Multimedia Messaging Service: System Description and Performance Analysis

Majid Ghaderi and Srinivasan Keshav
School of Computer Science
University of Waterloo, Waterloo, ON N2L 3G1, Canada
{mghaderi,keshav}@uwaterloo.ca

***Abstract*— Following the success of short messaging service (SMS), multimedia messaging service (MMS) is emerging as a natural but revolutionary successor to short messaging. MMS allows personalized multimedia messages containing content such as images, audio, text and video to be created and transferred between MMS-capable phones and other devices. By using IP and its associated protocols, MMS is able to interwork with other messaging systems such as Internet messaging services. An important feature of MMS is the guaranteed delivery of messages via a store-and-forward mechanism which temporarily stores messages in the network until successfully delivered. Unlike SMS, multimedia messaging service does not mandate any maximum size for a multimedia message. This enhanced flexibility of MMS requires a careful design of the network in order to avoid excessive message delays and losses. This paper develops a mathematical model for evaluating the performance of an MMS system. Using the model, closed-form expressions for major performance parameters such as message loss, message delay and expiry probability have been derived. Furthermore, a simple algorithm is presented to find the optimal temporary storage size for a given set of system parameters. The accuracy of the presented analysis is evaluated through simulations which shows a close agreement between analytic and simulation results.**

I. Introduction

Short Message Service (SMS) [1] is a globally accepted wireless service that allows mobile subscribers to send and receive alphanumeric messages of up to 140 bytes in length. A distinguishing characteristic of the service is the guaranteed delivery of short messages by the network via a store-and-forward mechanism. Temporary failures are identified, and the short message is stored in the network until the destination becomes available. Despite the enormous popularity of SMS, the content that can be transmitted is limited to short text messages, ring tones, and small graphics.

Due to recent developments in wireless communications, building more flexible and more capable messaging services has become the reality. The Multimedia Messaging Service (MMS), a revolutionary successor to SMS [2], has emerged as the result of research efforts primarily by the Third Generation Partnership Project (3GPP) [3] and Open Mobile Alliance (OMA) [4]. MMS will extend the revenue opportunities for network operators and manufacturers, and lead to lower costs for customers.

To the end user MMS is very similar to SMS as it provides automatic and fast delivery of multimedia messages (MMs) between capable phones and other devices. However, there are important technical differences between SMS and MMS. MMS supports richer content types such as text, graphics, music, video clips and more [5]. The MMS specifications do not mandate any specific content format for MMs. Instead the MMs are encapsulated in a standard way, so that the recipient can identify those content formats it does not support and handle them properly. The standard does not specify a maximum size for an MM either in order to avoid the SMS message size limitation.

With the increasing size and volume of messages being transmitted, the fast and robust delivery of messages becomes a challenging problem. As we will see later, the critical factor affecting the MMS system performance in terms of message delay and loss is the temporary storage of messages at server nodes. Therefore, an important problem in designing an MMS system is the proper sizing of the temporary storage in order to achieve a desirable performance. This requires the modeling of the end-to-end path which will be shown to reduce to modeling the behavior of a single MMS server. However, a simple application of M/M/1 model in this context is not appropriate due to the limited patient time of queued messages.

Unfortunately, the literature on the analysis of messaging systems is quite rare. We are able only to mention the work by Haung [6] on the analysis of optimal buffer size for SMS. To the best of our knowledge, the present paper is the first to address the message delay and loss probability of a multimedia messaging system. Indeed, the analytical method presented in this paper is applicable to a general class of queueing systems with reneging customers and that includes SMS as well. Using the analysis presented in this paper, MMS service providers can choose the right system settings, e.g. storage size, to achieve the best performance, i.e. message loss and delay, for a multimedia messaging system. The main contributions of the present paper are:

1) description of MMS and comparison with SMTP,
2) a simple approach to compute message delay which enables the derivation of closed-form expressions for both virtual and actual message delay, and,
3) a simulation study of the system sensitivity to different parameters such as load, message expiry and service rate.

The rest of the paper is organized as follows. In Section II an overview of the multimedia messaging service is presented

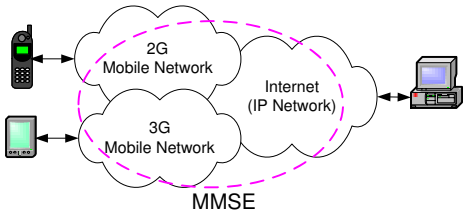

Fig. 1. General MMS architecture integrating different networks.

which covers the network architecture, operations and protocols involved in an MMS. Section IV is dedicated to the modeling and analysis of a multimedia messaging system. To investigate the accuracy of the presented analysis, simulation and analytical results are presented in Section V. Finally, Section VI reviews some related works and Section VII concludes the paper.

II. The Multimedia Messaging Service

A. Network Architecture

Fig. 1 shows a generalized view of the MMS architecture [7]. The architecture consists of different networks and integrates existing messaging systems within these networks. Mobile stations operate with the Multimedia Messaging Service Environment (MMSE) which provides all the necessary service elements, e.g. delivery, storage and notification functionality under the control of a single administration. Connectivity between these different networks is provided by the Internet Protocol (IP) and its associated set of messaging protocols. This approach enables messaging in 2G and 3G wireless networks to be compatible with messaging systems found on the Internet, i.e. SMTP-based email.

Fig. 2 shows the MMS network architecture consisting of all the elements required for providing a complete MMS to a user [7]. At the heart of this architecture is the MMS Relay/Server (MMS-RS) which is responsible for storage and reliable delivery of messages between possibly different messaging systems, akin to an SMTP mail transfer agent (MTA). The MMS-RS temporarily stores messages until they are successfully delivered. The MMS-RS may be a single logical element or may be separated into MMS Relay and MMS Server elements.

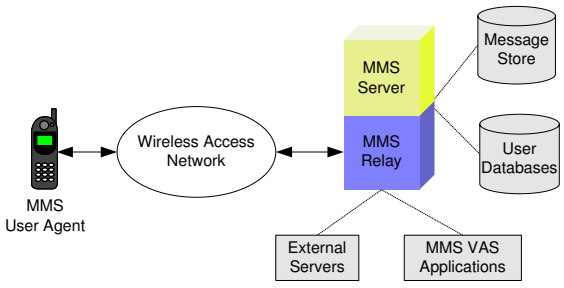

Fig. 2. MMS network architecture.

The MMS User Agent (MMS-UA) exists within a mobile station. This application akin to an email reader application lets user view, compose and handle (e.g. submit, receive, delete) multimedia messages. The retrieval of MMs from MMS-RS can be either automatic or manual. In automatic mode, an MM is retrieved without user involvement. In manual mode, the user is informed by a notification message and is allowed to make a decision whether to download the MM or not.

The MMS-RS has access to several User Databases, e.g. user profile database, subscription database and home location register (HLR). An optional feature of MMS is the support of persistent network-based storage called an MMBox. The MMS-RS has access to an MMBox in order to store, retrieve or delete messages. Depending on the operator configuration, each subscriber may configure his MMBox to automatically store incoming and submitted messages, or, manually request that specific messages be persistently stored.

The MMS VAS Applications provide value added services to the MMS users. Several External Servers may be included within or connected to an MMSE, e.g. E-Mail server, SMS server and Fax server. The MMS-RS is responsible for providing convergence functionality between External Servers and MMS-UAs. Thus mobile phone users can use an MMS-RS to access email, multimedia attachments, SMS or faxes.

III. MMS Operation

A. Transmission of Multimedia Messages

1) Sending Messages: A user sends a message by having its MMS-UA submit the message to its home MMS-RS. A message must have the address of the recipient and a MIME content type. Several other parameters may be set for a message including the desired time of expiry for the message and the message priority. Upon reception of a message from an originator MMS-UA, the originator MMS-RS assigns a message identification to the message and sends this message identification to the originator MMS-UA. If an MMBox is supported and enabled for the sender, MMS-RS automatically stores a copy of the message into the sender MMBox, then routes the message towards the recipients.

2) Receiving Messages: Upon reception of a message, the recipient MMS-RS verifies the recipient profile and generates a notification to the recipient MMS-UA. It also stores the message at least until one of the following events happens:

- the associated time of expiry is reached,
- the message is delivered,
- the recipient MMS-UA requests the message to be forwarded,
- the message is rejected.

If it has been requested, MMS-RS will also store the message in an MMBox, if the MMBox is supported and enabled.

When the recipient MMS-UA receives a notification, it uses the message reference in the notification to reject or retrieve the message, either immediately or at a later time, either manually or automatically, as determined by the operator configuration and user profile. If MMBoxes are supported, the MMS-UA may request retrieval of a message from the

user MMBox, based on a message reference received from a previous MMBox operation.

3) Message Adaptation: Within a request for delivery of a message, the recipient MMS-UA can indicate its capabilities, e.g. a list of supported media types and media formats, for the recipient MMS-RS. On getting a delivery request, the recipient MMS-RS uses the information about the capabilities of the recipient MMS-UA to prepare the message for delivery to the recipient MMS-UA. This preparation may involve the deletion or adaptation of unsupported media types and media formats [8]. Depending on the configuration and the capability of the recipient MMS-UA and the recipient MMS-RS, the MM-UA may use streaming for the retrieval of message contents.

4) Delivery Reports: Unlike SMPT, if a delivery report has been requested by the originator MMS-UA and if the recipient MMS-UA did not request a delivery report not to be generated, the recipient MMS-RS generates a delivery report and delivers the delivery report to the originator MMS-RS. The recipient MMS-RS stores delivery reports in the network until the originator MMS-RS becomes reachable or until the delivery report expires. A delivery report contains information such as:

- The identification of the original message for which the delivery report has been generated,
- Status information on how the message was handled (e.g. expired, rejected, delivered, forwarded or indeterminate),
- A time stamp showing when the message was handled.

The originator MMS-RS, in turns, stores delivery reports until the originator MMS-UA becomes reachable or until the delivery report expires.

B. MMS Interworking

Fig. 3 shows the message routing mechanism in MMS. From an end-to-end perspective, multimedia messages are always routed via both the sender and the recipient's home MMS-RSs. The MMS-RS provides access to the MMSE via the MM1 interface which can be implemented using WAP [9] or applications conforming to MExE [10] (e.g. Java and TCP/IP) as indicated by the 3GPP specification [7]. Whenever an MMS-RS receives a message whose recipient belongs to another MMSE, the originator MMS-RS must forward the message to the recipients MMS-RS. Reference point MM4 between MMS-RSs belonging to different MMSEs is used to transfer messages between them. Interworking between MMS-RSs is based on SMTP. Resolving the destination address to find the recipient MMS-RS IP address is the responsibility of the originator MMS-RS.

The MMS-RS is also connected to External Servers such as email servers via an IP network. This connectivity works in both directions to perform three operations [4]:

1) Sending messages to email servers: After converting the message to standard Internet MIME format, the MMS-RS submits the message to the recipient using the SMTP protocol. The MMS specific header fields will be converted into appropriate headers by appending an 'X-Mms-' to the header name. This permits MMS-aware systems to understand

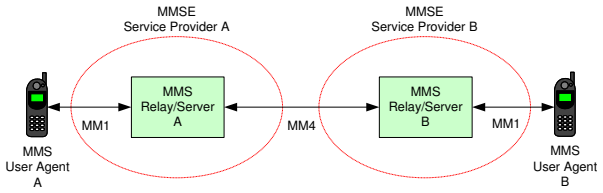

Fig. 3. Interworking of different MMSEs.

the fields while not being problematic for non-MMS-aware systems.

2) Receiving messages from email servers: Received messages will be similarly converted. The MIME part of the message is converted to the MMS format. Similarly, any headers found with a prefix of 'X-Mms-' can be converted back to the associated MMS header.

3) Retrieving messages from email servers: This is normally done through the use of the POP or IMAP protocols. Such retrievals are performed by the MMS-RS, which will then convert the data into an appropriate MMS format.

C. MMS Addressing

MMS supports the use of email address or MSISDN (E.164) or both to address the recipient of a message. Since MMS interworking across different networks (MMSEs) is provided based on SMTP, each MMSE is assigned a unique DNS domain name.

1) Addressing at the MM1 Interface: The message addressing on MM1 consists of three addresses: the address of the originator MMS-RS, the address of the recipient and the address of the sender. The address of the originator MMS-RS is the Uniform Resource Identifier (URI) of the MMS-RS given by the service provider. The sender address could be either a user address or a terminal address (e.g. terminal IP addresses). The recipient address can be a user address, a terminal address, or a short code. The user address can be either an MSISDN or email address.

2) Addressing at the MM4 Interface: For those recipients that appear in a message and belong to an external MMSE, the originator MMS-RS has to send the message to each of the recipients MMS-RS. The MMS-RS has to resolve the recipient MMS-RS domain name to an IP address based on the recipient address. In case of MSISDN addressing, the originator MMS-RS should translate the address to an email address using DNS-ENUM protocol [11].

IV. Multimedia Messaging Service Analysis

In this section, we analyze the message handling performance of an MMS-RS in isolation as the central element in a multimedia messaging system. We will later discuss end-to-end system performance. The performance parameters of interest are the queueing delay induced by the temporary storage of messages in the MMS-RS, and message loss probability due to storage overflow and message expiration while messages are temporarily waiting in storage. Indeed, the messaging system performance is dramatically affected by the

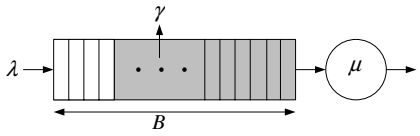

Fig. 4. A queueing model of the MMS-RS.

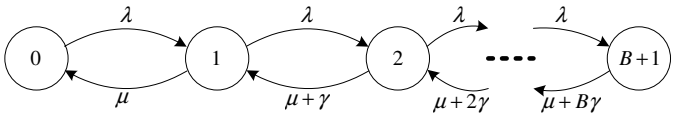

Fig. 5. A Markov chain representation of the MMS-RS.

size of the temporary storage. Having a small storage avoids long message delay. However, if the storage is too small then message loss due to storage overflow increases and the MMS-RS utilization deteriorates. Thus, it is crucial to determine an optimal storage size to achieve maximum utilization and minimum message loss probability, and thus prevent messages from being excessively delayed, which is especially needed for multimedia messages.

A conceptual model of the MMS-RS is shown in Fig. 4, where the MMS-RS is modeled as a queueing system. The assumptions and parameters involved in this model are stated below:

1) The new message arrival into the MMS-RS is Poisson distributed with rate λ.
2) The timeout period of the temporary stored messages in the MMS-RS is assumed to be exponentially distributed with mean $1/\gamma$. A stored message is removed from the temporary storage if it can not be transmitted within its timeout period. A message expiry that happens during the actual transmission of the message is ignored by the MMS-RS.
3) The transmission time of a message is assumed to be exponentially distributed with mean $1/\mu$. Stored messages are served according to FIFO scheduling.
4) A finite storage with capacity $B > 0$ messages is provided in the MMS-RS for temporary message storage.

In the real world, the message expiry and transmission times may not be exponential but exponential distributions allow mean value analysis, which indicates the performance trend of the system. The goal of the present paper is to develop a tractable and yet reasonably accurate model rather than trying to apply exact but intractable models that do not necessarily capture all the impact of complex system interactions. We believe our model is rich and analyzable enough to provide information that is practically important for MMS service providers.

Performance analysis of the MMS-RS can be accomplished by describing the system as a Markov chain corresponding to the system dynamics. Fig. 5 shows the transitions among different system states where state i indicates that there are i messages in the system (either in temporary storage waiting for delivery or being transmitted). Let p_i denote the steady-state probability of being in state i. Using balance equations, it is clear that

$$p_i = \prod_{j=1}^{i} \left(\frac{\lambda}{\mu + (j-1)\gamma} \right) p_0, \qquad 1 \le i \le B+1 \quad (1)$$

where p_0 can be found using the normalizing condition $\sum_{i=0}^{B+1} p_i = 1$. In this model, the service rate at state i is $\mu + (i-1)\gamma$.

Throughout this paper we will use the steady-state probability distribution of the system state as seen by an arriving message, i.e. system state at arrival epochs. Let a_i denote the probability of having i messages in the system just before a message arrives to the system and gets accepted by the MMS-RS. Such a message never sees the system in state $(B+1)$, i.e. blocking state. It can be shown that a_i is given by (for example refer to [12])

$$a_i = \frac{p_i}{1 - p_{B+1}}, \qquad 0 \le i \le B\,. \quad (2)$$

A. Message Delay Analysis

Define the message delay as the time between the acceptance of a message in the MMS-RS and the time its transmission starts. We are interested in finding the *actual message delay* denoted by W, and defined as the message delay experienced by a message that is successfully transmitted, i.e. did not expired before transmission. In the following discussion, we use notations $F_z(t)$ and $f_z(t)$ to denote the distribution and density function of a random variable z, where

$$f_z(t) = \frac{d}{dt} F_z(t)\,. \quad (3)$$

Let random variable T denote the *virtual message delay* defined as the message delay experienced by a message which has infinite expiry time. Such a message is referred to as a *virtual message* and remains in the system until it is transmitted. Using the conditional probability formulation, the relation between the actual message delay, W, and the virtual message delay, T, can be expressed as

$$F_W(t) = \mathbb{P}\{W \le t\} = \mathbb{P}\{T \le t \,|\, T \le X\}\,, \quad (4)$$

where X is a random variable denoting the expiry time of a typical message. We assume that message expiry times are independent and exponentially distributed with the parameter γ, hence

$$f_X(t) = \gamma e^{-\gamma t}\,. \quad (5)$$

Assume that a virtual message arrives to the system when there are i messages in the system, i.e. system is in state i. Let m_i denote this message. If $0 \le i \le B$ then m_i is accepted and the system state will change to $i+1$. If $i = 0$ then m_i will be immediately served, otherwise it must temporarily wait in the storage for i message departures (either transmission or expiry). Suppose we temporarily view the system as consisting of those i messages which are ahead of m_i. Let t_j denote the time required for the message population to decrease from j

to $j-1$ ($1 \leq j \leq B$). Then, t_j is exponentially distributed with rate parameter $\mu + (j-1)\gamma$, i.e.

$$f_{t_j}(t) = [\mu + (j-1)\gamma] e^{[\mu+(j-1)\gamma]t}. \tag{6}$$

Let T_i denote the virtual message delay of m_i, i.e. the amount of time m_i must wait before its transmission starts given that m_i is infinitely patient. Then

$$\begin{aligned} T_i &= t_1 + \cdots + t_i, \\ &= T_{i-1} + t_i, \qquad \text{for } i \geq 1. \end{aligned} \tag{7}$$

According to our definition of virtual message delay, it is clear that $T_0 = 0$. Therefore, we have

$$f_{T_i}(t) = \begin{cases} \mu e^{-\mu t}, & i = 1 \\ f_{T_{i-1}} * f_{t_i}(t), & i > 1 \end{cases} \tag{8}$$

where $f_{T_{i-1}} * f_{t_i}(t)$ is the convolution of $f_{T_{i-1}}(t)$ and $f_{t_i}(t)$ which can be expressed as

$$f_{T_{i-1}} * f_{t_i}(t) = \int_0^t f_{T_{i-1}}(x) f_{t_i}(t-x)\, dx. \tag{9}$$

By solving the recursive definition (8) using (6) and (9), we find that

$$f_{T_i}(t) = \mu \prod_{j=1}^{i-1} \left(\frac{\mu + j\gamma}{j\gamma}\right) \left(1 - e^{-\gamma t}\right)^{i-1} e^{-\mu t}. \tag{10}$$

Using (7), the mean virtual message delay, $\mathrm{E}[T_i]$, is given by

$$\mathrm{E}[T_i] = \sum_{j=1}^{i} \mathrm{E}[t_j] = \sum_{j=1}^{i} \frac{1}{\mu + (j-1)\gamma}. \tag{11}$$

Having obtained the virtual message delay distribution, we now turn our attention to actual message delay. Let m_i denote a message that is accepted in the system in state i. Let β_i denote the probability that m_i does not expire before it is being transmitted, i.e.

$$\beta_i = \mathbb{P}\{T_i \leq X\}. \tag{12}$$

Let β_{ij} denote the probability that m_i does not expire during t_j, the interval of time required to drive the message population from j to $j-1$. Then, β_{ij} can be obtained as follows

$$\beta_{ij} = \mathbb{P}\{t_j \leq X\} = \frac{\mu + (j-1)\gamma}{\mu + j\gamma}. \tag{13}$$

Therefore, β_i is given by

$$\beta_i = \prod_{j=1}^{i} \beta_{ij} = \frac{\mu}{\mu + i\gamma}. \tag{14}$$

Using the conditional probability given in (4), $F_{W_i}(t)$ is expressed as

$$\begin{aligned} F_{W_i}(t) &= \mathbb{P}\{T_i \leq t \,|\, T_i \leq X\} \\ &= \frac{\mathbb{P}\{T_i \leq t,\, T_i \leq X\}}{\mathbb{P}\{T_i \leq X\}}. \end{aligned} \tag{15}$$

From (12), we get

$$F_{W_i}(t) = \frac{1}{\beta_i} \int_0^t f_{T_i}(x)(1 - F_X(x))\, dx. \tag{16}$$

Equivalently,

$$\begin{aligned} f_{W_i}(t) &= \left(\frac{\mu + i\gamma}{\mu}\right) f_{T_i}(t)(1 - F_X(t)) \\ &= (\mu + i\gamma) \prod_{j=1}^{i-1} \left(\frac{\mu + j\gamma}{j\gamma}\right) \left(1 - e^{-\gamma t}\right)^{i-1} e^{-(\mu+\gamma)t}, \end{aligned} \tag{17}$$

where we substitute $f_{T_i}(t)$ from (10). Using (17), the mean actual message delay, $\mathrm{E}[W_i]$, is given by

$$\begin{aligned} \mathrm{E}[W_i] &= \int_0^\infty t f_{W_i}(t)\, dt \\ &= (i\gamma) \left[\prod_{j=1}^{i} \left(\frac{\mu + j\gamma}{j\gamma}\right)\right] \left[\sum_{j=1}^{i} \binom{i-1}{j-1} \frac{(-1)^{j-1}}{(\mu + j\gamma)^2}\right]. \end{aligned} \tag{18}$$

Furthermore, the steady-state performance parameters can be computed with respect to the steady-state probability distributions given by (2). In particular, the average steady-state actual message delay, W, is expressed as

$$W = \sum_{i=1}^{B} q_i W_i, \tag{19}$$

where q_i is the steady-state probability that a non-reneging message finds i messages in the system upon arrival. Please refer to the Appendix for the derivation of q_i.

Finally, our results can be represented in terms of special functions by noting that

$$\prod_{j=1}^{i} (\mu + j\gamma) = \frac{\Gamma(1 + \mu/\gamma + i)}{\Gamma(1 + \mu/\gamma)} \gamma^i, \tag{20}$$

$$\sum_{j=1}^{i} \frac{1}{(\mu + j\gamma)} = \frac{1}{\gamma}\Psi(1 + \mu/\gamma + i) - \frac{1}{\gamma}\Psi(1 + \mu/\gamma), \tag{21}$$

where, Γ and Ψ denote the gamma and digamma function [13] respectively.

B. Message Loss Probability

A message is lost if upon its arrival to the MMS-RS, the temporary storage is full, or, although accepted and waiting in the storage, fails to be transmitted within its timeout period and so is removed from the storage. Therefore, the message loss probability L is given by

$$L = p_{B+1} + (1 - p_{B+1})\alpha, \tag{22}$$

where, α denotes the probability that a typical message accepted in the system will expire before being transmitted and is given by

$$\alpha = \sum_{i=0}^{B} (1 - \beta_i) a_i = \frac{N}{1 - p_{B+1}} \left(\frac{\gamma}{\lambda}\right), \tag{23}$$

where, N is the average number of messages in storage.

C. Average Number of Queued Messages

We are interested to find a closed-form expression for the average number of queued messages in the temporary storage. Let β denote the probability that a typical accepted message will be transmitted. Then, β is expressed as

$$\beta = \sum_{i=0}^{B} a_i \beta_i = \frac{\lambda}{\mu} \left(\frac{1 - p_0}{1 - p_{B+1}} \right) . \tag{24}$$

Given that $\alpha + \beta = 1$, it is obtained that

$$N = \frac{1}{\gamma} \Big[\lambda(1 - p_{B+1}) - \mu(1 - p_0) \Big] . \tag{25}$$

D. End-to-End Performance

Message queueing may happen in three different locations before a message reaches it destination: 1) inside the originator MMS-UA, 2) in the originator MMS-RS, or 3) in the recipient MMS-RS. In previous subsections, the message delay and loss probability were analyzed for the recipient MMS-RS. In a typical interworking scenario, two MMS-RSs, namely the originator and the recipient, are involved. Given that for the originator MMS-RS, the incoming wireless link is the bottleneck not the outgoing wireline link, it can be assumed that the message loss and delay due to temporary storage in the originator MMS-RS are effectively zero. Considering the user behavior in generating multimedia messages, no queueing is expected to happen inside the user equipment, i.e. MMS-UA, either. Therefore, message loss in MMS-UA is also zero assuming that message expiry times are comparable to message transmission times (otherwise all the messages expire before entering the MMS system!).

Consequently, the end-to-end message loss probability is determined by the loss probability at the recipient MMS-RS. However, for computing the end-to-end message delay, the message transmission time in the originator MMS-UA must be considered. The impact of such a delay may be well captured by the fact that we assumed message expiry time is a random variable (with exponential distribution) not a deterministic value. Therefore, to have more accurate analysis, the impact of message transmission time in the originator MMS-UA must be included in the expiry time characterization and must be added to the end-to-end message delay as well.

V. Simulation Results

An event-driven simulation was developed to verify the accuracy of the analysis. The simulation considered a single MMS-RS in isolation. All the simulation parameters are relative to the message transmission rate μ. For the basic set of simulations, the average message size is considered to be 8 KBytes and the output link is a GPRS carrier transmitting at 10 Kbps. Simulation results are obtained by averaging over 8 independent samples. Only average points are plotted since the 95% confidence intervals were very close to the average value, and hence are not shown for the sake of having clear plots. The simulation length was long enough to avoid rare event problem for the simulation scenarios with small values of γ.

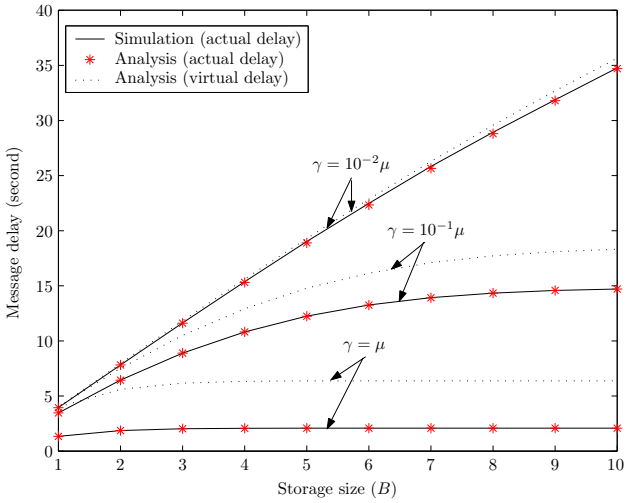

Fig. 6. Message delay for $\lambda = \mu$.

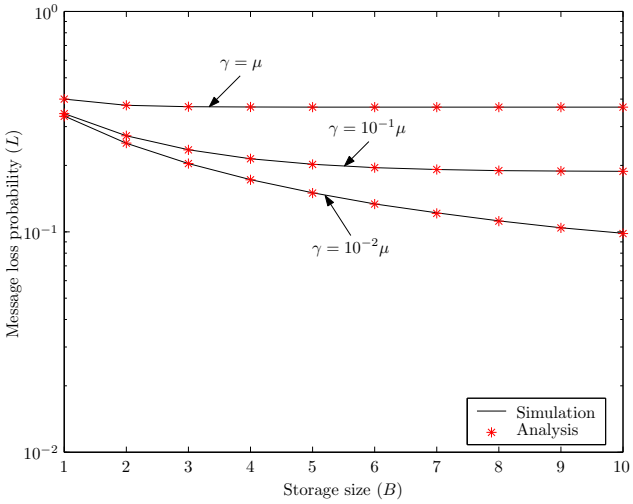

Fig. 7. Message loss probability for $\lambda = \mu$.

Overall, in all the cases simulated, analytic and simulation results match with almost no discrepancy.

1) Effect of the expiry time: Figs. 6 and 7 show the average message delay and message loss probability for different message expiry rates when the offered load is set to 1, i.e. $\lambda = \mu$. As expected, message loss increases by increasing the expiry rate but message delay decreases by increasing the expiry rate. In addition to the actual message delay, the virtual message delay computed using (11) is depicted in Fig. 6 as well. As the message expiry rate increases the difference between actual and virtual message delay increases too.

It is observed from Fig. 7 that as the expiry rate increases the message loss probability increases despite the fact that blocking probability has decreased. Referring to (22), the message loss probability depends on both blocking and expiry probability. Although, the blocking probability decreases by increasing the expiry rate, this decrease is not proportional to the increase in expiry probability. More formally, by substi-

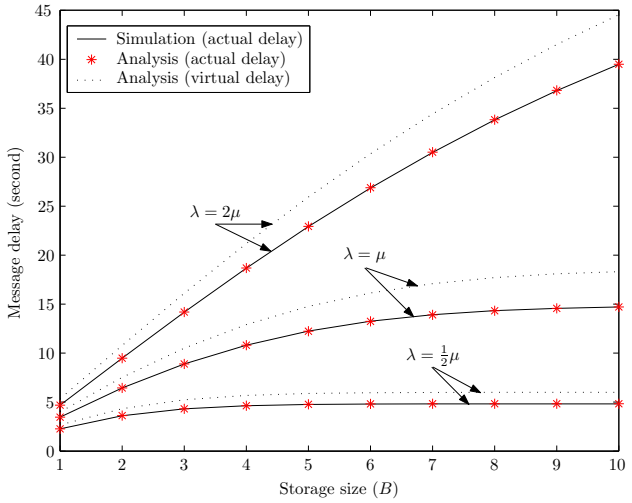

Fig. 8. Message delay for $\gamma = 10^{-1}\mu$.

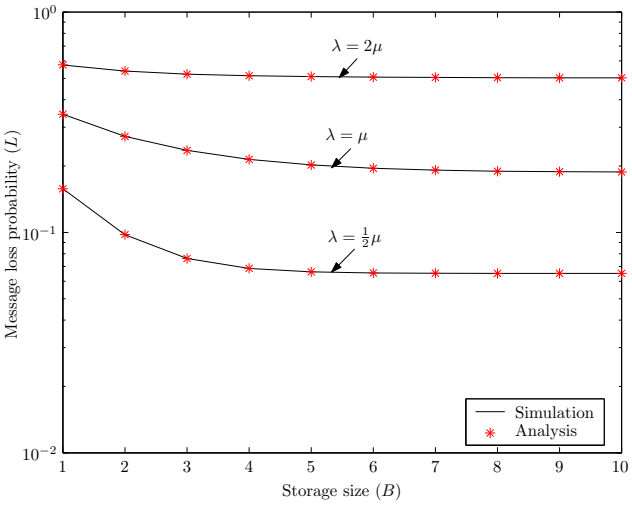

Fig. 9. Message loss probability for $\gamma = 10^{-1}\mu$.

tuting (25) in (23) and using (22), it can be shown that

$$L = p_{B+1} + (1 - p_{B+1}) \left[1 - \left(\frac{\mu}{\lambda}\right)\left(\frac{1 - p_0}{1 - p_{B+1}}\right)\right], \\ = 1 - \left(\frac{\mu}{\lambda}\right)(1 - p_0)\,. \tag{26}$$

In this case, since $\lambda = \mu$, it is obtained that $L = p_0$. Increasing the message expiry rate results in larger p_0 which consequently means larger message loss probability.

2) Effect of the offered load: To investigate the effect of the offered load on the system performance, we ran the simulations for the case of having $\gamma = 10^{-1}\mu$ with different arrival rates. Three arrival rates $\lambda = 2\mu$, $\lambda = \mu$ and $\lambda = \frac{1}{2}\mu$ corresponding to loads 2, 1 and $\frac{1}{2}$, respectively, were simulated. All other system parameters are the same as before. Figs. 8 and 9 show the message delay and message loss probability with respect to the storage size.

As expected, both message delay and loss probability grow by increasing the offered load. As shown in the figures, the loss probability is rather insensitive to the storage size specially for high loss rates (e.g. $\lambda = 2\mu$).

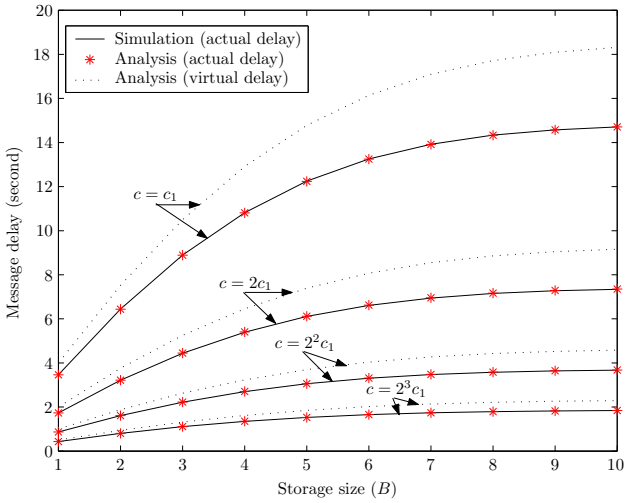

Fig. 10. Message delay for $\lambda = \mu$ and $\gamma = 10^{-1}\mu$.

3) Effect of the service time: A GPRS carrier with transmission rate 10 Kbps is perhaps too slow to be a good candidate for building a practical MMS system. To investigate the impact of service rate, equivalently the message service time, on the system performance, we did the simulations for different service rates. Fig. 10 show the message delay for four different service rates in terms of basic service rate $c_1 = 10$ Kbps with respect to the storage size. In these simulations, $\lambda = \mu$ and $\gamma = 10^{-1}\mu$. As the service rate increases, the message transmission time ($1/\mu$) decreases and consequently, message delay decreases too. It can be seen from the figure that as the service rate increases, the discrepancy between virtual and actual message delay decreases which can be also verified from the analysis. Changing the service rate does not affect the loss probability as the loss probability is a function of quantities λ/μ and γ/μ which are fixed in this case.

4) Optimal storage size: As shown in Figs. 7 and 9, message loss probability decreases by increasing the storage size up to a certain threshold. Increasing the storage size beyond that threshold will not significantly change the loss probability. However, if the storage size exceeds the threshold, message delay will continue to increase as depicted in Figs. 6, 8 and 10. This threshold is referred to as the *optimal storage size*. An iterative approach similar to the one proposed in [6] can be used to find the optimal buffer size for a given system configuration. The iterative algorithm follows the pseudo-code represented in Fig. 11, where ϵ is the desired precision for the convergence of message loss probability.

VI. Related Work

In queueing theory terminology, this type of system is usually referred to as queue with impatient or reneging customers. The literature on queueing systems with reneging is moderate. This includes classical works such as [14]–[16] and recent works such as [17]–[20]. Among them, Barrer [14]

$B \leftarrow 1$;
$L \leftarrow 0$;
repeat
  $L' \leftarrow L$;
  $L \leftarrow$ Message loss using (22);
  $B \leftarrow B + 1$;
until $(|L - L'| > \epsilon)$;

Fig. 11. Iterative algorithm for computing B.

obtained the reneging probability for deterministic patient time customers. Baccelli and Hebuterne [15] considered a queue with general patient time distribution. However, their analysis involves inverse Laplace transformations which does not provide a closed-form expression for the performance parameters. Queueing systems with state-dependent arrival/service rate have been studied in [17], [18]. References [12], [16] mostly focused on steady-state probability distributions of the queue length with reneging customers. To avoid complexity and numerical instability associated with exact analytical techniques, approximate solutions have been also investigated [19], [20].

Although queueing systems with impatient customers have been studied by several researchers, our contribution is a simpler convolution-based technique to find closed-form expressions for both the virtual and actual message delay for the Markovian system depicted in Fig. 5. This technique avoids complicated transformations and differential equations applied in previous papers by formulating the virtual message delay as a recursive convolution tailored to the MMS-RS model.

VII. Conclusion

This paper studied the architecture, operation and performance of a multimedia messaging system. Various components involved in the architecture and their functionalities were described. Then, a mathematical model developed to study the performance of the MMS-RS as the central component of a multimedia messaging system. The presented analysis models an MMS-RS as a finite capacity queueing system with reneging customers (multimedia messages). Using the Markovian behavior of the system, closed-form expressions describing the message delay distribution and message loss probability were presented. The analytical results were compared with those obtained from the simulation which confirmed the accuracy of the analysis.

As the future work, we would like to investigate the effect of having different priorities for different messages. Although the current MMS specification has provided a mechanism to specify the message priority, the network itself does not take any action with respect to the priorities. The sole purpose of these priorities is to inform the end user about the importance of the received messages. Another important aspect of the system that requires more investigation is the scaling with respect to load and service rate. Experiments show that scaling is not trivial and basically just increasing the storage may not be the right solution.

Appendix

Let q_i denote the steady-state probability that a non-reneging message finds i messages in the system upon arrival. A non-reneging message is a message that will receive service before expiration. Using the Bayes's theorem, q_i can be determined as follows:

$$\begin{aligned} q_i(t) &= \lim_{\delta \to 0} \mathbb{P}\{N(t) = i \,|\, \text{a NRA at } t+\delta\} \\ &= \lim_{\delta \to 0} \frac{\mathbb{P}\{N(t)=i\}\,\mathbb{P}\{\text{a NRA at } t+\delta \,|\, N(t)=i\}}{\sum_{j=0}^{B} \mathbb{P}\{N(t)=j\}\,\mathbb{P}\{\text{a NRA at } t+\delta \,|\, N(t)=j\}} \\ &= \lim_{\delta \to 0} \frac{p_i(t)\beta_i\lambda\delta}{\sum_{j=0}^{B} p_j(t)\beta_j\lambda\delta} = \frac{p_i(t)\beta_i}{\sum_{j=0}^{B} p_j(t)\beta_j}, \end{aligned} \tag{27}$$

where NRA stands for *non-reneging arrival*. Therefore,

$$q_i = \lim_{t\to\infty} q_i(t) = \frac{p_i\beta_i}{\sum_{j=0}^{B} p_j\beta_j} = \frac{\lambda/\mu}{1-p_0}\beta_i p_i\,. \tag{28}$$

References

[1] 3GPP TS 23.040, "Technical realization of the short message service (SMS); release 5," v5.2.0, 2001.
[2] Nokia, "MMS technology tutorial." [Online]. Available: http://www.nokia.com/support/tutorials/MMS/en/mms.html
[3] 3GPP TS 22.140, "Multimedia messaging service (MMS); stage 1; release 6," v6.6.0, June 2004.
[4] OMA, "Multimedia messaging service; architecture overview," v1.2, Dec. 2003.
[5] 3GPP TS 26.140, "Multimedia messaging service (MMS); media formats and codecs; release 6," v6.0.0, Sept. 2004.
[6] Y.-R. Haung, "Determining the optimal buffer size for short message transfer in a heterogeneous GPRS/UMTS network," *IEEE Trans. Veh. Technol.*, vol. 52, no. 1, pp. 216–225, Jan. 2003.
[7] 3GPP TS 23.140, "Multimedia messaging service (MMS); functional description; stage 2; release 6," v6.6.0, June 2004.
[8] S. Coulombe and G. Grassel, "Multimedia adaptation for the multimedia messaging service," *IEEE Commun. Mag.*, vol. 42, no. 7, pp. 120–126, July 2004.
[9] WAP Forum, "Wireless application protocol architecture specification; release 2.0," July 2001.
[10] 3GPP TS 23.057, "Mobile execution environment (MExE); functional description; stage 2; release 6," v6.2.0, 2003.
[11] P. Faltstrom, "E.164 number and DNS," IETF, RFC 2822, Sept. 2000.
[12] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 3rd ed. New York, USA: John Wiley & Sons, Inc., 1998.
[13] MathWorld. [Online]. Available: http://mathworld.wolfram.com
[14] D. Y. Barrer, "Queueing with impatient customers and ordered service," *Operations Research*, vol. 5, no. 5, pp. 650–656, Oct. 1957.
[15] F. Baccelli and G. Hebuterne, "On queues with impatient customers," in *Performance'81*, F. J. Kylstra, Ed. Oxford, UK: North-Holland Publishing Company, 1981, pp. 159–179.
[16] B. V. Gnedenko and I. N. Kovalenko, *Introduction to Queueing Theory*, 2nd ed. Boston, USA: Birkhauser, 1989.
[17] A. Movaghar, "On queueing with customer impatience until the beginning of service," *Queueing Systems*, vol. 7, no. 3, pp. 15–23, June 1998.
[18] J. Bae, S. Kim, and E. Y. Lee, "The virtual waiting time of the M/G/1 queue with impatient customers," *Queueing Systems*, vol. 38, no. 4, pp. 485–494, Aug. 2001.
[19] A. Brandt and M. Brandt, "Asymptotic results and a Markovian approximation for the M(n)/M(n)/s+GI system," *Queueing Systems*, vol. 41, no. 1-2, pp. 73–94, June 2002.
[20] A. R. Ward and P. W. Glynn, "A diffusion approximation for a Markovian queue with reneging," *Queueing Systems*, vol. 43, no. 1-2, pp. 103–128, June 2003.