A Study of the Short Message Service of a Nationwide Cellular Network

Petros Zerfos
Deutsche Telekom Laboratories
10587 Berlin, Germany
petros.zerfos@telekom.de

Xiaoqiao Meng, Starsky H.Y Wong, Vidyut Samanta, Songwu Lu
UCLA Computer Science Dept.
Los Angeles, CA, 90095, USA
{xqmeng,hywong1,vids,slu}@cs.ucla.edu

In recent years, cellular networks have experienced an astronomical increase in the use of Short Message Service (SMS), making it a popular communication means for interpersonal as well as content provider-to-person usage. Yet little is known about the traffic and message user behavior in real SMS systems. In this paper, we present a measurement study of SMS based on traces collected from a nationwide cellular carrier during a three-week period. We characterize message traffic at both the message level and the conversation thread level. We also examine the "store-and-forward" mechanism of SMS and present initial measurements on how messages are actually delivered.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations

General Terms

Measurement

Keywords

Short Message Service, Network Measurements, Traffic Characterization

1. INTRODUCTION

The Short Message Service (SMS) is arguably the most popular data service over cellular networks nowadays. Though it was originally conceived as a paging mechanism for voice mails as part of the GSM specification in 1992, SMS has evolved into one of the most successful wireless data services in recent years. According to [13], SMS accounts for about 80% of data revenue generated for Western European mobile operators in 2005. In the largest twenty European countries, over 200 billion messages are exchanged in each month [19]. In Asia, the two major cellular carriers, China Mobile and China Unicom, report 304.14 billion message volume in 2005. SMS allows for users to exchange short alphanumeric messages with other users globally. Moreover, cellular carriers have opened their SMS interfaces to Internet-based content providers to deliver a wide variety of services to cell phone users, e.g., ringtone download, news, weather forecast, voting and casting preferences, mobile search, etc.

The SMS method of communication exhibits a number of characteristics that have contributed to its increased popularity. First, device support for sending and receiving short messages is almost ubiquitous, ranging from low-end mobile phones to Web-interface gateways that can be accessed via ordinary PCs over the Internet. Moreover, delivery and routing of SMS messages are supported by most cellular networks around the world. Second, it follows a "push" model of operation and short messages are delivered to mobile devices in near real-time, making SMS the wireless counterpart of Internet-like instant messaging applications such as ICQ and AIM. Third, it follows the store-and-forward communication paradigm similar to emails, in which a message is not discarded if it cannot be immediately received by the destination, and instead is stored at the server temporarily and retransmitted later. Therefore, its communication is more immune to intermittent connectivity scenarios that are frequently encountered in wireless networks. Considering the above main features of the SMS service and their suitability for operating in wireless environments, it is not surprising that it is recognized as one of the most successful communication services, especially in Europe and Asia.

In this paper, we present a preliminary analysis of SMS data traffic based on logs of Charging Data Records (CDRs[1]) [4] that were obtained from the billing subsystem of a national cellular carrier in India. The logs were collected over a period of three weeks, during which over 59 million short messages were exchanged by more than 10 million mobile users.

We make three contributions through our trace analysis. First, we characterize short messages by examining the message size distribution and message arrivals, and try to explain them through user activity. Despite the conventional understanding that short messages are generated by person-to-person communications, 7.2% of total messages sent by mobile users are requests to SMS services, and at least 10.1% of total received messages are sent by content providers instead of mobile users. Second, we examine how messages are "stored-and-forwarded" in SMS. Although 94.9% of messages are successfully delivered, the other 5.1% are not, due

IMC'06, October 25–27, 2006, Rio de Janeiro, Brazil.

[1]In the traditional circuit domain, CDR has been used to denote "Call Detail Record", which was subsumed by the "Charging Data Record" term.

to either expiration or denial of delivery. Regarding the delivered messages, 73.2% of them reach their recipients within 10 seconds, 17% of them need more than one minute, and another 5% require more than an hour and a half. Third, we characterize the message traffic at the conversation thread level. A conversation thread consists of messages continuously exchanged between two mobile users. Our study shows that 87.1% of threads are static threads, whereby neither user changes her associated Mobile Switching Center (MSC). The thread duration can be modeled by either the Weibull or the Lognormal distribution, and the thread interarrival times can be modeled by either the Weibull or the Gamma distribution.

Our measurement results can serve as a benchmark for protocol design, performance evaluation, planning and provisioning of SMS or other messaging systems. For example, the authors in [12] analyzed the optimal buffer size issue in SMS, but their results are based on assumptions about message arrivals and the service time. For a similar SMS study, web traffic traces (instead of models from real SMS traces) are used in [15]. Clearly our models herein can be applied to evaluate their design. As Short Messaging represents a major form of computer-mediated communications, our derived user behavior models reveal truly “mobile behavior” of users, since the service is available anywhere and anytime, unlike traffic over Wi-Fi hotspots [14], where most users are essentially stationary. Finally, although SMS does not generate a large traffic volume, the next-generation wireless messaging system, Multimedia Messaging Service (MMS) [6], is expected to significantly surpass SMS in terms of both traffic demands and popularity. Our study of SMS traffic may provide useful insights and guidelines for the planning and provisioning of MMS systems.

The rest of this paper is organized as follows. In Section 2, we provide a brief introduction to SMS and describe general characteristics of the collected traces. In Section 3, we present a preliminary analysis of the SMS traces, including message-level characteristics, the store-and-forward behaviors, and thread-level characteristics. We conclude the paper in Section 4 by outlining future work.

2. BACKGROUND AND TRACES

Before presenting our analysis on the SMS traffic, we briefly introduce the network architecture for SMS, as well as the network elements that provide the service. We also provide information on the traces that were collected.

2.1 SMS network architecture

In current cellular networks, SMS messages are transmitted over the Common Channel Signaling System 7 (SS7), which is the digital signaling control network used by network elements of wireline and wireless telephone carriers to exchange control information for call setup, routing, mobility management, etc. Figure 1 shows the typical network elements and architecture employed for handset-to-handset communication (Point-to-Point SMS).

Conceptually, the network architecture is split into two segments that are central to the SMS philosophy of operations (Figure 1). The elements at the sender, i.e., the Mobile Originating (MO) part, include the Mobile Station (MS) of the sender, the Base Station (BS) that provides the radio infrastructure for wireless communications, and the Originating Mobile Switching Center (MSC) that manages routes

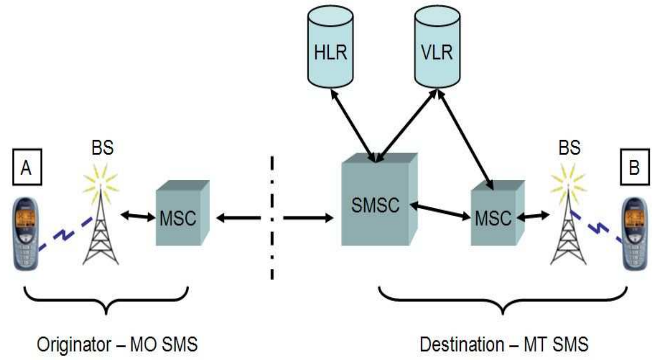


Figure 1: Typical handset-to-handset network architecture for SMS

and switches all traffic into and out of the cellular system on behalf of the mobile device of the sender. The elements at the destination of the message, more often known as the Mobile Terminating (MT) part, also feature a base station and an MSC (Terminating MSC) for the receiver. In addition, an SMS Center (SMSC) acts as a centralized, store-and-forward server that is responsible for accepting, storing, retrieving subscriber information, and forwarding messages to the intended recipients of the messages. It is assisted by two databases, namely the Home Location Register (HLR) and the Visitor Location Register (VLR), in which location information is kept regarding the subscribers and their mobile devices (such as the address of the MSC that the device is associated with). For more details on the SMS network architecture and its operations, as well as other scenarios with fixed entities that are capable of sending and receiving short messages (e.g., application servers), we refer the interested reader to the tutorial [17] and the specification [3].

2.2 Traces

We obtained permission from a cellular carrier in India to collect anonymized SMS logs for a period of about 3 weeks. The carrier operates a nationwide network, which covers almost the entire country except from a small area in the far east. The number of phone numbers appearing in our logs represents around 10% of the total mobile subscribers in the country. Due to the size of the set of mobile subscribers and the coverage of the network that was studied, we believe that the analysis carried out in this work provides useful insights on the SMS traffic characteristics[2].

The traces are logs of Charging Data Records (CDRs) [4]. For every SMS that is sent from one handset to another, two records are logged in the trace for billing purposes: one (MO-CDR) for the Mobile Originating part of the communication (from the originator mobile device to the SMSC), and another (MT-CDR) for the Mobile Terminating segment (from the SMSC to the destination handset). Each record logs information such as the time stamp of the event (with second-level granularity), mobile identification (and directory) numbers of the sender and the receiver, the number of retransmissions, the address of the MSC with which the

[2]As a side note, only sending a message is charged at a cost comparable to that of a 1-minute local voice call.

sender/receiver device is associated, the size of the message, the result of the delivery attempt, and a reference value that can be used to associate the MO and the MT records of an SMS.

The traces were made available to us in a proprietary, binary format used by the billing subsystem of the carrier, which we had to reverse-engineer to automate the retrieval of ASCII values from the records. The automated tool that we implemented for extracting the values was developed with the help of a GUI-based utility from the network analyzer package of the carrier to guide our reverse-engineering efforts. After the conversion we ended up with 20GB of data, consisting of 48,573,312 MO and 59,612,388 MT records. Overall, 10,854,135 SMS-capable entities sent 5.206G characters in short messages and received 5.198G characters.

The SMS logs used in the analysis span a period that starts on 04/04/2005, 15:01:02 and ends on 04/26/2005, 00:00:04. Unfortunately, due to the collection process in the premises of the carrier, certain time periods are not logged, which amount in a total of 544,521 seconds ($\approx$ 6.3 days) missing. Additionally, the overload incident reported in Figure 9 was based on a separate set of traces collected around the period of New Year's Eve of 2005.

3. RESULTS

In this section, we present a preliminary analysis of the SMS traces obtained from the cellular carrier. We first analyze the message-level statistics and then examine the store-and-forward aspects of the SMS. Finally, we group messages into conversational threads and characterize these threads.

3.1 Message-level characteristics

We first present message-level characteristics such as the distribution of message among users, the distribution of message size, and the distribution of messages at different locations.

Figure 2 plots the number of messages sent/received by each phone number (y-axis) versus the *rank* of the phone number (x-axis), in log-log scale. Phone numbers are ranked in a descending order according to the number of messages they send or receive.

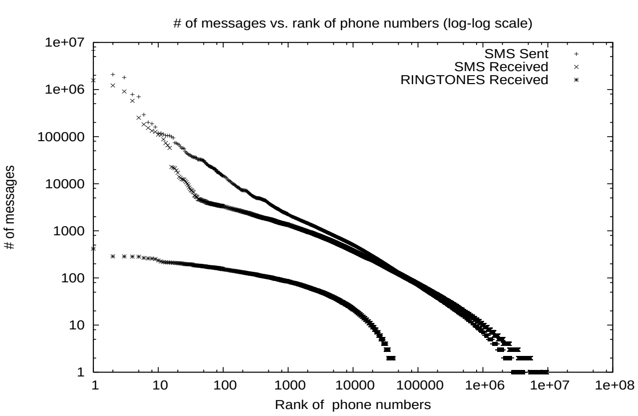


Figure 2: Messages per phone number versus the rank of phone numbers (log-log scale)

Figure 2 shows that the majority of phone numbers send only one or two messages; this can be verified by the tail of the curves in the figure. However, quite a few phone numbers send (and receive) thousands of messages within the 22 days, which cannot be explained by normal human activity. For example, the most active 28 phone numbers send more than 10,000 messages each, and 3573 out of the approximately 10 million subscribers sent more than 1000 each. We conjecture that these high-volume activities are due to value-added service and content providers, e.g., news, stock quotes, notifications for entertainment and lifestyle, horoscope, dating and others. Since service and content providers have distinguishing 4-digit phone numbers, we are able to identify almost 13,000 such providers in the trace. Furthermore, from the most active phone numbers, we identify 55 4-digit numbers of various SMS services that the cellular carrier offers, as listed in the carrier's website. As SMS becomes more widely used as a generalized notification and control mechanism in the future [19], such application-to-person (or even application-to-application) traffic might easily surpass the message volume generated by person-to-person communications.

The distribution of the message size (excluding message header) depends on the type of messages. Specifically, for those messages created in person-to-person communications, the length seems to evenly span the whole range of the allowable message size, for which the maximum value depends on the encoding that is used for each message [5]. This is also evident from the large number of messages with sizes 140 characters (6.12% of mobile originated messages and 6.18% of mobile terminated) and 160 characters (2.58% of mobile originated and 2.18% of mobile terminated messages), which are the fragmentation limits for messages encoded in 7-bit and 8-bit uncompressed format respectively. Person-to-application communication, namely mobile users requesting service by sending an MO SMS to the application server, tends to exhibit small message sizes; more than 96% of such requests have less than 50 characters. The reason might be that only limited information (such as a code or ID number) needs to be conveyed in the request. In the application-to-person communication, the messages exhibit a different pattern: such replies are usually customized versions of pre-formatted messages, and their sizes tend to concentrate around specific values. For example, more than 40% of these messages have a size between 67 and 106 bytes.

The messages in our trace are sent by 91 MSCs. By examining the message arrival processes at these MSCs, we find that the message arrival rates vary between 0 and 84 msg/minute. In order to derive a simple model for the arrival process, we tested several analytical distributions, such as the Exponential, Weibull, Lognormal, Gamma, Extreme-value, and the Pareto models. For those MSCs with low message arrival rates (less than 3 msg/minute), their message interarrival times are well modeled by the Exponential distribution. On the other hand, for MSCs with large arrival rates, no distribution can pass any statistical test.

3.2 Storing-and-forwarding short messages

To understand the "store-and-forward" operation in the SMS system, we examine the process of message delivery to the intended recipients. We seek to answer the following questions: (1) how much time do the messages stay in the messaging system before being delivered? (2) how many delivery attempts are needed for a message to be received by a mobile user?

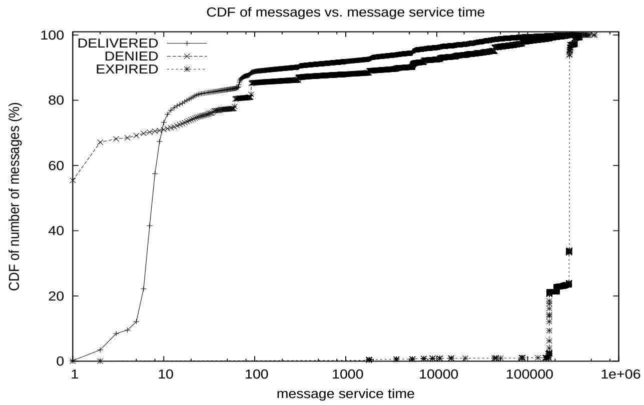


Figure 3: CDF of messages vs. message service time

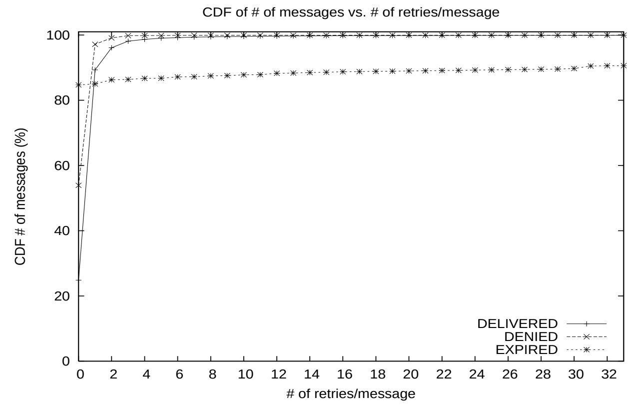


Figure 4: CDF of messages vs. number of retries per message

By comparing the timestamps of the mobile originating (MO) and the mobile terminating (MT) CDR of a message, we can estimate the time a message spent in the SMS system, which we call *service time*. In total, there are three types of messages: messages that were successfully delivered, messages denied, and messages that expired during the delivery process. Their percentages are 94.9%, 3.5%, and 1.6%, respectively. Figure 3 plots the cumulative distribution of the service times for all three types. The figure shows that, among the successfully delivered messages, 73.2% reach their destination in less than 10 seconds, which justifies the conventional characterization of SMS as "near real-time". On the other hand, 17% of delivered messages need more than a minute, and a significant 5% require more than an hour and a half.

While we can attribute the service times to factors such as processing latency at the SMS center, address resolution from the HLR/VLR server, network and radio resource shortage [2], the extraordinarily large service times are more likely caused by the mobile user behavior: a cell phone might be unreachable due to mobility, or the phone might simply be out of service (e.g., power off, sleep, inactive). We have collected more detailed traces from the operation of SMS over the SS7 stack that we intend to use in order to identify the exact cause that contributes to the service time.

For the 1.6% of messages that expired before they manage to reach their destinations, their service time tends to get two specific values: around 2 days, and 3.3 days. In principle, the expiration time for a message is manually configurable before the message is sent out [1]. However, our observation implies that either such an option is not allowed by mobile devices, or if it is allowed, users do not extensively use it; instead, they prefer to use some preset values. This fact should be taken into account when deriving analytical models for message delay [11].

Cumulative distributions of retries required by delivered, denied and expired short messages are shown in Figure 4. In [2], reasons for message delivery failure are classified into four categories: network, terminal, radio interface, and general problems. As part of the future work, we plan to extract the exact reasons for delivery failures by analyzing the traces from the operation of SMS over SS7.

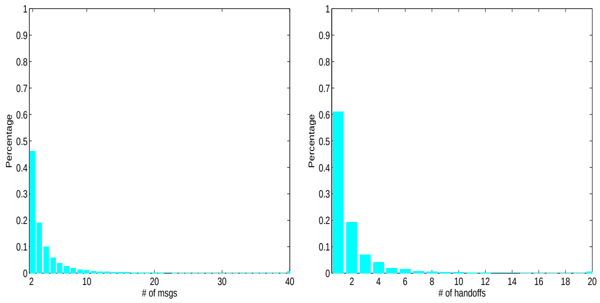
3.3 Thread-level characteristics

In person-to-person, 2-way instant communication, mobile phone users send messages to each other to say "hello", to prompt for something, to arrange a meeting, etc. Equally often a response is sent back to the user who initiates the communication, which further prompts a response, thus users essentially engage in a conversation through the exchange of short messages, in a way similar to Instant Messaging over the Internet.

Unfortunately, SMS does not incorporate the concept of threads or sessions: messages that are exchanged between two users are not associated in any way with each other, even though they could be grouped into a conversational thread. As wireless immediate messaging [7] is being integrated with Internet-style instant messaging systems [9] that incorporate the notion of threads, the design of a unified messaging system will need to characterize this behaviour. Such a characterization could be used as input to performance evaluation of any thread- or session-based protocol design for instant messaging [8].

We group messages into conversational threads by following the approach of [10]. Specifically, we merge consecutive, mobile-originated messages that are exchanged between the same two users into a thread, if each of these messages was sent within a predefined timeout period from the previous one. We experiment with relatively large values for the timeout, 1 minute and 10 minutes, to account for the fact that users typically need some time to reply, and to type the message using the somewhat inconvenient keypad of cell phones. We also do not consider in our analysis threads that fall near the gaps in our traces to protect the validity of the results.

Before presenting the results, we would like to explicitly state the limitations of the analysis. First, it is based only on traces collected from the SMS message center of a single cellular carrier. Clearly, cross-validation with logs from other locations is necessary before generalizing the conclusions. Second, the identification of threads is based on the somewhat arbitrary assumption of grouping messages that are exchanged between two users using a timeout value, but this hypothesis cannot be verified with the current logs. In fact, it cannot even be verified with traces from the underlying SS7 protocol, since no notion of thread (or session)



5.1: PDF for number of messages in each thread

5.2: PDF for number of handoffs in each roaming thread

Figure 5: Statistics on threads

exists in the system. Therefore we would like to caution the interested reader in interpreting the results about threads. Nevertheless, we believe that the following represents an initial step towards understanding SMS traffic, even though more work is admittedly needed.

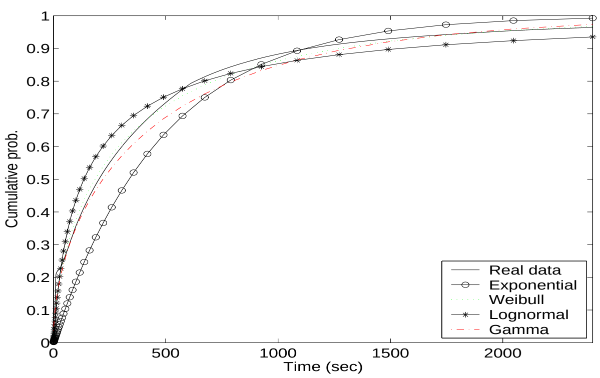


Figure 7: Empirical CDF for thread duration as well as CDFs for fitted analytical distributions

In the following, we present the analysis for the timeout value of 10 minutes. We identify 2.47 million threads in total. In Figure 5.1, we plot the distribution for the number of messages in each thread. On average, each thread consists of 4.9 messages. Of the 2.47 million identified threads, 87.1% are static, meaning that neither of the two users participating in the thread changes the associated MSC, throughout the thread. This also implies that for 87.1% of all the threads, thread mobility can be handled at the BTS/BTC (base station/base station controller) level. We further study the remaining 12.9% of the identified threads, which are roaming threads. In roaming threads, at least one user switches to another MSC during the thread duration. Figure 5.2 plots the distribution for the number of handoffs that occurred during the lifetime of roaming threads. The figure shows that more than half of the roaming threads, i.e., 61.1%, experience only one handoff at the MSC level.

Next, we model the thread duration against simple analytical distributions. Figure 7 gives the empirical CDF as well as the CDFs fitted by analytical distributions. The Weibull model turns out to give the minimum deviation, followed by the Lognormal model. On average the thread duration is 8 minutes.

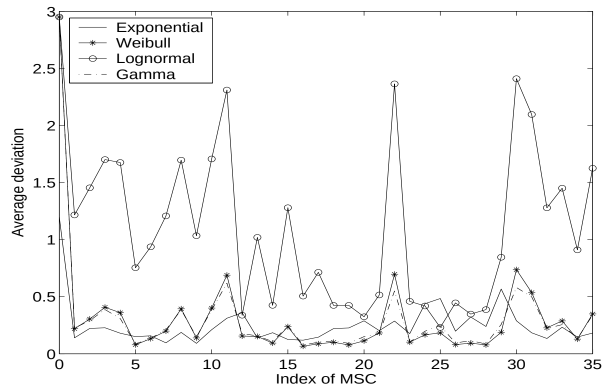


Figure 8: Average deviation of thread interarrival times from various analytical distributions at each MSC

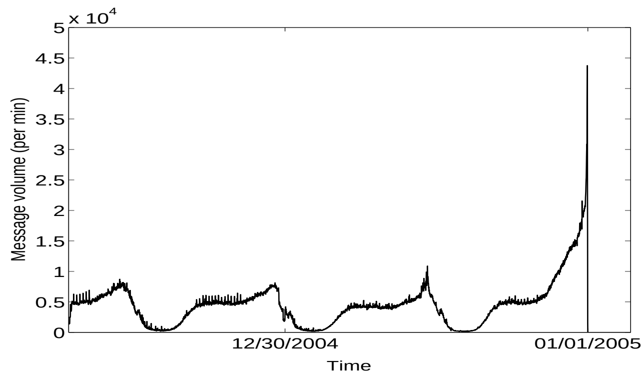


Figure 9: Message arrival rate in the period around December 31st, 2004

Last, we model the thread arrival process at individual MSCs by comparing the thread interarrival times against analytical distributions. To make the test statistically meaningful, we only consider the 35 MSCs which have generated more than 50 threads. In Figure 8, we give the average deviation of the analytical distributions when they are applied to each of the 35 MSCs. The figure shows that the thread interarrival times are well modeled by either the Weibull or the Gamma distribution. Specifically, for either of these distributions, the interarrival times at 34 MSCs can be well modeled with the average deviation less than 0.5. We further verify that this result holds when we set the timeout to be 1 minute.

4. CONCLUSION AND FUTURE WORK

In this paper we study the SMS traffic in a nation-wide cellular network. We touch on aspects of the service such as characteristics of short messages, message size distributions and their arrival process. We examine the "store-and-forward" mechanism of SMS, a central factor to its success as a convenient communication method. By grouping messages exchanged between users into conversation threads, we seek to characterize the traffic at the thread/session level.

To the best of our knowledge, similar studies have not been presented in the research community before, mostly because of the proprietary nature of cellular networks.

As part of ongoing work, we plan to systematically assess the reliability of SMS system. This becomes particularly important as the short message service is also being considered for mission-critical applications such as emergency alerts [16] and notifications for natural disasters [18]. For example, as it is shown in Figure 9 that plots the message arrival rate around the New Year's Eve of 2005 based on another set of traces obtained from the cellular network around that period, the SMS system is clearly overloaded[3] with the message rate increasing to more than eight times its normal value. Such an overload incident poses a significant risk for the reliability of the SMS service during special events such as holidays, especially if the system is not provisioned appropriately.

[3]The condition has been checked with the cellular operator.

5. REFERENCES

[1] Short Message Service/SMS tutorial. Technical report. http://www.developershome.com/sms/.

[2] TIA/EIA-41-D, Cellular Radiotelecommunications Intersystem Operations. Telecommunications Industry Association, July 1997.

[3] TIA/EIA-637-A, Short Message Service. Telecommunications Industry Association, December 1999.

[4] 3GPP. Ts 32.205, charging data description for the circuit switched (cs) domain; release 5. v.5.6.0, March 2003.

[5] 3GPP. TS 23.040, Technical Realization of the Short Message Service (SMS); Release 6. v6.5.0, September 2004.

[6] 3GPP. TS 22.140, multimedia messaging service (mms); stage 1; release 6. v6.7.0, March 2005.

[7] 3GPP2. S.r0061, wireless immediate messaging. v1.0, October 2002.

[8] B. Campbell and R. Mahy and C. Jennings. The Message Session Relay Protocol. Internet Draft, draft-ietf-simple-message-sessions-15.txt (Work in Progress), June 2006.

[9] M. Day, J. Rosenberg, and H. Sugano. A model for presence and instant messaging. IETF Request For Comments (RFC 2778), February 2000.

[10] C. Dewes, A. Wichmann, and A. Feldmann. An analysis of internet chat systems. In *Proceedings of the Internet Measurement Conference*. ACM Press, October 2003.

[11] M. Ghaderi and S. Keshav. Multimedia messaging service: System description and performance analysis. In *Proceedings of First International Conference on wireless Internet*. IEEE, July 2005.

[12] Y.-R. Huang. Determining the optimal buffer size for Short Message Transfer in a heterogeneous gprs/umts network. *IEEE Trans. Vehicular Technology*, 52(1), January 2003.

[13] B. Lagretteria. R u there?, February, 14th 2005. The Wall Street Journal Europe.

[14] X. Meng, S. H. Wong, Y. Yuan, and S. Lu. Characterizing flows in large wireless data networks. In *Proceedings of the Tenth Annual International conference on Mobile Computing and Networking*. ACM Press, September 2004.

[15] Z. Naor. An efficient Short Messages Transmission in cellular networks. In *Proc. INFOCOM*, Hong Kong, March 2004.

[16] O. of the Manager. SMS over SS7. Technical Report NCS TID 03-2, National communications system, December 2003.

[17] G. Peersman, S. Cvetkovic, P. Griffiths, and H. Spear. The global system for mobile communications Short Message Service. *IEEE Personal Communications*, June 2000.

[18] H. Rheingold. SMS disaster warning system?, December 2004. http://www.smartmobs.com/archive/2004/12/31/sms_disaster_wa.html.

[19] N. S.A. The netsize guide 2005 edition - the mobile is open for business, February 2005.