PERFORMANCE EVALUATION OF AN SMS-BASED EMAIL AND VOICEMAIL NOTIFICATION ARCHITECTURE

K. Koumpis, S. Cvetkovic and G. Peersman

Department of Computer Science
The University of Sheffield
Regent Court, 211 Portobello St.
S1 4DP, Sheffield UK
{K.Koumpis, S.Cvetkovic, G.Peersman}@dcs.shef.ac.uk

ABSTRACT

In this paper real time notification services based on low bandwidth messaging provided by the GSM Short Message Service (SMS) are studied. Our focus is on the performance evaluation of email submission to SMS in terms of message type/length, device drivers and protocols. We investigate the causes of the bottlenecks and discuss the effectiveness of the gateway architecture. The results have shown that the gateway presents a stable performance over the various types and size of incoming messages.

1. INTRODUCTION

The emphasis in cellular networks is changing from voice-only communication to a rich combination of voice and messaging. The message format is rapidly becoming more important than the content, as it determines how the information can be accessed. The two non-voice services widely available on GSM networks [7] are SMS and Data. SMS is used primarily for the transfer of text, whereas the Data service can be used to transmit data such as video and graphics from sources such as the Internet. A brief comparison of the above is given in table 1. GSM Data is beyond of the scope of this paper as we are interested in applications where delivering the information in the actual form is not easiest or best communication method.

GSM SMS	GSM Data
Limited in length	Limited in speed
Store and forward	Dedicated connection
Phone only	Phone plus data converter
Teleservice	Bearer service
Signal path	Radio channel

Table 1: Comparison between GSM SMS and Data

SMS was developed as part of the GSM phase 2 specification [3] and the messages can be sent and received simultaneously with GSM voice, data and fax calls. This is possible because voice, data and fax calls take over a dedicated radio channel for the duration of a call, whereas short messages (SMs) travel over and above the radio channel using the signaling path. The signaling path is always active monitoring phones and passing information back and forth across the mobile network therefore it rarely, if ever, gets congested on peak network usage times.

The two types of SMS currently provided are Cell Broadcast and Point-to-Point. This paper concerns the last one, which refers to the exchange of messages between two Short Message Entities (SME), i.e. two mobile phones, or a PC and a mobile phone. Necessary for the transmission of the messages is the existence of a Short Message Service Center (SMSC), which provides storage, routing, and confirmed delivery of SMs. Within the GSM standard, network entities such as Home Location Register (HLR) are incorporated to track the location and status of mobile phones [5]. As such, international roaming is possible with SMS. The above properties make SMS suitable for capturing and distributing information quickly and efficiently, reducing delays to important decision-making [2].

The remainder of the paper is structured as follows. Section 2 is concerned with the components of the gateway and the applications designed for it. A brief comparison of the protocols used to access SMSCs is given in Section 3. The configuration, experimental results and additional features are presented in Sections 4 and 5. The paper is concluded in Section 6.

2. SMS GATEWAYS IN SHEFFIELD

The potential of using SMS as a transport layer protocol to support applications in the area of electronic commerce over heterogeneous wide area networks has been investigated by our group as part of the EU GAIA project in collaboration with Dialogue Communications Ltd. (e.g. [6]). The main contributions include the establishment and performance evaluation of a fully developed architecture and a number of completed applications such as email to SMS, SMS to email gateways, subscription mechanisms and database querying on the move. The capability of SMS to relational database querying has been demonstrated by an on-line registration system to University courses. The interoperability of SMS with the Z39.50 Information Retrieval Protocol has also been examined [4]. Other research projects are studying the possibility of using the SMS as an alternative layer to TCP/IP, as non-text based SMS (for example in binary format) are also supported. Our generic message architecture crosses the gap of messaging prior to the 3rd generation mobile communication systems, by providing mobility using existing equipment and protocols.

The SMS testbed (Figure 1) consists of PC-based hardware running Linux operating system interconnected via a 10Mbps Ethernet hub. In the experiments presented here we use a X.25 Radio PAD, a modem and a V.DOT.

The latter are self-contained GSM phones with voice, data, fax and SMS capabilities controlled by standard and extended AT commands from the RS 232 serial port of a PC. For testing purposes we use combinations of SIM cards from the four UK GSM network operators. The gateway architecture can accommodate any number of hardware interfaces in order to submit short messages in parallel and achieve higher throughput (e.g. use three modems to make three phone calls to the same SMSC). Most of the SMSC interfaces also allow submission of more than one SMs in the same session, reducing the impact of call set-up and service access time.

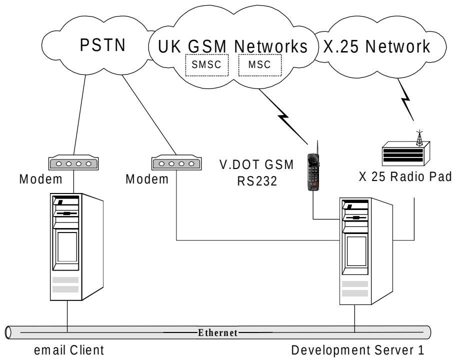

Figure 1: Overview of the SMS testbed used.

Figure 2 describes the protocol stack of the gateway and its integration within the OSI model. It contains all the necessary network drivers to communicate with the SMSC, whereas the routing algorithm is based on a set of rules describing the different submission mechanisms for a particular network.

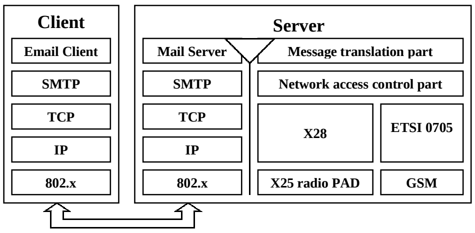

Figure 2: Gateway protocol stack

The gateway operates at the application layer when the SMTP server running on the gateway passes the email to the Message Translation Part (MTP). We use the alias features provided in Sendmail to map an email address to the input of a process. Consequently every time a new email is received by the gateway, the mail server starts a new instance of the message translation process, allowing parallel processing of incoming emails. There is one queue for each of the hardware interface connected to the gateway, with one independent submission process per queue (Figure 3). The messages are fragmented into 160 character chunks prior to submission. Although not used in the experiments reported herein, the gateway allows users to create special filters in order to decide for which incoming email messages should be sent an SMS notification and which are to be simply ignored.

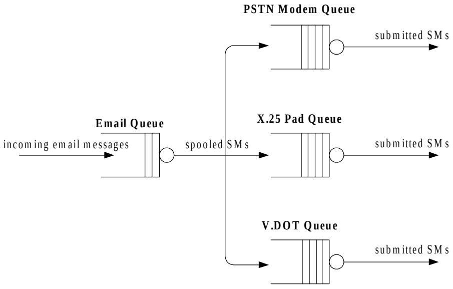

Figure 3: The gateway queuing model.

As a result from the way the gateway is implemented, each time an email arrives, the Sendmail server triggers a MTP process instance. For that reason it is expected that there will be at most one message at the Email Queue at any time. Depending on the routing decision, the message is placed into the appropriate queue. For each queue a dedicated Kermit process serves the incoming messages. The Kermit server controls the hardware device (modem, PAD, V.DOT) and handles the setup, connection and disconnection phase with the respective SMSC.

The MTP is divided into five sub-layers (figure 4), each taking care of one step in the overall short message creation process.

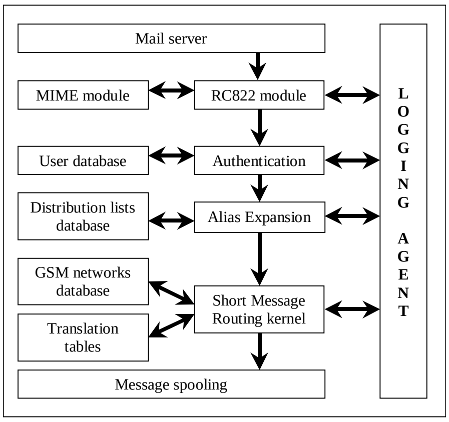

Figure 4: Message Translation Part sub-layers

The *RFC822 module* processes the incoming email and extracts the relevant information from the header. If necessary, the *MIME module* [1] is used to decode the body of the message, discarding attachments, and removing non-textual information and formatting tags when applicable. The module looks for recipients in the From:, To:, Cc:, and Bcc: headers fields. It also extracts the Subject header and passes the information to the next sub-layer. It was established that most of the character sets pages are machine dependent. This fact can cause the corruption of the received message. In order to solve the problem, an application was implemented which filters a character and converts it to an appropriate code point. Specifically, in our design a mapping between the ISO8859-1 and the default SMS character set is provided. This mechanism can easily be extended to cover any character set mapping. The message *Authentication* part provides security to the system by restricting access to registered users. This feature encrypts the Subject of the email using the standard UNIX crypt encoder and compares the resulting encrypted password with the one from its database. The next sub-layer checks whether any of the recipients is present in its distribution list database and expands them. This feature allows end users to send a SM to a group of people, in the same way that email mailing lists operate. Each step in the translation process is logged for statistical, invoicing or debugging purposes, by a separate process as shown by the *Logging agent*.

The *Short Message Routing Kernel* is the core of the system. Using routing tables it makes decisions on which SMSC to use so as to ensure efficient submission. The routing algorithm is rather complex and relies on a database containing network prefixes, SMSC access numbers, submission protocols, type of hardware interfaces connected, and other such relevant information. It does not use any least cost routing as at the moment the price of sending a SM is roughly the same among different operators. Using the same SMSC to submit all messages is feasible, however in many cases this would prove to reduce efficiency (i.e. minimum overall delay from reception of email to SM being delivered to the mobile).

3. SME-SMSC PROTOCOLS

During this study we have implemented all five protocols that are used to submit messages to the four UK GSM networks: TAP, Direct Access, Telenote, SMS2000 and ETSI-0705. Each provider uses a different protocol implementation, which conforms to the minimum requirements of the protocol specification. In this paper we report results from Vodafone (SMSC by Sema) TAP and Telenote and Orange (SMSC by Logica Aldiscon) TAP and Direct Access. A brief comparison of the protocols is given in table 2.

	Network Protocol	
Orange	TAP	Direct Access (DA)
	Bandwidth: 9600bps Max SMs / session: 9	Bandwidth: 9600bps Max SMs / session: 3
Vodafone	TAP	Telenote
	Bandwidth: 9600bps Max SMs / session: 2	Bandwidth: 2400bps (modem) Bandwidth: 4800bps (X.25) Max SMs / session: 2

Table 1: SME-SMSC Protocols

4. PERFORMANCE RESULTS

Emphasis is given on the several processing stages by the time an email is sent until the delivery of the SM. The primary metric is the time consumed in each processing stage, in order to find which parts constitute bottlenecks in the system. The following parameters were studied: the content type and length of the email message body, the length of the short message, the arrival rate of messages at the gateway, the number of SMs sent during a protocol session, and the available bandwidth for the transmission of the short messages.

For each pair of these values 50 email messages were sent to the gateway with a standard interarrival time of 5 sec, (12 emails per min). Mobile originated submission has the main advantage of not adding any overhead by waiting to access the network. The different scenarios that were examined to define the upper and lower limits of the processing delays of the gateway follow.

4.1 Email Content Type

In this experiment four different email content types were submitted to the gateway: Plain Text (ASCII), HTML (plain text plus HTML tags), Attachment (ASCII plus attachments in text, image or audio format) and Alternative (the same body is encoded in plain text and HTML format). The following table summarizes the delays of processing stages in MTP (in ms) for different size of these email content types.

		email content type			
	Processing Stage	Plain Text	Text / HTML	Text & HTML.	Text & Attachment
email = 50 bytes	Reading	5.0	2.2	2.0	1.6
	Authentication	9.2	10.2	15.8	12.2
	Expansion	0.0	0.2	0.2	0.6
	Routing	18.4	17.6	14.6	16.4
	Translation	0.4	0.2	1.4	1.2
	Spooling	1.2	1.8	3.6	3.2
	Total delay	*34.2*	*32.2*	*37.6*	*35.2*
email = 500 bytes	Reading	2.6	5.4	3.6	1.2
	Authentication	13	10.8	13.2	12.6
	Expansion	0.0	0.2	0.6	0.4
	Routing	16.2	19.2	16.2	18.6
	Translation	2.6	3.4	4.8	3.2
	Spooling	1.8	1.0	0.8	1.2
	Total delay	*36.2*	*45.4*	*39.2*	*37.2*
email = 5K bytes	Reading	7.8	23.4	3.8	4.8
	Authentication	11.2	12.8	13.2	14.2
	Expansion	0.0	0.2	0.4	1.2
	Routing	18.4	17.2	17.4	16.6
	Translation	1.6	4.8	4.0	3.2
	Spooling	1.6	1.6	0.4	3.8
	Total delay	*40.6*	*60.0*	*39.2*	*43.8*
email = 50K bytes	Reading	48	1445.4	25.6	23.2
	Authentication	10.4	11.8	11.8	12.8
	Expansion	3	2.0	4.4	3
	Routing	18.4	19.6	19.4	21.4
	Translation	3.6	3.6	2.4	3.6
	Spooling	2.0	1.0	1.6	1.8
	Total delay	*85.4*	*1483.4*	*65.2*	*131*

Table 2. Delays of Processing Stages in Message Translation Part (in ms)

The relatively bad performance of the *Authentication* and *Routing* routines can be explained by the fact that important databases have to be loaded from the hard disk when the mails server starts the process. A quick solution to this problem would be to implement a caching mechanism where the table is stored in memory and shared by all the concurrent processes. The messages of text/plain and multipart content types need about the same time period for each message length separately. On the other hand, the text/HTML formatted email messages require greater amount of time in compare with the rest message content types. The content type/subtype and length of the email messages primarily affect the processing delay, showing an increase in time as the message length increases.

4.2 Service Time

The following test shows the setup, submission and service times for different networks and protocols. The *setup time* is the time needed to establish a connection between the gateway and the SMSC of the respective GSM network. The *submission time* shows the duration of the short message transmission over the network link. For each network protocol 50 email of plain text content type were sent in total with a rate of one message per session. This is the worst case scenario, as the Email Queue has no more than one messages at any time and a connection made for the transmission of just one message, independently from the maximum number of messages supported in one session. For that reason, the calculated *service time* is equal to the time needed to serve one message, and it is given by the sum of setup time and submission time. The average delays are shown in figure 5.

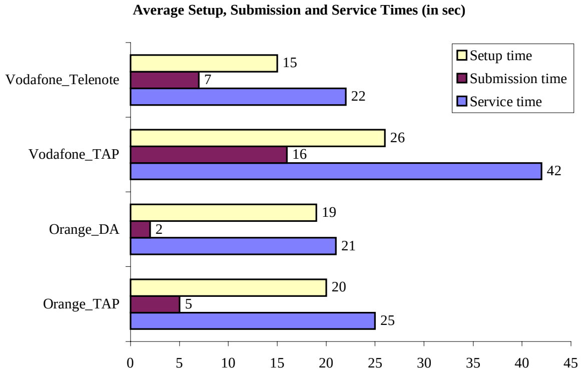

Figure 5: Average setup, submission and service times

According to the results, *submission times* vary among the four protocols for the same SMs. The longest setup time is for Vodafone TAP, which means that there is extra control information carried between the two SMEs during the setup phase. This increase can also be the result of any retransmissions of control data carried out during this phase. The maximum available bandwidth does not seem to affect the results under the conditions this test was performed. Indeed, Vodafone Telenote with maximum available bandwidth of 2400bps gives values closer to Orange protocols, which have 9600bps bandwidth. Although Orange TAP gives the second best average delay it does show variance, while the results for each of the other protocols can be grouped in a particular range of values.

4.3 Message Length

In this test it is examined how the short message body length affects the submission time (figure 6). The email content type send during this test was plain text. The values for the SM body length are chosen to be 10, 80 and 160 characters (plus the sender's signature). It was shown that the transmission times are almost constant for each protocol. There is a small increase of 1 sec for SMs with body length of 160 characters. Vodafone TAP gives again the longest submission times, while Orange Direct Access has the lowest ones showing a stable performance in the transmission time. These results coincide with the ones shown in figure 5, where the SMs were transmitted with their maximum length of 160 characters.

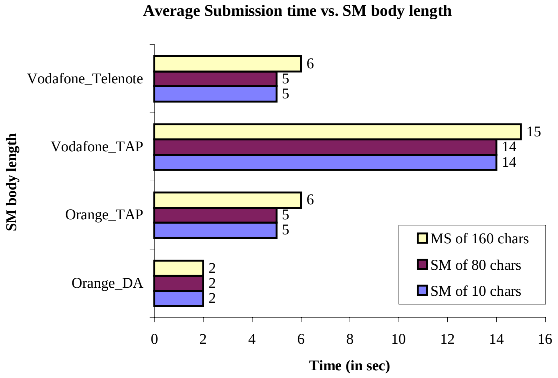

Figure 6: Average submission time vs. SM body length

5. DISCUSSION AND FURTHER WORK

Considering the different protocols the followings can be stated for each one of them:

Orange TAP has the highest throughput, up to 9 SMs/min, allowing 9 messages to be served per session. For that reason a SM waits (the greatest time is encountered for the maximum arrival rate of 60 SMs/min) a small amount of time (up to 158secs) before being served. But the service time, under the worse case scenarios was calculated to be 25secs, which is the second greater amount of time. Also the service time showed the greater deviation for Orange TAP, making its performance unstable regarding this characteristic. On the other hand, the percentage of successfully submitted SMs was about 99%.

Orange Direct Access is the second in the performance scale because its throughput can be up to 6 SMs/min, allowing 3 messages to be served per session, thus 2 sessions per minute. Also it has the second lower waiting time after Orange TAP, which can be up to 304secs (≅5mins). Under the worse case scenarios examined, it has the lowest service time of 21secs, and a percentage of successfully submitted SMs of almost 99%.

Vodafone Telenote a text based protocol, like Direct Access, can be considered to have a moderate performance. It has a low throughput, up to 4 SMs/min, and high delays that can be up to 452secs (≅7mins). Although it has a considerable small service time of 22secs, the maximum number of messages served per session is only 2, increasing the delays of the system. The percentage of successfully serviced SMs was about 84%, which is a rather low.

Vodafone TAP does not have the same performance with Orange TAP as it would be expected. The two different implementations of the same protocol present great differences. Vodafone TAP is the last one in the performance scale with a maximum throughput of 2msgs/min and delays that can be up to 1079secs (or about 17min.). The service time is the greatest of all (42secs) and it has a percentage of successfully serviced SMs of 64%.

Apart from protocol study, we are currently adding speech-processing capabilities to the gateway. More specifically we expect to make the voicemail retrieval on the move scheme more efficient by sending the caller id and a summary of voicemail messages as text on a mobile display. Users of existing voicemail systems on the receipt of notification have to call their answering machine and listen to the recorded messages. However, they are likely to wish to receive their voicemail as a text summary on cellular phones - especially for messages taken by answering services other than the one provided by the network operator (e.g. home answering machine or corporate voicemail system). Figure 7 shows the functional overview of this architecture.

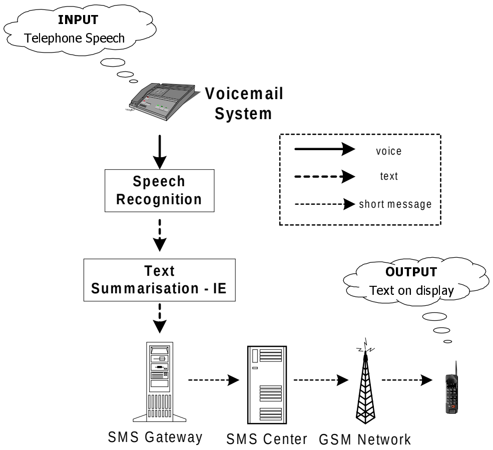

Figure 7: Voicemail retrieval on mobile display

Multiple platforms such as Acceptor HMM acoustic modeling speech recognizers, information extraction techniques and SMS gateways are incorporated into the system's architecture. The ABBOT system [8] is being optimized for the tasks of telephony and voicemail speech using standard international reference databases (such as the Nynex Phonebook and IBM Voicemail). Future work will be on developing statistical and grammar-based models for extracting specific types of information from messages and evaluating the performance of submitting the summary of voicemail messages through the SMS gateway.

6. CONCLUSIONS

This paper has presented a number of experiments for evaluation of a networking platform for notification services. It was concluded that the gateway presents a stable performance regarding its processing stages over various types and sizes of the incoming messages. Evaluation of four different message submission protocols was also reported. We are currently working on a service that transcribes and summarizes voicemail messages in order to deliver them through the gateway to mobile phones.

ACKNOWLEDGENENT

Part of the experimental results presented herein are based on I. Rousohatzaki's MSc. Dissertation "Performance Analysis of SMS Submission Protocols using an Email to SMS Gateway", Department of Computer Science, University of Sheffield.

REFERENCES

[1] N. Borenstein and N. Freed. MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies, *RFC 1521*, Bellcore, Innosoft, September 1993.

[2] S. Buckingham, *Data on SMS*, available from http://www.dataonsms.com, Sept. 1998.

[3] ETSI GSM 3.40 Digital Cellular Telecommunications System (Phase 2+) Technical Realization of the Short Message Service Point-to-Point, *European Telecommunications Standards Institute* TC SMG, Version 4.13.0, May. 1996.

[4] J. Moore, G. Peersman, S. R. Cvetkovic and K. El-Malki. GAIA Performance Testing of Z39.50 and SMS based Notification Services, Internal Deliverable, Ref. 1005X, *ACTS Project* Ref. AC221, May 1998.

[5] M. Mouly and M. B. Pautet. *The GSM System for Mobile Communications*, Telecom Publishing, 1992.

[6] G. Peersman, S. R. Cvetkovic, C. Smythe, P. Griffiths and H. Spear. Interworking of GSM Short Message Service and Interactive Voice Technology, *Proceedings of the IEE Colloquium on Telecommunication Networks and Services*, Jun. 1997.

[7] M. Rahnema. Overview of the GSM System and Protocol Architecture, *IEEE Communications Magazine*, Vol. 31, No. 4, pp. 92-100, Apr. 1993.

[8] T. Robinson, M. Hochberg and S. Renals. The Use of Recurrent Neural Networks in Continuous Speech Recognition, in *Advanced Topics in Speech and Speaker Recognition*, in C. H. Lee, K. K. Paliwal and F. K. Soong (Eds), pp. 233-258, Kluwer Academic Pub., Jun. 1996.