

# One-Node-Based Mobile Architecture for a Better QoS Control

Khadija Daoud, Philippe Herbelin

Orange Labs

Issy Les Moulineaux, France

{khadija.daoud, philippe.herbelin}@orange-ftgroup.com

Noël Crespi

Institut TELECOM SudParis

Evry, France

Noel.crespi@it-sudparis.eu

**Abstract—** Although IMS (IP Multimedia Subsystem) brings a large set of new services and offers service convergence, it introduces complexity on network procedures and architecture. This paper focuses on IMS session establishment procedure and highlights the long delay introduced by this procedure. It also raises failure problems encountered when only partial radio resources are available for the requested service. To solve these issues, we propose a new architecture, called UFA (Ultra Flat Architecture) which is an IMS distributed and flat architecture composed of only base stations. We define session establishment procedure for UFA and describe the way it handles access network resource information to adapt the service accordingly. We compare UFA and IMS classical architectures in terms of session establishment delay. Evaluation uses queuing theory and shows that UFA enables a considerable gain.

**Keywords-component:** network architecture, IMS, PCC, QoS.

## I. INTRODUCTION

IMS (IP Multimedia Subsystem) [1] is being defined within 3GPP to provide a rich set of integrated multimedia services such as instant messaging, video conferencing, VoIP, application sharing. It offers service convergence, meaning that service platforms and their related control functions can be shared between different access networks. This role is achieved thanks to the separation of IMS service control layer and access network layer. IMS introduces independent and dedicated network components based on SIP protocol. Its main elements are the proxy call service control function (P-CSCF), the interrogating CSCF (I-CSCF), and the serving CSCF (S-CSCF). The P-CSCF acts as a SIP proxy between the MN and the I-CSCF/S-CSCF. The S-CSCF performs user authentication and implements the actual SIP registrar functionality and session control, including service triggering.

Separating IMS service layer and access layer enables service convergence but introduces complexity on network procedures. In this paper we are interested in the performance of session establishment procedure. Since service and access network levels are separated, PCC (Policy Control and Charging) architecture [2] has been introduced by 3GPP to perform the interaction between the two levels during session establishment, modification and release procedures. PCC ensures that the bearer established on the access network uses the resources corresponding to the session negotiated at the service level and allowed by the operator policy and user

subscription. Session establishment procedure consumes a long delay and does not interact tightly with the dynamic resource information in the access network. For example it does not handle the cases when only partial radio resources are available for supporting the requested service. This paper analyses these limitations and proposes a new mobile access architecture, called Ultra Flat Architecture (UFA), that reduces session establishment delay and enhances the interaction between radio resource information and session layer. The reminder of this paper is organised as follows: section II describes PCC architecture components and session establishment procedure as defined in IMS standards, section III carries out a qualitative analysis of IMS and PCC limitations. Then, section IV and V describe UFA architecture and compare its performances with IMS architecture in terms of session establishment delay. Finally sections VI and VII give the related work and conclude the paper.

## II. 3 GPP IMS SESSION ESTABLISHMENT PROCEDURE

PCC [2] introduces mainly three entities: the AF (Application Function), the PCRF (Policy and Charging Rules Function) and the PCEF (Policy and Charging Enforcement Function). The AF is the P-CSCF. The PCRF acts as a policy decision point to authorize bearer resources based on the AF QoS inputs and on operator policies. The PCEF is a Gateway (GW) on the user plane that enforces QoS and gating policies received from the PCRF. It is the GGSN in the case of UMTS.

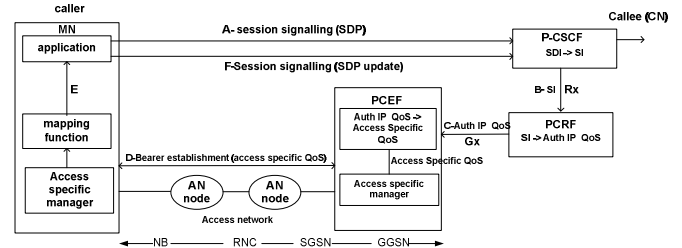


Figure 1. Session establishment phases and PCC architecture

In PCC architecture, the AF/P-CSCF is linked to the PCRF via Rx interface, and the PCRF is linked to the PCEF via Gx interface; both interfaces are based on Diameter protocol. In Figure 1, we illustrate the main phases of a normal session establishment procedure. Detailed messages [3], [4] involved in these phases are given in figure 2 considering UMTS as an

example of access network. Only message arguments of phases D, E and F are specific to the case 1 presented in section III.A.

During phase A, a caller invites a callee to initiate a video call and negotiate the session characteristics. The caller's mobile node (noted as MN) sends a SIP INVITE message to the callee's mobile node (noted as CN). This message contains an SDP (Session Description Protocol) [5] that describes the application components (voice+video) and the proposed codecs. The P-CSCF receives SIP INVITE message and consults its local policies to authorize the session request. Afterwards, SDP offer/answer exchanges continue with the callee to agree on the media and their related codec that will be used for the session. Based on these exchanges, the P-CSCF deduces in phase B the Service Information (SI) and relays it to the PCRF. In phase C, the PCRF deduces from the received SI the *authorized IP QoS* information based on other policies and sends it to the PCEF (e.g. GGSN). The PCEF computes the *access specific QoS* and establishes with the MN in phase D a bearer with a bitrate equal to *access specific QoS*. In theory *authorized IP QoS* and *access specific QoS* can be calculated per IP flow i.e. per media; however because of the limited PCEF capacity in terms of simultaneous active bearers (e.g; number of PDP contexts in the GGSN), only one global bearer with *access specific QoS* calculated for all IP flows negotiated within the session will be established. At the MN, when this global bearer is established, access specific layer will notify the higher SIP layer in phase E. In phase F, MN SIP layer informs the CN that resources are established using SIP UPDATE message. If the resources are also established on the callee side, it will be informed of the incoming call (via SIP RINGING message). The aim of performing phase D before phase F is to check in advance the resource availability on the access network before informing the callee of the incoming call avoiding thus ghost calls. To reflect resource reservation state on the SIP level, QoS preconditions have been defined by the IETF in RFC 3312 [7]. QoS preconditions are defined for each media line in the SDP and give the desired and the current state of the resources needed for the media, as detailed in the following example:

```

m=audio 3456 RTP/AVP 97 96
b=AS:25.4
a=curr:qos local none
a=curr:qos remote none
a=des:qos mandatory local sendrecv
a=des:qos none remote sendrecv
a=rtpmap:97 AMR
a=fmtp:97 mode-set=0,2,5,7; mode-change-period=2
a=rtpmap:96 telephone-event
a=maxptime:20

```

"a=des:qos mandatory local sendrecv" indicates the desired (des) quality of service (qos) precondition at the side of the message sender (local). It means that resources need to be reserved and the session cannot take place (mandatory) if the indicated resources are not reserved. "a=curr:..." indicates the current state (curr) of resources. "a=curr:qos local none" means that no (none) local resources are reserved on the side of the message sender. As defined in 3GPP, QoS preconditions are set by the MN. SIP signalling is exchanged between the caller and the callee until the current state of the QoS precondition reaches the desired state on both sides. In our case, when

resources are established in phase D, the MN sets the QoS preconditions and then performs phase F.

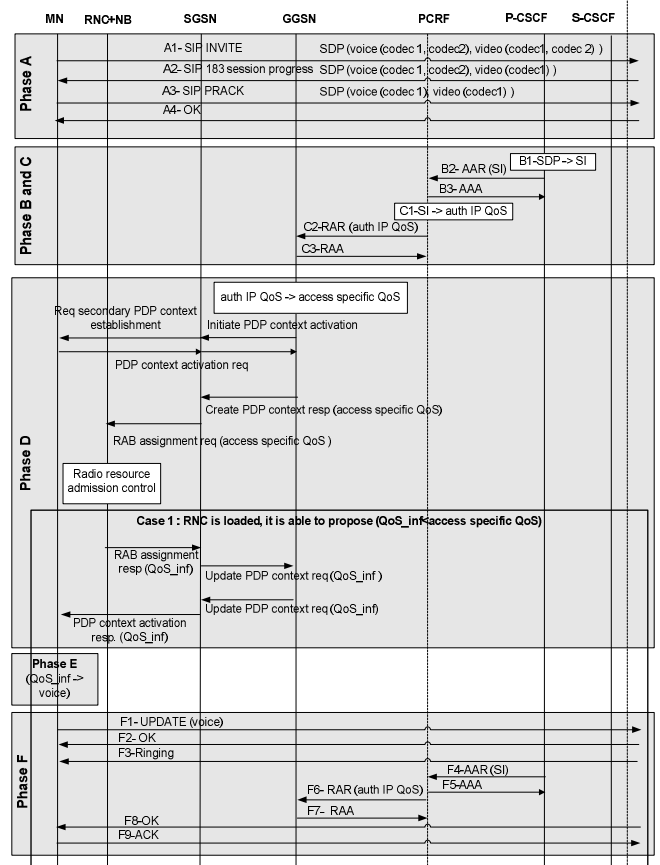


Figure 2. Flow chart for IMS session establishment over UMTS (case 1: resource problem in the RNC)

### III. ANALYSIS AND LIMITATIONS OF IMS AND PCC MODELS

#### A. Poor interaction with access network resources

PCC handles only the impact of session level on the bearer level and considers that authorised resources expressed by *access specific QoS* are statically available for the global bearer. Emergent services offer the possibility of mixing different components in the same service (e.g. voice + video + withebord), therefore it should be feasible to adapt the service according to the available resources. This is not possible in the current mobile architectures since resource information is localised in other network elements and are not reported to PCC. Let's consider the case of a video call session in case of an UMTS access network. If the access network cannot offer the required resources (*access specific QoS*) necessary to support the global bearer related to these two media, two cases arise.

Case 1: this case is illustrated in Figure 2. If the RNC has the capability to propose an alternate QoS ( $QoS_{inf} < access\ specific\ QoS$ ), it sends a response to the SGSN (UPDATE PDP CONTEXT) and the terminal (PDP CONTEXT ACTIVATION RESPONSE) about the proposed QoS ( $QoS_{inf}$ ). Assuming this RNC capability, when the MN receives PDP CONTEXT ACTIVATION RESPONSE message, if it implements advanced cross layering mechanism,

it will be able to find a new mapping (phase E) between  $QoS\_inf$  and a new service composition (voice only). Then, in phase F it sends a SIP UPDATE message indicating that resources are reserved for voice only and that video media will be inactive. The call can be thus established with voice only.

Case 2: generally, conditions of case 1 cannot be reached. In that case, the call will fail and the user has to re-initiate the session.

### B. A long session establishment delay

Session establishment procedure necessitates the exchange of a high number of messages which results in a high delay. This delay is also impacted by IMS elements load (P-CSCF, PCRF) especially that they are centralised.

## IV. PROPOSAL OF A NEW ARCHITECTURE

The above limitations come from the centralised IMS nature and from the separation of the control and the access network layers. As it can be extrapolated from case 1 analysis, to take into account radio resource information, PCC architecture decision point (the PCRF) should be closer to this information to interact more tightly with it. In view of this, we propose a new distributed IMS architecture called UFA (Ultra Flat architecture) that represents the ultimate step towards network flat architectures. With UFA (Figure 3), the network is no more composed of an access network and of P-CSCF and PCRF, but of BSs encompassing all the functions of the stated nodes and directly linked to the internet. PCRF function is not really needed as its mediation role between the P-CSCF and the PCEF can be achieved via cross layering mechanism inside the BS. By gathering session information and network resource information in the same node, it becomes possible to simultaneously establish sessions, check network resource availability, and adapt the service within the same procedure. This leads to a more efficient establishment procedure with a reduced delay. A SIP B2BUA is implemented in the BS to make the coordination between resources and services, autonomously orchestrated by the network.

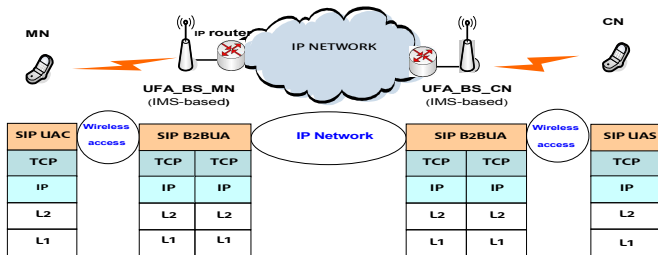


Figure 3. UFA architecture

In the next section, we will detail session establishment procedure for UFA. As UFA is a mobile architecture, it shall support mobility. The solution for that is already proposed in our recent work [6]. It is a network-controlled mobility executed by the BS via SIP protocol.

### A. Session establishment within UFA architecture

In classical IMS architectures, QoS preconditions are set by the terminal to reflect the resource reservation state. Within the proposed architecture, SDP and QoS preconditions are modified by the BS based on the knowledge of resource information. By modifying QoS preconditions, the BS pushes implicitly the end users to choose only among the media and the codecs for which it can offer resources. Therefore, there is no more need to rely on terminal intelligence as in case 1.

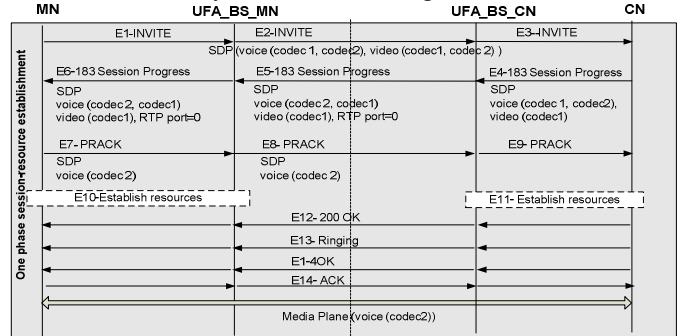


Figure 4. Flow chart for UFA architecture

The detailed messages for UFA establishment procedure are given in Figure 4. To initiate a video call, MN sends an INVITE message indicating the desired media and their codec (SDP (voice (codec 1, codec2), video (codec1, codec 2))). The two base stations UFA\_BS\_MN and UFA\_BS\_CN forward the message to the CN. Preconditions are included in this message and the current state is set to "none". CN answers in E4 with the supported media and codecs (SDP (voice (codec1, codec2), video (codec1))). When receiving this message, UFA\_BS\_CN has resources only for the codec2 of the voice media and has no resources for the video media. Consequently, it puts the codec2 in the first position before codec1 and modifies precondition current state to "sendrecv" meaning that resources are available for codec2. For video it puts the port number to 0. Then it sends E4 message with SDP (voice (codec2, codec1), video (RTP port=0)). UFA\_BS\_MN looks whether it has resources for voice (codec2), if it is the case it sends E6 message to the MN indicating that local resources are available (SDP voice (codec2, codec1), video (RTP port=0)). Since UFA\_BS\_MN has the knowledge of MN capabilities and state, it can determine its configuration (e.g. bearer and resource configuration) and include in E6. When receiving E6, MN configures itself and replies using E7. E10 may be executed to finalise resource establishment phase if not done in E6 and E7. In all cases resource establishment cannot fail since the UFA\_BS\_MN has already guaranteed the availability of its resources in E6. In E9, UFA\_BS\_CN reproduces the same process as UFA\_BS\_MN in step E6. After E11, resources are established on both sides and the call is notified to the CN.

## V. SESSION ESTABLISHMENT DELAY COMPARISON

In this section we compare IMS and UFA architectures regarding session establishment delay. For IMS architecture we consider the flow chart given in figure 2 which is applicable for normal and radio resource problem cases. To be generic, we suppose that the MN and the CN are situated on two different networks (domains). The difference between UFA and IMS delays is due to the number of exchanged messages and to the

number of the crossed network nodes and their load. Session establishment delay is composed of 1) the delay of resource establishment procedure (phase D in IMS and step E10 in UFA) and 2) the E2E delays of all exchanged SIP and Diameter messages. The E2E delay of a given SIP message is shown in Figure 5 and defined in Table I, it is composed of a) the queuing delays in the crossed nodes due to the processing delays and to the waiting times, especially in case of high load, b) the transmission delay over the wireless interface, c) the transmission delay over the other network interfaces.

Diameter messages are exchanged between the P-CSCF and the PCRF and between the PCRF and the GGSN. For these messages only queuing delay in the PCRF is considered.

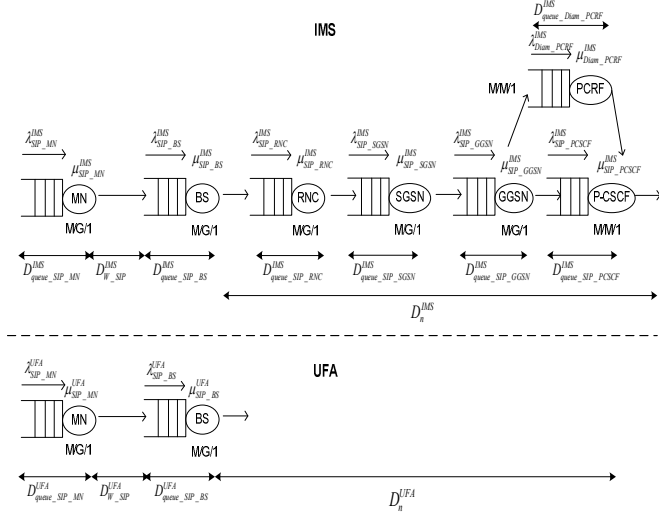


Figure 5. Delay components (IMS vs. UFA)

TABLE I. LIST OF SYSTEM PARAMETERS

$\lambda_{proto\_X}^{archi}$	The arrival rate of the messages belonging to the protocol "proto" in the network element "X" within the architecture "archi"
$1/\mu_{proto\_X}^{archi}$	The processing delay of the messages belonging to the protocol "proto" in the network element "X" within the architecture "archi". $\mu_{proto\_X}^{archi}$ is the processing rate.
$\rho_{p\_X}^{archi}$	The load of the traffic "p" processed in priority regarding traffic SIP "np" in the network element "X" within the architecture "archi"
$D_{queue\_proto\_X}^{archi}$	The queuing delay of messages belonging to the protocol "proto" in the network element "X" within the architecture "archi"
$D_{W\_SIP}^{archi}$	The transmission delay over the wireless interface for transmitting one SIP message within the architecture "archi"
$D_n^{archi}$	Delay transmission on the network interfaces from the BS within one domain for the architecture "archi"
$D_{estab\_res}^{archi}$	Delay for establishing resources in architecture "archi"

- "archi"  $\in$  {IMS, UFA}
- "X"  $\in$  {MN, CN, BS, RNC, SGSN, GGSN, P-CSCF, PCRF} when archi = IMS  
"X"  $\in$  {MN, CN, UFA\_BS} when archi = UFA
- "proto"  $\in$  {SIP, Diam}, Diam refers to diameter

TABLE II. SYSTEM PARAMETER VALUES

		X=						
		MN	BS	RNC	SGSN	GGSN	P-CSCF	PCRF
$\lambda_{SIP\_X}^{archi}$	IMS	1 ... 14	15 * $\lambda_{SIP\_MN}^{IMS}$	40 * $\lambda_{SIP\_BS}^{IMS}$	5 * $\lambda_{SIP\_RNC}^{IMS}$	2 * $\lambda_{SIP\_SGSN}^{IMS}$	1 * $\lambda_{SIP\_GGSN}^{IMS}$	
	UFA	1... 14	15 * $\lambda_{SIP\_MN}^{UFA}$					
$\lambda_{Diam\_X}^{IMS}$	IMS							$\lambda_{SIP\_GGSN}^{IMS}$
1	IMS	10	5	5	2	8	20	20
	UFA	10	40					
$\rho_{u\_X}^{archi}$	IMS	0.6	0.6	0.6	0.6	0.6		
	UFA	0.6	0.6					
$D_n^{archi}$	IMS	40						
	UFA	40						
$D_{estab}^{archi}$	IMS	2500						
	UFA	1000						

### A. Queuing delay

#### 1) Assumptions on processing and queuing delays

Processing and queuing delay parameters are shown in Figure 5 and defined in Table I and Table II for both UFA and IMS architectures. As SIP is an application layer, processing time of SIP messages in IMS-based equipments (MN, CN, P-CSCF) is high compared to non-IMS equipment. In UFA\_BS, SIP processing time is higher than in classical IMS-based equipments since additional L2 tasks are performed when receiving SIP messages. In the access network nodes (BS, RNC, SGSN, GGSN) the processing delays of SIP messages is on the contrary less significant, but their queuing delay may be high since they are less prior than other traffic types "u" (e.g conversational traffic).

#### 2) Queuing delay modeling

Rough estimates of queuing delays of SIP and Diameter messages can be obtained using classical queuing theory and waiting time-based formulas. M/M/1 queuing model is considered for the P-CSCF and the PCRF, assuming that they perform dedicated jobs for SIP and Diameter messages respectively. For the other network nodes (MN, CN, BS, RNC, SGSN, GGSN, UFA\_BS) we assumed a priority based M/G/1 model since these nodes may be busy with a most prior traffic "p" other than the less prior SIP traffic "np". Our modelling is inspired from [8] and checked in [9]. We give waiting time formulas using generic parameters  $\lambda$ ,  $\mu$  and  $\rho$  that represent respectively the arrival rate, the processing rate and the load of a given traffic  $\rho = \lambda/\mu$ .

1) In M/M/1 queue, the queuing delay for a given traffic is

$$D_{queue} = \frac{1}{\mu - \lambda} \quad (1)$$

2) In M/G/1 queue with a pre-emption, the queuing delay for a non prior traffic "np" in presence of higher prior traffic "p" is:

$$D_{queue - np} = \frac{(1 - \rho_p - \rho_{np}) + R}{(1 - \rho_p) \times (1 - \rho_p - \rho_{np})} \quad (2)$$

$$\text{Where } R = \lambda_p \overline{S_p^2} + \lambda_{np} \overline{S_{np}^2} \quad (3)$$

$\overline{S^2}$  is the second moment of  $\frac{1}{\mu}$ , it is given by

$\overline{S^2} = V(S) + [E(S)]^2 = [E(S)]^2 \times (1 + c_v^2)$  where  $V(S)$  is the variance and  $c_v$  is the deviation coefficient. We assume that  $c_v = 0.05$  which gives

$$R = 0.501 \times \left( \frac{\rho_p}{\mu_p} + \frac{\rho_{np}}{\mu_{np}} \right) \quad (4)$$

## B. Numerical results

The arrival rate of SIP INVITE messages in the MN is the rate of session launching. The arrival rate of SIP INVITE messages in the BS is proportional to the one in the MN given the number of users that can be attached simultaneously to the BS. The same assumption applies for SIP arrival rate in the other network elements. We used typical values (Table 2) for the system parameters and derived session establishment delay values using matlab tool. Figure 6 and Figure 7 show the variation of establishment delay components as a function of the number of sessions launched by the user and for a radio bitrate equal to 128 kbps. We notice that the establishment delay within IMS architecture grows rapidly from 7.42s to 12.58s, whereas it remains almost constant (4.49s) for UFA architecture, meaning a gain of 8.09s with UFA in high load situation. The increase of the session number per user impacts only the queuing delay which varies from 2.92s to 8.08s in IMS architecture and remains almost constant (2.01s) in UFA architecture. This is due to the fact that SIP messages quantity treated in UFA\_BS is largely smaller the one treated in the P-CSCF and PCRF given their centralised nature

## VI. RELATED WORK

[10] propose a solution to report resource information to the PCRF by the introduction of a new functional entity (QIF). QIF is linked to the PCRF and keeps track of the available resources on access networks to pre-reserve them on requests coming from the PCRF. However the interaction of this new function with existing network elements is unfeasible without introducing consequent modifications on the network. [11] covers mainly the issue of the long IMS session establishment delay in normal cases and propose to reduce it by executing some of the steps of session establishment procedure in parallel.

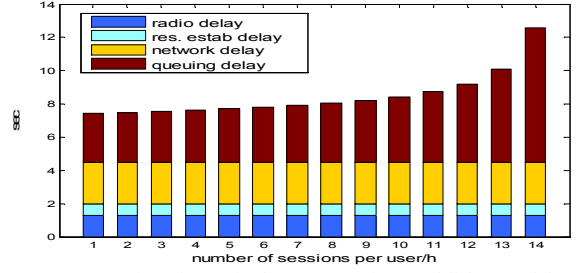


Figure 6. Impact of session arrival rate on session establishment delay in IMS

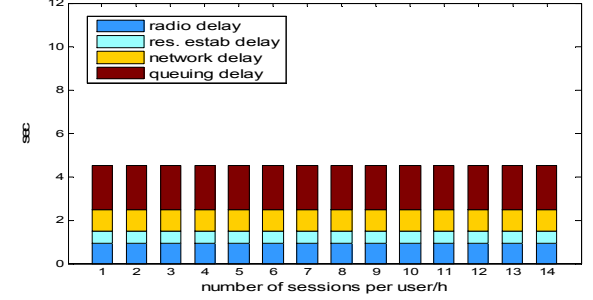


Figure 7. Impact of session arrival rate on session establishment delay in UFA

## VII. CONCLUSION AND FUTURE CHALLENGES

Based on the limitations of IMS and PCC architectures, we proposed in this paper a new mobile architecture (UFA). The key idea of UFA is an integrated session and resource negotiation procedure controlled by the network. Results show that this architecture enables a significant gain regarding session establishment delay. Future works have to develop the remaining functions of this architecture. UFA and IMS comparison shall be performed regarding all network functions.

## REFERENCES

- [1] 3GPP TS 23.228, "IP Multimedia Subsystem (IMS)".
- [2] 3GPP TS 23.203, "Policy Control and charging architecture".
- [3] 3GPP TR 24.930, "Signalling flows for the session setup in the IP multimedia core network subsystem (IMS) based on Session initiation protocol (SIP) and Session Description Protocol (SDP)".
- [4] 3GPP TS 23.060, "General Packet Radio Service (GPRS); Service description".
- [5] IETF RFC 2327, "SDP: Session Description Protocol", April 1998.
- [6] K. Daoud, P. Herbelin, Noel Crespi, "UFA: an ultra flat architecture for high bitrate services in mobile networks", PIMRC 2008.
- [7] IETF RFC 3312, "Integration of Resource Management and Session Initiation Protocol", October 2002.
- [8] N. Banerjee, K. Basu, S. K. Das, "Hand-off Delay Analysis in SIP-based Mobility Management in Wireless Networks", Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS'03).
- [9] L. Kleinrock, QUEUING SYSTEMS volume I: theory, John Wiley Sons, 1975.
- [10] M. I. Corici et al, "A network controlled QoS model over the 3GPP system architecture evolution", second international on wireless broadband and ultra wideband communications, 2007.
- [11] S. Zaghoul et al, "Extending QoS from radio access to an all-IP core in 3G networks: an operator's perspective", IEEE communications magazine, September 2007.