Measurement of SIP Signaling Performance for Advanced Multimedia Services

A. Brajdic, M. Suznjevic, M. Matijasevic
University of Zagreb, Faculty of Electrical Engineering and Computing Unska 3, HR-10000 Zagreb, Croatia
{agata.brajdic, mirko.suznjevic, maja.matijasevic}@fer.hr

Abstract— **The purpose of this paper is to provide an insight in the process of obtaining and interpreting values of the collection of metrics that are intended for performance evaluation of the Session Initiation Protocol (SIP). Three of the analyzed metrics are defined in a recently proposed Internet Engineering Task Force (IETF) draft, however, additional two metrics addressing negotiation and renegotiation procedures are introduced in this paper for the first time. The testing was performed upon emulated IP Multimedia Subsystem (IMS) network that incorporates a specialized network entity supporting dynamic quality of service (QoS) adaptation of SIP session parameters. Obtained results provided us with an overview of a typical behavior of SIP metrics' values. Furthermore, by comparing results obtained for different network scenarios, we were able to isolate several weaknesses of the testbed and deduce potential solutions to them.**

I. Introduction

Signaling performance of Session Initiation Protocol (SIP) is an important factor affecting overall quality of service (QoS) in next generation networks such as the IP Multimedia Subsystem (IMS) [1]. Figure 1 depicts simplified IMS architecture [2]. The key multimedia session control element is the Call Session Control Function (CSCF). CSCF may play several roles: handling signalling traffic related to session establishment, modification and release (Serving CSCF, S-CSCF) and acting as a proxy server performing user registration and resource authorization procedures (Proxy CSCF, P-CSCF). Home Subscriber Server (HSS) is a database containing all user subscription data, while the Application Servers (AS) offer various functionalities and may be located in either the home or external network.

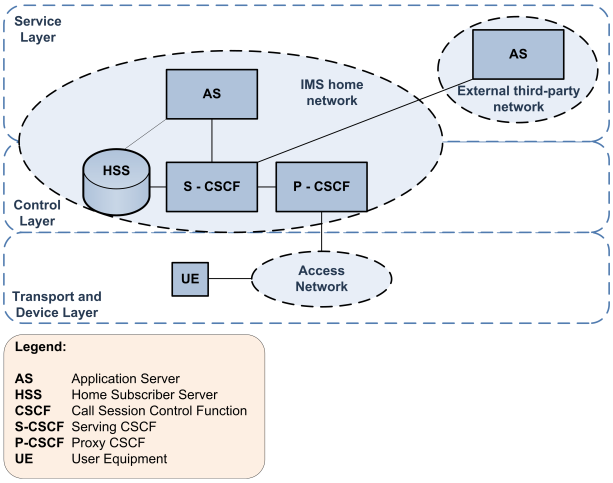

Figure 1. IMS architecture overview

Although there are many existing standards for performance evaluation of various signaling protocols, none of them specifically addresses SIP. Consequently SIP performance in the literature is described in various ways and on many levels, varying from Network Interface Card (NIC) driver, kernel, and SIP stack performance to overall network and application performance. The current research in network performance area produced a collection of definitions of SIP metrics [3] that evaluate a communication system during SIP session establishment, its duration and termination. In our research we aim to examine in detail the SIP performance of our testbed implementation of Dynamic Service Adaptation Model (DSAM) [4], which provides functionality of dynamic adaptation of QoS parameters to changes occurring in either network resources, client preferences, terminal/access type or service requirements. In addition, we developed new metrics in order to better describe the specific behavior of our system. Measurements have been performed upon laboratory implementation of DSAM model mapped on IMS through a series of simulations of different network scenarios. Numerical values were analyzed and interpreted in the context of the utilized test environment. Results are used in determining overall SIP performance, for identification of time consuming system parts and potential system bottlenecks, and to identify potential improvements that can be implemented in next software versions.

II. Related work

Although multiple papers deal with assessment of SIP server performance in various network scenarios, the common set of metrics, used for purposes of evaluation and comparison of similar SIP systems, remains undetermined. In order to numerically characterize SIP server performance, authors typically develop their own metrics to quantify session setup delay. Authors in [5] hence propose a methodology for performance measurement of SIP systems hosting virtual machine based applications. In their work, they identify various tuning parameters ranging from number of servers and CPUs per server to garbage collector and thread configuration. Depending on determined parameters, they evaluate defined performance metrics such as application latency, SIP message throughput and SIP node capacity. This type of evaluation results in identification of parameters to be tuned in order to enhance system performance.

Furthermore, authors in [6] explore dependency of initial call establishment time delay on SIP thread architecture, memory allocation and SIP message length. Similarly, authors in [7] express SIP message parsing time as the percentage of total message processing time depending on several parameters such as underlying operating system, SIP server configuration and SIP message format. Performance evaluation of SIP signaling scenarios in Universal Mobile Telecommunications System (UMTS), focused on selected SIP based services, has been presented by authors in [8]. Results from their simulations showed that the large SIP message sizes and the number of messages contribute significantly to call setup delays in voice based services. Finally, work in [9] addresses influence of utilized transport protocol, user authorization procedures and SIP proxy working configuration (stateful/stateless) on overall call establishment delay. Moreover, authors evaluated latency metric defined as the time interval between the initial SIP INVITE request and corresponding 200 OK (INVITE) response. Their results proved this metric's value is highly influenced by the load, thus, increasing percentage of unsuccessful call attempts and the average user waiting time.

In our work, instead of evaluating system specific metrics, we focused on evaluation of metrics proposed in [3] and assessment of their suitability for our system. Similarly to aforementioned works, our evaluation of metrics' values resulted in identification of time consuming system parts.

III. OVERVIEW OF DSAM MODEL

The system upon which testing is performed is an implementation of Dynamic Service Adaptation Model (DSAM), described in detail in [4].

The DSAM model provides end-to-end support for QoS signaling for advanced multimedia services at the session layer. The support is realized throughout the initial negotiation of QoS parameters at the session startup and mechanisms for their dynamic adaptation as a result of various changes occurring in the system.

The session establishment is conducted through negotiation of session parameters that are supported by the client and required by the requested service. These parameters are composed into the client and service profile, respectively. Furthermore, DSAM specifies two procedures for calculation of optimal service configuration, namely parameters matching and optimization. Matching process determines a series of feasible service configurations that conform to restrictions imposed by requested service, client's terminal and access network and meet all user personal preferences. The optimization process performs a suitable optimization of resources that are to be allocated for the purpose of service provision towards maximization of user perceived quality.

Once the session is established, renegotiation process is invoked in case one of the following changes occurs in the system:

- Change in service requirements (e.g. adding or removing application multimedia components which entail modification of allocated network resources)
- Change in network resource availability (referring to increase/decrease of available bandwidth of previously authorized and reserved network resources)
- Change in client profile (occurs in case of significant modification in user preferences or switch to another access network or terminal)

Renegotiation process may invoke reoptimization procedure if previously allocated resources are not substantial to serve new service configuration.

IV. SIP PERFORMANCE METRICS APPLIED TO DSAM MODEL

In this section, we highlight metrics selected from [3] that we believe to be the most suitable for our specific SIP system. Moreover, it is illustrated how these metrics were applied to our test system in the terms of SIP messages exchanged throughout different phases of the SIP session. Moreover, we also introduce additional SIP metrics, defined in conformance with the template for SIP metric design.

DSAM model is independent of underlying network topology. In case of mapping onto IMS network as shown in [10], matching and optimization procedures are conducted by specialized SIP Application Server called QoS Matching and Optimization Application Server (QMO AS). This network node serves as a generic and reusable building block of the IMS network providing adaptive QoS support to the numerous different multimedia services developed by third-party providers. Figure 2 and Figure 3 present signaling occurring in the IMS network with end-to-end QoS support during different phases of multimedia session.

A. Session establishment and termination signaling

Figure 2 depicts defined protocol for multimedia session establishment and termination in the IMS. The exchanged signaling is independent of the actual service content (Networked Virtual Reality - NVR scene, audio-video content) that is being delivered once the session has been established.

Client initiates session establishment process by sending SIP INVITE request carrying client profile towards AS hosting desired service. This initial SIP transaction involves the negotiation process within which all relevant session parameters are agreed by both sides involved. Negotiation is further carried out through exchange of SIP 183 Session Progress response and PRACK and UPDATE requests utilized for the purpose of delivery of matched and optimized final service profile to both end-points.

In accordance with the definition in [3], the metric Session Request Delay (SRD) was calculated at the client side as the time interval between sending the INVITE request and receiving the 180 Ringing response which indicates the beginning of processing of received request at the other end. The SRD metric is used to represent a time delay for receiving initial response from the other

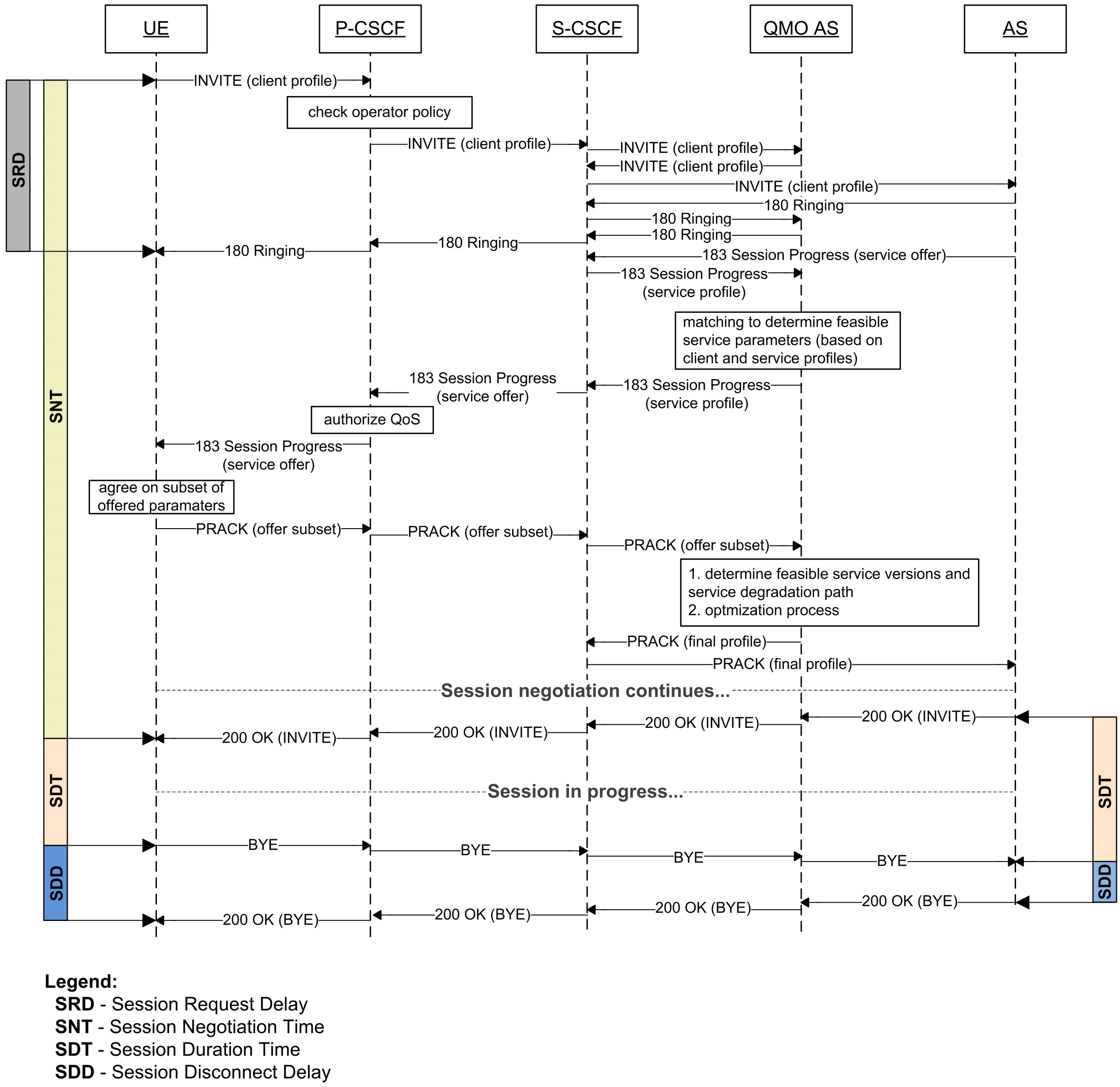
UE
P-CSCF
S-CSCF
QMO AS
AS
SRD
SNT
SDT
SDD
INVITE (client profile)
check operator policy
INVITE (client profile)
INVITE (client profile)
INVITE (client profile)
INVITE (client profile)
180 Ringing
180 Ringing
180 Ringing
180 Ringing
180 Ringing
183 Session Progress (service offer)
183 Session Progress (service profile)
matching to determine feasible service parameters (based on client and service profiles)
183 Session Progress (service offer)
183 Session Progress (service profile)
authorize QoS
183 Session Progress (service offer)
agree on subset of offered paramaters
PRACK (offer subset)
PRACK (offer subset)
PRACK (offer subset)
1. determine feasible service versions and service degradation path
2. optmization process
PRACK (final profile)
PRACK (final profile)
Session negotiation continues...
200 OK (INVITE)
200 OK (INVITE)
200 OK (INVITE)
200 OK (INVITE)
Session in progress...
BYE
BYE
BYE
BYE
200 OK (BYE)
200 OK (BYE)
200 OK (BYE)
200 OK (BYE)
Legend:
SRD - Session Request Delay
SNT - Session Negotiation Time
SDT - Session Duration Time
SDD - Session Disconnect Delay

Figure 2. Definition of SIP metrics during session establishment and termination

end of the communication. It is, thus, insufficient to provide an accurate session setup assessment due to the complexity of the test system and conducted negotiation protocol. Inasmuch as the negotiation procedure of session characteristics is specified by 3GPP and based on Internet Engineering Task Force (IETF), it is expected that many other similar systems implement an analogue negotiation protocol. Therefore, we find it essential to introduce another SIP metric that would characterize a complete negotiation phase and serve as the indicator of average initial user waiting time as well as the mean for comparing the similar systems.

Formally, we have defined the proposed Session Negotiation Time (SNT) metric as the elapsed time between the SIP INVITE request that initiates negotiation procedure and its corresponding 200 OK response. Moreover, this metric can be used to detect problems causing a failure of negotiation procedure, and consequently, a failure in the session establishment. Metric definition may be expanded to address other application protocols, which might utilize different SIP messages for conveying negotiation information. In our model, the SIP message in question is the INVITE request. Moreover, this metric correlates with latency metric defined in [9] and exhibits same behavior under increased number of session requests, as it will be demonstrated in Section VI-B. It should be further noted that the negotiation is assumed to take place before or during session establishment. The process of repeated negotiation or renegotiation, however, is characterized with another similar metric.

Additionally, the 200 OK (INVITE) response represents begin time of another SIP metric, the Session Duration Time (SDT). This metric represents a typical session duration time and can be utilized to detect problems

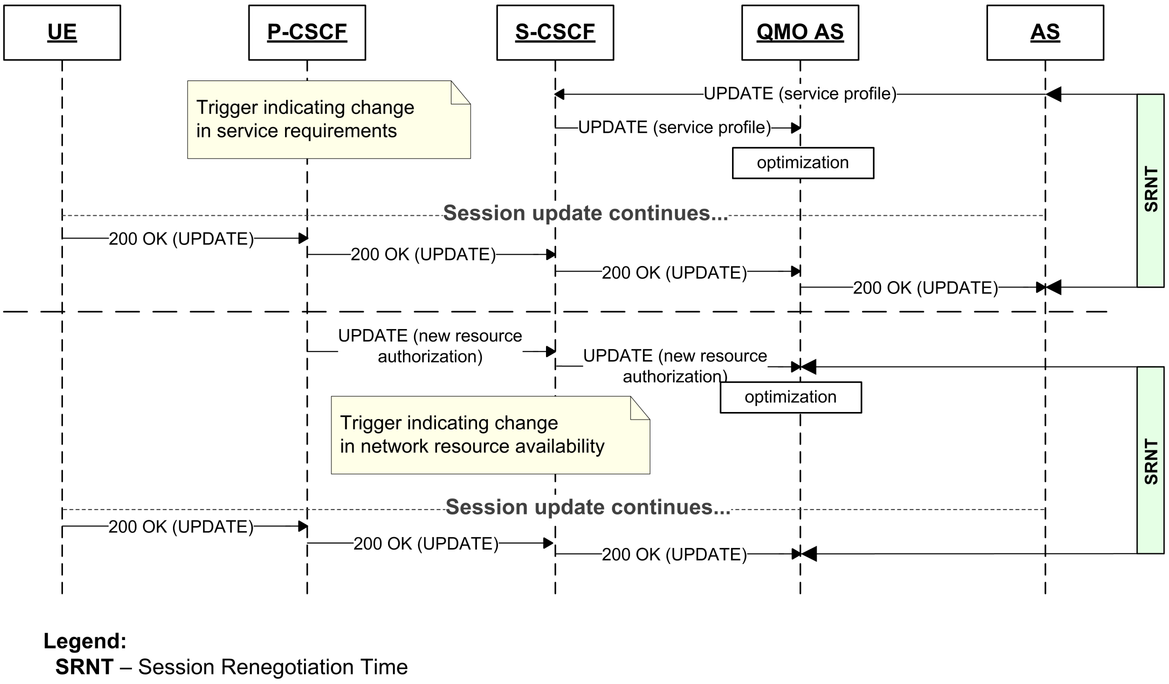
UE
P-CSCF
S-CSCF
QMO AS
AS
Trigger indicating change in service requirements
UPDATE (service profile)
UPDATE (service profile)
optimization
SRNT
Session update continues...
200 OK (UPDATE)
200 OK (UPDATE)
200 OK (UPDATE)
200 OK (UPDATE)
UPDATE (new resource authorization)
UPDATE (new resource authorization)
optimization
Trigger indicating change in network resource availability
SRNT
Session update continues...
200 OK (UPDATE)
200 OK (UPDATE)
200 OK (UPDATE)
Legend:
SRNT – Session Renegotiation Time

Figure 3. Definition of SIP metrics in case of changes occurring in the system

causing exceptionally short session durations [3]. The termination of multimedia session conforms to the generic model defined in SIP specification [12]. Therefore, the definition of the Session Disconnect Delay (SDD) and the ending time of SDT metric applied to signaling in our test system are rather straightforward. Measurement of SDT value ends when the user decides to terminate the session by sending SIP BYE request, whereas the same request represents the begin time of the SDD metric which ends once the request is confirmed.

B. Signaling indicating a change in the system

SIP metrics proposed in [3] are not sufficient to address the signaling that is being exchanged throughout the session duration. In order to precisely assess renegotiation process that is triggered by the change in the system and used for updating session parameters on both end points of the communication, we have introduced another SIP metric called Session Renegotiation Delay (SRNT). SRNT is used to characterize renegotiation process that is typically initiated by the SIP UPDATE or re-INVITE message carrying updated information and terminated by an adequate SIP 200 OK response. The metric is being measured on the network entity that initiates renegotiation process, which varies depending on the nature of the change. The value of this metric indicates time delay that needs to pass for session to be updated and can further be utilized to detect failures or impairments causing delays in renegotiation process.

1) Change in service requirements: The change in service requirements may be simulated by adding additional media object (e.g. additional video stream) into the current service configuration. This event would trigger sending SIP UPDATE request from the service content AS with updated service profile. SIP UPDATE request further invokes the recalculation of the application operating point in the repeated optimization process on QMO AS element. In this type of change in the system, the SRNT metric partially coincides with the time spent by the user waiting on the service, namely, the reproduction of video content. Consequently, by detecting and improving the process in the system that contributes to the overall value of the SRNT metric the most, we would be able to improve user perception of service quality. Referring to the thresholds defined in [11], user waiting should not be longer than 1s, since the user is more sensitive to service delay once the service has started than in the session establishment phase.

2) Change in network resources: This type of system event is typically simulated by alteration of the amount of available network bandwidth (e.g. reduction of downlink bandwidth in user access network). Once the resource manager has detected this reduction, P-CSCF composes an adequate SIP UPDATE request comprising new resource reservation parameters and sends it over to the AS hosting the service. On its way to the service content server, the request would be intersected by QMO AS element which would again perform the optimization of parameters needed for the resource reservation. The client side is not being involved in this renegotiation process till the very end, when it is being informed about the updated final profile. However, the user will typically experience short suspension of streamed multimedia content until the service configuration is updated (e.g. new codec for the video stream determined). Length of this interruption is also included in the SRNT metric value. In DSAM model, SRNT value in this type of network scenario is measured at the QMO AS element, where, strictly speaking, the renegotiation process begins. Similar to the previous scenario, the SRNT metric value again includes optimization process as well as the transmission of the updated final profile through the network. Therefore, it would be of interest to compare the concrete numerical values of the same metric for different network scenarios in the following section.

V. CASE STUDY

As a case study we used two prototype applications in the laboratory testbed. Applications, testbed and measurement scenarios are described next.

Our prototype applications incorporate audio and video streaming with 3D virtual worlds realized in Virtual Reality Modeling Language (VRML). The first application is Virtual Auto Gallery (VAG). Users in VAG can explore the virtual world and retrieve video and audio commercials of certain cars by clicking on the advertising booths. The second application is Inheritance Chase (IC), a web-based interactive multiplayer game. IC virtual world consists of two separate environments: island and interactive chessboard. To win a game a player needs to find a testament hidden somewhere on the island by following certain clues. Clues are realized trough audio and video streaming depending on the service configuration.

A. Measurement methodology

Test scenarios that were conducted throughout the measurement process differed in the number of users that simultaneously requested the service.

A typical user scenario used for the purpose of measurement procedure consisted of invoking VAG/IC service establishment through matching and optimization of client and service parameters, and fetching the 3D virtual scene that is displayed in user's browser. Users can explore and interact with the virtual world. User interaction triggers adding of additional streaming multimedia components to the session and hence initiating signaling of new service requirements and negotiation of corresponding QoS parameters (DSAM scenario Change in service requirements). Change in authorized network resources (DSAM scenario Change in resource availability) was carried out on the audio channel during ongoing audio transmission. Its bandwidth was decreased from the value of 11.7 kbit/s to the value of 3.9 kbit/s.

Equivalent procedure was repeated for increased number of simultaneous users. We performed measurements with three different users which did not have previous experience with this type of application. In a multi-user scenario each user can follow his path and pace, so the duration of the session can differ significantly between different users.

Metric's values evaluation was performed by inspecting the arrival time of SIP signaling messages at measuring points. Packets are being captured using Wireshark network analyzer [13]. Values obtained throughout the evaluation process and their relation served us as a quantitative indicator of system's behavior in a multi-user environment. This way we were able to determine side effects of increase of number of users.

We performed 20 test iterations for one and two concurrent users and VAG 10 for three users.

B. Testbed configuration

Our testbed consists of 6 PCs connected by a 100 Mbit/s Ethernet LAN. The test bed configuration consisted of hardware and software (I) and our own software deployed as shown in Figure 4. While our testbed uses primarily MS Windows platform, it would be interesting to compare results on a more diverse (e.g. Linux) platforms.

TABLE I
LABORATORY CONFIGURATION

HOST	HARDWARE	SOFTWARE
PC 1	Pentium IV, 1.6 GHz, 512 MB RAM, table superscripts	Fedora Core; NISTNet emulator
PC 2	Pentium IV, 2.4 GHz, 512 MB RAM	Windows XP; Apache Tomcat 5.5; JMF 2.1.1
PC 3	Pentium IV, 1.7 GHz, 512 GB RAM	Windows XP
PC 4	Pentium IV, 2.4 GHz, 1 GB RAM	Windows XP; Cortona 3.0 VRML plug-in; JMF 2.1.1
PC 5	Pentium IV, 1.6 GHz, 512 MB RAM	Windows XP; Cortona 3.0 VRML plug-in; JMF 2.1.1
PC 6	Pentium III, 1 GHz, 384 MB RAM	Windows XP; Cortona 3.0 VRML plug-in; JMF 2.1.1

We used network emulator NIST Net to emulate different network parameters. Changes in the parameters of the network simulated by NIST Net triggered the change of the service by renegotiating session parameters using SIP protocol. Furthermore, Apache Tomcat 5.5 was used for hosting our virtual world and Java Media Framework (JMF) 2.1.1 for streaming audio and video media. All network applications (client application, service AS, QMO AS, CSCF nodes) have been developed under Java SE 6 environment.

Furthermore, Apache Tomcat 5.5 was used for hosting our virtual world and JMF 2.1.1 for streaming audio and video media. All network applications (client application, service AS, QMO AS, CSCF nodes) have been developed under Java SE 6 environment.

VI. MEASUREMENT RESULTS

The following section provides obtained numerical results and their interpretation in the context of tested system.

A. Single user scenarios

As the Figure 5 demonstrates, values of SRD and SDD metrics coincide for both applications. Applications are, however, characterized with different session duration times and, consequently, different values of SDT metric. This type of metric's regularity is highly desirable since it certifies the invariance of delay corresponding to the processing of session request and the one corresponding to the session termination time. When comparing values

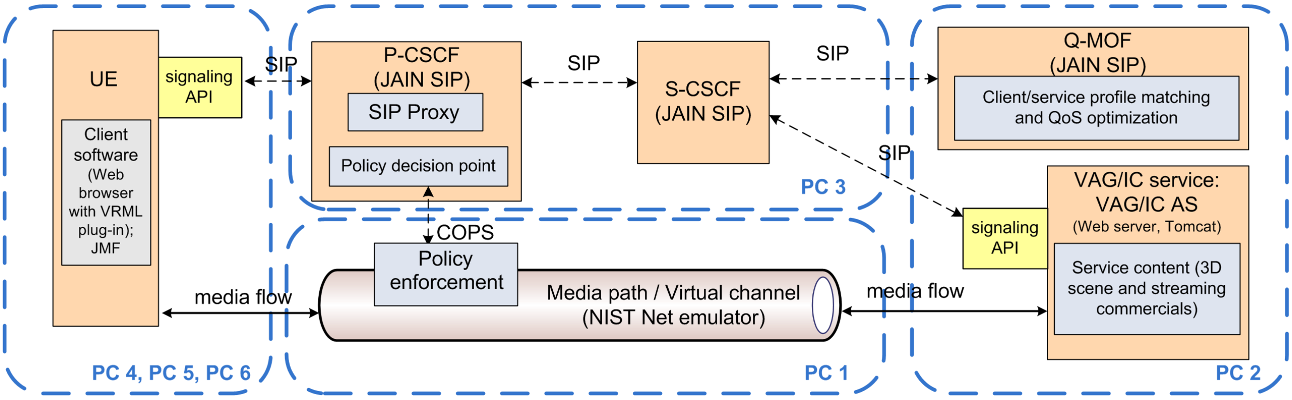
UE
signaling API
Client software (Web browser with VRML plug-in); JMF
SIP
P-CSCF (JAIN SIP)
SIP Proxy
Policy decision point
SIP
S-CSCF (JAIN SIP)
SIP
Q-MOF (JAIN SIP)
Client/service profile matching and QoS optimization
SIP
VAG/IC service: VAG/IC AS
(Web server, Tomcat)
signaling API
Service content (3D scene and streaming commercials)
PC 3
COPS
Policy enforcement
Media path / Virtual channel (NIST Net emulator)
media flow
media flow
PC 4, PC 5, PC 6
PC 1
PC 2

Figure 4. DSAM entities in mapped to laboratory testbed

obtained for SRD and SNT SIP metrics, it is easy to notice their disproportion. If we take into account that in our case the SNT metric directly reflects the user waiting time on a service, its average numerical value significantly exceeds limit value of approximately 2 to 5 seconds proposed in [11]. This was the main motivation for measuring the duration of separate processes included in the negotiation procedure in order to determine the most time consuming one. We have ascertained that the optimization process does not contribute to the value of the SNT metric with as big proportion as it was initially expected. The efficiency of the initial optimization process amounts to 4.5% for VAG and 5.5% for IC application. Similarly, the efficiency of the optimization process in case when the change in service requirements has been detected amounts to 13.5% and 8%, and when the decrease in network resources has occurred to 9% and 8.5% for VAG and IC applications, respectively. We believe that it is reasonable to expect optimization process to be consuming 10-15% of overall time. Therefore, the complexity of utilized system protocol that consists of many transactions among nodes involved in the communication has the main impact on the SNT and SRNT metric values. Furthermore, every SIP message is being routed through QMO AS element creating an additional delay. Finally, many transmitted SIP messages contain profile (client, service or final) having a size is between 9000 bytes and 17600 bytes. Therefore, one of possible adaptations in the next version of system implementation would be storing predefined profile information on the QMO AS element. In this case, instead of conveyance of complete profiles, only references to it would be transmitted through the SIP signalization. Consequently, this would reduce time needed for transmitting a single SIP message and, finally, SNT and SRNT metric values.

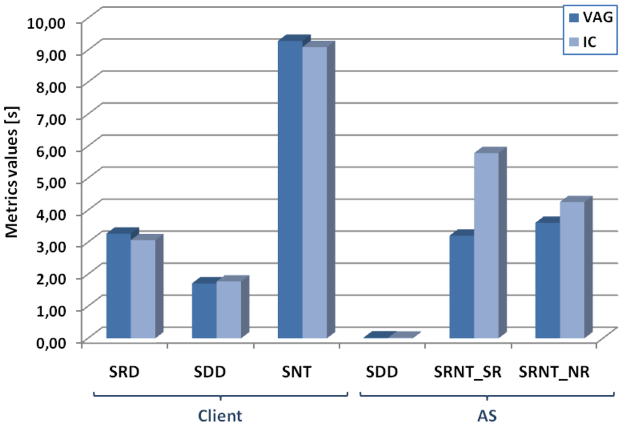

Figure 5. Illustration of metrics' values for two different NVR applications

Similar conclusions can be derived from the comparison of the optimization process efficiency. Due to differences of service content, the optimization time varies between applications, but relative proportions of the parameter calculated for different session phases remain the same. Since fewer SIP messages are being exchanged during processing the change in service requirements, in this case, the optimization impacts SNT metric the most.

B. Multiple users scenarios

The next testing phase consisted of performing same scenarios simultaneously on two or three separate client applications installed on separate machines. Figure 6 presents SIP metric values comparison obtained for different number of user sessions.

The service that is being requested in all further scenarios is VAG. All applications were configured equally and, consequently, were characterized with the identical client profile. We were interested in general conclusions about system scalability and SIP metrics' values that can be derived from comparing metrics' values obtained from measurements involving a varying number of concurrent users requesting the same network service.

1) Two user scenario: As expected, the obtained metrics' values for each of the users mutually coincide. If we now compare the metrics' values with the ones from the single user environment, we are able to notice an interesting anomaly. Every SIP metric that was measured, except SRD, has relatively increased in its value with respect to the single user case. On the other hand, the SRD metric's value is not only reduced in this scenario, but also significantly oscillates in different iterations of the measurement process. The proper explanation for this unexpected system behavior can be found in the

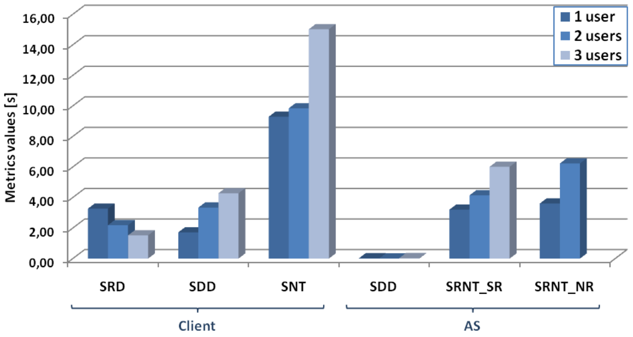

Figure 6. Metrics' values for different number of network users

system implementation. Namely, both client and server applications are developed in Java programming language. Before executing Java class files, they need to be loaded and then translated into a language of a machine on which the Java Virtual Machine (JVM) is run. This procedure may take some time to be performed. Once the Java class file gets translated, it remains in the system's memory for some period of time. Thus, if the next incoming request is being served by invoking the same class file, its repeated translation is unnecessary and the processing of the request may begin almost immediately. This explains our observations that the service request that is received first in the system is being regularly served longer than the following ones. Finally, the average SDT metric per user is being reduced in its value.

Expected increase in values of SNT, SRNT, and SDD metrics can be explained as follows. The decrease in a metric's value that was explained earlier occurs only if parallel processing of two concurrent session requests do not consume much of the CPU time and can be performed in parallel threads in the server application. Once the processing becomes too complex and time consuming, like in the cases of optimization process affecting values of SNT and SRNT metrics and resource release procedures affecting SDD metric value, threads handling parallel requests start to compete for CPU processing time. By measuring CPU activity on QMO AS using Microsoft Process Explorer [14] tool, we notice a significant increase in time interval in which CPU load was over 60 percent. Consequently, end-users will experience increase in time delay during session establishment, update and termination.

The average time period needed for performing the optimization in case of two users has decreased in the phase of session establishment (when the users were simultaneously attempting to establish session). However, in case of change in service requirements, the delay caused by the optimization is not altered significantly. This can be explained by the fact that users were not attempting to start the video stream at the very same time. Therefore, due to the increase of the SRNT metric value, the efficiency of the optimization process is reduced to 5%-11.5%. The conclusion derived from this observation is that the system lacks proper methods for storing and reapplying past results of the optimization process in future calculations. In case they were implemented, there would be no need for executing optimization for users who are characterized with same client profile and are requesting the same service. This improvement will be taken into consideration in the next steps of system development.

2) Three users scenario : Since the behavior of metrics' values in this scenario is similar to the case with two users, the discussion from that case can be applied here as well. In addition, we will just point out that the increase in metrics' values, although noticeable, is still not that drastic. For this application, progressive degradation of metrics' values of approximately 10% per user is expected.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we have applied the subset of SIP metrics proposed in [3] for the purpose of performance assessment of our current implementation of DSAM model. Numerical results obtained throughout measurement procedure helped us identify several deficiencies in current software version in terms of time delay for session establishment/update and user perceived quality. Furthermore, the optimization process, that was primarily suspected to be a potential bottleneck, however, proved to perform within acceptable thresholds.

Moreover, it has been proved that the information provided by proposed SIP metrics is often insufficient for getting a precise conclusion about system performance. Therefore, considering the negotiation procedure as a common feature present in a session establishment and update, we have extended the existing set of metrics with new ones addressing negotiation and renegotiation session phases.

For future work we plan to undergo same measurement procedure using provisionary SIP load generator (e.g. SIPp) in order to simulate alternative network scenarios with dynamic network users behavior.

REFERENCES

[1] G. Camarillo and M.A. Garcia-Martin, *The 3G IP Multimedia Subsystem, Merging The Internet and Cellular Worlds*, Wiley, 2004.

[2] 3GPP TS 23.228: IP Multimedia Subsystem (IMS); Stage 2, Release 8, 2008-06

[3] D. Malas, *SIP End-to-End Performance Metrics* (draft-ietf-pmol-sip-perf-metrics-02.txt), IETF Working Group, Work in Progress, October 2008.

[4] L.Skorin-Kapov and M.Matijasevic, *End-to-end QoS signaling for Future Multimedia Services in the NGN*, Proc. Next Generation Teletraffic and Wired/Wireless Advanced Networking, 6th Int'l Conf., LNCS, vol. 4003, May/June 2006, pp. 408-419.

[5] C. Hrischuk, G. DeVal, *A Tutorial on SIP Application Server Performance and Benchmarking*, in Proc. 32nd International Computer Measurement Group Conference, Nevada, USA, Dec. 2006

[6] M. Cortes, J.E. Ensor, J.O. Esteban, *On SIP Performance*, Bell Labs Technical Journal, 9(3), 155-172, Nov. 2004

[7] S. Wanke, M. Scharf, S. Kiesel, S. Wahl, *Measurement of the SIP Parsing Performance in the SIP Express Router*, Lecture Notes in Computer Science, Vol. 4606/2007, August 20007, pp. 103-110

[8] D. Pesch, I. M. Pous, and G. Foster, *Performance evaluation of SIP-based multimedia services in UMTS*, Computer Networks, vol. 49, no. 3, October 2005, pp. 385-403.

[9] E.M. Nahum, J. Tracey, C.P. Wright, *Evaluationg SIP Proxy Performance*, Proc. of 17th International workshop on Network and Operating Systems Support for Digital Audio and Video, June 2007

[10] L. Skorin-Kapov, M. Mosmondor, O. Dobrijevic, and M. Matijasevic, *Application-level QoS Negotiation and signaling for Advanced Multimedia Services in the IMS*, IEEE Communications Magazine, Vol. 45 No. 7, 2007, pp. 108-116.

[11] T. Guenkova-Luy, A. J. Kassler and D.Mandato , *End-to-End Quality-of-Service Coordination for Mobile Multimedia Applications*, IEEE Journal on Selected areas in Communications, Vol. 22, No 5, June 2004.

[12] J. Rosenberg et al., *SIP: Session Initiation Protocol*, IETF RCF 3261, June 2002.

[13] Wireshark 1.0.6 network protocol analyzer, http://www.wireshark.org/

[14] Microsoft Process Explorer v11.33, http://technet.microsoft.com/en-us/sysinternals/bb896653.aspx