Cost Analysis of Short Message Retransmissions

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

Abstract—Short Message Service (SMS) is the most popular mobile data service today. In Taiwan, a subscriber sends more than 200 short messages per year on average. The huge demand for SMS significantly increases network traffic, and it is essential that mobile operators should provide efficient SMS delivery mechanism. In this paper, we study the short message retransmission policies and derive some facts about these policies. Then, we propose an analytic model to investigate the short message retransmission performance. The analytic model is validated against simulation experiments. We also collect SMS statistics from a commercial mobile telecommunications network. Our study indicates that the performance trends for the analytic/simulation models and the measured data are consistent.

Index Terms—Mobile telecommunications network, Short Message Service (SMS), retransmission policy, delivery delay.

1 Introduction

Short Message Service (SMS) provides a non-real-time transfer of messages with low-capacity and low-time performance. Because of its simplicity, SMS is the most popular mobile data service today. Many mobile data applications have been developed based on the SMS technology [5], [6], [8], [9], [10], [11], [13].

Fig. 1 illustrates the SMS network architecture for *Universal Mobile Telecommunications System* (UMTS) [1], [2], [3]. In this architecture, when a *User Equipment* (UE) (called the originating UE; see Fig. 1a) sends a short message to another UE (called the terminating UE; see Fig. 1b), the short message is first sent to the *Originating Mobile Switching Center* (MO-MSC; Fig. 1d) through the originating *UMTS Terrestrial Radio Access Network* (UTRAN; Fig. 1c). The MO-MSC forwards the message to the *Inter-Working MSC* (IWMSC; Fig. 1e). The IWMSC passes this message to a *Short Message-Service Center* (SM-SC; Fig. 1f). Following the UMTS roaming protocol, the *Gateway MSC* (GMSC; Fig. 1g) forwards the message to the *Terminating MSC* (MT-MSC; Fig. 1h). Finally, the short message is sent to the terminating UE via the terminating UTRAN. Due to user behavior and mobile network unavailability (e.g., the user moves to a weak signal area such as a tunnel, an elevator, a basement, and so on), a short message may not be successfully delivered at the first time. If a short message transmission fails, the SM-SC retransmits the short message to the terminating UE after a waiting period. Retransmission may repeat several times until the short message is successfully delivered or the SM-SC gives up delivering the short message.

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. To address this issue, we propose analytic and simulation models to investigate the performance of SMS retransmission in terms of the expected number of retransmissions and the message delivery delay. We also collect measured data from a commercial UMTS system to further analyze the performance trends on SMS retransmission policies.

2 Some Facts on Short Message Retransmission

This section defines SMS retransmission policies, and derives some facts on short message retransmission.

Definition 1. *An SMS retransmission policy is an order set* $\mathrm{s} = \{a_i | 0 \le i \le I\}$, *where* I *is the maximum number of retransmissions and* a_i *is the time of issuing the* i*th retransmission. By convention,* $a_0 = 0$ *is the first time when the short message is sent to the terminating UE. If the short message delivery fails at the* $(i-1)$*th retransmission, then it is resent after a waiting period* $t_{r,i} = a_i - a_{i-1}$, *for* $i \ge 1$.

Example 1. If $\mathrm{s} = \{0$, 5 mins, 10 mins, 20 mins, 1 hr, 3 hrs, 6 hrs, 12 hrs, 24 hrs$\}$, then $I = 8$, and $t_{r,3} = 10$ minutes (=20-10). We note that after I retransmissions, the short message will not be retransmitted, and the delivery fails.

Definition 2. *Consider a short message delivered under policy* s. *Denote* $T(\mathrm{s})$ *as the last time when the short message is retransmitted, and* $N(\mathrm{s})$ *as the number of retransmissions.*

- *S.-I. Sou is with the Department of Electrical Engineering, National Cheng Kung University, University Road 1, Tainan, Taiwan 701, R.O.C. E-mail: sisou@mail.ncku.edu.tw.*
- *Y.-B. Lin is with the Department of Computer Science, National Chiao Tung University, University Road 1001, Hsinchu, Taiwan 30010, R.O.C., and the Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan. E-mail: liny@cs.nctu.edu.tw.*
- *C.-L. Luo is with the Department of Computer Science, National Chiao Tung University, University Road 1001, Hsinchu, Taiwan 30010, R.O.C., and the Multimedia Applications Laboratory, Chunghwa Telcom Co., Ltd., R.O.C. E-mail: lojely@cht.com.tw.*

Manuscript received 16 Dec. 2008; revised 1 May 2009; accepted 12 May 2009; published online 27 May 2009.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2008-12-0499. Digital Object Identifier no. 10.1109/TMC.2009.104.

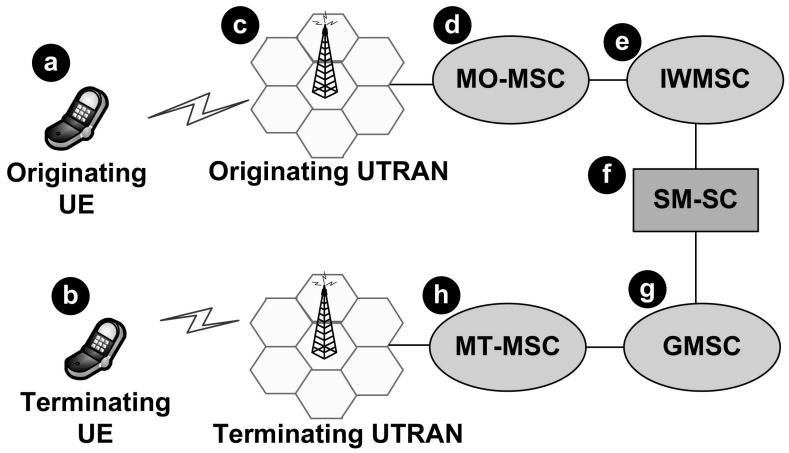


Fig. 1. The SMS delivery architecture.

In Example 1, if a short message is successfully delivered at hour 1, then $T(\mathrm{s}) = 1$ hr and $N(\mathrm{s}) = 4$.

Definition 3. *A policy* s' *is a subset of another policy* s *(i.e.,* $\mathrm{s}' \subseteq \mathrm{s}$*) if for all* $b \in \mathrm{s}'$*, we have* $b \in \mathrm{s}$.

To make a fair comparison of the retransmission policies, we consider the *same SMS network availability condition* concept (i.e., different policies see the same connected/disconnected periods of a UE). In Fig. 2, the time line shows the connected and disconnected periods of a terminating UE, where the short message is successfully delivered in the connected periods and cannot be delivered in the disconnected periods. Let $\mathrm{s} = \{0, a_1, a_2, \ldots\}$ and $\mathrm{s}' = \{0, b_1, b_2, \ldots\}$, where $\mathrm{s}' \subseteq \mathrm{s}$, $b_1 = a_1$, $b_2 = a_4$, and so on. In this figure, $T(\mathrm{s}) = a_4$ and $N(\mathrm{s}) = 4$ for policy s. For policy s', $T(\mathrm{s}') = b_2$ and $N(\mathrm{s}') = 2$.

Under the same SMS network availability condition, we derive the following facts:

Fact 1. If $\mathrm{s}' \subseteq \mathrm{s}$, then for every short message, we have

$$T(\mathrm{s}) \leq T(\mathrm{s}').$$

Proof. Due to the same SMS network availability condition, if a short message is delivered under s', then it can also be successfully delivered at b under policy s. Furthermore, it is possible that the message is successfully delivered at some time $a \in \mathrm{s}$, where $a \leq b$. Therefore, $T(\mathrm{s}) = a \leq b = T(\mathrm{s}')$. □

Definition 4. *Consider two policies* $\mathrm{s}' \subseteq \mathrm{s}$*. For any time point* $a \in \mathrm{s}$*, define the subset* $\sigma(a) = \{c | c < a, c \in \mathrm{s}, c \notin \mathrm{s}'\}$*. Denote* $|\sigma(a)|$ *as the size of* $\sigma(a)$.

Fact 2. For $\mathrm{s}' \subseteq \mathrm{s}$, if $T(\mathrm{s}) = T(\mathrm{s}')$, then $N(\mathrm{s}) \geq N(\mathrm{s}')$.

Proof. Since $|\sigma(T(\mathrm{s}))| \geq 0$, s' retransmits same or more times than s before $T(\mathrm{s})$, and $N(\mathrm{s}) \geq N(\mathrm{s}')$. □

Fact 3. For $\mathrm{s}' \subseteq \mathrm{s}$ and $T(\mathrm{s}) < T(\mathrm{s}')$, let $b_j = T(\mathrm{s}')$. There exists $b_i \in \mathrm{s}'$ such that $b_i \leq T(\mathrm{s}) < T(\mathrm{s}') = b_j$. Then, $N(\mathrm{s}) \geq N(\mathrm{s}')$ if and only if $|\sigma(T(\mathrm{s}))| \geq j - i$.

Proof. $|\sigma(T(\mathrm{s}))|$ is the number of retransmissions performed in s but not in s', and $j - i$ is the number of retransmissions performed in s' but not in s. Therefore, $N(\mathrm{s}) \geq N(\mathrm{s}')$ if and only if $|\sigma(T(\mathrm{s}))| \geq j - i$. □

Note that the condition $|\sigma(a)| \leq j - i$ holds when the SMS network experiences very short periods of outage (and therefore, a policy that quickly retransmits the short message is preferred). The condition $|\sigma(a)| \geq j - i$ holds when the receiver of the short message turns off the UE for a long period, but the status is not known to the mobile network.

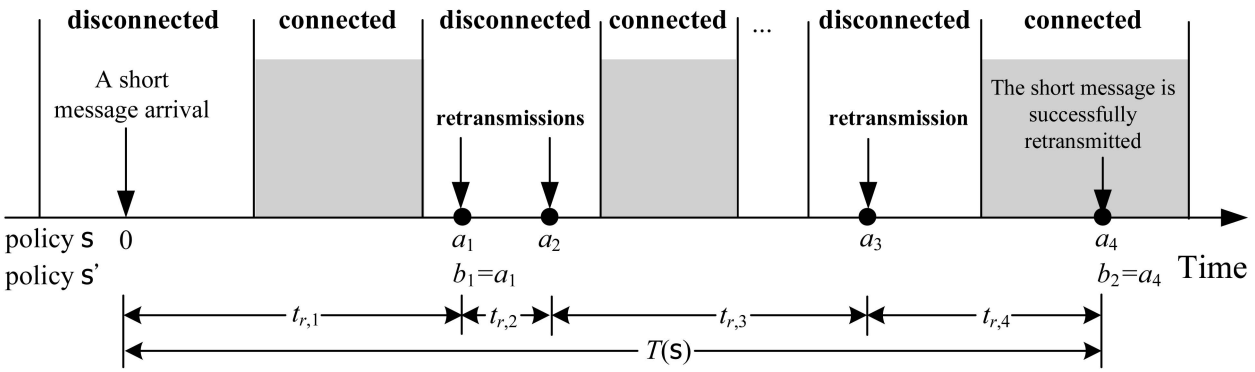


Fig. 2. Timing diagram for the connected and disconnected periods ("•" represents a short message retransmission).

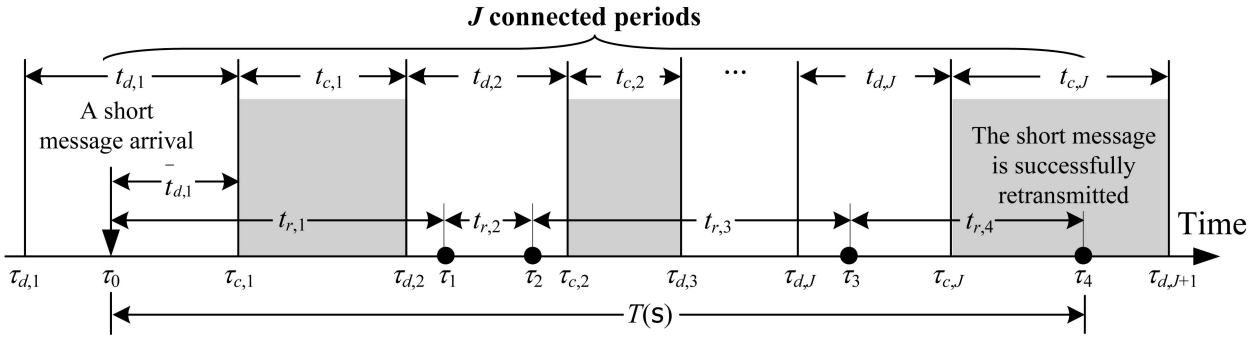


Fig. 3. Timing diagram for the short message retransmissions ("•" represents a short message retransmission).

Let $p_f(\mathrm{s})$ be the probability that s fails to deliver a short message. If a short message is not successfully delivered within I retransmissions, then the SM-SC stops resending this short message, and the delivery fails.

Fact 4. Suppose that $\mathrm{s}' \subseteq \mathrm{s}$, then $p_f(\mathrm{s}) \leq p_f(\mathrm{s}')$.

Proof. Directly from Fact 1, that is, if s' can successfully deliver a short message at retransmission time b, then under the same SMS network availability condition, s can also successfully deliver the message at b. On the other hand, s may successfully deliver a message at time $a \in \mathrm{s}$ where $a \notin \mathrm{s}'$ (i.e., $|\sigma(b)| \geq 0$). Therefore, $p_f(\mathrm{s}) \leq p_f(\mathrm{s}')$. □

3 Analytic Modeling

This section describes an analytic model for investigating short message retransmission. Specifically, we derive the expected number of retransmissions for a message delivery and the expected delivery delay. As described in Fig. 2, the terminating UE cannot be accessed from the mobile network in the disconnected periods, and the UE fails to receive the short message. Fig. 3 redraws Fig. 2 to illustrate the disconnected and the connected periods for a terminating UE. In this figure, $t_{d,j} = \tau_{c,j} - \tau_{d,j}$ is the jth disconnected period, and $t_{c,j} = \tau_{d,j+1} - \tau_{c,j}$ is the jth connected period, where $j \geq 1$. Let $t_{d,j}$ and $t_{c,j}$ be random variables with density functions $f_d(t_{d,j})$ and $f_c(t_{c,j})$, and the Laplace transforms $f_d^*(s)$ and $f_c^*(s)$, respectively. In Fig. 3, a short message is sent to the UE at τ_0 ($\tau_{d,1} < \tau_0 < \tau_{c,1}$) and the UE fails to receive the message. For a retransmission policy $\mathrm{s} = \{0, a_1, a_2, \ldots\}$, the SM-SC retransmits this short message to the UE at times $\tau_1 = \tau_0 + a_1$, $\tau_2 = \tau_0 + a_2$, and so on. The interval between the $(i-1)$th and the ith SMS retransmissions is $t_{r,i} = a_i - a_{i-1}$, which is Exponentially distributed with mean $1/\lambda$.

Suppose that the short message (re)transmissions occur at times τ_0, τ_1, τ_2, τ_3 and τ_4. The terminating UE is disconnected from the network at τ_0, τ_1, τ_2, and τ_3. At τ_4, the UE has reconnected to the network and successfully receives the short message. From Definition 3, $T(\mathrm{s}) = \tau_4 - \tau_0$ is the delivery delay of the short message delivery, and $N(\mathrm{s}) = 4$ is the number of retransmissions. In this section, we derive the expected values $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$.

We consider the scenario where the network retransmits the short message until the UE can successfully receive it; that is, $I \to \infty$. We first compute the probability mass function for $N(\mathrm{s})$. Since the short message arrival at τ_0 can be considered as a random observer of connected and disconnected periods for the terminating UE, $\Pr[N(\mathrm{s}) = 0]$ is the probability that the UE is connected to the network at a random observation point. Let the expected values for $t_{c,j}$ and $t_{d,j}$ be m_c and m_d, respectively. From the alternative renewal theory [12], $\Pr[N(\mathrm{s}) = 0]$ is expressed as

$$\Pr[N(\mathrm{s}) = 0] = \frac{m_c}{m_c + m_d}. \tag{1}$$

The probability $\Pr[N(\mathrm{s}) = n]$ can be derived using the recursive method described below. Let J be the number of connected periods occurring in period (τ_0, τ_4). For $N(\mathrm{s}) > 0$ and from (1), we have

$$\begin{aligned}
&\Pr[N(\mathrm{s}) = n] \\
&\quad = (1 - \Pr[N(\mathrm{s}) = 0]) \\
&\qquad \times \sum_{j=1}^{\infty} \Pr[N(\mathrm{s}) = n | J = j] \Pr[J = j] \\
&\quad = \left(\frac{m_d}{m_c + m_d}\right) \\
&\qquad \times \sum_{j=1}^{\infty} \Pr[N(\mathrm{s}) = n | J = j] \Pr[J = j].
\end{aligned} \tag{2}$$

The SM-SC performs the ith short message retransmission after an Exponential random time $t_{r,i}$ with mean $E[t_{r,i}] = 1/\lambda$. Denote $N(\mathrm{s}, t)$ as the number of retransmissions occurring in period t. Then, $\Pr[N(\mathrm{s}, t) = n]$ follows the Poisson distribution

$$\Pr[N(\mathrm{s}, t) = n] = \left[\frac{(\lambda t)^n}{n!}\right] e^{-\lambda t}. \tag{3}$$

In (2), the probability mass function $\Pr[J = j]$ can be derived as follows: For $j > 1$, we first compute the probability that no SMS retransmission occurs in the first $j-1$ connected periods $t_{c,k}$, where $1 \leq k \leq j-1$, and then compute the probability that an SMS retransmission occurs in the jth connected period. For $j = 1$, we only need to compute the probability that an SMS retransmission occurs in the $t_{c,1}$ period. From (3), $\Pr[J = j]$ is expressed as

$$\begin{aligned}
&\Pr[J=j] \\
&= \prod_{k=1}^{j-1}\left[\int_{t_{c,k}=0}^{\infty} \Pr[N(\mathrm{s},t_{c,k})=0] f_c(t_{c,k}) dt_{c,k}\right] \\
&\quad \times \left[1-\int_{t_{c,j}=0}^{\infty} \Pr[N(\mathrm{s},t_{c,j})=0] f_c(t_{c,j}) dt_{c,j}\right] \\
&= \left\{\prod_{k=1}^{j-1}\left[\int_{t_{c,k}=0}^{\infty} e^{-\lambda t_{c,k}} f_c(t_{c,k}) dt_{c,k}\right]\right\} \\
&\quad \times \left[1-\int_{t_{c,j}=0}^{\infty} e^{-\lambda t_{c,j}} f_c(t_{c,j}) dt_{c,j}\right] \\
&= [f_c{}^*(\lambda)]^{j-1}[1-f_c{}^*(\lambda)].
\end{aligned} \tag{4}$$

In the right-hand side of (2), the term $\Pr[N(\mathrm{s})=n|J=j]$ is derived below. As shown in Fig. 3, $\bar{t}_{d,1}=\tau_{c,1}-\tau_0$ is the interval between the time when the short message arrives and when the UE is connected to the network again. Let $r_d(\bar{t}_{d,1})$ and $r_d^*(s)$ be the density function and the Laplace transform of $\bar{t}_{d,1}$, respectively. From the renewal theory [4], $\bar{t}_{d,1}$ is the residual time of $t_{d,1}$. Therefore, we have

$$r_d^*(s)=\left(\frac{1}{sm_d}\right)[1-f_d^*(s)]. \tag{5}$$

For $J=1$, $\Pr[N(\mathrm{s})=n|J=1]$ is the probability that $n-1$ message retransmissions occur in the $\bar{t}_{d,1}$ period. From (3), $\Pr[N(\mathrm{s})=n|J=1]$ is computed as

$$\begin{aligned}
&\Pr[N(\mathrm{s})=n|J=1] \\
&= \int_{\bar{t}_{d,1}=0}^{\infty} \Pr[N(\mathrm{s},\bar{t}_{d,1})=n-1] r_d(\bar{t}_{d,1}) d\bar{t}_{d,1} \\
&= \int_{\bar{t}_{d,1}=0}^{\infty} \left[\frac{(\lambda \bar{t}_{d,1})^{n-1}}{(n-1)!}\right] e^{-\lambda \bar{t}_{d,1}} r_d(\bar{t}_{d,1}) d\bar{t}_{d,1} \\
&= \left[\frac{(-\lambda)^{n-1}}{(n-1)!}\right]\left[\frac{d^{n-1} r_d^*(s)}{ds^{n-1}}\Big|_{s=\lambda}\right].
\end{aligned} \tag{6}$$

For $J=j>1$ and $0\le l\le n-1$, we first consider that there are l message retransmissions occurring in the jth disconnected period $t_{d,j}$, and $n-l-1$ retransmissions occur in the first $j-1$ disconnected periods. From (3), $\Pr[N(\mathrm{s})=n|J=j]$ is expressed as

$$\begin{aligned}
&\Pr[N(\mathrm{s})=n|J=j] \\
&= \sum_{l=0}^{n-1}\left\{\int_{t_{d,j}=0}^{\infty} \Pr[N(\mathrm{s},t_{d,j})=l] f_d(t_{d,j}) dt_{d,j}\right\} \\
&\quad \times \Pr[N(\mathrm{s})=n-l|J=j-1] \\
&= \sum_{l=0}^{n-1}\left\{\int_{t_{d,j}=0}^{\infty} \left[\frac{(\lambda t_{d,j})^{l}}{l!}\right] e^{-\lambda t_{d,j}} f_d(t_{d,j}) dt_{d,j}\right\} \\
&\quad \times \Pr[N(\mathrm{s})=n-l|J=j-1] \\
&= \sum_{l=0}^{n-1}\left\{\left[\frac{(-\lambda)^{l}}{l!}\right]\left[\frac{d^l f_d{}^*(s)}{ds^l}\Big|_{s=\lambda}\right]\right. \\
&\quad \left.\times \Pr[N(\mathrm{s})=n-l|J=j-1]\right\}.
\end{aligned} \tag{7}$$

Equations (6) and (7) can be used to iteratively compute $\Pr[N(\mathrm{s})=n|J=j]$. Finally, $\Pr[N(\mathrm{s})=n]$ is computed by substituting (4) and (7) into (2). For $N(\mathrm{s})=1$, (7) is simplified to

$$\begin{aligned}
&\Pr[N(\mathrm{s})=1|J=j] \\
&\quad = f_d^*(\lambda)\Pr[N(\mathrm{s})=1|J=j-1] \\
&\quad = r_d^*(\lambda) f_d^*(\lambda)^{j-1}.
\end{aligned} \tag{8}$$

Substituting (5) and (8) into (2) to yield

$$\begin{aligned}
&\Pr[N(\mathrm{s})=1] \\
&= \left(\frac{m_d}{m_c+m_d}\right) \\
&\quad \times \sum_{j=1}^{\infty} \Pr[N(\mathrm{s})=1|J=j]\Pr[J=j] \\
&= \left(\frac{m_d}{m_c+m_d}\right)\sum_{j=1}^{\infty}\left\{\frac{[f_d^*(\lambda) f_c{}^*(\lambda)]^{j-1}}{\lambda m_d}\right\} \\
&\quad \times [1-f_d^*(\lambda)][1-f_c{}^*(\lambda)] \\
&= \left[\frac{1}{\lambda(m_c+m_d)}\right] \\
&\quad \times \left\{\frac{[1-f_c{}^*(\lambda)][1-f_d{}^*(\lambda)]}{1-f_c{}^*(\lambda) f_d^*(\lambda)}\right\}.
\end{aligned} \tag{9}$$

For $N=2$, (7) is rewritten as

$$\begin{aligned}
&\Pr[N(\mathrm{s})=2|J=j] \\
&= \sum_{l=0}^{1}\left\{\left[\frac{(-\lambda)^{l}}{l!}\right]\left[\frac{d^l f_d{}^*(s)}{ds^l}\Big|_{s=\lambda}\right]\right. \\
&\quad \left.\times \Pr[N(\mathrm{s})=n-l|J=j-1]\right\} \\
&= -\lambda\left\{[f_d^*(\lambda)]^{j-1}\left[\frac{dr_d^*(s)}{ds}\Big|_{s=\lambda}\right]\right. \\
&\quad \left.+ (j-1) r_d^*(\lambda)[f_d^*(\lambda)]^{j-2}\left[\frac{df_d^*(s)}{ds}\Big|_{s=\lambda}\right]\right\}.
\end{aligned} \tag{10}$$

Substituting (5) and (10) into (2), $\Pr[N(\mathrm{s})=2]$ is derived as

$$\begin{aligned}
&\Pr[N(\mathrm{s})=2] \\
&= \left(\frac{m_d}{m_c+m_d}\right)\sum_{j=1}^{\infty}\Pr[N(\mathrm{s})=2|J=j]\Pr[J=j] \\
&= \left\{\frac{-\lambda m_d[1-f_c{}^*(\lambda)]}{m_c+m_d}\right\} \\
&\quad \times \left\{\left[\frac{dr_d^*(s)}{ds}\Big|_{s=\lambda}\right]\sum_{j=1}^{\infty}[f_c^*(\lambda) f_d{}^*(\lambda)]^{j-1}\right. \\
&\quad + \left[\frac{r_d^*(\lambda)}{f_d^*(\lambda)}\right]\left[\frac{df_d^*(s)}{ds}\Big|_{s=\lambda}\right] \\
&\quad \left.\times \sum_{j=1}^{\infty}(j-1)[f_c^*(\lambda) f_d{}^*(\lambda)]^{j-1}\right\} \\
&= \left(\frac{1}{m_c+m_d}\right)\left[\frac{1-f_c^*(\lambda)}{1-f_c^*(\lambda) f_d^*(\lambda)}\right] \\
&\quad \times \left\{\frac{1-f_d^*(\lambda)}{\lambda}+\left[\frac{df_d^*(s)}{ds}\Big|_{s=\lambda}\right]\right. \\
&\quad \left.\times \left[1-\frac{f_c^*(\lambda)(1-f_d^*(\lambda))}{1-f_c^*(\lambda) f_d^*(\lambda)}\right]\right\}.
\end{aligned} \tag{11}$$

TABLE 1
Comparison of the Analytic and the Simulation Results (Where $E[t_{r,i}] = 1/\lambda$, $t_{c,j}$ and $t_{d,j}$ are Exponentially Distributed, and $m_c = m_d$)

λ	$E[N(\mathrm{s})]$			$E[T(\mathrm{s})]$			$p_f(\mathrm{s})$ $(I = 5)$		
(Unit: $1/m_c$)	Ana.	Sim.	Diff.*	Ana.	Sim.	Diff.	Ana.	Sim.	Diff.
0.1	1.0500	1.0495	0.05%	10.5000	10.4945	0.05%	1.970%	1.974%	0.21%
0.25	1.1250	1.1239	0.10%	4.5000	4.4966	0.08%	2.650%	2.656%	0.21%
0.5	1.2500	1.2498	0.01%	2.5000	2.4991	0.04%	3.890%	3.901%	0.28%
0.75	1.3750	1.3747	0.02%	1.8333	1.8324	0.05%	5.220%	5.218%	-0.03%
1	1.5000	1.4984	0.11%	1.5000	1.4988	0.08%	6.580%	6.569%	-0.17%
2.5	2.2500	2.2538	-0.17%	0.9000	0.9015	-0.17%	14.230%	14.206%	-0.17%
5	3.5000	3.4984	0.05%	0.7000	0.6998	0.03%	23.130%	23.192%	0. 27%
7.5	4.7500	4.7252	0.52%	0.6333	0.6323	0.16%	28.670%	28.681%	0.04%
10	6.0000	5.9363	1.06%	0.6000	0.6014	-0.23%	32.360%	32.329%	-0.10%

* Ana: Analysis data; Sim: Simulation data; Diff: Difference

If $t_{c,j}$ and $t_{d,j}$ are Exponentially distributed, then (2) can be simplified to

$$\begin{aligned}&\Pr[N(\mathrm{s}) = n]\\ &\quad = \left[\frac{m_c(1+\lambda m_c)^{n-1}}{m_c + m_d}\right]\\ &\qquad \times \left(\frac{m_d}{m_c + m_d + \lambda m_c m_d}\right)^n, \quad n > 0,\end{aligned} \tag{12}$$

and $E[N(\mathrm{s})]$ is expressed as

$$\begin{aligned}&E[N(\mathrm{s})]\\ &\quad = \sum_{n=1}^{\infty} n \Pr[N(\mathrm{s}) = n]\\ &\quad = \left(\frac{m_c}{m_c + m_d}\right) \sum_{n=1}^{\infty} n \left[\frac{m_d^n (1+\lambda m_c)^{n-1}}{(m_c + m_d + \lambda m_c m_d)^n}\right]\\ &\quad = \left(\frac{m_d}{m_c}\right)\left[\frac{m_c + m_d(1+\lambda m_c)}{m_c + m_d}\right].\end{aligned} \tag{13}$$

The expected delivery delay can be expressed as

$$E[T(\mathrm{s})] = E[t_{r,i}]E[N(\mathrm{s})]. \tag{14}$$

If $t_{r,i}$ is Exponentially distributed with mean $1/\lambda$, then from (13) and (14), we have

$$E[T(\mathrm{s})] = \left(\frac{m_d}{\lambda m_c}\right)\left[\frac{m_c + m_d(1+\lambda m_c)}{m_c + m_d}\right]. \tag{15}$$

Based on (1) and (12), the probability $p_f(\mathrm{s})$ that a short message is not successfully delivered within I retransmissions can be derived as

$$p_f(\mathrm{s}) = 1 - \sum_{n=0}^{I} \Pr[N(\mathrm{s}) = n]. \tag{16}$$

If $t_{c,j}$ and $t_{d,j}$ are Exponentially distributed, (16) is rewritten as

$$\begin{aligned}&p_f(\mathrm{s})\\ &\quad = \frac{m_d}{m_c + m_d} - \sum_{n=1}^{I}\left[\frac{m_c(1+\lambda m_c)^{n-1}}{m_c + m_d}\right]\\ &\qquad \times \left(\frac{m_d}{m_c + m_d + \lambda m_c m_d}\right)^n\\ &\quad = \left(\frac{1}{m_c + m_d}\right)\left[m_d - \left(\frac{c m_c}{1+\lambda m_c}\right)\left(\frac{1-c^I}{1-c}\right)\right],\end{aligned} \tag{17}$$

where $c = \frac{m_d(1+\lambda m_c)}{m_c + m_d + \lambda m_c m_d}$.

The above analytic model is validated against simulation experiments. Details of the simulation model are described in Appendix A. Table 1 lists $E[N(\mathrm{s})]$, $E[T(\mathrm{s})]$, and $p_f(\mathrm{s})$ values when the connected periods and the disconnected periods have Exponential distributions. Table 2 lists the $\Pr[N(\mathrm{s}) = n]$ values when the connected periods and the disconnected periods have Gamma distributions. Both tables indicate that the discrepancies between the analytic results (specifically, (1), (9), (12), (13), (15), and (17)), and the simulation data are within 2 percent.

4 Numerical Examples

This section uses numerical examples to investigate the performance of the SMS retransmission policies. We show how the retransmission rate λ, the expected connected (disconnected) periods $m_c(m_d)$, and the variances $v_c(v_d)$ affect the expected number $E[N(\mathrm{s})]$ of the SMS retransmissions (using (13)) and the expected delivery delay $E[T(\mathrm{s})]$

TABLE 2
Comparison of the Analytic and the Simulation Results (Where $E[t_{r,i}] = 1/\lambda$, $t_{c,j}$ and $t_{d,j}$ are Gamma Distributed, and $m_c = m_d = \frac{1}{2\lambda}$)

$v_c = v_d$	$\Pr[N(\mathrm{s}) = 0]$			$\Pr[N(\mathrm{s}) = 1]$			$\Pr[N(\mathrm{s}) = 2]$		
(Unit: $1/\lambda^2$)	Ana.	Sim.	Diff.*	Ana.	Sim.	Diff.	Ana.	Sim.	Diff.
0.01	0.50000	0.50001	-0.003%	0.24260	0.24267	-0.027%	0.12730	0.12724	0.051%
0.1	0.50000	0.50006	-0.012%	0.22400	0.22399	0.003%	0.12510	0.12508	0.014%
1	0.50000	0.50002	-0.003%	0.13650	0.13647	0.020%	0.09560	0.09558	0.024%
10	0.50000	0.49828	0.344%	0.03800	0.03804	-0.103%	0.03210	0.03200	0.302%
100	0.50000	0.50029	-0.058%	0.00660	0.00659	0.227%	0.00600	0.00597	0.517%

* Ana: Analysis data; Sim: Simulation data; Diff: Difference

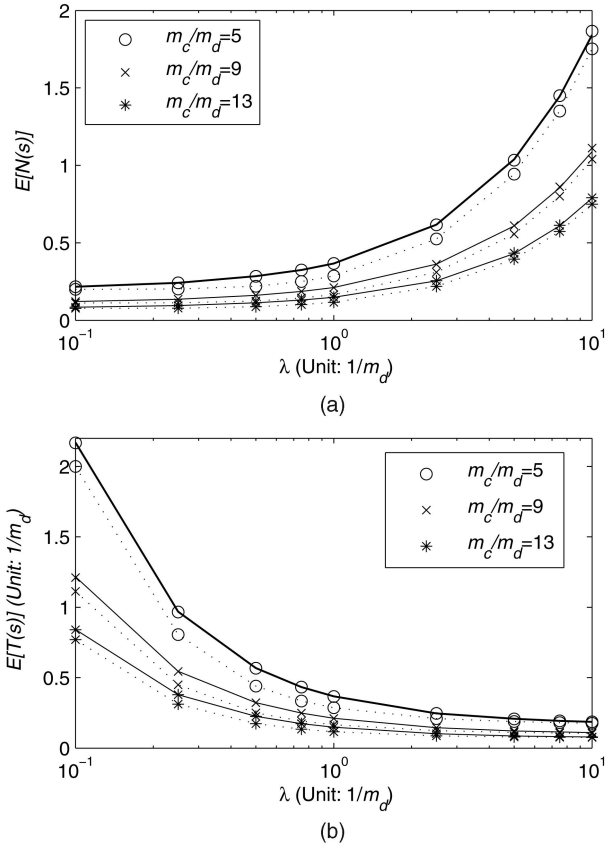


Fig. 4. Effects of λ and the $t_{r,i}$ distribution ($v_c = m_c^2$ and $v_d = m_d^2$; Solid lines: Exponential $t_{r,i}$; Dashed lines: Fixed $t_{r,i}$). (a) Effects on $E[N(\mathrm{s})]$. (b) Effects on $E[T(\mathrm{s})]$.

(using (15)). For demonstration purposes, we consider Gamma-distributed $t_{c,j}$ and $t_{d,j}$. Gamma distribution is selected because of its flexibility in setting various parameters and can be used to fit the first two moments of the sample data. From [7], the Laplace transforms of $f_c(.)$ and $f_d(.)$ are

$$f_c^*(s) = \left[\frac{1}{1+(v_c/m_c)s}\right]^{m_c^2/v_c}, \qquad f_d^*(s) = \left[\frac{1}{1+(v_d/m_d)s}\right]^{m_d^2/v_d}. \tag{18}$$

In (18), m_c^2/v_c and m_d^2/v_d are the shape parameters, and v_c/m_c and v_d/m_d are the scale parameters. The effects of the input parameters are described as follows.

Effects of the $t_{r,i}$ distribution. Fig. 4 plots $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$ for Exponential and Fixed $t_{r,i}$ distribution, where $v_c = m_c^2$ and $v_d = m_d^2$. This figure indicates that the performance of Fixed $t_{r,i}$ (the dashed curves) is slightly better than that of Exponential $t_{r,i}$ (the solid curves).

Effects of the retransmission rate λ. Fig. 4 shows that $E[N(\mathrm{s})]$ is an increasing function of λ, which means that $|\sigma(T(\mathrm{s}))| > j - i$ in Fact 3. $E[T(\mathrm{s})]$ is a decreasing function of λ (Fact 1). When $\lambda < 1/m_d$, $E[N(\mathrm{s})]$ is not significantly affected by the change of λ, and $E[T(\mathrm{s})]$ significantly decreases as λ increases. Conversely, when $\lambda > 1/m_d$, increasing λ significantly increases $E[N(\mathrm{s})]$, but only has insignificant effect on $E[T(\mathrm{s})]$. Therefore, when λ is small, increasing λ can significantly improve the $E[T(\mathrm{s})]$ at the cost of slightly degrading $E[N(\mathrm{s})]$. When λ is large, decreasing λ can significantly improve $E[N(\mathrm{s})]$ at the cost of slightly degrading $E[T(\mathrm{s})]$. Fig. 4 indicates that $0.5/m_d < \lambda < 5/m_d$ is the range that both $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$ do not degrade significantly when λ changes.

Effects of the availability ratio m_c/m_d. Fig. 5 plots $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$ against m_c/m_d, where $v_c = m_c^2$, $v_d = m_d^2$, and $t_{r,i} = 1/\lambda$ is fixed. This figure shows trivial results that both $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$ decrease when m_c/m_d increases. The nontrivial observation is that when $m_c/m_d < 7.5$, the

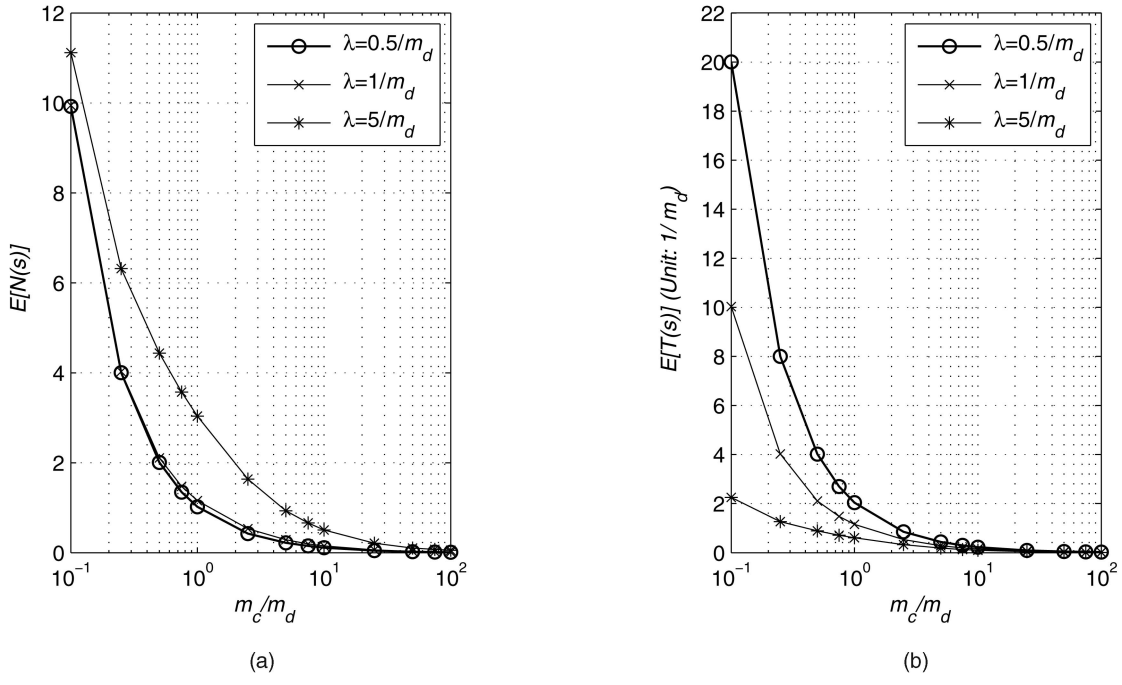


Fig. 5. Effects of m_c/m_d and λ ($v_c = m_c^2$, $v_d = m_d^2$, and $t_{r,i} = 1/\lambda$ is fixed). (a) Effect on $E[N(\mathrm{s})]$. (b) Effect on $E[T(\mathrm{s})]$.

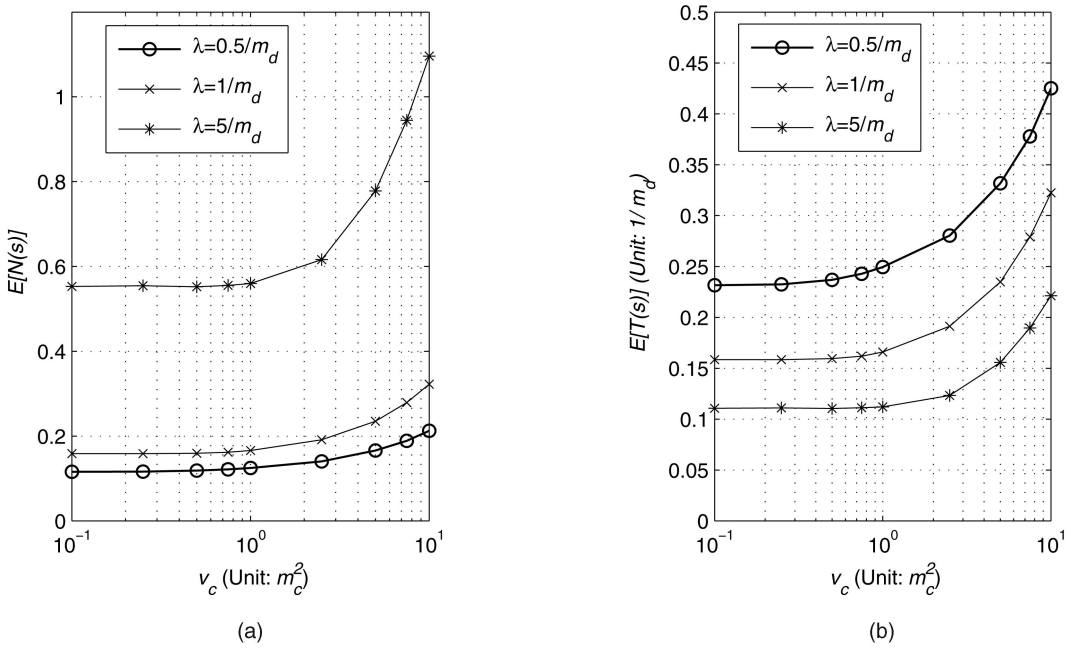


Fig. 6. Effects of v_c ($m_c/m_d = 9$ and $t_{r,i} = 1/\lambda$ is fixed). (a) Effect on $E[N(\mathrm{s})]$. (b) Effect on $E[T(\mathrm{s})]$.

$E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$ performances are significantly degraded as m_c/m_d decreases. In a commercial SMS network, the base station deployment should meet the requirement $m_c/m_d > 10$ to achieve good SMS performance.

Effects of variances v_c and v_d. Fig. 6 plots $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$ as functions of the variance v_c, where $m_c/m_d = 9$, $v_d = m_d^2$, and $t_{r,i} = 1/\lambda$ are fixed. When v_c is large (e.g., $v_c > 2.5m_c^2$), both $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$ increase as the variance v_c increases. This phenomenon is explained as follows: As the variance v_c increases, the number of short connected periods are far more than the number of long connected periods. Therefore, in an arbitrary connected/disconnected cycle, the disconnected time is more likely longer than the connected time [12]. When the SM-SC sends the short message to the terminating UE at τ_0, if the UE is disconnected from the network, message retransmissions are more likely to fall in the disconnected periods than the connected periods. When v_c is larger than $2.5m_c^2$, the performances of both $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$ significantly degrade as v_c increases. When $v_c < m_c^2$, $E[N(\mathrm{s})]$, and $E[T(\mathrm{s})]$ are not affected by the change of v_c. Furthermore, for larger λ, the effect of v_c on $E[N(\mathrm{s})]$ is more significant. On the other hand, the effects of changing v_c on $E[T(\mathrm{s})]$ are similar for different λ values.

Effects of v_d are similar to v_c, and the results are not presented. From Fig. 6, it is important that a mobile operator maintains low v_c and v_d values for good SMS performance.

Effects of I on $p_f(\mathrm{s})$. Fig. 7 plots the $p_f(\mathrm{s})$ curves against retransmission rates λ and the maximum number I of retransmissions. It is trivial that $p_f(\mathrm{s})$ decreases as I increases from $\lceil 5\lambda \rceil$ to $\lceil 10\lambda \rceil$. We observe that $p_f(\mathrm{s})$ decreases as λ increases. This effect is consistent with Fact 4 (i.e., if $\mathrm{s}' \subseteq \mathrm{s}$, then $p_f(\mathrm{s}) \leq p_f(\mathrm{s}')$). This figure also indicates that when λ is sufficiently large (e.g., $\lambda > 3/m_d$ for $I = \lceil 5\lambda \rceil$), increasing λ does not improve $p_f(\mathrm{s})$ performance.

Performance trends observed from the measured data. From the commercial UMTS system of Chunghwa Telecom (CHT), we obtained the output measures $\Pr[N(\mathrm{s}) = n]$, $E[N(\mathrm{s})]$, and $E[T(\mathrm{s})]$ for several retransmission policies. Define $\mathrm{s_k}$ as a policy where a short message is retransmitted for every k minutes. For $k = 5$, 10, 20, and 30, it is clear that $\mathrm{s_{30}} \subseteq \mathrm{s_{10}} \subseteq \mathrm{s_5}$ and $\mathrm{s_{20}} \subseteq \mathrm{s_{10}} \subseteq \mathrm{s_5}$. We have collected the statistics for more than 400,000 SMS deliveries (100,000 deliveries for each policy). Fig. 8 shows $E[N(\mathrm{s_k})]$ and $E[T(\mathrm{s_k})]$ in the "$\circ$" curves. To compare the measured data and the analytic results, we consider a policy $\mathrm{s_e}$ with Exponential $t_{r,i}$, where $E[t_{r,i}]$ is the same as $\mathrm{s_k}$ for various k values. We use (13) and (15) to analytically compute $E[N(\mathrm{s_e})]$ and $E[T(\mathrm{s_e})]$, and the results are shown in the "$\triangleleft$" curves in Fig. 8.

When λ is small, the analytic results are the lower bounds for the measured data. The reason why both results do not exactly match is due to the fact that we are not able

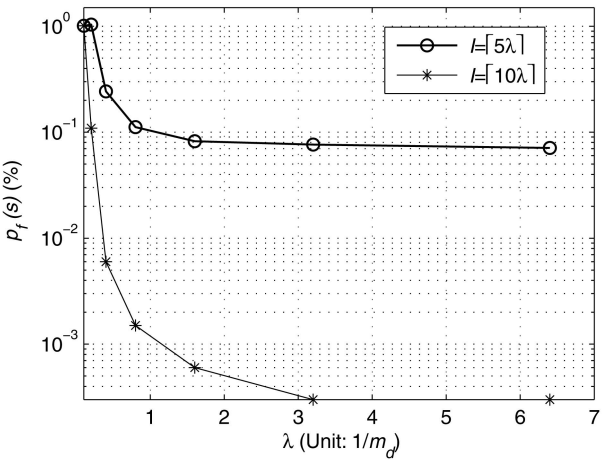


Fig. 7. Effects of I on $p_f(\mathrm{s})$ ($m_c/m_d = 9$ and $t_{r,i} = 1/\lambda$ is fixed).

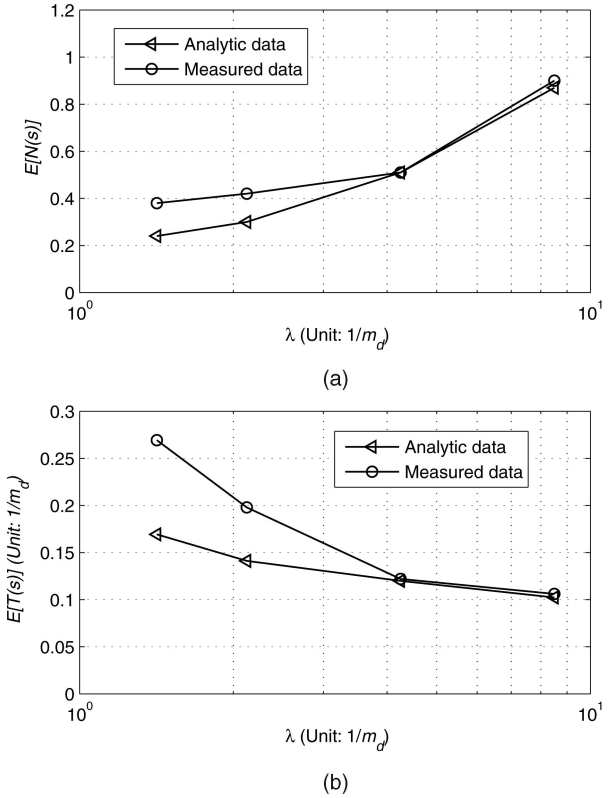


Fig. 8. Performance trends of the analytic data and the measured data ($m_c/m_d = 9$; the parameters for $\mathrm{s_e}$ are $v_c = v_d = m_d^2$ and $E[t_{r,i}] = 1/\lambda$). (a) Effect on $E[N(\mathrm{s})]$. (b) Effect on $E[T(\mathrm{s})]$.

to measure the connected/disconnected distributions. However, the figure shows that the trends are consistent, which follows Fact 1. Fig. 8a shows that $|\sigma(T(\mathrm{s}))| > j - i$, which implies that short periods of outage seldom occur in CHT's network (see the comments on Fact 3).

Fig. 9 plots the probability mass function $\Pr[N(\mathrm{s}) = n]$. From the measurements of the CHT's commercial SMS network, $\Pr[N(\mathrm{s}_{20}) = 0] \approx 0.9$, which means that $m_c/m_d \approx 9$ (see (1)). We also note that $\Pr[N(\mathrm{s}_{20}) = 1] \approx 0.042$. For policy $\mathrm{s_e}$ (with Exponential retransmission delays) where $m_c/m_d = 9$ and $\Pr[N(\mathrm{s_e}) = 1] = 0.042$, we obtain $\lambda m_c = 11.43$ from (12).

With the computed λm_c value, we can derive $\Pr[N(\mathrm{s_e}) = n]$ for $N(\mathrm{s}) > 1$. Fig. 9 shows that $\Pr[N(\mathrm{s_e}) = n]$ has the same trend as $\Pr[N(\mathrm{s}_{20}) = n]$. Therefore, our study indicates that the performance trends for the analytic/simulation models and the measurements from the CHT commercial SMS network are consistent. Note that exact match of analytic and measured data is not possible at this point because we are not able to produce the same SMS network availability condition for the commercial SMS data collection, and as previously mentioned, the exact connected and disconnected periods cannot be measured in commercial operations. A useful conclusion is that (13) and (15) can be used to quickly and roughly estimate the SMS network performance for network planning.

5 Conclusions

In this paper, we studied the short message retransmission policies and derived some facts on short message

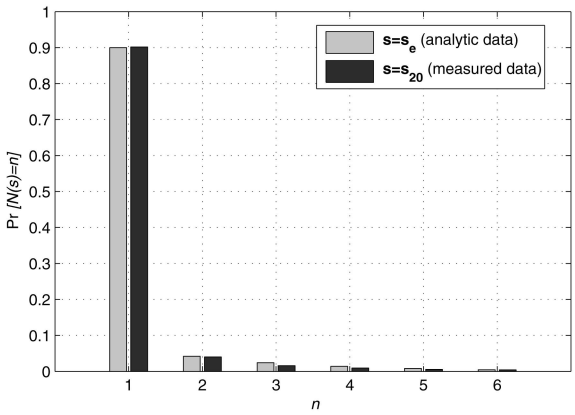


Fig. 9. Performance trends of the analytic data and the measured data.

retransmissions. Then, we proposed an analytic model to investigate the short message retransmission performance in terms of the expected number of short message retransmissions $E[N(\mathrm{s})]$ and the message delivery delay $E[T(\mathrm{s})]$. The analytic model was validated against simulation experiments. We showed how the retransmission rate, the expected values, and the variances of the connected/disconnected period distributions affect $E[N(\mathrm{s})]$ and $E[T(\mathrm{s})]$. The following observations are made in our study:

- The performance of fixed retransmission period is slightly better than that of Exponential retransmission approach. Therefore, it is appropriate that an operator chooses fixed retransmission period value in the delivery policy.
- The retransmission rate should be less than five times of the handset disconnected period and more than half times of that period so that the SMS performance is not significantly affected when the retransmission rate changes.
- In a commercial SMS network, the operator should maintain at least 9 to 1 ratio for the connected/disconnected periods to achieve good SMS performance.
- For each message delivery, it suffices to set the maximum number of retransmissions to be less than 10.

Our study indicates that by selecting appropriate retransmission policy (in particular, the retransmission frequency), the SMS delivery cost can be significantly reduced. We also collected measured data from a commercial mobile telecom network, and observed that the performance trends of the commercial operation are consistent with our analytic/simulation models. Based on our study, a mobile operator can predict the performance trends and select the appropriate parameter values for the SMS retransmission policy in various network conditions.

Appendix

- $f_c(.)$: the density function of $t_{c,j}$.
- $f_c^*(.)$: the Laplace transform of $f_c(.)$.
- $f_d(.)$: the density function of $t_{d,j}$.
- $f_d^*(.)$: the Laplace transform of $f_d(.)$.

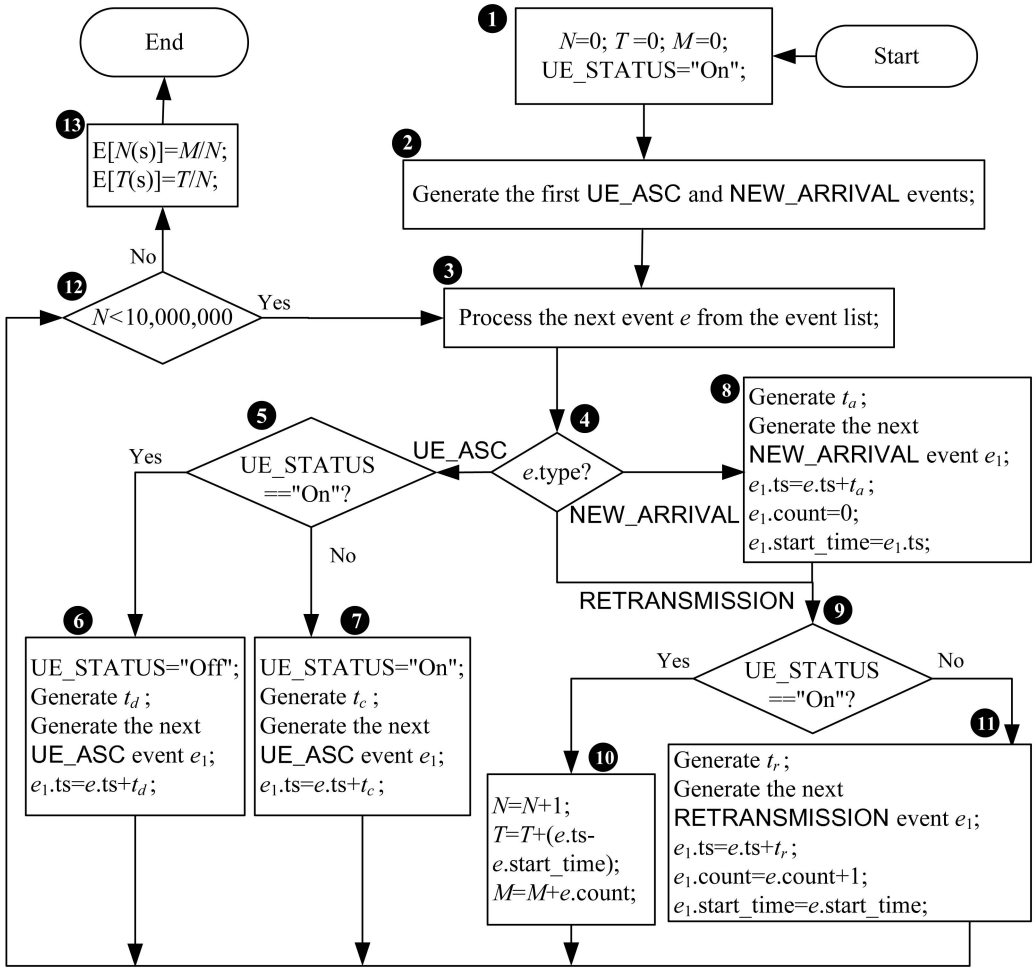


Fig. 10. Simulation flow chart.

- I: the maximum number of short message retransmissions.
- J: the number of connected periods occurring in period (τ_0, τ_4) in Fig. 3.
- m_c: the expected value of $t_{c,j}$.
- m_d: the expected value of $t_{d,j}$.
- $N(\mathrm{s})$: the number of retransmissions of a short message delivery for policy s.
- $N(\mathrm{s}, t)$: the number of message retransmissions occurring in period t for policy s.
- $1/\lambda$: the expected value of the retransmission period $t_{r,i}$.
- $p_f(\mathrm{s})$: the probability of failed SMS delivery for policy s.
- $\Pr[N(\mathrm{s}) = n]$: the probability that a short message is retransmitted for n times in policy s, where $n \geq 0$.
- $r_d(.)$: the density function of $\bar{t}_{d,1}$.
- $r_d^*(.)$: the Laplace transform of $\bar{t}_{d,1}$.
- $t_{c,j}$: the jth connected period, where $j \geq 1$.
- $t_{d,j}$: the jth disconnected period, where $j \geq 1$.
- $\bar{t}_{d,1}$: the residual time of $t_{d,1}$.
- $t_{r,i}$: the period between the $(i-1)$th (re)transmission and the ith retransmission of a short message delivery, where $i \geq 1$.
- $T(\mathrm{s})$: the short message delivery time for policy s or the elapsed time between when the SM-SC starts and when it stops delivering a short message.
- v_c: the variance of $t_{c,j}$.
- v_d: the variance of $t_{d,j}$.

In this paper, we developed a discrete event simulation model for short message retransmission. In this simulation model, three types of events are defined:

- UE_AVAILABILITY_STATUS_CHANGE (UE_ASC): the switching of the UE availability status. These events occur at $\tau_{c,j}$ and $\tau_{d,j}$ in Fig. 3.
- NEW_ARRIVAL: the arrival of a short message (occurring at τ_0 in Fig. 3).
- RETRANSMISSION: retransmission of a short message (occurring at τ_1, τ_2, τ_3, and τ_4 in Fig. 3).

To simulate the status of the UE availability, a variable UE_STATUS is maintained in the simulation. When UE_STATUS is set to "On," the UE is connected to the mobile network (in the $t_{c,j}$ periods) and can receive the short message successfully. When UE_STATUS is set to "Off," the UE is disconnected from the mobile network (in the $t_{d,j}$ periods) and the message transmission fails. Several random number generators are used to produce the connected period t_c, the disconnected period t_d, and the retransmission waiting period t_r. In an experiment, we

simulate N short messages (excluding retransmissions) delivered to the terminating UE. The short message interarrival time is t_a. The reader should note that t_a and t_r are different.

In a UE_ASC event, the following fields are maintained:

- the "type" field identifies the UE_ASC event.
- the "ts" field records the time stamp of the event (when the event occurs).

Besides the "type" and the "ts" fields, the following fields are also maintained in a NEW_ARRIVAL and a RETRANSMISSION events:

- The "start_time" field records the time when the short message is first transmitted (i.e., τ_0 in Fig. 3).
- The "count" field records the number of retransmissions. Note that it is always true that "count=0" in a NEW_ARRIVAL event and "$\text{count} > 0$" in a RETRANSMISSION event.

The events are inserted in an event list, and are deleted/processed from the event list in the nondecreasing time stamp order. In each experiment, $N = 10{,}000{,}000$ short messages are simulated to ensure that the results are stable. The output measures of the simulation are T and M, where

- T is the sum of message delivery delays of the N short messages simulated in the experiment;
- M is the number of retransmissions for the N short messages simulated in the experiment.

These output measures are used to compute $E[N(\text{s})]$ and $E[T(\text{s})]$ as follows:

$$E[N(\text{s})] = \frac{M}{N} \quad \text{and} \quad E[T(\text{s})] = \frac{T}{N}. \tag{19}$$

The simulation flowchart is shown in Fig. 10. Step 1 initializes the variables N, T, M, and UE_STATUS. Step 2 generates the first UE_ASC and NEW_ARRIVAL events, and inserts these events into the event list. In Steps 3 and 4, the next event e in the event list is processed based on its type.

If e.type==UE_ASC, Step 5 checks if UE_STATUS is "On." If so, Step 6 sets UE_STATUS to "Off," and generates the disconnected period t_d and the next UE_ASC event e_1, where the next switching time is e_1.ts=e.ts+t_d. If UE_STATUS=="Off" at Step 5, Step 7 sets UE_STATUS to "On," generates the connected period t_c and the next UE_ASC event e_1, where the next switching time is e_1.ts=e.ts+t_c. The simulation proceeds to Step 12.

If e.type==NEW_ARRIVAL at Step 4, the simulation proceeds to Step 8. Step 8 generates the next NEW_ARRIVAL event e_1 with the interarrival time t_a, where e_1.ts=e.ts+t_a, e_1.count=0 and e_1.start_time=e_1.ts. At Step 9, if UE_STATUS is "On," the message is successfully delivered to the UE, and Step 10 updates the variables N, T, and M. Otherwise (UE_STATUS is "Off" at Step 9), the UE is disconnected from the network, and Step 11 generates the next RETRANSMISSION event e_1 with the waiting period t_r for the next message retransmission, where e_1.ts=e.ts+t_r, e_1.count=e.count+1 and e_1.start_time= e.start_time. The simulation proceeds to Step 12.

If e.type==RETRANSMISSION at Step 4, the simulation proceeds to execute Steps 9-11.

At Step 12, if $N = 10{,}000{,}000$ short messages are successfully delivered, the simulation terminates and the output measures are calculated at Step 13.

ACKNOWLEDGMENTS

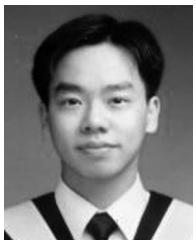
The authors would like to thank the editor and the anonymous reviewers. Their valuable comments have significantly improved the quality of this paper. Their efforts are highly appreciated. The work of S.-I. Sou was supported in part by the Taiwan National Science Council under Contract NSC 97-2218-E-006-020, and in part by the Chung Hwa Telecom project. The work of Y.-B. Lin was sponsored in part by NSC 97-2221-E-009-143-MY3, NSC 97-2219-E009-016, Far Eastone Telecom, Chung Hwa Telecom, ITRI/NCTU Joint Research Center, and MoE ATU.

REFERENCES

[1] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Mobile Application Part (MAP) Specification (Release 8)," Technical Specification 3G TS 29.002 V8.7.0 (2008-09), 2008.
[2] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Technical Realization of the Short Message Service (SMS) (Release 8)," Technical Specification 3G TS 23.040 Version 8.3.0 (2008-09), 2008.
[3] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; Network Architecture (Release 8)," Technical Specification 3G TS 23.002 Version 8.3.0 (2008-09), 2008.
[4] R.G. Gallager, *Discrete Stochastic Processes.* Kluwer Academic Publishers, 1999.
[5] C.-C. Huang-Fu, Y.-B. Lin, and C.-H. Rao, "iP2P: A Peer-to-Peer System for Mobile Devices," to be published in *IEEE Wireless Comm.*
[6] R.S. Ivanov, "Controller for Mobile Control and Monitoring via Short Message Services," *Proc. Sixth Int'l Conf. Telecomm. in Modern Satellite, Cable and Broadcasting Service,* vol. 1, pp. 108-111, Oct. 2003.
[7] L. Kleinrock, *Queueing Systems: Vol. I—Theory.* John Wiley & Sons, 1976.
[8] P. Lin, S.-H. Wu, C.-M. Chen, and C.-F. Liang, "Implementation and Performance Evaluation for a Ubiquitous and Unified Multimedia Messaging Platform," *Springer Wireless Networks,* vol. 15, pp. 163-176, 2009.
[9] C. Markett et al., "Using Short Message Service to Encourage Interactivity in the Classroom," *Computers & Education,* vol. 46, no. 3, pp. 280-293, 2006.
[10] M.A. Mohammad and A. Norhayati, "A Short Message Service for Campus Wide Information Delivery," *Proc. Fourth Nat'l Conf. Telecomm. Technology,* pp. 216-221, Jan. 2003.
[11] C.H. Rao, D.-F. Chang, and Y.-B. Lin, "iSMS: An Integration Platform for Short Message Service and IP Networks," *IEEE Network,* vol. 15, no. 2, pp. 48-55, Mar./Apr. 2001.
[12] S.M. Ross, *Stochastic Processes.* John Wiley & Sons, 1996.
[13] S.I. Sou, Y.-B. Lin, Q. Wu, and J.-Y. Jeng, "Modeling of Prepaid Mechanism of VoIP and Messaging Services," *IEEE Trans. Vehicular Technology,* vol. 56, no. 3, pp. 1434-1441, 2007.

Sok-Ian Sou received the BS, MS, and PhD degrees from the National Chiao Tung University (NCTU), Taiwan, in 1997, 2004, and 2008, respectively. She is an assistant professor in the Department of Electrical Engineering, National Cheng Kung University (NCKU), Taiwan. Her current research interests include design and analysis of mobile communication services, mobile computing, and performance modeling. She is the coauthor with Yi-Bing Lin of the book *Charging for Mobile All-IP Telecommunications* (Wiley, 2008). She is a member of the IEEE.

Yi-Bing Lin is the dean and chair professor of the College of Computer Science, National Chiao Tung University (NCTU). He is a senior technical editor of *IEEE Network*. He serves on the editorial boards of the *IEEE Transactions on Wireless Communications* and the *IEEE Transactions on Vehicular Technology*. He has been a general or program chair for prestigious conferences including ACM MobiCom 2002. He is a guest editor for several first-class journals including the *IEEE Transactions on Computers*. He is the author of the books *Wireless and Mobile Network Architecture* (Wiley, 2001), *Wireless and Mobile All-IP Networks* (Wiley, 2005), and *Charging for Mobile All-IP Telecommunications* (Wiley, 2008). He is listed in ISIHighlyCited.Com among the top 1 percent most cited computer science researchers. He received numerous research awards including the 2005 NSC Distinguished Researcher award and the 2006 Academic Award from the Ministry of Education. He is a fellow of the IEEE, ACM, the AAAS, and the IET.



Chao-Liang Luo received the MSCSIE degree from the National Tsing Hua University (NTHU), Taiwan, in 2001. He is currently working toward the PhD degree in the Department of Computer Science and Engineering, National Chiao Tung University. In 2001, he joined the Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., and was involved in the implementation of value-added services in mobile networks. In 2005, he was with the short message service team. Since then, he has been involved in the design of the Next Generation Network (NGN), mobile packet switched data and multimedia services, and the study of mobile network evolution. His research interests include the design and analysis of personal communications services network, 3G networks, wireless Internet, mobile computing, and performance modeling.