# Performance of Linear-Type Mobile Data Transmission

Chen-Yuan Chuang, Yi-Bing Lin, *Fellow, IEEE,* and Yu-Tien Yeh

***Abstract*—In Next Generation Network (NGN), IP Multimedia Subsystem (IMS) plays an important role to offer IP-based multimedia services. Based on Chunghwa Telecom's NGN/IMS, we develop linear-type data transmission applications such as Bulletin Board System (BBS) or forum-based websites to deliver new articles to the mobile users. We propose the push-*N* method that pushes every *N* articles from the BBS server to the user. When the user actually connects to the BBS, the not-yet-pushed messages are pulled by the user. Our study indicates that by selecting appropriate *N* values, push-*N* can balance against the push and the pull operations, and therefore improves the user access experience without wasting too much network resources.**

***Index Terms*—Forum-based website applications, mobile data transmission, IP multimedia subsystem (IMS).**

## I. Introduction

IN *Next Generation Network* (NGN), *IP Multimedia Subsystem* (IMS) supports IP-based broadband multimedia services through mobile devices [1], [2]. Chunghwa Telecom (CHT), Taiwan, has deployed one of the largest commercial NGN/IMS in Asia to provision enhanced voice, video, and Internet-based multimedia services. The current CHT NGN/IMS capacity can accommodate commercial operation for about five hundred thousand subscribers. Figure 1 illustrates a simplified NGN/IMS network architecture. In this figure, the NGN/IMS network (Figure 1 (a) and (b)) provides a standard approach for access to the application network (Figure 1 (d)) from *User Equipment* (UE) such as a smart phone (Figure 1 (3)) via the *Packet-Switched* (PS) domain of *Public Land Mobile Network* (PLMN; Figure 1 (c)). An example of PLMN-PS network is *Universal Mobile Telecommunications System* (UMTS) with the *General Packet Radio Service* (GPRS) core network [1]. In NGN/IMS, the *Call Session Control Function* (CSCF; Figure 1 (1)) is responsible for call control. The *Session Border Controller* (SBC; Figure 1 (2)) supports the functions of topology hiding and NAT/firewall traversal. Details of a more complete NGN/IMS architecture can be found in [3], [4].

Manuscript received December 16, 2010; revised February 24 and April 14, 2011; accepted April 16, 2011. The associate editor coordinating the review of this letter and approving it for publication was S. Mao.

C.-Y. Chuang is with Chunghwa Telecom, Taipei, Taiwan. He is also with the Department of Computer Science, National Chiao Tung University, Taiwan. (e-mail: zzy@cht.com.tw).

Y.-B. Lin is with the Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan. He is also with the Institute of Information Science and the Research Center for Information Technology Innovation, Academia Sinica, Nankang, Taipei, Taiwan, and a visiting professor at King Saud University (e-mail: liny@csie.nctu.edu.tw).

Y.-U. Yeh is with the Department of Computer Science, National Chiao Tung University, Taiwan (e-mail: ytyeh@cs.nctu.edu.tw).

Y.-B. Lin's work was supported in part by NSC 98-2219-E-009-016, 98-2221-E-009-059-MY2, 97-2221-E-009-143-MY3, Chunghwa Telecom, IBM, ITRI and NCTU joint research center, and MoE ATU plan.

Digital Object Identifier 10.1109/TWC.2011.060711.102227

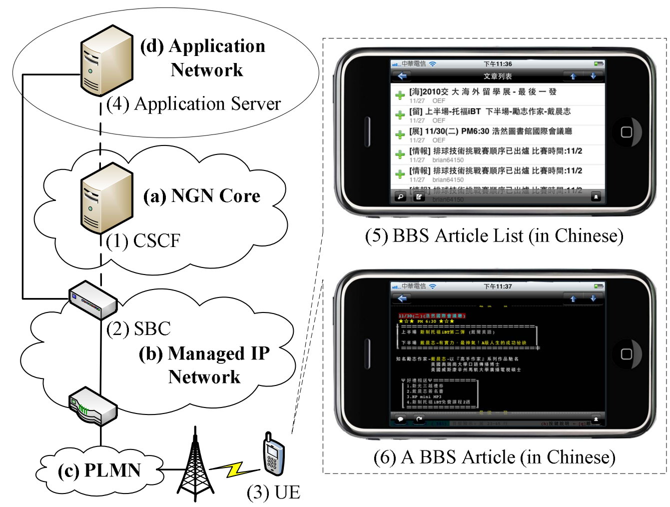



Fig. 1. A simplified NGN/IMS architecture.

Based on CHT's NGN/IMS, we have developed mobile data transmission applications that allow pull or push of information from an application server (Figure 1 (4)) to a UE (Figure 1 (3)). Basically, there are two types of wireless data access. In the *locality-type access*, a user may access data with locality, and the cache mechanisms are typically used to speed up the data access [5], [6]. In the *linear-type access*, the data are sequentially pushed to the UE. Examples of applications for linear-type access are Bulletin Board System (BBS) [7], forum-based websites or e-book applications [8]. Note that multimedia data such as large video clips do not fit the linear-type access model and is out of the scope of this paper.

In a campus BBS, information (such as campus news) is pulled by a user when he/she connects to the BBS. Alternatively, the application server may push the information to the user immediately when an article is published. We have developed a mobile campus bulletin board system called iNCTU BBS that posts Chinese articles. With 3G service in Figure 1, the iNCTU BBS articles can be accessed through iPhone (Figure 1 (5) and (6)).

To our knowledge, performance of linear-type access has not been investigated in the literature. For this type access, the performance of a mobile data transmission application is significantly affected by the operation of "delivering articles to the user". Under the NGN/IMS architecture, the articles are delivered through protocols such as telnet and http. These protocols are first established between the UE and the application server through path (3)-(c)-(2)-(1)-(4) in Figure 1. Then the articles are delivered through path (4)-(2)-(c)-(3). In the http-based pull operation, all new articles are packed in one message and then delivered to the UE when the user attempts to access these articles. In this operation, the user may need to wait for a long time before all articles are actually transmitted to the UE. In the push operation, a new article is delivered to the UE immediately after it is published. Since all articles were already delivered to the UE in advance, the user does not need to wait to access the articles, and therefore, is more user-friendly. However, the push operation requires one message delivery per article. From the network aspect, the push operation consumes more resources; that is, more time and bandwidth consumption for path setup and message header transmission are required for same amount of content delivery. To balance against the consumption of network resources and user waiting time, this paper proposes the *push-N* method, and develops an analytic model to study the performance of this method.

## II. Data Transmission of the Push-*N* Method

This section describes the push-*N* method. We assume that every article is downloaded to the UE for exactly one time. If a BBS article is modified, then it is considered as a new article. In the push-*N* method, the server delivers articles to the UE in two events:

**Event 1 (Periodic Push):** When the server receives *N* new articles, it pushes these articles to the UE.

**Event 2 (Pull):** The user attempts to access the articles. All new articles that have not been pushed in Event 1 are delivered to the user.

Figure 2 illustrates the timing diagram of push-*N* (where *N*=3) in two consecutive user accesses at $t_{a,1}$ and $t_{a,2}$ (marked by ●). The application server is updated when new articles arrive at $t_{u,0}$, $t_{u,1}$, $t_{u,2}$ and $t_{u,4}$ (marked by ○). When the *N*-th article arrives at $t_{u,3}$, the *N* articles collected at the server are pushed to the user (marked by $\Delta$). In Figure 2, $t_{u,0} < t_{a,1} < t_{u,1} < t_{u,2} < t_{u,3} < t_{u,4} < t_{a,2}$. When a UE attempts to access the articles (at $t_{a,2}$ in Figure 2), the not-yet-pushed article (i.e., article (d)) is pulled by the UE. Therefore, the UE only needs to wait for the transmission of less than *N* articles. By periodically pushing new articles in advance, the user does not need to wait for a long access time. Furthermore, the user may read those already downloaded articles (articles (a), (b) and (c)) while the UE is pulling the rest of articles, which also reduces user waiting time.

In the push-*N* method, it is clear that the size of *N* affects the performance of linear-type mobile data transmission. When $N = 1$, push-*N* is the same as the push operation. On the other hand, $N \to \infty$ results in the pull operation. For a specific data transmission application, it is essential to select an appropriate *N* value such that the net cost of user waiting time and extra network resource consumption is optimized.

## III. An Analytic Model for Push-*N*

This section investigates the performance of the push-*N* method. From Figure 2, the inter-access interval is $\tau_a = t_{a,2} - t_{a,1}$. Let the user accesses be a Poisson process, then $\tau_a$ is an exponential random variable with the mean $1/\lambda_a$. The inter-update intervals are $\tau_{u,i} = t_{u,i+1} - t_{u,i}$ $(i = 0, 1, 2 \ldots)$, which are assumed to be an i.i.d. random variable with the density function $f_u(\cdot)$, the mean $1/\lambda_u$, the variance $V_u$, and the Laplace transform $f_u^*(\cdot)$. Since the user accesses are a Poisson process, the previous access at $t_{a,1}$ is a random observer of the inter-update interval $\tau_{u,0} = t_{u,1} - t_{u,0}$. From the residual life theorem [9], the residual life of the inter-update interval is $\tau_r = t_{u,1} - t_{a,1}$ with the Laplace transform $r^*(\cdot)$ expressed as

$$r^*(s) = \left(\frac{\lambda_u}{s}\right)[1 - f_u^*(s)] \tag{1}$$

Let $n$ be the number of article arrivals during the inter-access period $\tau_a$ (in Figure 2, $n$ = 4). If $n \geq N$, then Event 1 occurs. In Figure 2, $N = 3$, and Event 1 occurs at $t_{u,3}$; i.e., the application server pushes three new articles to the user at $t_{u,3}$. Let $m$ be the number of the pushed messages during $\tau_a$ (in Figure 2, $m = 1$). Let $n_a$ be the number of new articles pulled by the user at $t_{a,2}$; i.e., Event 2 occurs (in Figure 2, $n_a = 1$). It is clear that

$$\begin{aligned} m &= \left\lfloor \frac{n}{N} \right\rfloor \\ n_a &= n - mN \end{aligned} \tag{2}$$

Based on the above description, we now elaborate on the cost of the push-*N* method. In this method, the push cost $C_1$ is experienced by the network, and the pull cost $C_2$ is mainly experienced by the user. These two costs are typically weighted by a factor $\alpha$, $0 \leq \alpha \leq 1$. That is, for the period $\tau_a$, the net cost $C$ of the push-*N* method is expressed as

$$C = \alpha C_1 + C_2 \tag{3}$$

In (3), a smaller $\alpha$ value means that fast real-time transmission (for Event 2) is more important to reduce user waiting time. The costs $C_1$ and $C_2$ can be measured by the message delay or the transmission bandwidth. To make things simple, we consider the delays that can be actually measured from the commercial mobile telecom network. The transmission for a message delivery consists of two delays:

**Delay 1:** the elapsed time to set up path (4)-(2)-(c)-(3) in Figure 1 and to deliver the message header from the server to the UE. Note that after the BBS application has been executed at the UE, path (4)-(2)-(c)-(3) is established. However, radio link (c)-(3) may be released after a short timeout period (while path (4)-(2)-(c) is still maintained); radio link (c)-(3) is re-established at Events 1 or 2, which contributes to a part of Delay 1 at push or pull operations.

**Delay 2:** the delay to deliver new articles (packaged as the payload in one message) from the server to the UE.

Let $\tau_h$ be Delay 1 and $\tau_c$ be the delay to deliver one article. If there are $k$ articles delivered in one message, then Delay 2 is $k\tau_c$. For a push message, the transmission delay is $\tau_h + N\tau_c$, and the push cost $C_1$ between two consecutive user accesses (i.e., in period $\tau_a$) is

$$C_1 = E[m](\tau_h + N\tau_c) \tag{4}$$

For a pull message in period $\tau_a$, the cost $C_2$ is

$$C_2 = \tau_h + E[n_a]\tau_a \tag{5}$$

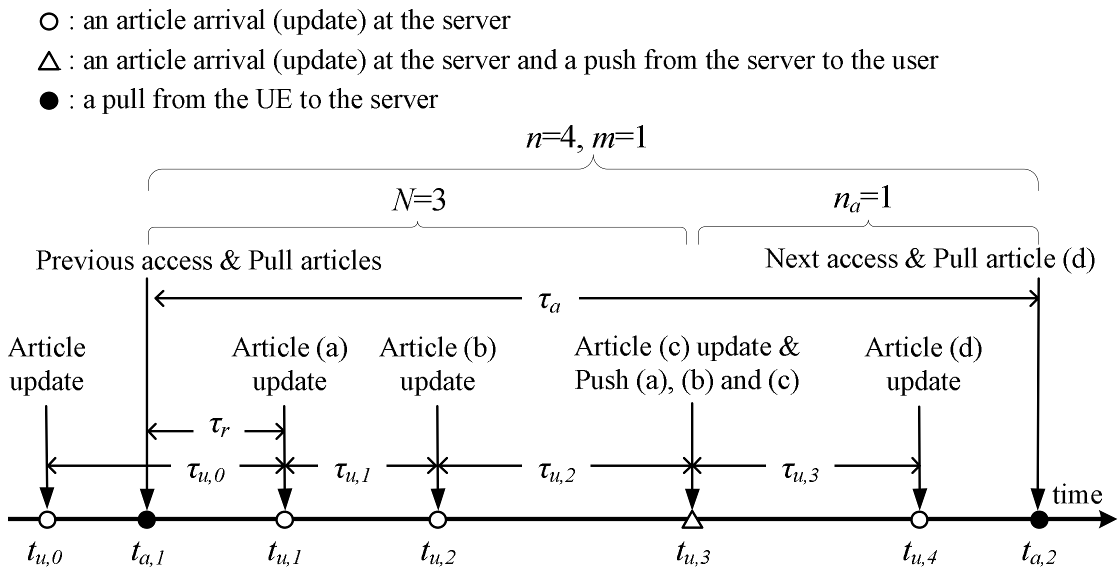



Fig. 2. Timing diagram for Push-*N* (pushes between two consecutive pulls).

$E[n_a]$ and $E[m]$ are derived as follow. We first define a random variable

$$T_{u,k} = \tau_r + \sum_{i=1}^{k-1} \tau_{u,i} \tag{6}$$

with the density function $f_{u,k}(\cdot)$ and the Laplace transform $f^*_{u,k}(\cdot)$. From (1), (6) and the convolution property of the Laplace transform, we have

$$\begin{aligned} f^*_{u,k}(s) &= r^*(s) \prod_{i=1}^{k-1} f^*_u(s) \\ &= \left(\frac{\lambda_u}{s}\right) [f^*_u(s)]^{k-1} [1 - f^*_u(s)] \end{aligned} \tag{7}$$

To obtain $E[n_a]$, we first derive the probability $\Pr[n \geq k]$ that at least $k$ new articles arrive during $\tau_a$. From (6), it is clear that for $k \geq 1$, $\Pr[n \geq k] = \Pr[\tau_a > T_{u,k}]$, which is derived as

$$\begin{aligned} \Pr[\tau_a > T_{u,k}] &= \int_{T_{u,k}=0}^{\infty} f_{u,k}(T_{u,k}) \\ &\quad \times \int_{\tau_a = T_{u,k}}^{\infty} \lambda_a e^{-\lambda_a \tau_a} d\tau_a dT_{u,k} \\ &= \int_{T_{u,k}=0}^{\infty} f_{u,k}(T_{u,k}) e^{-\lambda_a T_{u,k}} dT_{u,k} \\ &= \left(\frac{\lambda_u}{\lambda_a}\right) [f^*_u(\lambda_a)]^{k-1} [1 - f^*_u(\lambda_a)] \end{aligned} \tag{8}$$

From (8) and for $k \geq 1$, $\Pr[n = k]$ is computed as

$$\begin{aligned} \Pr[n = k] &= \Pr[n \geq k] - \Pr[n \geq k+1] \\ &= \left(\frac{\lambda_u}{\lambda_a}\right) [f^*_u(\lambda_a)]^{k-1} [1 - f^*_u(\lambda_a)]^2 \end{aligned} \tag{9}$$

From (2), the probability that $n_a = k$ is

$$\Pr[n_a = k] = \sum_{m=0}^{\infty} \Pr[n = k + mN], \, k \geq 1 \tag{10}$$

From (9) and (10), we have

$$\Pr[n_a = k] = \left(\frac{\lambda_u}{\lambda_a}\right) \left\{ \frac{[f^*_u(\lambda_a)]^{k-1} [1 - f^*_u(\lambda_a)]^2}{1 - [f^*_u(\lambda_a)]^N} \right\}, \quad k \geq 1 \tag{11}$$

From (11), the expected value of $n_a$ is

$$\begin{aligned} E[n_a] &= \sum_{k=1}^{N-1} k \Pr[n_a = k] \\ &= \left(\frac{\lambda_u}{\lambda_a}\right) \Bigg\{ 1 - N [f^*_u(\lambda_a)]^{N-1} \\ &\quad \times \left\{ \frac{1 - f^*_u(\lambda_a)}{1 - [f^*_u(\lambda_a)]^N} \right\} \Bigg\} \end{aligned} \tag{12}$$

From (2), the probability that $m = k$ is expressed as

$$\Pr[m = k] = \sum_{n_a=0}^{N-1} \Pr[n = kN + n_a], \, k \geq 1 \tag{13}$$

Substitute (9) into (13) to yield

$$\begin{aligned} \Pr[m = k] &= \left(\frac{\lambda_u}{\lambda_a}\right) [f^*_u(\lambda_a)]^{kN-1} [1 - f^*_u(\lambda_a)] \\ &\quad \times \left\{ 1 - [f^*_u(\lambda_a)]^N \right\}, \, k \geq 1 \end{aligned} \tag{14}$$

From (14), we derive the expected value of $m$ as

$$\begin{aligned} E[m] &= \sum_{k=1}^{\infty} k \Pr[m = k] \\ &= \left(\frac{\lambda_u}{\lambda_a}\right) \left\{ \frac{[f^*_u(\lambda_a)]^{N-1} [1 - f^*_u(\lambda_a)]}{1 - [f^*_u(\lambda_a)]^N} \right\} \end{aligned} \tag{15}$$

Assume that $\tau_u$ is a Gamma random variable with the mean $1/\lambda_u$, the variance $V_u$, and the Laplace transform

$$f^*_u(s) = \left(\frac{1}{V_u \lambda_u s + 1}\right)^{\frac{1}{V_u \lambda_u^2}} \tag{16}$$

We consider the Gamma distribution because this distribution is widely used in telecom modeling; see [10] and the references there in. From (12) and (16), $E[n_a]$ is re-written as

$$E[n_a] = \left(\frac{\lambda_u}{\lambda_a}\right)\left\{1 - N\left[\frac{1-(V_u\lambda_u\lambda_a+1)^{\frac{1}{V_u\lambda_u^2}}}{1-(V_u\lambda_u\lambda_a+1)^{\frac{N}{V_u\lambda_u^2}}}\right]\right\} \quad (17)$$

Similarly, substitute (16) into (15) to yield

$$E[m] = \left(\frac{\lambda_u}{\lambda_a}\right)\left\{\frac{1-(V_u\lambda_u\lambda_a+1)^{\frac{1}{V_u\lambda_u^2}}}{1-(V_u\lambda_u\lambda_a+1)^{\frac{N}{V_u\lambda_u^2}}}\right\} \quad (18)$$

Equations (17) and (18) are validated against the discrete event simulation experiments. In the simulation, the arrival of articles and user accesses are generated as discrete events following the same methodology as the one developed in [11]. The discrepancies between the analytic and simulation results are within 0.8%.

## IV. Numerical Examples

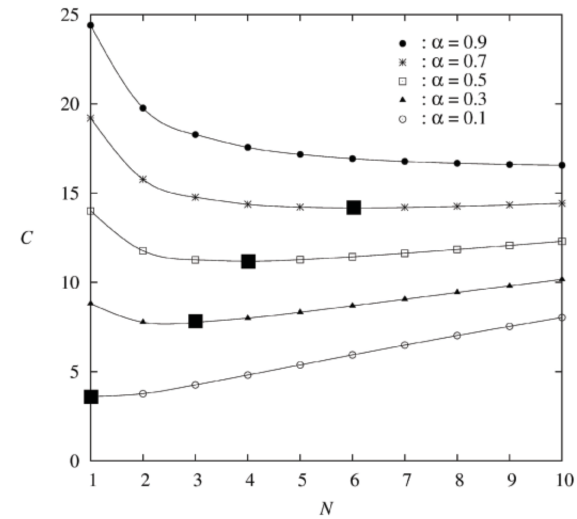
This section uses numerical examples to investigate how parameters $N$ (the number of articles in a push message), $\lambda_a$ (the user access rate), $\lambda_u$ (the update or article arrival rate), $V_u$ and $\tau_c/\tau_h$ (the article size) affect the push-*N* method.

The parameters $\tau_h$ and $\tau_c$ are actually measured from the commercial 3G service. From 84 measurements, the average $\tau_h$ value is 444.6ms. The average $\tau_c$ for the 25K-byte and the 50K-byte articles are 747.05ms and 1457.99ms, respectively. To simplify our discussion, $\tau_c$ is normalized by $\tau_h$ with a new notation $\tau_c^* = \tau_c/\tau_h$. Therefore $\tau_c^* \approx 1.6$ for the 25K-byte article and $\tau_c^* \approx 3.2$ for the 50K-byte article. The effects of the input parameters are discussed as follows.

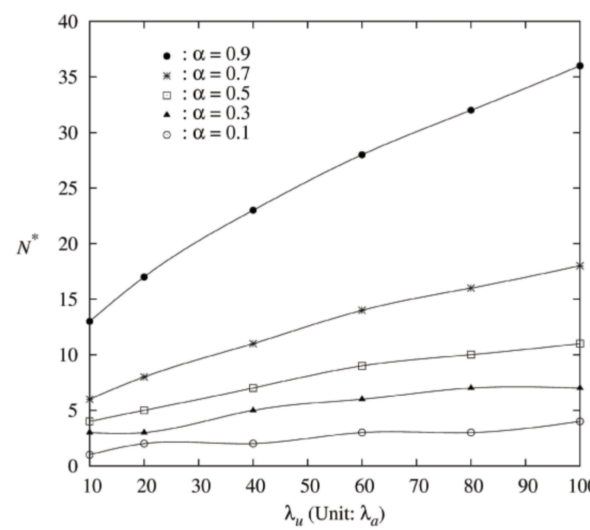
**Effects of $\alpha$ and $N$:** Figure 3(a) plots the net cost $C$ (see Eq. (3)) as a function of $N$ and $\alpha$, where $\lambda_u = 10\lambda_a$, $\tau_c^* = 1.6$ and $V_u = 1/\lambda_u^2$. This figure indicates when $\alpha$ is large (e.g., $\alpha > 0.9$), the net cost $C$ decreases as $N$ increases. When $\alpha$ is small (e.g., $\alpha \leq 0.1$), $C$ increases as $N$ increases. When $\alpha$ is in the medium range $[0.3, 0.7]$, $C$ decreases and then increases. In this figure, a ■ mark represents the minimal net cost $C$ for a specific $\alpha$ value, which occurs at $N^*$ (called the optimal $N$ value). It is apparent that $N^*$ increases as $\alpha$ increases. That is, if reducing user waiting time is "not so important", then a large $N$ should be chosen.

**Effects of article arrival rate $\lambda_u$:** Figure 3(b) plots $N^*$ as a function of $\lambda_u$ and $\alpha$, where $\tau_c^* = 1.6$ and $V_u = 1/\lambda_u^2$. The figure shows the trivial result that $N^*$ increases as $\lambda_u$ increases. The non-trivial observation is that when $\lambda_u$ is large (e.g., $\lambda_u = 100\lambda_a$), $N^*$ significantly increases as $\alpha$ increases. That is, if the BBS articles are frequently generated, we should select a large $N$, especially when $\alpha$ is large.

**Effects of $\tau_c^*$ or artical size:** Figure 4(a) plots $N^*$ as a function of $\lambda_u$ and $\tau_c^*$, where $\alpha = 0.7$ and $V_u = 1/\lambda_u^2$. Since $\tau_c^*$ proportionally increases as the article size increases, $\tau_c^*$ can be used to investigate the effect of article size. This figure indicates that $N^*$ decreases as the article size increases. As the article size increases, the path setup and header transmission cost for pushing per-byte content decreases. Therefore, larger-size BBS articles should be pushed to the user more frequently (i.e., a small $N$ should be selected when the article size is large). A non-trivial observation is that when $\lambda_u$ is larger (i.e., new articles are more frequently generated), $N^*$ is more sensitive to the change of $\tau_c^*$ (i.e., the article size).



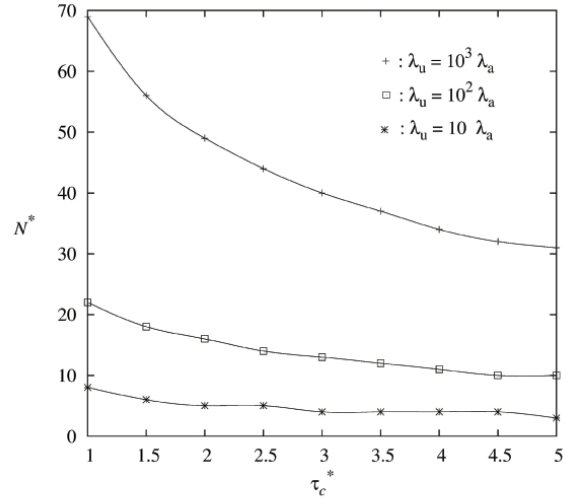

(a) Effect of $N$ ($\lambda_u = 100\lambda_a$)



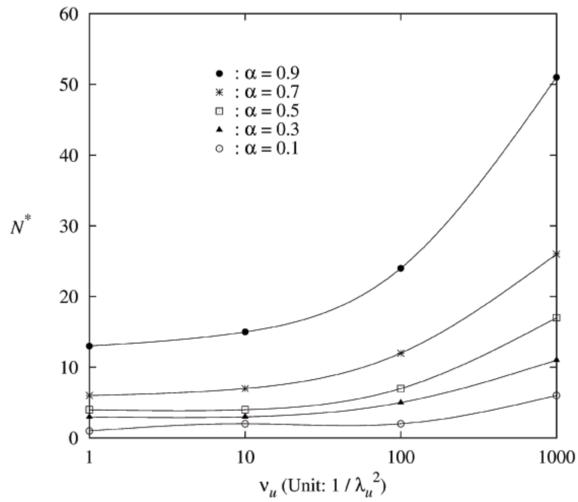

(b) Effect of $\lambda_u$

Fig. 3. Effects of $N$, $\alpha$ and $\lambda_u$ ($\tau_c^* = 1.6$ and $V_u = 1/\lambda_u^2$).

**Effects of $V_u$:** Figure 4(b) plots $N^*$ as a function of $V_u$ and $\alpha$, where $\lambda_u = 10\lambda_a$ and $\tau_c^* = 1.6$. When the variance $V_u$ is larger, the article arrivals become more irregular. This figure indicates that $N^*$ increases as $V_u$ increases. When $V_u$ becomes large, more long $\tau_{u,i}$ intervals and short $\tau_{u,i}$ intervals are observed. Consider an extreme case where $V_u$ is huge. That is, between two consecutive user accesses, we find either many updates ($\lambda_u$ is large) or no updates ($\lambda_u \approx 0$). In the latter case, $C = 0$ no matter what $N$ value is chosen. In the former case, a larger $N$ results in smaller $C$ as shown in Figure 3(b). Therefore, if the BBS articles are generated irregularly, we should select a large $N$.




(a) Effect of $\tau_c^*$ and $\lambda_u$ ($\alpha = 0.7$ and $V_u = 1/\lambda_u^2$)




(b) Effect of $V_u$ and $\alpha$ ($\tau_c^* = 1.6$ and $\lambda_u = 10\lambda_a$)

Fig. 4. Effects of $\tau_c^*$ , $\lambda_u$, $V_u$ and $\alpha$.

## V. Conclusions

In linear-type mobile data transmission applications such as mobile BBS, the articles can be pushed to or pulled by a user. Real-time push operations provide best user access experience at the cost of consuming extra network resources. This paper proposed the push-$N$ method that balances against the push and the pull operations to optimize the performance of linear-type mobile data transmission. We utilized an analytic model to investigate how to select the optimal $N$ values to minimize the net cost. It is clear that if reducing user waiting time (improving user experience) is important, then a small $N$ should be selected. Our study further indicated that

1) When new articles are frequently generated, a large $N$ should be selected.
2) If the BBS publishes larger-size articles, a small $N$ should be selected.
3) If the BBS articles are generated irregularly, a large $N$ should be selected.

Based on the above guidelines, our study can quantitatively suggest how to select the optimal $N$ value for specific input parameter values.

## References

[1] Y.-B. Lin and A.-C. Pang, *Wireless and Mobile All-IP Networks*. John Wiley & Sons, Inc., 2005.

[2] ETSI, Telecommunications and Internet Converged Services and Protocols for Advanced Networking (TISPAN); NGN Functional Architecture; ETSI ES 282 001, Release 3.4.1; ETSI: Sophia Antipolis, France, 2009.

[3] Y.-C. Ho, Y.-B. Lin, R.-H. Liou, and Y.-K. Tu, "Implementing value added applications in next generation networks," *Future Internet*, vol. 2, pp. 282-294, 2010.

[4] G.-D. Fan, C.-C. Huang, Y.-B. Lin, C.-S. Tang, C.-Y. Twu, and Y.-H. Wen, "Enhanced video phone services for NGN/IMS," *Wireless Commun. Mobile Comput.*, 2010.

[5] Y.-B. Lin, W.-R. Lai, and J.-J. Chen, "Effects of cache mechanism on wireless data access," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 1247-1258, Nov. 2003.

[6] Y. Fang, Z. J. Haas, B. Liang, and Y.-B. Lin, "TTL prediction schemes and the effects of inter-update time distribution on wireless data access," *ACM Wireless Netw.*, vol. 10, no. 5, pp. 607-619, Sep. 2004.

[7] K. Sung, S. Srinivasan, and H. Schulzrinne, "BBS-ONE: bulletin board and forum system for mobile opportunistic networks," in *Proc. IEEE International Conf. Wireless Commun., Netw. Inf. Security*, pp. 371-376, June 2010.

[8] N. Shiratuddin, M. Landoni, F. Gibb, and S. Hassan, "E-Book technology and its potential applications in distance education," *J. Digital Inf.*, vol. 3, 2003.

[9] Y.-C. Lai, P. Lin, and S.-M. Cheng, "Performance modeling for application-level integration of heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 2426-2434, June 2009.

[10] S.-R. Yang, "Dynamic power saving mechanism for 3G UMTS system," *ACM/Springer Mobile Netw. Appls.*, vol. 12, no. 1, pp. 5-14, Jan. 2007.

[11] Y.-B. Lin, "Eliminating tromboning mobile call setup for international roaming users," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 320-325, Jan. 2009.