is not feasible even if all parts of the communication path are operator-owned. This is because access networks are mostly nondeterministic, e.g. WLAN and Ethernet. The resource availability in the end-devices may also become the bottleneck in multimedia communication and it becomes impossible to *guarantee* end-to-end resource availability and consequently, service quality [5].

For maximizing user satisfaction, QoS needs to be adapted according to system resource availability. As resource availability in the system varies, both in networks and end devices, it has to be signaled to an end-device or an Application Server (AS) *during a single session*. The current IMS architecture does not provide specifics of such on-the-fly QoS negotiation. This work provides the details of such resource availability signaling for IMS, without altering the existing IMS architecture. The proposed system requires monitoring of available network and end-device resources, hence the need for Resource Managers (RM) at the communicating devices.
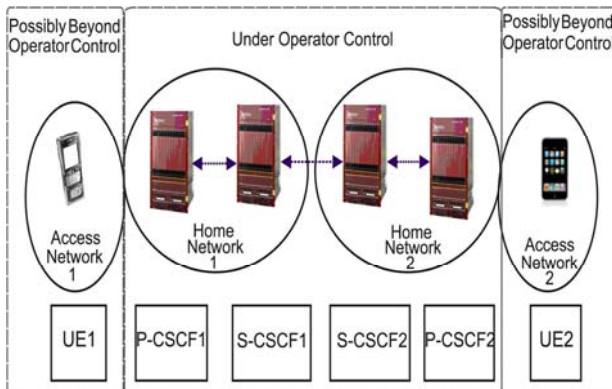

Figure 1. The end-to-end IMS network including access networks

The rest of this paper is organized as follows: The existing QoS mechanisms in IMS and related work in the literature are explained in Section 2. Our proposed solution with resource management and QoS adaptation is described in Section 3. The experimental results are given in Section 4. Finally, conclusions are drawn in Section 5.

2. Existing QoS Signaling Mechanisms for IMS

The IMS layered architecture is divided into three planes with different functionalities [5]. The transport and connectivity layer is separated from the application and service layer, by means of the signaling and control layer, whose job is to carry out call session control. In this plane, there is a standard set of control mechanisms valid for all services provisioned. Ideally, if there is a policy agreement among the networks that lie along the end-to-end path, the serving operator(s) can use these control mechanisms in order to improve the end-to-end QoS of the provisioned services. Due to such agreement, the operator would be aware of the available resources in these networks and the networks would apply the QoS decisions of the operator regarding call admission and resource reservation.

However, this scenario has a drawback. Even if such an agreement is present, a QoS policy translation problem across different networks still exists as the access networks outside the CN may have their own QoS models and semantics [9]. For example, there exist four QoS classes in UMTS framework (i.e. conversational, stream
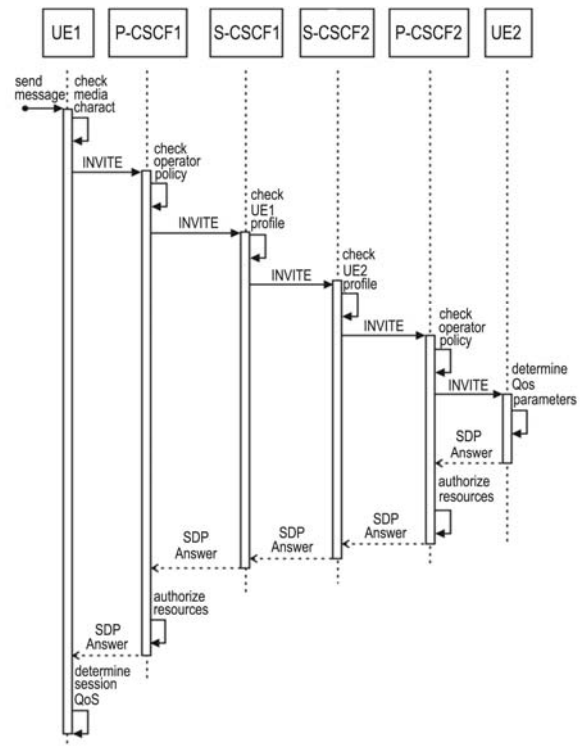

Figure 2. M**essage Sequence Diagram showing pre-session QoS negotiation in an IMS network**

JSSM

ing, interactive, background) [10], two QoS classes in Differentiated Services (i.e. expedited forwarding, assured forwarding) [11], and three QoS classes in Integrated Services (i.e. guaranteed, controlled-load, best-effort) [12].

In the IMS standard, QoS parameters can be negotiated between two user equipments prior to the session establishment using SIP. Once the QoS parameters have been negotiated between user devices (checked against fixed device capabilities) and been approved/modified by the Call Session Control Functions (CSCF) (checked against user subscription credentials) associated with both users, the IMS network asks the CN and the access network to reserve resources for this session. The SIP INVITE and SIP UPDATE messages [7] are used for this purpose as shown in Figrue. 2. The first INVITE message from the caller user equipment (UE1) to the call receiver (UE2) carries the QoS proposal, which is checked against the subscription levels of users at the Serving-Call Session Control Functions (S-CSCF) in both of the home networks. The QoS parameters are modified at these locations if there is a mismatch.UE2 puts her own QoS proposal in the answer and this is again checked and modified at the associated S-CSCF's of the users. Finally, UE1 can accept this counter QoS proposal and start the session or try to renegotiate with a SIP UPDATE message. Within the body of these SIP messages the session data is passed using Session Description Protocol (SDP) [13].

Maniatis *et al*. [9] have tried to tackle the QoS model translation problem across different networks by an intelligent mapping algorithm for end-to-end QoS negotiation such that the best suiting QoS class is selected in each network along the session path. Similarly, the operator can employ AC by using the Policy Decision Function (PDF) at the access network border and by making the access network physical bearer enforce a Service Based Local Policy (SBLP), assuming the bearer is listening to the commands from the IMS signaling and control layer. Even inter-network policy agreement would not be enough for access networks with unlicensed frequency band and anonymous usage rights, since already existing IMS sessions can still be jeopardized by new non-IMS services initialized in the access network. Furthermore, the access network can be nondeterministic and non-dedicated with interferences from the outer world, which is the case for Ethernet and WiFi. In this case, IMS sessions are bound to suffer unless the session QoS parameters are modified to fit the resource availability when there is a shortage or a boost.

There are several RFCs published by the Internet Engineering Task Force (IETF) that propose resource/ capability signaling among end-devices. These signaling schemes mainly concentrate on multimedia services, whereas our method can be applied to any real-time service over IMS. In [14], a method that integrates resource management (specifically RSVP) and SIP signaling is introduced in order to make network resource reservation before the session is established, i.e. before the called end-device is alerted such that session establishment failures are avoided. However, this RFC proposes no signaling during a single session. In our proposed solution it is assumed that reservations may not be possible in the access networks, and therefore the relation with reservation protocols, like RSVP, is not considered in this paper. Furthermore, signaling is done during a single session. Another difference is that, we propose an architecture in which local resources of the end-devices can also be transmitted to other interested parties, whereas [14] gives information about network resources only.

In [15], Internet Media Guides (IMG), i.e. multimedia session descriptions that can use SDP format are introduced. However, it is denoted in [15] that SDP syntax causes a huge amount of overhead in delivering IMG metadata over the network and SDP can carry only a small subset of IMG metadata in practical cases (e.g. codec type).

The bandwidth modifier of [16] notifies the receiving end-device on the maximum media codec rate to be used and the communication bit-rate required for the bit stream. Thus, [16] aims to convey bit-rate information only, without conveying any information about end-device resource availability.

In [17], bandwidth modifiers for RTP Control Protocol (RTCP) are introduced to SDP such that the amount of bandwidth allocated to RTCP in an RTP session is adapted (typically kept below 5% of the overall data rate). We envision that SIP resource availability signaling is preferable for protecting the privacy of resource availability data compared to transport layer protocols (e.g. RTCP), which lack to provide means for authentication, encryption and billing.

An extended SDP protocol for capability declaration (e.g. codec) amongst end-devices to be used in multimedia sessions is introduced in [18]. It is declared that such capability declarations can be intended for session negotiation, but such session negotiation mechanisms are not described.

In the IMS, it is envisioned that the end-to-end QoS negotiation and resource allocation should be reevaluated during the session depending on requests from the application, network load and link quality [5]. On the other hand, the implementation specifics of such a *QoS renegotiation* mechanism are not provided.

3. Resource Availability Signaling During A Session

As explained in the previous section, IMS allows communication sessions for which session QoS guarantees cannot be given [5]. For example, if the user is connected through a non-dedicated access network, e.g. Ethernet or

WiFi, it may not be possible to guarantee the requested end-to-end QoS due to scarcity of both end-device and network resources deteriorating the user experience. The deterioration effect is especially apparent while using resource demanding services such as multimedia streaming. The best approach to solve this problem is to introduce service quality adaptation (SQA) according to system resource availability (local, remote and network). The design constraint for the solution is to avoid possible alterations to the IMS architecture.

In order to perform such adaptation, the decision units for service quality adaptation at the end-devices, i.e. Service Quality Management (SQM) modules, must be aware of the availability of local and remote resources (e.g. battery, memory, storage CPU etc.) and network resources (e.g. throughput, buffer space) during the session and adapt their service quality level/type accordingly, hence the need for *resource availability monitoring* and *signaling*. In this section, we introduce i) the Resource Management (RM) module, ii) a Resource Availability Server (RAS) as an application server, and iii) a resource availability signaling mechanism for real-time adaptation of multimedia communication streaming and data streaming (e.g. video-on-demand) services in the IMS network. The operation of the SQM modules is beyond the scope of this paper.

The RM module is a crucial part of the proposed resource availability signaling framework. It is responsible for tracking local and network resources available to the end-user device in real-time. At the receiving device, the RM module publishes this information to the RAS server to be accessed by the remote transmitting device. At the transmitting device, the RM is responsible for gathering the resource availability data of the remote receiving device from the RAS.

It is the proposed RAS server that is responsible for collecting resource availability information from the receiving end-devices and delivering it to the transmitting end-device. Note that in a multimedia communication scenario, e.g. video-conferencing, an end-device can be transmitting multimedia, receiving multimedia, or both. In order to be IMS compliant, it is appropriate to employ the existing SIP call session control protocol of IMS for resource availability signaling in the proposed architect-
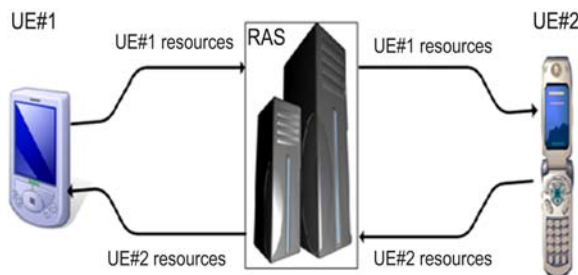
ture. In Figrue. 3, resource availability signaling is shown, where it is assumed that user equipments UE1 and UE2 have registered to each other's resource availability information at the RAS server.

Resource availability data is carried from end-devices to RAS and back in SIP event notification messages, such that a resource update is signaled whenever the local resources (e.g. memory, CPU, storage etc.) or network resources (e.g. bandwidth, jitter etc.) at one end crosses a critical boundary threshold. It is beyond the scope of this paper to propose methods for determining those thresholds. For example, suppose that a given real-time service offers three ways in which it can operate: multimedia, audio only and text. The communicating parties can select either one of these modes based on their resource availability. For instance, they can choose text when 90% or more of the times the CPU is occupied with other tasks in one of the end-devices, they can chose multimedia when the CPU is occupancy is less than 10%, or they can choose audio only in between. Therefore, the amount of end-device and network resources spent on resource monitoring and signaling is negligible in the proposed framework. A worst case scenario is investigated in the next section.

The proposed message flow diagram from the end-device to RAS for resource availability signaling is depicted in Figure 4 and our additions to the SIP/SDP pa-
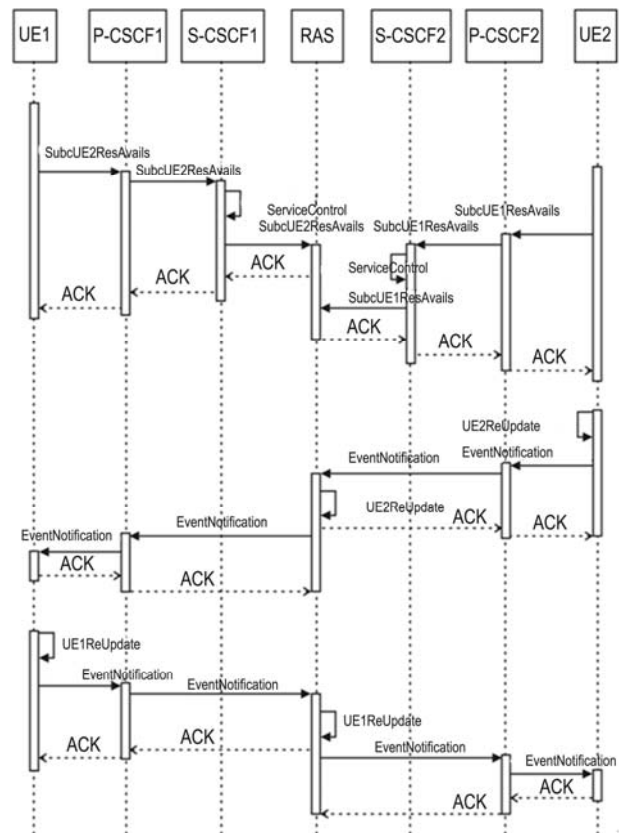

Figure 3. The proposed system architecture for mid-session SIP-based resource availability signaling

Figure 4. Message Sequence Diagram showing SIP based resource availability update signaling

Table 1. Proposed Additional Resource Data in SDP

r ::= "memory" | "CPU" | "storage" | "throughput" | "battery"
t ::= "Mbytes" | "Kbytes" | "kbps" | "seconds" | "percentage"
a ::= <resource availability measure>

Table 2. Example SDP with Resource Data

```
NOTIFY sip:abc.somename.com SIP/2.0
Via: SIP/2.0/UDP abc.tue.nl:5060;branch=z9hG4bKnashds7
Max-Forwards: 70
To: Bob <sip:server.somename.com>
From: Bob <sip:abc.somename.com>;tag=456248
Call-ID: 843817637684230@998sdasdh09
CSeq: 1826 NOTIFY
Contact: <sip:abc@192.0.2.4>
Expires: 7200
Content-Type: application/sdp
Content-Length: 131


v= RM/1.45
o= abc 5876768686868 7698798797979 IP 1.2.3.4
s= 123456789
i= Resource update to presence server
c= IN IP4 1.2.3.4
b= 100 kbps
k= none


r= memory
i= free memory status
t= kbytes
a= 12450


r= battery
i= battery charge remaining
t= percentage
a= 86
  --msg ends--
```

rameters as resource indicators are shown in Table 1.

After the addition of the proposed resource availability parameters, an example resource update SDP message looks as shown in Table 2. Here, both users will be aware of each other's local and network resources and SQM can use this information to perform its functionality.

4. Experimental Results

In our experiments, the RAS server (i.e. a PC) and the RM modules at both end-devices (i.e. two PDA's with WinCE operating system) communicate with each other during multimedia streaming between the two PDA's. The PDA applications start communicating using SIP INVITE. The media flow is started after the ACK is received from the caller. The session ends with a SIP BYE message which terminates the media session. Resource availability data from each PDA is reported to the RAS module using our SIP NOTIFY messages with the new header fields introduced in Table 1.

The multimedia session between PDA's is carried over Real-time Transport Protocol (RTP) and the PDA's are
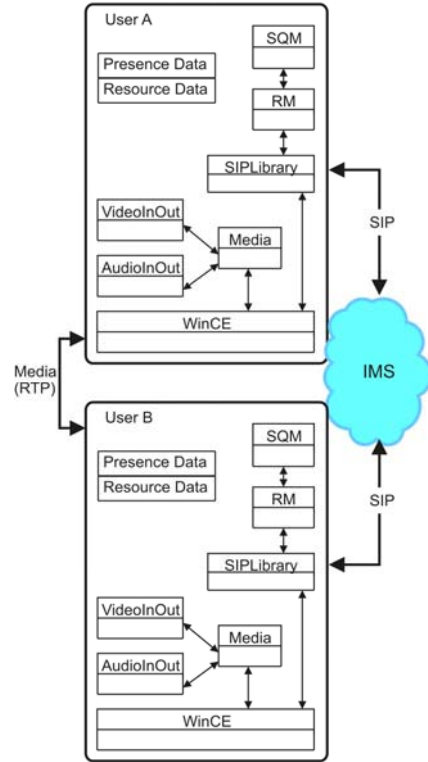
Figure 5. Deployment view of the user devices and IMS core
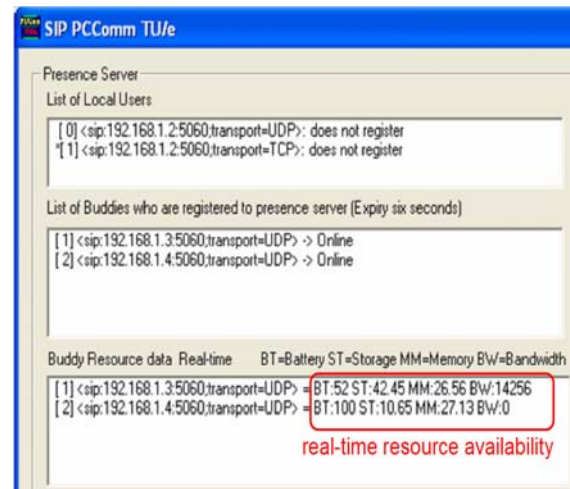
Figure 6. A snapshot of the RAS log interface

subscribed to each other's resource information at the

RAS server when a call takes place. The real-time resource availability signaling is done in parallel with the session using SIP/SDP as shown in Figure 5.
Figure 7 shows the PDA User Interface (UI) of the client test application. The top left menu is used to make a call or to exit the application. The first line shows the local IP address and port. The local and the remote resource availability data is displayed in the second line and the last line respectively.

In our experiments, we assumed that the real-time service that is going to make use of the proposed architect-
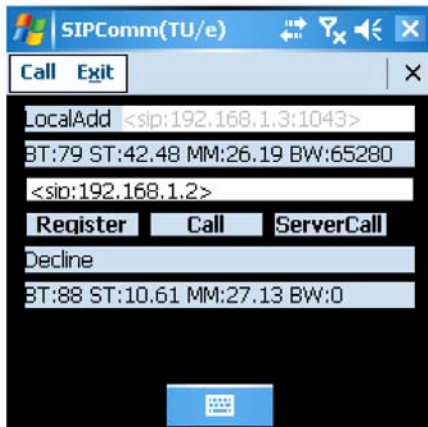

Figure7. A snapshot of a user interface of the PDA

Figure 8. Resource availability signaling channel over-head in the worst case scenario (signaling period: 1 sec.)

ture is a multimedia service. Multimedia services are the driving services that are offered by the IMS. In order to enable those services, availability of considerable amount of resources in the IMS network and end-devices is needed during a single session. Since the amount of resources available varies, quality of the provided services drops. One solution to this problem is introducing quality adaptation based on real-time resource availability, which can be achieved by multimedia bit rate adaptation.

However, such adaptation cannot be done arbitrarily, since multimedia codecs have their own limitations in changing the encoding bit rate on-the-fly even in the case of scalable codecs or bit stream switching. Therefore, in multimedia service quality adaptation, adaptation speed should not be higher than that of the multimedia codec. In [19], it is argued that the video adaptation algorithms in the literature need up to 3 groups of pictures (GOP) in order to converge to a target bit rate every time the video is adapted. Here a GOP is defined as a frame sequence of a given structure in a video stream, whose first frame is an intra-coded (I) frame. Furthermore, it is also denoted in [19] that the size of a GOP has to be kept large in an encoded video bit stream in order to attain reasonable compression efficiency and 1 GOP per second is taken as a rule of thumb, which would allow 1 adaptation in every 3 seconds for the other rate controllers in the literature and 1 adaptation per second for the advanced rate controller of [19]. Therefore, we assume that the maximum video adaptation frequency is 1 adaptation per second for typical videos. Updating resource availability at the speed higher than the adaptation speed would result in no quality improvements. Moreover, resource consumption in the network and end-devices will be higher.

The maximum overhead caused by the proposed architecture will be in case the resource availability is done at the maximum adaptation speed of the multimedia codec used (i.e. 1 adaptation per second). To analyze this, we used periodic resource availability update with a period of 1-second, for which the resource availability signaling overhead is measured to be 8 kbps as shown in Figrue. 8.
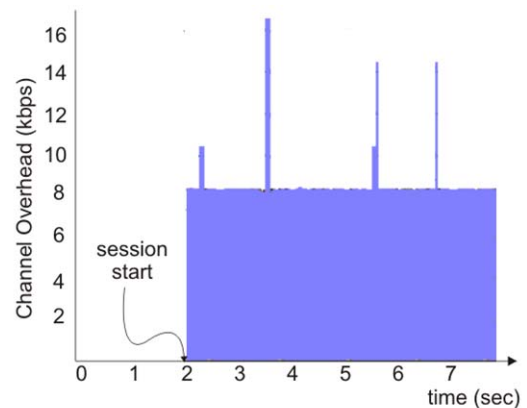
The signaling overhead is zero before the session starts, and it increases to 8 kbps on average per session after the session is initiated.

In [20], the design and implementation of an Open IMS core has been done and the load on the IMS core network and the proxies due to SIP message flow is investigated. In an IMS network with 100,000 subscribers, 1/3 of the subscribers are assumed to be online and registered simultaneously at a typical instant and 1/16 of these subscribers are assumed to be engaged in a multimedia session with an average duration of 180 seconds. In this case, the system would have to support 11,57 calls per second and the Open IMS core would have to process around 81 SIP messages per second (7 SIP messages for each multimedia call setup). In their experimental results, it was shown that a simple Intel Pentium 4 processor running at 3GHz (HyperThreaded) is enough to do the tasks of all IMS core components at once (i.e. I-CSCF, S-CSCF, P-CSCF and HSS) and still handle 120 SIP messages per second (around 17 calls per second). Considering the above data, in a worst case scenario of the proposed architecture, i.e. when each and every one of the active users has to adapt their multimedia within a given second, around 2000 SIP messaging events would need to be handled by the IMS core. This is quite realizable in a real-life deployment of the IMS core network since i) all components (CSCF's and HSS's) of the system normally reside on different hardware nodes in a deployed IMS core, ii) using multiple instances of the same component (e.g. multiple S-CSCF's) is very common for load-balancing, and iii) the state-of-the-art processors of today (e.g. multi-core processors) are much more powerful than a 3GHz Intel Pentium 4 processor. Clearly, in a more realistic case, the resource availability signaling overhead decreases even further when the resource signaling is done based on critical thresholds as described in the previous section.

The RAS is an additional server unit that can be implemented as an AS and it is independent of the IMS CSCF. Therefore, the existence of RAS does not put any

computational overhead on the CSCF's.

5. Conclusion

The emerging IMS framework allows mobile users to benefit from highly resource consuming services (e.g. multimedia streaming) at anytime and anyplace. In contrary to the traditional practice in telecom networks, the IMS access networks can be non-dedicated and/or nondeterministic, and the availability of the network and the end-device resources may show severe oscillations within a single session. The currently available mechanisms in the IMS standard are insufficient to provide guaranteed QoS. Service quality adaptation is needed in order to enhance user experience, which can be realized using resource availability monitoring and SIP-based end-to-end resource availability signaling as proposed in this paper.

The main contribution of this paper is an extended IMS architecture for real-time resource availability monitoring and signaling. The novelty of the proposed architecture comes from the fact that it takes into consideration uncontrolled IMS access networks with no resource reservation on the one hand, and end-devices with limited resources on the other. The available resources of a user's end-device are collected at the RM and advertised to the users that are subscribed to her resource status at the RAS server through SIP signaling in real-time.

Experiments show that resource availability signaling introduces negligible overhead to the overall network traffic even in a worst case scenario for real-time multimedia services.

6. Acknowledgement

REFERENCES

[1] R. Yavatkar, D. Pendarakis and R. Guerin, "A Framework for Policy-based Admission Control," *RFC* 2753, IETF Network Working Group, http://www.ietf.org/rfc/rfc2753.txt, January 2000.

[2] R. Braden, L. Zhang, S. Berson, S. Herzog and S. Jamin, "Resource ReSerVation Protocol (RSVP), " *RFC* 2205, IETF Network Working Group, http://www.ietf.org/rfc/rfc2205.txt, September 1997.

[3] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja and X. Xiao, "Overview and Principles of Internet Traffic Engineering," *RFC* 3272, IETF Network Working Group, http://www.ietf.org /rfc/rfc3272.txt, May 2002.

[4] http://www.itu.int/ITU-T/ngn/index.phtml.

[5] Digital Cellular Telecommunications System (Phase 2+), *Universal Mobile Telecommunications System (UMTS), IP Multimedia Subsystem (IMS)*, Stage 2, V7.6.0, TS 23.228, 3GPP, December 2006.

[6] http://www.3gpp.org.

[7] J. Rosenberg, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley and E. Schooler, "SIP: Session Initiation Protocol," *RFC* 3261, IETF Network Working Group, http://www.ietf.org/rfc/rfc3261.txt, June 2002.

[8] http://www.comet-consortium.org.

[9] S.I. Maniatis, E.G. Nikolouzou and I.S. Venieris, "End-to-End QoS Specification Issues in the Converged All-IP Wired and Wireless Environment, " *IEEE Communications Magazine*, vol. 42, no. 6, 2004, pp. 80-86.

[10] S. Chia,"The Universal Mobile Telecommunication System, " *IEEE Communications Magazine*, vol. 30, no. 12, 1992,pp. 54-62.

[11] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, "An Architecture for Differentiated Services, " RFC 2475, IETF DiffServ Working Group, http://www.ietf.org /rfc/rfc2475.txt, December 1998.

[12] R. Braden, D. Clark and S. Shenker, "Integrated Services in the Internet Architecture: an Overview, "RFC 1633, IETF Network Working Group, http://www.ietf.org/rfc/rfc1633.txt, June 1994.

[13] M. Handley and V. Jacobson, "SDP: Session Description Protocol," *RFC* 2327, IETF Network Working Group, http://www.ietf.org/rfc/rfc2327.txt, April 1998.

[14] G. Camarillo, W. Marshall and J. Rosenberg, "Integration of Resource Management and Session Initiation Protocol (SIP)", *RFC* 3312, October 2002.

[15] Y. Nomura, R. Walsh, J-P. Luoma, H. Asaeda and H. Schulzrinne, "A Framework for the Usage of Internet Media Guides (IMGs)", *RFC* 4435, April 2006.

[16] M.Westerlund, "A Transport Independent Bandwidth Modifier for the Session Description Protocol (SDP)", *RFC* 3890, September 2004.

[17] S.Casner, "Session Description Protocol (SDP) Bandwidth Modifiers for RTP Control Protocol (RTCP) Bandwidth", *RFC* 3556, July 2003.

[18] F. Andreasen, "Session Description Protocol (SDP) Simple Capability Declaration", *RFC* 3407, October 2002.

[19] F. De Vito, T. Ozcelebi, O. Sunay, M. Tekalp, R. Civanlar and J. C. De Martin, "Per-GOP Bitrate Adaptation for H.264 Compressed Video Sequences," in L. Atzori *et al*. (Eds.): *LNCS* 3893, pp. 198-206, Springer-Verlag Berlin Heidelberg 2006.

[20] D. Vingarzan, P. Weik, T. Magedanz, "Design and Implementation of an Open IMS Core," in T. Magedanz *et al*. (Eds.) *LNCS* 3744, pp. 284-293, Springer-Verlag Berlin Heidelberg 2005.

---

[i] In this paper, the terms "end-device", "user-device" and "user-equipment" are used interchangeably.
[ii] A session is a durable connection between two user devices or a user device and a server.