This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2457914, IEEE Transactions on Vehicular Technology

Transmission Policies for Multi-Segment Short Messages

Yi-Bing Lin, *Fellow, IEEE,* Sok-Ian Sou, and Chao-Liang Luo

Abstract—Performance of single-segment Short message service (SMS) (i.e., the message sizes are smaller than 140 octets) has been intensively investigated. On the other hand, multi-segment messages (with sizes larger than 140 octets) are seldom evaluated in the literature. This paper proposes analytical models to study two transmission policies for multi-segment short messages. We show how to improve the performance of the multi-segment short message delivery by selecting appropriate transmission parameters: the retransmission interval and the maximum number of transmissions. The proposed models are validated by measured data collected from the largest telecom operator in Taiwan over a period of six months. We provide useful guidelines to configure parameters for SMS transmission policy. Numerical examples showed that the performance can be improved by over 20% in terms of successful delivery ratio.

***Index Terms*—delivery delay, mobile telecommunications network, short message service, transmission policy, wireless transmission**.

1 Introduction

Portio Research [1] reported that short message service (SMS) traffic in the Asia Pacific region will surpass 4 trillion messages annual in 2014. By the end of 2017, there will be more than 6.4 billion SMS users to generate $11.7 trillion US dollars in worldwide service revenues. Therefore, SMS is considered as an important mobile data application, and has been intensively studied for several years [8-11]. Fig. 1 shows the SMS architecture for Universal Mobile Telecommunications System (UMTS) [3,6,12]. A person-to-person (P2P) message issued from a (UE, Fig. 1 (a)) is first sent to the Short Message-Service Center (SM-SC; Fig. 1 (e)) through the originating UMTS Terrestrial Radio Access Network (UTRAN; Fig. 1 (b)), the Mobile-Originating Mobile Switching Center (MO-MSC; Fig. 1 (c)) and the Inter-Working MSC (IWMSC; Fig. 1 (d)). The SM-SC then sends the short message to the Gateway MSC (GMSC; Fig. 1 (f)). The GMSC interrogates the Home Subscriber Server (HSS; Fig. 1 (g)) to identify the Mobile Terminating MSC (MT-MSC; Fig. 1 (h)) of the destination UE and forwards the message to the MT-MSC. Finally, the message is delivered to the destination User Equipment (UE; Fig. 1 (j)) via the terminating UTRAN (Fig. 1 (i)).

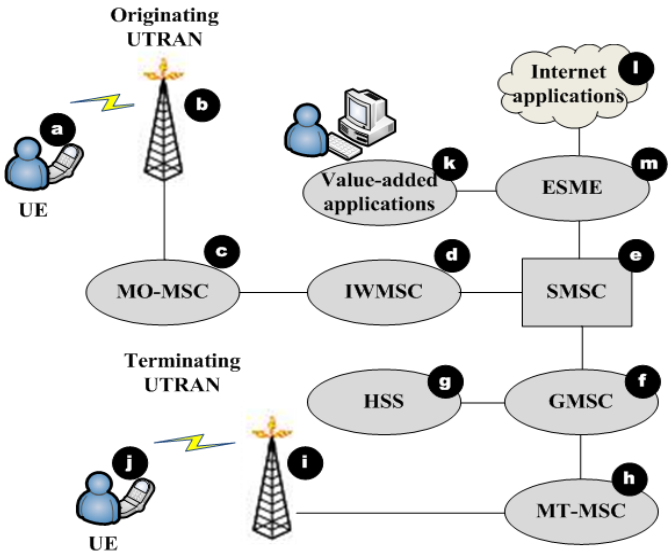


Fig. 1. SMS architecture

By using Short Message Peer-to-Peer Protocol (SMPP) [7], SMS delivery is intensively used as a bearer channel by many external value-added applications (Fig. 1 (k)) or Internet applications (Fig. 1 (l)) to transmit text contents to end-users. The application-to-person (AP2P) messages usually contain long content. In SMPP, payload length is limited by the constraints of the signaling protocol to precisely 140 octets. When the length of message content exceeds 140 octets, the SM-MC divides the content and processes the SMS with multiple segments SMS [2]. There is an upward trend of multiple segments SMS, especially those originated from

Manuscript received December 3, 2014; revised May 12, 2015; accepted July 13, 2015. The work was sponsored in part by the National Science Council (NSC) of Taiwan (R.O.C) under the contract number NSC 102-2221-E-006 -112 -MY3, MOST 103-2622-E-009-012 and MOST 104-2218-E-009-009, by Academia Sinica of Taiwan (R.O.C.) under the contract number AS-102-TP-A06. Y.-B. Lin is with the Department of Computer Science, National Chiao Tung University, Hsinchu 30010, Taiwan, R.O.C. and also with the Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C. (e-mail:liny@cs.nctu.edu.tw). S.-I. Sou is with the Department of Electrical Engineering, Institute of Computer and Communication Engineering, National Cheng Kung University, Tainan, 701, Taiwan, R.O.C. (email: sisou@mail.ncku.edu.tw). C.-L. Luo is with the Telecommunication Laboratories, Chunghwa Telecom Co., Taiwan (email: lojely@cht.com.tw).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2457914, IEEE Transactions on Vehicular Technology

internet applications. We call an SMS is an *I*-segment message when the SMS is partitioned into *I* segments. In Taiwan, there is more than 20% of SMS are delivered with multiple segments SMS in the SM-SC. The distribution of the number of segments in Chunghwa Telecom is as follows: *I*=1 (77%), *I*=2 (18%), *I*=3 (3%), *I*=4 (1%), *I*=5 (0.8%), and $I > 5$ is less than 1% of all short messages issued.

When a SMS delivery fails, the SM-SC receives the response code [3] issued from the MT-MSC. There are several reasons (error codes) for delivery failure. For example, if a network problem causes a short message failure, the error code indicates the reason "System Fail". If the user is temporarily unavailable to access the base station (e.g., moving through a tunnel), the error code indicates the reason of "Unreachable". SMS retransmission may result in huge mobile network signaling traffic and long elapsed times of short message delivery. Therefore, it is essential to exercise an efficient SMS retransmission policy to determine when and how many times to retransmit a short message to the terminating UE. In case of the delivery failure for a single segment SMS, the SM-SC simply retransmits the SMS after a pre-configured waiting period [4]. However, the retransmission policy for configuring the maximum number of attempts and the waiting period of multi-segment SMS is more complicate. For a multi-segment SMS containing several segments, the SM-SC needs to retransmit each part of multi-segment SMS in sequence. When a transmission of the first segment fails, the SM-SC waits for a retransmission interval, and then retransmits this segment again. Note that the other segments are stored in the SM-SC and cannot be sent before the terminating UE receives the first segment. The delivery of a multi-segment SMS is success only when the terminating UE receives all segments.

The huge demand for multi-segments SMS significantly increases and it is essential that mobile operators should provide efficient SMS delivery mechanism. In particular, it is important to select the transmission parameters (including the retransmission interval and the maximum number of transmissions) to improve the performance of the SMS delivery. In this paper, we investigate the transmission policy for multi-segments SMS. Specifically, we validate the performance based on SMS statistics (i.e., the logs of Charging Data Records (CDR)) that are obtained from a commercial mobile telecommunications network. The anonymized logs with size 140GB were collected from the largest telecom operator in Taiwan over a period of six months that starts on 01/15/2013, 00:00:12 and ends on 07/16/2013, 23:59:40, during which over 1.4 billion short messages were exchanged by more than 12 million mobile users. The carrier operates a nationwide 2G/3G/4G network including more than 30 MSCs, which covers most the entire country. Each record logs include the timestamps of the short messages, caller numbers, called numbers, the number of retransmissions, the sizes of the message, the numbers of segmentation, and the results of the delivery attempts.

The remaining of the paper is organized as follows: Section 2 investigates the performance trends on SMS retransmission and proposes analytic models to study two transmission policies for multi-segment short messages. Section 3 presents the multi-segment-SMS retransmission analysis. Finally, Section 4 concludes this study and outlines the future work.

2 MULTI-SEGMENT-SMS TRANSMISSION POLICY

This section investigates the transmission policy for delivering multi-segment SMS. We first investigate the performance trends on SMS retransmission policies based on the collected six-month SMS statistics from a commercial SMS operation in a six-month period. Then we propose analytic models to study two transmission policies for multi-segment short messages, which are validated by measured data.

For an *I*-segment message, let $p_{I,i,j}$ be the probability that the *j*th transmission of the *i*th segment is successful. From the measured data, we have the following observation.

Observation 1. Consider the *j*th transmission of the *i*th segment for an *I*-segment and a *K*-segment messages for all $I, K \geq i$. We have $p_{I,i,j} = p_{K,i,j}$.

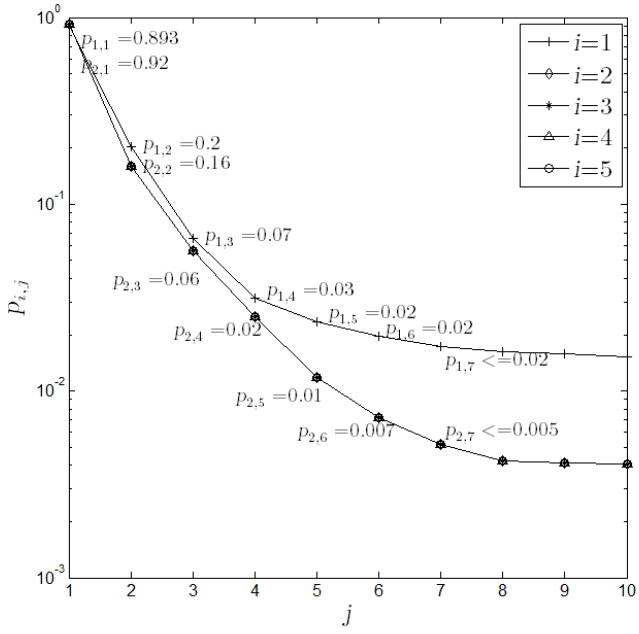


Fig. 2. The $p_{i,j}$ values ($t_r = 20$ minutes)

Observation 1 says that the availability of the mobile network for the *j*th transmission of any *i*th segments are independent of *I*. Therefore $p_{I,i,j} = p_{K,i,j}$, and we can replace the notation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2457914, IEEE Transactions on Vehicular Technology

$p_{I,i,j}$ by $p_{i,j}$ for all multiple-segment messages. Fig. 2 plots $p_{i,j}$ curves against j, where the retransmission interval t_r=20 minutes. Based on the results given in this figure, we further make **Observations 2-4**.

Observation 2. For all i and j, $p_{i,j} \geq p_{i,j+1}$.

This observation says that when the jth transmission fails, then it is more likely that the $j+1$st transmission also fails because it is very likely that the cause for network unavailability may still exist during the retransmission interval.

Observation 3. For $i,k \geq 2$ and for all j values, $p_{i,j} = p_{k,j}$.

This observation is important because it implies that the network environment conditions are the same for transmission of the ith segment, where $i \geq 2$. The transmission conditions for the first segment is different from that for the subsequent segments as described in the next observation.

Observation 4. For $i,j \geq 2, p_{1,1} < p_{i,1}$, and $p_{1,j} > p_{i,j}$.

Our measurements show that $p_{1,1} = 0.893 < 0.92 = p_{i,1}$. A segment transmission fails due to one of two reasons. First, the destination UE is turned off. Second, the destination UE is not available due to temporarily bad or weak radio coverage. For a multi-segment message, transmission of the first segment may fail due to both reasons. On the other hand, for $i \geq 2$, transmission for the ith segment is unlikely to experience turn-off of the destination UE because the UE was on when the first segment was delivered. Therefore, $p_{1,1} < p_{i,1}$. In addition, we observe that $p_{1,j} > p_{i,j}$ for $i \geq 2, j \geq 2$ in Fig. 2.

Let $p_i(j_i)$ be the probability that the ith segment is delivered with j_i transmissions. For $j_i = 1$, It is clear that $p_i(j_i) = p_{i,1}$. For j_i>1, the message fails at the first j_i-1 attempts (with probability $\prod_{j=1}^{j_i-1}(1-p_{i,j})$) and is success at the j_i-th attempt (with probability p_{i,j_i}). Then

$$p_i(j_i) = \begin{cases} p_{i,1} & \text{, for } j_i = 1 \\ p_{i,j_i}\prod_{j=1}^{j_i-1}(1-p_{i,j}) & \text{, for } j_i > 1 \end{cases} \quad (1)$$

2.1 Proposed transmission policy

Consider the ith segment of an I-segment message (where $1 \leq i \leq I$). If this segment is delivered with j_i transmissions (where $j_i \geq 1$), it means that the first j_i-1 transmissions of the segments fail, and the segment is successfully received by the destination UE at the j_i th transmission. After the ith segment is successfully received, the SM-SC begins to transmit the (i+1)th segment. When a transmission of the ith SMS segment fails, the SM-SC waits for a retransmission interval t_r, and then retransmits the segment again [3]. The SM-SC stops to transmit the following kth (where $i < k \leq I$) SMS segment before the ith segment is successfully received by the terminating UE. Suppose that the first i segments are delivered with J_i transmissions, then it is clear that

$$J_i = j_1 + j_2 + \cdots + j_i \quad (2)$$

Suppose that this I-segment message is finally delivered with J_I transmissions. In any commercial SMS operation, the number of transmissions allowed to deliver an I-segment message is limited to a value J_I^*. In other words, the message can be successfully delivered only if $J_I \leq J_I^*$. We describe two major policies to perform multiple-segment SMS transmission. Policy B is actually used in the existing commercial mobile operation. Policy A is a new approach proposed in this paper.

Policy A. The number of transmissions for a segment is not limited as long as $J_I \leq J_I^*$.

Policy B. The number of transmissions for the ith segment is limited to a number j_i^* such that $J_I^* = \sum_{i=1}^{I} j_i^*$.

Two output measures are defined to evaluate the performance of the SMS transmission policy: The probability $p_{A,I}$ ($p_{B,I}$)that an I-segment message for Policy A (B) is delivered, and the expected number $\mathrm{E}_A[J_I]$ ($\mathrm{E}_B[J_I]$)of transmissions performed for a short message delivery (either a success or a failure) with Policy A (B). In the following sections, we evaluate the performance of Policies A (B) in terms of $p_{A,I}$ ($p_{B,I}$) and $\mathrm{E}_A[J_I]$ ($\mathrm{E}_B[J_I]$). Then we study the effect of the retransmission interval t_r.

2.2 Analysis of Policy A

In Policy A, for $1 \leq i \leq I$, the number of transmissions for the ith segment is not limited as long as $J_I \leq J_I^*$, Since $j_1 \geq 1$, from (2), it implies that

$$J_i \leq J_I^* - (I - i) \quad (3)$$

By considering all possible combinations of the j_i values, and from (2) and (3), and by convention $j_0 = 0$, we have

$$j_i \leq J_I^* - j_1 - j_2 - \cdots - j_{i-1} - (I - i) \quad (4)$$

From (1) and (4), the probability $p_{A,I}$ that an I-segment message is delivered with Policy A can be expressed as

$$p_{A,I} = \sum_{j_1=1}^{J_I^*-(I-1)} \sum_{j_2=1}^{J_I^*-j_1-(I-2)} \cdots \times \sum_{j_I=1}^{J_I^*-j_1-j_2-\cdots-j_{I-1}} \times \prod_{i=1}^{I} p_i(j_i) \quad (5)$$

We compute the $p_{A,I}$ values based on (5) and compare them with the measurement. The discrepancies are within 2%. The solid curves in Fig. 3 plots the $p_{A,I}$ curves for $5 \leq J_I^* \leq 20$ where the retransmission interval t_r is 20 minutes. The figure shows that by increasing J_I^* from 5 to 20, the $p_{A,I}$ performance is improved. The expected number $\mathrm{E}_A[J_I]$ of transmissions performed for a short message delivery can be derived in two parts. Let $\mathrm{E}_{A,S}[J_I]$ be the expected number of transmissions performed for a successful I-segment short message. Then

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2457914, IEEE Transactions on Vehicular Technology

$$\mathrm{E}_{A,S}[J_I]p_{A,I} = \sum_{j_1=1}^{J_I^*-(I-1)} \sum_{j_2=1}^{J_I^*-j_1-(I-2)} \cdots$$

$$\cdots \times \sum_{j_I=1}^{J_I^*-j_1-j_2-\cdots-j_{I-1}} \times (j_1+j_2+\cdots+j_I) \prod_{i=1}^{I} p_i(j_i)$$

Let $\mathrm{E}_{A,F}[J_I]$ be the expected number of transmissions performed for a failed *I*-segment short message. Because the maximum number of transmissions for a multi-segment SMS allowed in the SM-SC is J_I^*, we have $\mathrm{E}_{A,F}[J_I] = J_I^*$. Also, a message delivery either successes or fails, we have

$$\mathrm{E}_A[J_I] = \mathrm{E}_{A,S}[J_I]p_{A,I} + \mathrm{E}_{A,F}[J_I](1-p_{A,I})$$

Therefore, we obtain

$$\mathrm{E}_A[J_I] = \sum_{j_1=1}^{J_I^*-(I-1)} \sum_{j_2=1}^{J_I^*-j_1-(I-2)} \cdots$$

$$\times \sum_{j_I=1}^{J_I^*-j_1-j_2-\cdots-j_{I-1}} (j_1+j_2+\cdots+j_I) \prod_{i=1}^{I} p_i(j_i)$$

$$+ J_I^*(1-p_{A,I}) \qquad (6)$$

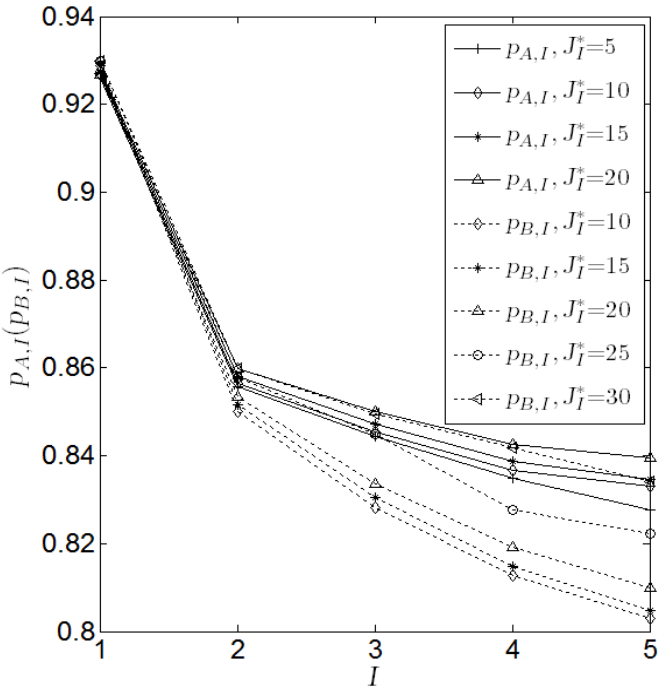


Fig. 3. The $p_{A,I}(p_{B,I})$ performance from measurements ($t_r = 20$ minutes)

Denote $\mathrm{E}_X[J_I|J_I^*]$ as $\mathrm{E}_X[J_I]$ with the maximum number of transmissions J_I^*. For Policy A with the retransmission interval $t_r = 20$ minutes, we observe that by increasing J_I^* from 5 to 20, the $\mathrm{E}_A[J_I]$ performance is significantly degraded for large *I*. In some commercial SMS operations, transmission costs may be cheap, and the operators are willing to allow 20% more retransmissions to improve 0.1% or 0.2% of the $p_{A,I}$ performance. Let $p_{X,I}(t_r, J_I^*)$ be the probability $p_{X,I}$ with retransmission interval t_r and the maximum number J_I^* of transmissions for *I*-segment SMS. Then Fig. 3 indicates that $p_{A,5}(t_r = 20, J_5^* = 20) > p_{A,3}(t_r = 20, J_3^* = 5)$ and $p_{A,3}(t_r = 20, J_3^* = 20) \approx p_{A,2}(t_r = 20, J_2^* = 5)$.

In other words, when $t_r = 20$ minutes, by increasing J_5^* from 5 to 20, the $p_{A,I}$ performance of 5-segment SMS is improved to be better than that of 3-segment SMS. On the other hand, our measurements indicated that

$$p_{A,1}(t_r = 20, J_1^* = 1) > p_{A,I}(t_r = 20, J_I^* = k)$$

for $I \geq 2$ and $k \geq 1$. That is, the $p_{X,I}$ performance of single-segment SMS always outperforms that of multi-segment SMS no matter how many re-transmissions are conducted.

2.3 Analysis of Policy B

In Policy B, the number of transmissions for the *i*th segment is limited to j_i^* such that $J_I^* = \sum_{i=1}^{I} j_i^*$. Let $Q_{i,j}$ be the probability that the first *j* transmissions fail for the *i*th segment. Then

$$Q_{i,j} = \prod_{k=1}^{j} (1-p_{i,k}) \qquad (7)$$

is the probability that the first *j* transmissions fail for the *i*th segment. By convention, $Q_{i,0}$ =1. From (7), Eq. (1) is re-written as

$$p_i(j_i) = p_{i,j_i} Q_{i,j_i-1} \qquad (8)$$

That is, the message fails at the first j_i-1 attempts (with probability $\prod_{j=1}^{j_i-1}(1-p_{i,j})$) and is successful at the j_i-th attempt (with probability p_{i,j_i}). The SM-SC successes to deliver the *i*-th segment with probability $1 - Q_{i,j_i^*}$. From (8), the probability $p_{B,I}$ that an *I*-segment message is delivered in Policy B can be expressed as

$$p_{B,I} = \prod_{i=1}^{I} \left[1 - Q_{i,j_i^*}\right] \qquad (9)$$

Most commercial SMS operations exercise Policy B with the following constraint:

$$j_i^* = \left\lfloor \frac{J_I^*}{I} \right\rfloor, 1 \leq i \leq I \qquad (10)$$

From **Observation 3** and (10), Eq. (9) is re-written as

$$p_{B,I} = \left(1 - Q_{1,J_I^*}\right)\left(1 - Q_{2,J_I^*}\right)^{I-1} \qquad (11)$$

We have compared (11) with the measured data, and the errors are within 1%. The dashed curves in Fig. 3 plot the $p_{B,I}$ values with $t_r = 20$ minutes. The figure shows trivial results that $p_{B,I} > p_{B,K}$ for $I < K$. Furthermore, if $m > n$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2457914, IEEE Transactions on Vehicular Technology

$p_{B,I}(J_I^* = m) > p_{B,I}(J_I^* = n)$. Fig. 3 also indicates that Policy A outperforms Policy B in terms of the $p_{X,I}$ performance. Specifically $p_{A,I}$ for $J_I^* = 20$ are larger (better) than $p_{B,I}$ for $J_I^* = 30$.

The expected number $\mathrm{E}_B[J_I]$ can be derived as follows. For $i \geq 1$, let N_i be the expected number of transmissions performed for a successful delivery of ith segment. This segment is delivered with probability $1 - Q_{i,j_i^*}$. From (8)

$$N_i = \left[\sum_{j=1}^{\left\lceil \frac{J_I^*}{I} \right\rceil} j\, p_i(j)\right]\left(1 - Q_{i,J}\right)^{-1}$$

$$= \left[\sum_{j=1}^{\left\lceil \frac{J_I^*}{I} \right\rceil} j\, p_{i,j} Q_{i,j-1}\right]\left(1 - Q_{i,J}\right)^{-1} \quad (12)$$

Let $\mathrm{E}_{B,S}[J_I]$ be the expected number of transmissions performed for successful delivery of an I-segment message. From **Observation 3**, $N_i = N_j$ for $2 \leq i, j \leq I$, and $\sum_{i=2}^{I} N_i = (I-1)N_2$. Therefore,

$$\mathrm{E}_{B,S}[J_I] = \sum_{i=1}^{I} N_i = N_1 + (I-1)N_2 \quad (13)$$

where N_1 and N_2 are expressed in (12). The expected number N_i^* of transmissions performed for an I-segment that fails at the ith segment is

$$N_i^* = \begin{cases} \left\lceil \frac{J_I^*}{I} \right\rceil & , i = 1 \\ N_1 + (i-2)N_2 + \left\lceil \frac{J_I^*}{I} \right\rceil & , i \geq 2 \end{cases} \quad (14)$$

In (14), if the SMS delivery fails at the first segment (i.e, $i = 1$), then there are $\left\lceil \frac{J_I^*}{I} \right\rceil$ unsuccessful transmissions. For $i \geq 2$, if the SMS delivery fails at the i th segment (with $\left\lceil \frac{J_I^*}{I} \right\rceil$ unsuccessful transmissions), then the first $i-1$ segments are successfully transmitted with $N_1 + (i-2)N_2$ transmissions. Let $P_F(i)$ be the probability that the multi-segment SMS delivery fails at the ith segment. Then from (7)

$$P_F(i) = \begin{cases} Q_{1,\left\lceil \frac{J_I^*}{I} \right\rceil} & , i = 1 \\ \left(1 - Q_{1,\left\lceil \frac{J_I^*}{I} \right\rceil}\right)\left(1 - Q_{2,\left\lceil \frac{J_I^*}{I} \right\rceil}\right)^{i-2} Q_{2,\left\lceil \frac{J_I^*}{I} \right\rceil}, & i \geq 2 \end{cases} \quad (15)$$

Let $\mathrm{E}_{B,F}[J_I]$ be the expected number of transmissions performed for a failed I-segment message. Then from (14) and (15), we have

$$\mathrm{E}_{B,F}[J_I]\left(1 - p_{B,I}\right) = \sum_{i=1}^{I} N_i^* P_F(i)$$

$$= \left\lceil \frac{J_I^*}{I} \right\rceil P_F(1) + \sum_{i=2}^{I} \times \ldots$$

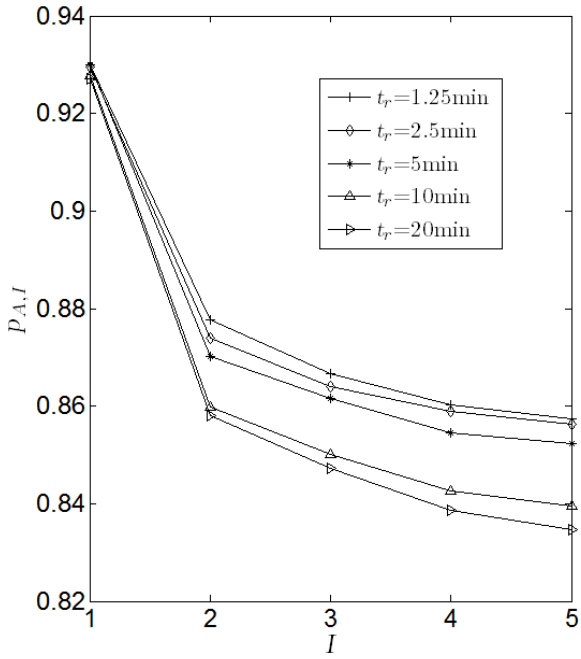
$$\times \left\{N_1 + (i-2)N_2 + \left\lceil \frac{J_I^*}{I} \right\rceil\right\} P_F(i) \quad (16)$$

From (10), (13), (15) and (16), we have

$$\mathrm{E}_B[J_I] = \mathrm{E}_{B,S}[J_I]p_{B,I} + \mathrm{E}_{B,F}[J_I]\left(1 - p_{B,I}\right)$$

$$= [N_1 + (I-1)N_2]p_{B,I} + \left\lceil \frac{J_I^*}{I} \right\rceil Q_{1,\left\lceil \frac{J_I^*}{I} \right\rceil} + \sum_{i=2}^{I} \left\{N_1 + (i-2)N_2 + \left\lceil \frac{J_I^*}{I} \right\rceil\right\} \times \left(1 - Q_{1,\left\lceil \frac{J_I^*}{I} \right\rceil}\right)\left(1 - Q_{2,\left\lceil \frac{J_I^*}{I} \right\rceil}\right)^{I-2} Q_{2,\left\lceil \frac{J_I^*}{I} \right\rceil} \quad (17)$$

Fig. 4. The $p_{A,I}$ performance ($J_I^* = 20$)



We compared (17) with the measured data and found that the errors are within 1.2%, which validates our analytic analysis against the measurement. Our study indicates that $\mathrm{E}_B[J_I]$ is insignificantly affected by J_I^*. Specifically, with $t_r = 20$ minutes, $\mathrm{E}_B[J_1|J_1^* = 10] = 3.29$, $\mathrm{E}_B[J_1|J_1^* = 20] = 3.31$, $\mathrm{E}_B[J_2|J_2^* = 10] = 5.41$, $\mathrm{E}_B[J_2|J_2^* = 20] = 5.48$, $\mathrm{E}_B[J_3|J_3^* = 10] = 7.83$, $\mathrm{E}_B[J_3|J_3^* = 20] = 7.91$, $\mathrm{E}_B[J_4|J_4^* = 10] = 10.21$, $\mathrm{E}_B[J_4|J_4^* = 20] = 10.41$, $\mathrm{E}_B[J_5|J_5^* = 10] =$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2457914, IEEE Transactions on Vehicular Technology

14.64 and $E_B[J_5|J_5^* = 20] = 15.11$, and the improvements of J_I by increasing J_I^* from 10 to 20 are within 3.1%. We also observe that $E_B[J_I] \approx E_A[J_I]$ for all I and J_I^*.

3. SMS RETRANSMISSION ANALYSIS

Consider the interval t_r between two consecutive transmissions of an SMS segment. This section studies how t_r affects the SMS performance. Under Policy A with $J_I^* = 20$, Fig. 4 plots $p_{A,I}$ against I with the following results: for a specific I, $p_{A,I}$ increases as t_r decreases, which implies that most failed SMS transmissions are due to temporary un-availability of communication between the UEs and the network, and if the SM-SC quickly re-tries, there is a good chance that SMS retransmission will be successful. This phenomenon become more significant for multi-segment SMS. For example, by decreasing t_r from 20 minutes to 1.25 minutes, $p_{A,I}$ is improved by 1.03% for $I = 1$, and 5.14% for $I = 5$.

Table 1. The $p_{A,I}$ performance

t_r	20	1.25					
I \ J_I^*	20	10	12	14	16	18	20
1	0.920	0.912	0.917	0.920	0.922	0.924	0.925
2	0.850	0.851	0.854	0.858	0.861	0.865	0.869
3	0.84	0.840	0.843	0.845	0.848	0.851	0.854
4	0.830	0.831	0.834	0.837	0.841	0.844	0.848
5	0.826	0.827	0.828	0.830	0.831	0.833	0.835

Table 2. The $E_A[J_I]$ performance

	$t_r = 20$		$t_r = 1.25$		Improvement
I	J_I^*	$E_A[J_I]$	J_I^*	$E_A[J_I]$	
1	20	3.32	15	2.62	21.1%
2	20	5.53	10	4.32	20.1%
3	20	8.1	10	6.52	19.5%
4	20	10.9	10	8.97	17.7%
5	20	15.2	10	12.85	15.6%

Based on the above discussion, we use a commercial operation example to show how we can select the parameters t_r and J_I^* to improve the performance of the SMS delivery. The baseline setups exercised by the commercial operation are $t_r = 20$ minutes and $J_I^* = 20$. From the observation in Fig. 4, we suggest to select $t_r = 1.25$ minutes. Table 1 lists the $p_{A,I}$ values for $t_r = 1.25$ minutes with J_I^* ranging from 10 to 20. The $p_{A,I}$ values for the current commercial SMS setup (i.e., $t_r = 20$ minutes and $J_I^* = 20$) are also listed. The table indicates to achieve the same $p_{A,I}$ performance as that for current commercial operation ($t_r = 20$ minutes, $J_1^* = 20$), it is appropriate to select $t_r = 1.25$ minutes, $J_1^* = 15$, and $J_I^* = 10$ for $I \geq 2$. Let $p_{A,I}(t_r, J_I^*)$ be $p_{A,I}$ where the retransmission interval is t_r and the maximum number of transmissions is J_I^*. Let $p_A(t_r, J_I^*)$ be the probability of successful delivery for an arbitrary short message with the retransmission interval t_r and the maximum transmissions number J_I^*; that is,

$$p_A(t_r, J_I^*) = \sum_{I=1}^{5} \alpha_I p_{A,I}(t_r, J_I^*) \quad (18)$$

where α_I is the percentage of I-segment SMS. From the measurements, we have $\alpha_1 = 0.77, \alpha_2 = 0.18, \alpha_3 = 0.03, \alpha_4 = 0.012$, and $\alpha_5 = 0.008$. Substitute the values of Table 1 into (18) to yield

$$p_A(t_r = 20, J_I^* = 20) = 0.910486 \quad (19)$$

and

$$p_A(t_r = 1.25, J_1^* = 15, J_I^* = 10 \text{ for } I \geq 2) = 0.911444 \quad (20)$$

Eqs. (19) and (20) quantitatively strengthen our previous observation that with a shorter t_r, it is possible to use a smaller J_I^* to achieve the same or larger $p_{A,I}$. Now we investigate the expected number $E_A[J_I]$ of transmissions executed in an I-segment SMS delivery. Table 2 lists $E_A[J_I]$ for the baseline setups ($t_r = 20$ minutes, $J_I^* = 20$) and the suggested new setups ($t_r = 1.25$ minutes, $J_1^* = 15$, $J_I^* = 10$ for $I \geq 2$). Table 2 shows that depending on the I values, the new setups can reduce the number of transmissions by 15.6%-21.1% as compared with the baseline setups.

Let $E_A[J|t_r, J_I^*]$ be $E_A[J_I]$ where the retransmission interval is t_r and the maximum number of transmissions is J_I^*. Let

$$E_A[J|t_r, J_I^*] = \sum_{I=1}^{5} \alpha_I E_A[J_I|t_r, J_I^*]$$

Then from Table 2,

$$E_A[J|t_r = 20, J_I^* = 20] = 4.05 \quad (21)$$

and

$$E_A[J|t_r = 1.25, J_1^* = 15, J_I^* = 10 \text{ for } I \geq 2] = 3.20 \quad (22)$$

Eqs. (21) and (22) indicate that with the same $p_{A,I}$ performance, the $E_A[J]$ performance can be improved by 21% if t_r is reduced from 20 minutes to 1.25 minutes by setting the appropriate J_I^* values. Although the above

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2457914, IEEE Transactions on Vehicular Technology

analysis indicates that a shorter t_r is better, one may question whether a short t_r will result in burst transmissions in the SMS network or not; i.e., many retransmission for a short message delivery are performed in a short interval of time. To investigate this issue we measure the number n_m of the short message issued during the *m*-th hour of a day (i.e., the time slot [*m*, *m*+1] where $0 \le m \le 23$). We also measure n_m^*, the number of first transmissions plus the number of retransmissions in the *m*-th hour. Fig. 5 plots the n_m^*/n_m curves for t_r ranging from 1.25 minutes to 20 minutes. The figure indicates that a short t_r does not cause more traffic jam than a long t_r does. The n_m^*/n_m value is between 1.77 and 1.68 for all t_r values.

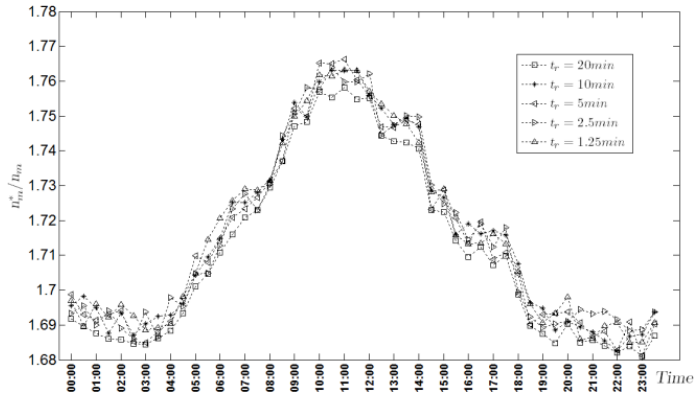


Fig. 5. The n_m^*/n_m values in 24 hours ($J_I^* = 20$)

We note that our study indicates that "the smaller the t_r value, the better the SMS delivery performance". However, the minimal operators are only allowed to adjust t_r value we can set is 1.25 minutes, and it is clear that optimal t_r occurs at a value smaller than 1.25 minutes but is not 0. We have asked the SM-SC vendor why the operator cannot set t_r less than 1.25 minutes. The reason is as follows: With $t_r \ge$ 1.25 minutes, it guaranteed that network (especially signaling) will not be congested (which is proven in Fig. 5). Therefore, the mobile operators are only allowed to adjust t_r larger than 1.25 minutes. With this constraint, our study is still general because all operators have to use mobile equipment following 3GPP specifications, and will have similar minimal t_r constraints.

4 Concluding Remarks

This paper proposed analytic models to investigate two multi-segment short message transmission policies. The analytic models were validated against by more than 100 millions measured data obtained from a 6-month commercial SMS operation. Our analytic model can effectively speed up network planning for commercial SMS operation. Specifically, we investigated how the setups for the retransmission interval t_r and the maximum number of transmissions J_I^* affect the output measures for two SMS transmission policies. These output measures include the probability $p_{A,I}$ ($p_{B,I}$) that an *I*-segment message is successfully delivered and the expected number $E_A[J_I]$ ($E_B[J_I]$) of transmissions performed for a short message delivery in Policy A (Policy B).

We make the following observations:

- The $p_{A,I}$ ($p_{B,I}$) performance of single-segment SMS always outperforms that of multi-segment SMS no matter how many retransmissions are conducted. Therefore, it is more desirable to optimize the transmission parameters for multi-segment SMS.
- Policy A (where the number of transmissions for a segment is not limited as long as $J_I \le J_I^*$) outperforms Policy B (where the number of transmissions for the *i*th segment is limited to a number j_i^* such that $J_I^* = \sum_{i=1}^{I} j_i^*$). Therefore the mobile operators are advised to utilize Policy A.
- By selecting a short retransmission interval t_r, the expected number of retransmissions for a short message delivery is less (and therefore better) than that for a long t_r under the same $p_{A,I}$ ($p_{B,I}$) performance.
- For a specific commercial SMS system, if $p_{i,j}$ (the probability that the *j*th transmission of the *i*th segment is successful) are measured, then our analytic model can be used to predict the SMS performance for both Policies A and B with various J_I^* values without conducting tedious measurements. In other words, our analytic model can effectively speed up network planning for commercial SMS operation.

To conclude, our study provided guidelines to adjust t_r and J_I^* of a SM-SC, the expected number of SMS transmissions can be reduced by over 20% at the same successful delivery probability.

5 References

[1] Portio Research. (July 4, 2013) Mobile Messaging Futures 2013-2017 [Online]. Available: http://www.portioresearch.com/en/major-reports/current-portfolio/mobile-messaging-futures-2013-2017.aspx

[2] 3GPP. Technical realization of the Short Message Service (SMS). TS 23.040 v12.2.0, 3rd Generation Partnership Project (3GPP), Dec. 2013.

[3] 3GPP. Technical Specification Group Core Network and Terminals; Mobile Application Part (MAP) specification (Release 12). Technical Specification3G TS 29.002 version 12.4.0 (2014-03), 2014.

[4] S-I Sou, Y-B Lin, and C-L Luo. "Cost analysis of short message retransmissions," *IEEE Transactions on Mobile Computing*, vol. 9, no. 2, pp. 215–225, 2009.

[5] P. Zerfos, X. Meng, and S. H.Y Wong "A Study of the Short Message Service of a Nationwide Cellular Network,"

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2457914, IEEE Transactions on Vehicular Technology

ACM Internet Measurement Conference, pp. 263 – 268, 2006.

[6] 3GPP. Technical Specification Group Services and Systems Aspects; Network Architecture (Release 12). Technical Specification 3G TS 23.002 version 12.4.0 (2014-03), 2014.

[7] SMS forum, Short Message Peer-to-Peer Protocol (SMPP) Specification version 5.0, 2003

[8] Tomi T. Ahonen (January 13, 2011). "Time to Confirm some Mobile User Numbers: SMS, MMS, Mobile Internet, M-News". Blog. Retrieved September 16, 2013.

[9] M. M. Ahmed, and W. L. Soo "Development of Novel Distribution Automation System (DAS) on Customer Side Distribution System," *Transmission and Distribution Conference and Exposition IEEE PES*, pp. 1-7, 2012.

[10] U. Goel "The personal SMS Gateway," *IEEE 3rd International Conference on Communication Software and Networks (ICCSN)*, pp. 717-621, 2011.

[11] K.-C. Lee, J.-H. Lee "SMS Transmission Mechanisms for Multi-Protocols on VoIP," *Advanced Communication Technology, The 9th International Conference on*, vol. 3, pp. 1697- 1701, 2007.

[12] A.-C. Pang, J.-C. Chen, Y.-K. Chen, and, P. Agrawal, "Mobility and Session Management: UMTS vs. cdma2000," *IEEE Wireless Communications*, vol. 1, no. 4, pp. 30-44, 2004.

Yi-Bing Lin (M'96-SM'96-F'03) received his Bachelor's degree from National Cheng Kung University, Taiwan, in 1983, and his Ph.D. from the University of Washington, USA, in 1990. From 1990 to 1995 he was a Research Scientist with Bellcore (Telcordia). He then joined the National Chiao Tung University (NCTU) in Taiwan, where he remains. In 2010, Lin became a lifetime Chair Professor of NCTU, and in 2011, the Vice President NCTU. Since 2014, Lin has been appointed as Deputy Minister, Ministry of Science and Technology, Taiwan.

Lin is also an Adjunct Research Fellow, Institute of Information Science, Academia Sinica, Research Center for Information Technology Innovation, Academia Sinica, and a member of board of directors, Chunghwa Telecom. He serves on the editorial boards of IEEE Trans. on Vehicular Technology. He is General or Program Chair for prestigious conferences including ACM MobiCom 2002. He is Guest Editor for several journals including IEEE Transactions on Computers. Lin is the author of the books *Wireless and Mobile Network Architecture* (Wiley, 2001), *Wireless and Mobile All-IP Networks* (John Wiley,2005), and *Charging for Mobile All-IP Telecommunications* (Wiley, 2008). Lin received numerous research awards including 2005 NSC Distinguished Researcher, 2006 Academic Award of Ministry of Education and 2008 Award for Outstanding contributions in Science and Technology, Executive Yuen, 2011 National Chair Award, and TWAS Prize in Engineering Sciences, 2011 (The Academy of Sciences for the Developing World). He is in the advisory boards or the review boards of various government organizations including Ministry of Economic Affairs, Ministry of Education, Ministry of Transportation and Communications, and National Science Council. Lin is AAAS Fellow, ACM Fellow, IEEE Fellow, and IET Fellow.

Sok-Ian Sou is an associate professor in the Institute of Computer and Communication Engineering and the Department of Electrical Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan. She received the BS, the MS and the PhD degrees in Computer Science and Information Engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1997, 2004, and 2008, respectively. She was a visiting scholar at Carnegie Mellon University (CMU) during July-August 2009. She was a recipient of the Investigative Research Award from the Pan Wen Yuan Foundation in 2009 and the Young Researcher in Service Science Award from the Sayling Wen Cultural & Educational Foundation in 2012. Her current research interests include design and analysis of mobile communication services, vehicular networking and performance modeling. She is the co-author of the book entitled Charging for Mobile All-IP Telecommunications (with Yi-Bing Lin; published by Wiley, 2008).

Chao-Liang Luo is a researcher of Telecommunication Laboratories, Chunghwa Telecom Co., Taiwan. He received the PhD degree from the National Chiao Tung University (NCTU), Taiwan, in 2014. In 2005, he was with the short message service and value-added service team. Since then, he has been involved in the design of the Long Term Evolution (LTE) network, mobile packet switched data network planning, multimedia services, and mobile network evolution. His research interests include the design and analysis of personal communications services, 4G networks, mobile service and performance modeling.