Abstract—IP Multimedia Subsystem (IMS), an architectural framework for delivering multimedia services standardized by Third Generation Partnership (3GPP/3GPP2), is deployed as part of 3G and beyond wireless networks. It facilitates wireless carriers to offer rich multimedia services such as Chat, Presence, and Video Conference with a high degree of flexibility. These services vary in terms of their delay sensitivity and cost. With increasing demand for IMS services, service providers are facing substantial new challenges to deliver services with the required Quality of Service (QoS) to end users. One of the main factors that negatively influences the QoS is session setup delay experienced by the users. This work focuses on the impact of session setup delay on the QoS experienced by users for each of the IMS services, and proposes a simulation model to optimize selection of services that can be offered with the required QoS so as to maximize revenue generated. The preliminary results show that the proposed approach maximizes the QoS, as well as the revenue for the service providers.

Index Terms—IP Multimedia Services; service selection optimization; Quality of Service; modeling and analysis.

I. INTRODUCTION

The IP Multimedia Subsystem (IMS) architecture framework was created by the Third Generation Partnership Project (3GPP/3GPP2) as part of the 3G-multimedia protocols set. IMS has now emerged as the preferred solution for service delivery in Next Generation Networks (NGN). It is designed to be integrated with 3G and beyond networks to offer a wide range of rich services, such as Push to Talk (PTT), Chat, Presence, and Video Conference. IMS is a complex system that contains several sets of servers collaborating with different protocols to support powerful mechanisms for services management. These mechanisms mainly include the creation, maintenance and termination of services and sessions setup among users. The services supported by IMS have different delay sensitivity and vary in their cost. One of the main challenges for the IMS network designers is to ensure that the system is capable of providing the expected QoS 3GPP2 classifies services into four classes: streaming class, conventional class, interactive class, and background class.

The review of the literature identified several papers that analyze the delays detected during session establishment processes and provided solutions to address them [1-5]. For example, reference [1] presents an approach to protect the performance of a Session Initiation Protocol (SIP) signaling network from degrading under overload conditions. Reference [2] provides a comprehensive analysis of the IMS architecture performance from an end to end delay perspective, whereas the reported work in [3-4] provides an analysis of session setup delay in IMS while deployed in a CDMA2000 network. Overall, the review of the literature confirmed that the IMS session setup delay problem was mainly addressed by considering a general session setup sequence. To the best of our knowledge, the impact of offering different services on the overall IMS delay has not been thoroughly studied. This paper presents a modeling approach for the registration delay as well as session setup delays for five different services, namely the IPTV, Watcher Presence, Publish Presence, Push To Talk, and Video Conference. The focus of the work is on the technique of selecting services to be processed by the IMS system. When different services arrive to the system, the system needs to process a set of services and these selected set needs to ensure the best QoS offered to the user. Based on the setup delay observed for the above services, we propose a simulation and optimization model that maximizes the service providers’ revenues and QoS in addition to reducing the setup delay experienced by users.

II. IMS ARCHITECTURE

The core of the IMS system consists of three Call Session Control Function (CSCF) servers that ensure the proper operation of the control and signaling functions. This work looks in detail at those servers since they have the largest involvement in the service setup process. The CSCF function is subdivided into the Proxy-, Interrogating-, and Serving-Call Session Control Functions (P-, I-, S-CSCF), as presented in
Figure 1 below. The first one, the Proxy-CSCF function, usually designed to be part of the visitor network and intended to serve as the first point of contact for users, is responsible for user security and for routing IP packets to the home network associated with that P-CSCF [6-10]. Also, P-CSCF secures and maintains the connection with resources required by the terminal. The second function, the Interrogating-CSCF function, located within the home network, re-routes the packets to different networks, servers, and databases and accesses IMS subscribers’ database, which is denoted as the Home Subscriber Server (HSS) [6-10]. The third function, Serving-CSCF, authenticates and registers users by accessing the HSS. The last subsystem presented in Figure 1 is the Application Server (AS), which is an essential part of the IMS as it manages and provides the required resources for the five different services considered in this study.

An IMS user, as the one depicted in Figure 1, should first register with the system prior to initiating requests to use the services. The S-CSCF checks with the database to create a profile and then authenticates the user. IMS has other components that take cater to the needs of multimedia applications such as controlling and allocating resources, and initiating SIP protocol messages to facilitate communication between servers.

### III. Modeling and Analysis

This work addresses the selection of the number of requests from different services, such that it provides the optimal QoS without neglecting services request with less delay sensitivity in the set. The work factor in the session setup delay of each service is used as the main factor in the selection. It is presumed that a buffer (located in the P-CSCF) holds the requested services for a short period of time. This buffer gives the system the chance to select services for processing based on the QoS classifications. Thus a given IMS system that offers 5 types of services (IPTV, Push to Talk, Presence, web Publish, video conference) will have different classifications of QoS for these services.

A simulation model was built to identify and analyze the total delay experienced by users when requesting IMS services, which includes the registration delay and the different services delays. The model considers one of the worst case scenarios for the network loading factor and outputs data related to the total delay experienced by the users. The loading factor worst case scenario addressed in the model considers two or more users located in different networks who request service sessions at the same time. Therefore, two sets of IMS systems located in different networks are modeled, where the first one represents the originating side (home network) and the second one represents the terminating end (visitor network). The two networks are considered to be of the WiMax and LTE network types.

To model the IMS system’s proxy server, usually located in the visitor network, the simulation model assumes 10 proxies spread over the entire coverage area. Since IMS services follow different flow diagrams, the services pose different loads on the main IMS servers. Although the registration process is not a part of the initially defined services, it is considered in the model since every user of the system needs to register with the IMS before requesting any service. The overall performance measure considered in this study, the Total Delay ($D_T$) experienced by users when requesting IMS services is defined as the summation of the Transmission Delay ($D_a$), Queuing Delay ($D_q$) and Processing Delay ($D_p$), as follows:

$$D_T = D_a + D_q + D_p$$  \hspace{1cm} (1)

Where, the Transmission Delay is defined as the delay time which occurs when sending the packets from one end to another, the Processing Delay is defined as the time it takes for the servers to process a request or query, and the Queuing Delay is defined as the time the IMS message spends in the queue of the system. Only queuing and processing delay are considered in this work. Each of the IMS servers is modeled as a network of M/M/1 queues with feedback representing the message transfer between those servers. The processing rate has exponential distribution with mean of 100. According to [2], the utility function of an IMS network can be modeled as in equation (2).

$$U = \text{Max} \left[ V - K(\rho - b\lambda^\beta) - \beta[T - K/(\rho - \lambda)] \right]$$  \hspace{1cm} (2)

Where, $V$ is the revenue obtained by serving $\lambda$ connections, $K$ is the number of servers, $a$ is the cost of each server based on its service rate $\mu$, $b$ is the sales volume, $\beta$ is a proportional constant, and $T$ is the total time for establishing the connection.

At the equilibrium state (steady-state), the Queuing Delay of the IMS session setup can be found as follows:

$$D_q = E[W_{P-CSCF}] + E[W_{P-CSCF}] + E[W_{S-CSCF}]$$  \hspace{1cm} (3)

Where, $E[W_{P-CSCF}]$, $E[W_{P-CSCF}]$, and $E[W_{S-CSCF}]$ are the expected waiting times for the IMS P-, I-, S-CSCF functions, respectively. The simulation model is used to obtain the maximum number of services requests the system can accommodate. To further improve the simulation model results, an optimization model is defined having the objective function given by equation (4) and its constraints given by equations (5, 6).

$$\text{Max } z = w_{IPTV}\lambda_{IPTV} + w_{PTT}\lambda_{PTT} + w_{VC}\lambda_{VC}$$  \hspace{1cm} (4)
\[ s.t \quad WT_{IPTV} \cdot WT_{PTT} \cdot WT_{VC} < \delta \]  
\[ WT_{P-CSCF_{orig}} + WT_{I-CSCF_{orig}} + WT_{S-CSCF_{orig}} + \]
\[ WT_{P-CSCF_{term}} + WT_{I-CSCF_{term}} + WT_{S-CSCF_{term}} < \delta \]  

Where, \( WT_{IPTV,PTT,VC} \) is the weight associated with IPTV, push to talk and video conference respectively; \( \lambda_{IPTV,PTT,VC} \) is the arrival rate of IPTV, PTT, video conference respectively; \( WT_{IPTV,PTT,VC} \) is the waiting time of the respected services; \( \delta \) is the amount of the delay that triggers request rejection, considered as 20 seconds [11]; \( WT_{P,I,S-CSCF_{orig}} \) \( WT_{P,I,S-CSCF_{term}} \) are the waiting times in IMS servers in the originating and terminating sides.

IV. EXPERIMENTAL RESULTS

Simulation experiments have been carried out for the purpose of identifying the capacity of the IMS system for each of the services being deployed. During the simulation experiments, performed using the Arena® simulation environment, ten simulation replications were run for 4 hours of simulated time each, plus an initial 30 minutes of warm-up period for each of the 10 replications in order to reach a steady-state. Using different arrival rates, the queuing delay and processing delay results obtained for video conference were recorded. Figure 2,3 demonstrates that push to talk and video conference services delays are mostly impacted by the I-CSCF and S-CSCF servers. Moreover, the simulation experiments were performed assuming that a service request is rejected if the service setup time exceeds 20 seconds [11].

An optimization of the simulation output was performed using the OptQuest® environment. Due to computing limitations, the optimization model considers only three services (IPTV, Push to Talk and Video Conference), since these services have the highest impact on the IMS due to the large amount of signaling required for establishment. The optimization model yields an optimized selection sets for these service requests. Different combinations of arriving services requests were considered here. These combinations give an insight on the type of selections associated with different arrivals, such as the case of having only push to talk and IPTV requests arriving, as represented by the (011) set. Figure 4 shows the selection sets for different combinations of the three services. For all these cases no weights were considered in the experiments. Figure 5 shows the optimized selected sets, when different combinations of weights are considered. In practice, weights can represent the revenue factor of the service providers, who may prefer a larger number of specific services to be processed to increase their revenue. Figure 5 shows the variety of options available to the service providers, such that they can implement different weight values based on their preferences. After obtaining the optimized set, an implementation was carried out in the IMS simulation model to prove that the optimized set reduces the delay. In the implementation, only the set (111) is considered, where all the three services arrives to the system. The simulation was implemented by taking a set of equal values for the different weights.
service requests and then compared with the optimized set. Figure 6 shows a clear reduction in the waiting time in the system when comparing the optimized set with the normal set having the same type and percentage of arrivals.

The implementation was carried out for the three services which were examined in the optimization. The rest of the services were considered as having a fixed arrival rate of 10 requests per second. The optimized set of services shows an important reduction in the delay influencing the IMS. In the implementation, the highest value of arrival rate in this case represents the upper bound of the services beyond which the services experience more than 20 seconds delay and the user is subjected to service rejection. Using the optimized scheme results in the reduction of the delays which means that the system can handle more service requests, and thus increase the revenue for the service providers. For the utilization function, it is assumed that the parameters of equation (2) have the following values: $a = 4$, $b = 0.0043$, $\beta = 0.01$, $T = 20$ seconds, $u = 100$ requests/second, and $k = 6$ servers. It is considered that the cost for video conference is $25$, the cost for IPTV is $1.5$ and the cost for push to talk is $1$. Table 1 shows the revenue ($V$) for every optimized set in case in which no weights were considered, while Table 2 shows the utilization for every selection set for the case in which weights were used.

Due to the increasing demands for IMS services, it has become more and more challenging for the service providers to provide the best QoS for their services. Delay is one of the factors that influences the QoS. To reduce the delay in the IMS system, an optimization technique is proposed. This paper addresses the identification of the optimum set of services request to be processed by the system. The technique lets the system select a number of services without degrading the QoS. It provides the system with an array of services for processing with the minimum effects of delay in a given time slot. Simulation and optimization environments were used to implement the IMS system with five services, and identify the upper bound of the services considered. Weighted services were implemented in the optimization model, such that service providers could tune their service preference based on the revenue. A comparison between the normal selection and the optimized use of the IMS simulation model is also performed. The model also provides an insight on the utility function for each selection. This proposed model can be implemented in IMS (P-CSCF) by introducing a table with all possible arrival rate combinations, in which case the system can follow the selections provided by the table and provide the optimum QoS.

### Table 1. The revenue in the case of all weights equal to one.

<table>
<thead>
<tr>
<th>VC</th>
<th>IPTV</th>
<th>PTT</th>
<th>Utilization ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>19.5</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>11.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12.5</td>
</tr>
</tbody>
</table>

### Table 2. The revenue in the case of different weights.

<table>
<thead>
<tr>
<th>VC</th>
<th>IPTV</th>
<th>PTT</th>
<th>Utilization ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>19.5</td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>11.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>0.5</td>
<td>13.5</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
<td>11.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.2</td>
<td>0.3</td>
<td>13</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>13</td>
</tr>
</tbody>
</table>

### V. Conclusion

Due to the increasing demands for IMS services, it has become more and more challenging for the service providers...

### REFERENCES


