Prefetching for Mobile Web Album

Yi-Bing Lin, *Fellow, IEEE*, and Po-Kao Huang

Department of Computer Science

National Chiao Tung University

{liny, pkhuang}@cs.nctu.edu.tw

Abstract

A web album service allows a user to publish photo albums on the web and view albums of other users. Through broadband mobile telecom, users can enjoy watching contents of web albums at any place in real time. User experience on mobile web album is affected by the transmission delay of mobile network, which determines whether the user needs to wait to view the images. We propose a prefetching mechanism that enhances user experience on accessing mobile web albums. A TCP-like sliding window protocol (of size N) is exercised, and when the buffer for the sliding window at the user equipment (UE) is full, prefetching is suspended. The buffer size N affects the prefetching performance. The larger the N value, the better the user experience. However, a large N value means that many images will be prefetched. If they are not actually viewed by the user, the network resources for transmitting these images are wasted. This paper proposes both analytic and

simulation models to select the smallest N (the optimal N value) so that the expected user experience can be achieved.

1. Introduction

Touch-screen based mobile devices with high-resolution image display features have become popular. In particular, large-size screen has become a trend in mobile device design. For example, HTC launched 4.7-inch smartphone HTC One to replace 4.3-inch HTC One SC in 2013 [1], and Samsung also released 4.8-inch smartphone GALAXY S III to replace 4.3-inch GALAXY S II in 2012 [2]. Convenience provided by these features encourages implementation of applications such as web albums (e.g., Facebook) in mobile devices. Web albums allow a user to publish photo albums on the web and view albums of other users. Through broadband mobile telecom, users can enjoy watching contents of web albums at any place in real time.

We have implemented a mobile web album called CloudPocket (**Dear Reviewer: Please see supplementary document for detail.**) with Industrial Technology Research Institute (ITRI) [3], Taiwan. CloudPocket was one of the award winners among more than 100 candidates in the 2012 Mobile App Competition of Ministry of Economic Affairs, Executive Yuan, Taiwan. User experience on CloudPocket indicates that many users access several images in sequential order (e.g.,

flipping the photos in facebook), and then jump forward to access an out-of the-sequential-order image (Note that jump backward does not involve accessing to the server). Therefore, it is essential to enhance user experience in sequential web album access. We note that "sequential access of images" does not mean that the received images are manipulated sequentially. When the user receives image i, she/he may zoom or comment on some images $j \leq i$, and then moves on to access image $i + 1$. One important issue is to reduce the waiting time for accessing images. A solution is prefetching that has been deployed by web browser plugins on laptops [4,5] and web content [18]. To speed up wireless transmission for images, data compression has also been proposed. With the improvement of the CPU computing power, mobile devices can simultaneously execute prefetching and compression of images.

From the network aspect, prefetching may waste resources if the prefetched images are not used (i.e., photos are not viewed by the users). To balance against the network bandwidth consumption and the user experience (i.e., the waiting time), this paper proposes a combined prefetching and compression method for mobile web albums access. Then we develop analytic and simulation models to study the performance of this mechanism.

2. The Prefetching Mechanism

This section uses CloudPocket as an example to describe the image prefetching

mechanism on mobile devices. As illustrated in Fig. 1, CloudPocket can be implemented in a 3G environment (WCDMA) that consists of a mobile core network (Fig. 1 (a)) and based stations (Fig. 1 (b)). Alternatively, it can be implemented in a Wi-Fi environment with Wi-Fi routers (Fig. 1 (c)). Like other web album applications, CloudPocket follows the client-server model where the user equipment (UE; i.e., mobile device; Fig. 1 (d)) is the client that accesses the images from CloudPocket in the server (Fig. 1 (e)) through the Internet (Fig. 1 (f)). In Fig. 1, the UE communicates with CloudPocket via the ftp protocol. The communication path is established between UE and the server through (e)-(f)-(a)-(b)-(d) for 3G service or (e)-(f)-(c)-(d) for Wi-Fi service.

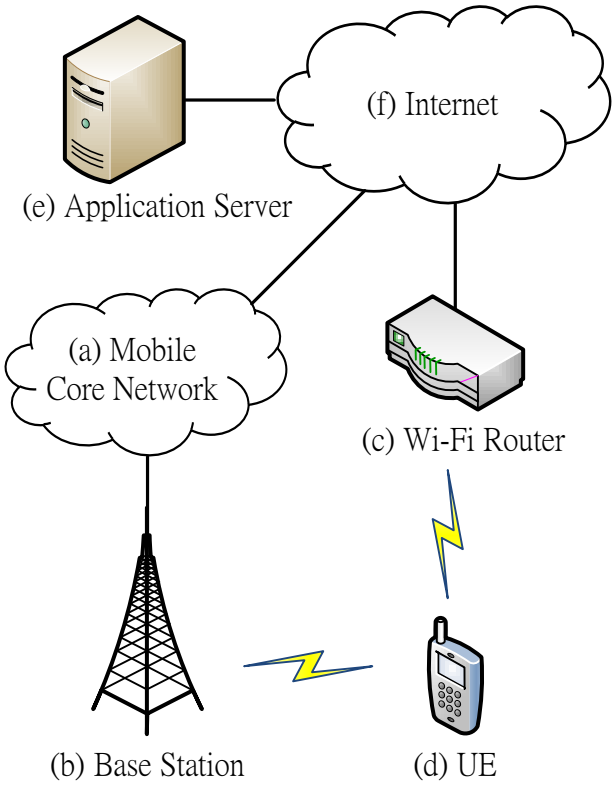

Fig. 1. A simplified architecture for wireless applications

Our album prefetching mechanism follows the TCP-like sliding window protocol [6] with an image buffer of size N implemented in the UE, where the buffer size is the window size. We also utilize ZipArchive for compression [7]. The images published in CloudPocket are compressed in advance. The server repeats transmitting compressed images to the UE until the image buffer of the UE is filled. Upon receipt of an image, the UE decompresses and stores it in the buffer. After an image is accessed by the user, it is removed or stored in other memory space of the UE, and the buffer can accommodate the next prefetched image. If all images in the buffer are viewed and the user attempts to view the next images, then the user must wait until these images are fetched. The prefetching mechanism can be evaluated by three measures.

- $E[n]$: the expected number of the images that the user can continue to access without waiting
- $E[w]$: the expected waiting time that the user has to wait for the arrival of the next image (excluding the initial waiting for the first image)
- $E[n^*]$: the number of wasted transmitted images

When a user stops viewing, let n^* be the number of the prefetched images in the buffer that are not viewed by the user. These images are considered wasted. Prefetching of these images consumes extra network resources (especially the

wireless bandwidth). We can quickly investigate the $E[n^*]$ performance through a worst case study. In the steady state, if the buffer of the UE is full, then from the view of the random process, the number of wasted image is $\frac{N-1}{2}$. Since the buffer may not be full when the user stops viewing, we have $E[n^*] \leq \frac{N-1}{2}$.

Both $E[n]$ and $E[w]$ are considered as output measures for the user experience. The user experience of prefetching is good if $E[n]$ is large and $E[w]$ is small, which are affected by three factors: the image transmission delay τ_T, the image viewing time τ_V, and the size N of the buffer. One of the following three cases may occur.

Case I: When $\tau_V \gg \tau_T$, the user only waits for transmission of the first image, the subsequent images can be viewed without waiting. In this case, the optimal buffer size is $N = 2$ because the user never waits for the arrival of the next image. Also, since only one image is prefetched, at most one image is wasted (i.e., not viewed).

Case II: When $\tau_T \gg \tau_V$, the user always waits for transmission of the next image. In this case, prefetching does not work.

Case III: The intervals τ_T and τ_V are of the same order of magnitude, the prefetching mechanism may effective with an appropriate image buffer size N.

In Case III, selection of N is important to balance against $E[n]$, $E[w]$ and $E[n^*]$. When $N = 1$, there is no prefetching. In the prefetching mode, the larger the

N, the better the $E[n]$ and the $E[w]$. On the other hand, if N is too large, many images may be prefetched without being viewed (i.e., $E[n^*]$ is large), and the valuable network resources for transmission are wasted. The maximum transmission resources wasted is $n^* = N - 1$ images, when the prefetching time is much faster than the viewing time. On the average, $E[n^*] \leq \frac{\mathrm{N}-1}{2}$. Therefore, it is essential to select an appropriate N value to balance against the user experience ($E[n]$ and $E[w]$) and network consumption ($E[n^*]$).

3. An Analytic Model for Prefetching with $N = 2$

From the viewpoint of the service provider, it is important to invest network resources to enhance user experience. In service planning (i.e., to set up the N parameter), we typically ask the following question: "If we want to engineer the service at $E[n] < x$ and $E[w] < y$, what is the smallest N value ?" Then the service is engineered at the selected buffer size N if N is reasonably small but suffices to achieve the user experience we expect. Therefore, it is important to conduct accurate derivation of $E[n]$ and $E[w]$ as functions of N. This section proposes an analytic model for prefetching with $N = 2$. Fig. 2 illustrates a timing diagram that describes the prefetching mechanism.

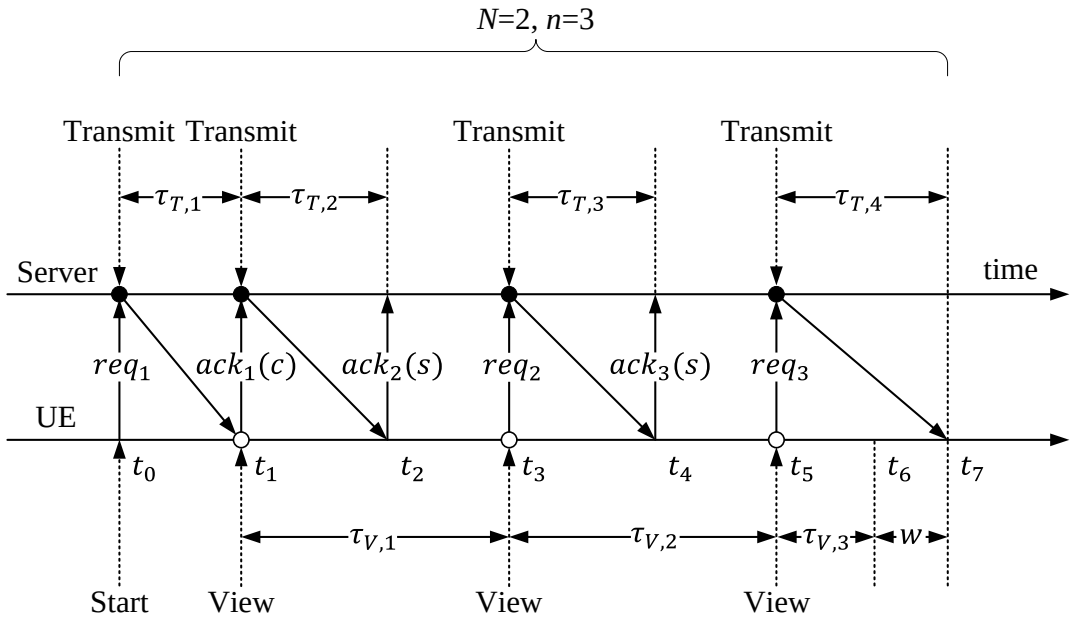

Fig. 2. Timing diagram for prefetching with $N = 2$

In this figure, the symbol "●" represents that the server starts to transmit an image, the symbol "○" represents that the user starts to view a received image. At t_0, the UE starts prefetching by sending an image access request (req_1 in Fig. 2) to ask the server to transmit the first image. Upon receipt of req_1, the server transmits the image. At t_1, the image arrives at the UE. The UE sends a continue-to-prefetch acknowledgement ($ack_1(c)$ in Fig. 2) to inform the server to transmit the next image. The user views the first image in period $[t_1, t_3]$. At t_2, the UE receives the second image. Since the user is still viewing the first image, the buffer (which contains the viewed and the prefetched images) is full. The UE sends a suspend-prefetching acknowledgement $ack_2(s)$ to inform the server to suspend image transmission. At t_3, the user continues to view the second image, and the first image is deleted or

stored in other memory space of the UE, and the buffer can accommodate the next prefetched image. The UE sends an image access request req_2 to inform the server to transmit the next image. At t_4, the UE receives the third image. Since the user is still viewing the second image, the buffer is full. The UE sends a suspend-prefetching acknowledgement $ack_3(s)$ to suspend transmission. At t_5, the user continues to view the third image, and the buffer storage for the second image is released. The UE sends the image access request req_3 to prefetch the next image. At t_6, all images in the buffer are viewed and the user attempts to view the next image which has not been received. This image arrives at t_7, and the user must wait for the period $[t_6, t_7]$ before she/he can view the next image. Let n be the number of images that are consecutively viewed without waiting. In this example, $n = 3$ in $[t_0, t_7]$. Then the user has to wait for the 4th image that will arrive at time t_7. Let w be the waiting time. In Fig. 2, $w = t_7 - t_6$ for the 4th image. Let $\tau_{T,i}$ be the transmission delay for image i. In Fig. 2, $\tau_{T,2} = t_2 - t_1$ for the second image. Let $\tau_{V,i}$ be the viewing time for image i. In Fig. 2, $\tau_{V,2} = t_5 - t_3$ for the second image. From a user experience test of 20 users, we observe that $\tau_{V,i}$ and $\tau_{T,i}$ are independent variables.

To derive $E[n]$ and $E[w]$, we define $p(n)$, the probability that after the user has received the first image (i.e., after t_1 in Fig. 2), she/he can continue to view exactly $n - 1$ subsequent images without waiting (and then waits for the arrival of

the nth subsequent image). In this section, we describe an analytic model for deriving $p(n)$, $E[n]$ and $E[w]$ under $N = 2$. In this scenario, whether the user has to wait for the arrival of the $(i+1)$th image only depends on the relationship of $\tau_{V,i}$ for viewing the ith image and $\tau_{T,i+1}$ for transmitting the $(i+1)$th image. If we assume that $\tau_{T,i}$ and $\tau_{V,i}$ are both $i.i.d.$ random variables, then the notation can be simplified as τ_T and τ_V respectively, and $\Pr[\tau_T > \tau_V]$ is the probability that the user has to wait for the arrival of the next image. Therefore, $p(n)$ can be expressed as

$$p(n) = (1 - \Pr[\tau_T > \tau_V])^{n-1}(\Pr[\tau_T > \tau_V]) \quad (1)$$

In (1), the term $(1 - \Pr[\tau_T > \tau_V])^{n-1}$ means that the user can continuously view n images without waiting (including the first image, which was viewed at t_1 in Fig. 2). If τ_T is an Erlang random variable with the shape parameter k, the rate parameter μ, and the mean k/μ, then its density function is

$$f_T(\tau_T) = \frac{\mu^k \tau_T^{k-1} e^{-\mu\tau_T}}{(k-1)!} \quad \text{for } k \geq 1 \quad (2)$$

From (2) and assume that τ_V has an arbitrary density function $f_V(\tau_V)$, we have

$$\begin{aligned}\Pr[\tau_T > \tau_V] &= \int_{\tau_V=0}^{\infty}\int_{\tau_T=\tau_V}^{\infty} f_V(\tau_V) f_T(\tau_T) d\tau_T d\tau_V \\ &= \int_{\tau_V=0}^{\infty}\int_{\tau_T=\tau_V}^{\infty} f_V(\tau_V)\left[\frac{\mu^k \tau_T^{k-1} e^{-\mu\tau_T}}{(k-1)!}\right] d\tau_T d\tau_V\end{aligned}$$

$$= \sum_{m=0}^{k-1} \left(\frac{\mu^m}{m!}\right) \left[\frac{(-1)^m d^m f_v^*(s)}{ds^m}\bigg|_{s=\mu}\right] \tag{3}$$

We consider Erlang τ_T distribution because any distribution can be represented by a mixture of Erlang distribution [8]. If τ_V is a Gamma random variable with the density function $f_V(\tau_V)$, the mean $1/\lambda$, the variance V_V, then its Laplace transform is

$$f_V^*(\mathrm{s}) = \left(\frac{1}{V_V s\lambda + 1}\right)^{\frac{1}{V_V\lambda^2}} \tag{4}$$

and

$$\frac{d^m f_v^*(s)}{ds^m} = (-V_V\lambda)^m \left(\frac{1}{V_V s\lambda + 1}\right)^{\frac{1}{V_V\lambda^2}+m} \prod_{l=1}^{m}\left(\frac{1}{V_V\lambda^2} + l - 1\right) \tag{5}$$

Note that the Gamma distribution can be used to describe a random variable with a large range of variance. That is, a Gamma random variable τ_V is appropriate to represent users with both regular and irregular viewing behaviors [9,10]. Substitute (5) into (3) to yield

$$\begin{aligned}\Pr[\tau_T > \tau_V] &= \sum_{m=0}^{k-1}\left[\frac{(-\mu)^m}{m!}\right]\left[(-V_V\lambda)^m \left(\frac{1}{V_V\mu\lambda + 1}\right)^{\frac{1}{V_V\lambda^2}+m} \prod_{l=1}^{m}\left(\frac{1}{V_V\lambda^2} + l - 1\right)\right] \\ &= \sum_{m=0}^{k-1}\binom{\frac{1}{V_V\lambda^2} + m - 1}{m} (V_V\mu\lambda)^m \left(\frac{1}{V_V\mu\lambda + 1}\right)^{\frac{1}{V_V\lambda^2}+m}\end{aligned} \tag{6}$$

From (1) and (6), $p(n)$ is expressed as

$$p(n) = \left\{1 - \left[\sum_{m=0}^{k-1} \binom{\frac{1}{V_V\lambda^2} + m - 1}{m} (V_V\mu\lambda)^m \left(\frac{1}{V_V\mu\lambda + 1}\right)^{\frac{1}{V_V\lambda^2}+m}\right]\right\}^{n-1}$$
$$\times \left[\sum_{m=0}^{k-1} \binom{\frac{1}{V_V\lambda^2} + m - 1}{m} (V_V\mu\lambda)^m \left(\frac{1}{V_V\mu\lambda + 1}\right)^{\frac{1}{V_V\lambda^2}+m}\right] \quad (7)$$

From (7), $E[n]$ can be expressed as

$$E[n] = \sum_{n=1}^{\infty} n \times p(n)$$
$$= \frac{1}{\Pr[\tau_T > \tau_V]}$$
$$= \left\{\sum_{m=0}^{k-1} \left(\frac{\mu^m}{m!}\right) \left[\frac{(-1)^m d^m f_v^*(s)}{ds^m}\bigg|_{s=\mu}\right]\right\}^{-1} \quad (8)$$

For Gamma random variable τ_V, (8) is rewritten as

$$E[n] = \left\{\sum_{m=0}^{k-1} \binom{\frac{1}{V_V\lambda^2} + m - 1}{m} (V_V\mu\lambda)^m \left(\frac{1}{V_V\mu\lambda + 1}\right)^{\frac{1}{V_V\lambda^2}+m}\right\}^{-1} \quad (9)$$

Since $w = \max(0, \tau_T - \tau_V)$, $E[w]$ is expressed as

$$E[w] = \int_{\tau_V=0}^{\infty} \int_{\tau_T=\tau_V}^{\infty} (\tau_T - \tau_V) f_V(\tau_V) f_T(\tau_T) d\tau_T d\tau_V$$
$$= \int_{\tau_V=0}^{\infty} f_V(\tau_V) \int_{\tau_T=\tau_V}^{\infty} (\tau_T - \tau_V) \left[\frac{\mu^k \tau_T^{k-1} e^{-\mu\tau_T}}{(k-1)!}\right] d\tau_T d\tau_V$$
$$= \left(\frac{k}{\mu}\right) \sum_{m=0}^{k} \left(\frac{\mu^m}{m!}\right) \left[\frac{(-1)^m d^m f_v^*(s)}{ds^m}\bigg|_{s=\mu}\right] - \left(\frac{1}{\mu}\right) \sum_{m=0}^{k-1} \left(\frac{\mu^{m+1}}{m!}\right) \left[\frac{(-1)^{m+1} d^{m+1} f_v^*(s)}{ds^{m+1}}\bigg|_{s=\mu}\right] \quad (10)$$

For Gamma random variable τ_V, we have

$$E[w] = \left(\frac{k}{\mu}\right)\sum_{m=0}^{k}\left[\binom{\frac{1}{V_V\lambda^2}+m-1}{m}(V_V\mu\lambda)^m\left(\frac{1}{V_V\mu\lambda+1}\right)^{\frac{1}{V_V\lambda^2}+m}\right]$$

$$-\left(\frac{1}{\mu}\right)\sum_{m=0}^{k-1}\left[\binom{\frac{1}{V_V\lambda^2}+m-1}{m}(V_V\mu\lambda)^{m+1}\left(\frac{1}{V_V\lambda^2}+m\right)\left(\frac{1}{V_V\mu\lambda+1}\right)^{\frac{1}{V_V\lambda^2}+m+1}\right] \quad (11)$$

4. An Analytic Model for Prefetching with $N \to \infty$

This section derives $p(n)$ when $N \to \infty$ where $\tau_{V,i}$ and $\tau_{T,i}$ are Exponentially distributed. The exponential distribution assumption provides the mean value analysis for a primary study on the trends of the parameters impact. Also, this paper uses the analytic results based on the exponential assumption to validate the simulation model. After validation, the simulation experiments can be extended to accommodate other distributions. In this scenario, the server always repeats transmitting images to the UE. Let $T_{V,i} = \sum_{j=1}^{i} \tau_{V,j}$ be the sum of the times for the first i images $(i \geq 1)$. Let $T_{T,i} = \sum_{j=2}^{i} \tau_{T,j}$ be the sum of the transmission delays of the second image to the ith image $(i \geq 2)$. It is clear that if the user can continue to view i images without waiting, the sum $T_{V,i-1}$ of image viewing times from the first image to the $(i-1)$th image must be larger than the sum $T_{T,i}$ of image transmission delays from the second image to the ith image; that is, for $i \geq 2$

$$\sum_{j=1}^{i-1} \tau_{V,j} \geq \sum_{j=2}^{i} \tau_{T,j} \tag{12}$$

For $i \geq 2$, let T_i be the period between when the ith image is prefetched and when the $(i-1)$th image is completely viewed under the condition that the user has viewed the first $i-1$ images without waiting. In other words,

$$T_i = T_{V,i-1} - T_{T,i} \quad \text{where } T_{V,j-1} \geq T_{T,j} \text{ for } 2 \leq j \leq i \tag{13}$$

Therefore $\Pr[T_i \geq 0]$ is the probability that the user can view i or more images without waiting. Then $p(n)$ can be expressed as

$$p(n) = \begin{cases} 1 - \Pr[T_{n+1} \geq 0] & \text{, for } n = 1 \\ \Pr[T_n \geq 0] - \Pr[T_{n+1} \geq 0] & \text{, for } n \geq 2 \end{cases} \tag{14}$$

We derive $\Pr[T_i \geq 0]$ assuming that both $\tau_{V,i-1}$ and $\tau_{T,i}$ are Exponentially distributed with the mean $1/\lambda$ and the mean $1/\mu$, respectively. Consider $i = 2$. The relationship among $\tau_{V,1}$, $\tau_{T,2}$, and T_2 in (13) can be re-arranged as

$$\tau_{V,1} = \tau_{T,2} + T_2 \tag{15}$$

Let $f_i(T_i)$ be the density function of T_i. From (15), $f_2(T_2)$ is derived as

$$f_2(T_2) = \int_{\tau_{T,2}=0}^{\infty} \lambda e^{-\lambda(\tau_{T,2}+T_2)} \mu e^{-\mu\tau_{T,2}} \, d\tau_{T,2}$$

$$= \left(\frac{\lambda\mu}{\lambda+\mu}\right) e^{-\lambda T_2} \tag{16}$$

From (14) and (16), $\Pr[T_2 \geq 0] = \frac{\mu}{\lambda+\mu}$ and

$$p(1) = 1 - \Pr[T_2 \geq 0] = \frac{\lambda}{\lambda+\mu} \tag{17}$$

Consider $i = 3$. From (13) and (15), we have

$$\tau_{V,2} = \tau_{T,3} - T_2 + T_3 \tag{18}$$

From (18), we derive $f_3(T_3)$ in two cases.

Case A: For $T_2 < T_3$, $f_3(T_3)$ can be represented as $f_{3A}(T_3)$ where

$$\begin{aligned} f_{3A}(T_3) &= \int_{\tau_{T,3}=0}^{\infty} \int_{T_2=0}^{T_3} f_2(T_2)\lambda e^{-\lambda(\tau_{T,3}-T_2+T_3)} \mu e^{-\mu\tau_{T,3}} dT_2 d\tau_{T,3} \\ &= \left(\frac{\lambda\mu}{\lambda+\mu}\right)^2 e^{-\lambda T_3} T_3 \end{aligned} \tag{19}$$

Case B: For $T_3 \leq T_2$, $f_3(T_3)$ can be represented as $f_{3B}(T_3)$ where

$$\begin{aligned} f_{3B}(T_3) &= \int_{T_2=T_3}^{\infty} \int_{\tau_{T,3}=T_2-T_3}^{\infty} f_2(T_2)\lambda e^{-\lambda(\tau_{T,3}-T_2+T_3)} \mu e^{-\mu\tau_{T,3}} \, d\tau_{T,3} dT_2 \\ &= \left(\frac{\lambda\mu}{\lambda+\mu}\right)^2 e^{-\lambda T_3} \left(\frac{1}{\lambda+\mu}\right) \end{aligned} \tag{20}$$

From (19) and (20), $f_3(T_3)$ can be expressed as

$$f_3(T_3) = f_{3A}(T_3) + f_{3B}(T_3) = \left(\frac{\lambda\mu}{\lambda+\mu}\right)^2 e^{-\lambda T_3} \left(T_3 + \frac{1}{\lambda+\mu}\right) \tag{21}$$

From (14) and (21), $\Pr[T_3 \geq 0] = \left[\frac{\mu^2}{(\lambda+\mu)^3}\right](2\lambda+\mu)$ and

$$p(2) = \Pr[T_2 \geq 0] - \Pr[T_3 \geq 0]$$

$$= \frac{\lambda^2 \mu}{(\lambda + \mu)^3} \tag{22}$$

Similarly, we have

$$f_4(T_4) = \left(\frac{\lambda\mu}{\lambda + \mu}\right)^3 e^{-\lambda T_4} \left[\frac{T_4^2}{2} + \frac{2T_4}{\lambda + \mu} + \frac{2}{(\lambda + \mu)^2}\right] \tag{23}$$

$$f_5(T_5) = \left(\frac{\lambda\mu}{\lambda + \mu}\right)^4 e^{-\lambda T_5} \left[\frac{T_5^3}{6} + \frac{3T_5^2}{2(\lambda + \mu)} + \frac{5T_5}{(\lambda + \mu)^2} + \frac{5}{(\lambda + \mu)^3}\right] \tag{24}$$

$$f_6(T_6) = \left(\frac{\lambda\mu}{\lambda + \mu}\right)^5 e^{-\lambda T_6} \left[\frac{T_6^4}{24} + \frac{2T_6^3}{3(\lambda + \mu)} + \frac{9T_6^2}{2(\lambda + \mu)^2} + \frac{14T_6}{(\lambda + \mu)^3} + \frac{14}{(\lambda + \mu)^4}\right] \tag{25}$$

$$f_7(T_7) = \left(\frac{\lambda\mu}{\lambda + \mu}\right)^6 e^{-\lambda T_7} \left[\frac{T_7^5}{120} + \frac{5T_7^4}{24(\lambda + \mu)} + \frac{7T_7^3}{3(\lambda + \mu)^2} + \frac{14T_7^2}{(\lambda + \mu)^3} + \frac{42T_7}{(\lambda + \mu)^4} + \frac{42}{(\lambda + \mu)^5}\right] \tag{26}$$

and so on. From (22) and (23), we obtain

$$\begin{aligned} p(3) &= \left[\frac{\mu^2(2\lambda + \mu)}{(\lambda + \mu)^3}\right] - \left[\frac{\mu^3}{(\lambda + \mu)^5}\right](5\lambda^2 + 4\lambda\mu + \mu^2) \\ &= \frac{\mu}{\lambda + \mu} - \left[\frac{\mu^3}{(\lambda + \mu)^5}\right](5\lambda^2 + 4\lambda\mu + \mu^2) - p(2) \\ &= \frac{2\lambda^3\mu^2}{(\lambda + \mu)^5} \end{aligned} \tag{27}$$

From (24) and (27), we obtain

$$\begin{aligned} p(4) = &\left[\frac{\mu^3}{(\lambda + \mu)^5}\right](5\lambda^2 + 4\lambda\mu + \mu^2) \\ &- \left[\frac{\mu^4}{(\lambda + \mu)^7}\right](14\lambda^3 + 14\lambda^2\mu + 6\lambda\mu^2 + \mu^3) \end{aligned}$$

$$= \left[\frac{\mu^2(2\lambda+\mu)}{(\lambda+\mu)^3}\right] - \left[\frac{\mu^4}{(\lambda+\mu)^7}\right](14\lambda^3 + 14\lambda^2\mu + 6\lambda\mu^2 + \mu^3) - p(3)$$
$$= \frac{5\lambda^4\mu^3}{(\lambda+\mu)^7} \tag{28}$$

Similar to the derivation for (25), (26), and (28), we have

$$p(5) = \frac{14\lambda^5\mu^4}{(\lambda+\mu)^9} \tag{29}$$

$$p(6) = \frac{42\lambda^6\mu^5}{(\lambda+\mu)^{11}} \tag{30}$$

With elaboration, we found that $p(n)$ can be expressed by a general form

$$p(n) = \frac{C_n\lambda^n\mu^{n-1}}{(\lambda+\mu)^{2n-1}} \tag{31}$$

where $C_n = \frac{(2n-2)!}{n!\,(n-1)!}$ for $n \geq 1$ is of the form of Catalan number [11].

From (31),

$$E[n] = \sum_{n=1}^{\infty} n \times p(n) = \sum_{n=1}^{\infty} \frac{[2(n-1)!\,\lambda^n\mu^{n-1}}{[(n-1)!]^2(\lambda+\mu)^{2n-1}} \tag{32}$$

5. Numerical Examples

We have implemented a discrete event simulation model (similar to the approach in [12-14]), which is validated against by (7), (9), (11) for $N = 2$, and (31) and (32) for $N \to \infty$. The discrepancies between the analytic and simulation results are within 1%. Based on the simulation model, this section uses numerical examples

to investigate how parameters N (the size of the image buffer), τ_V (the user viewing times), and τ_T (the image transmission delays) affect the prefetching mechanism.

We consider web albums in CloudPocket where the average image size is 922.68KB. The τ_V samples are measured from a technical trial with 10 users. With the fitting techniques [15], these τ_V samples are fit by the Gamma distribution, where $E[\tau_V] = 9.06$ seconds and $V_V = 0.04E[\tau_V]^2$. The τ_T samples are measured from the commercial 3G and Wi-Fi services [16,17], and are fit by the Erlang distribution. For 3G without compression, $E[\tau_T] = 3.283$ seconds and $V_T = 0.18E[\tau_T]^2$. For 3G with compression (including the decompression time at the UE), $E[\tau_T] = 1.995$ seconds and $V_T = 0.15E[\tau_T]^2$. For Wi-Fi without compression, $E[\tau_T] = 0.846$ seconds and $V_T = 0.24\,E[\tau_T]^2$. For Wi-Fi with compression, $E[\tau_T] = 0.623$ seconds and $V_T = 0.25E[\tau_T]^2$. These measured τ_T samples can be fit by the Erlang distribution with $k = 6$ for 3G, and $k = 4$ for Wi-Fi.

In our measurements, very good user experience is observed with the minimal image buffer size $N = 2$. For 3G with $N = 2$, $E[n] = 41.67$ and $E[w] = 0.007$ $E[\tau_T]$ without compression, and $E[n] = 1.38 \times 10^4$ and $E[w] = 1.7 \times 10^{-5}$ $E[\tau_T]$ with compression. For Wi-Fi without compression, $E[n] = 7.7 \times 10^7$ and $[w] = 1.29 \times 10^{-8}$ $E[\tau_T]$, and $E[n] = 1.01 \times 10^{10}$ and $E[w] = 9.9 \times 10^{-11}$ $E[\tau_T]$ with compression. In our experiments, most users sequentially access the

images, and regularly manipulate the received images (and small V_V variances are observed), therefore good user experience can be achieved with $N = 2$, where $E[n^*] \leq \frac{1}{2}$ is the network resources wasted in prefetching. Although the user behaviors observed in CloudPocket's technical trial are common, it is important that we also consider irregular viewing scenarios where V_V may be much larger than $0.04E[\tau_V]^2$. In the remainder of this section, we investigate irregular viewing behavior with larger variances; i.e., $0.1E[\tau_V]^2 \leq V_V \leq 10E[\tau_V]^2$. For example, when $V_V = 10E[\tau_V]^2$, the user will quickly flip many images and then stop to manipulate other operations before she/he accesses the subsequent images from the server. In such a case, an image buffer size larger than 2 is required. To see the effect of V_V, Fig. 3 plots the $E[n]$ and $E[w]/E[\tau_T]$ curves for V_V ranging from $0.1E[\tau_V]^2$ to $10E[\tau_V]^2$, where $N = 2$. Fig. 3 indicates that as V_V increases, $E[n]$ decreases and $E[w]/E[\tau_T]$ increases. For V_V =$10E[\tau_V]^2$, $E[n] = 1.33$ and $E[w] = 0.691E[\tau_T]$ for 3G without compression. For 3G with compression, $E[n] = 1.4$ and $E[w] = 0.658E[\tau_T]$. For Wi-Fi without compression, $E[n] = 1.53$ and $E[w] = 0.606E[\tau_T]$. For Wi-Fi with compression, $E[n] = 1.58$ and $E[w] = 0.587E[\tau_T]$. It is clear that the user experience is not good (i.e., $E[n] < 2$) for large viewing variance (e.g., $V_V = 10E[\tau_V]^2$).

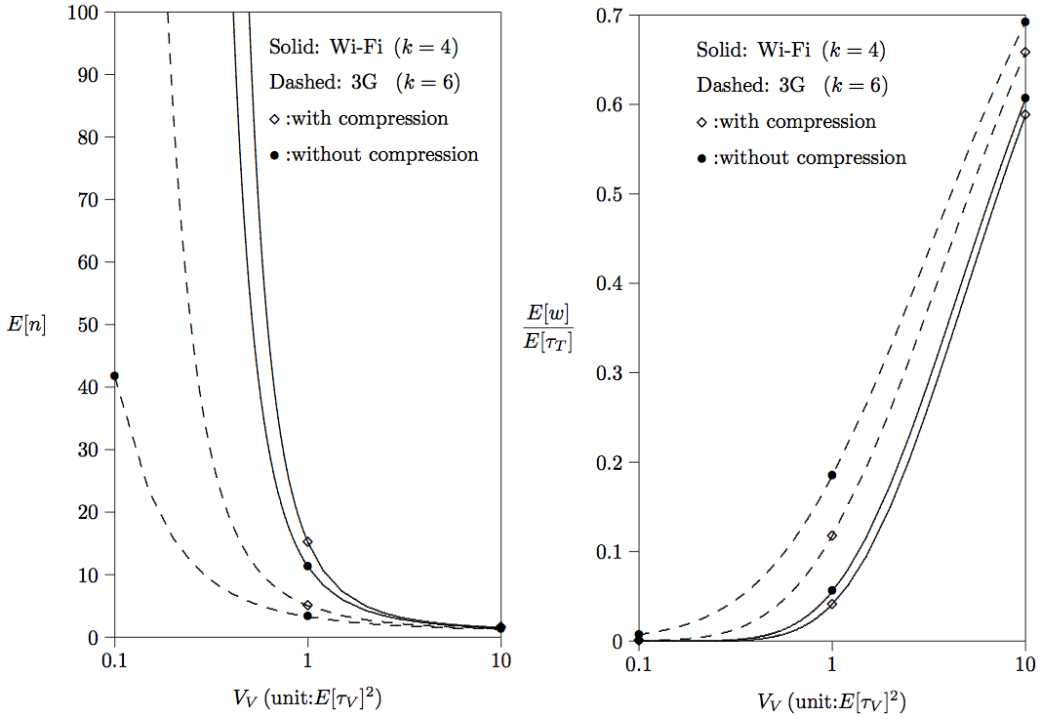

Fig. 3. Effects of V_V $(N = 2)$; the measured τ_T and τ_V distributions are used except that V_V ranges from $0.1E[\tau_V]^2$ to $10E[\tau_V]^2$.

Fig. 4 plots the $E[n]$ and $E[w]/E[\tau_T]$ curves against N, where measured τ_V and τ_T distributions are used as inputs except that we set $V_V = 10E[\tau_V]^2$. This figure indicates trivial results that as N increases, $E[n]$ increases and $E[w]/E[\tau_T]$ decreases. The non-trivial results are that the curves provide guidelines to select the appropriate N values. If we expect to achieve the user experience goals such that $E[n] > 100$ and $E[w]/E[\tau_T] < 0.1$, then $N = 26$ should be selected for 3G without compression and $N = 19$ for 3G with compression. For Wi-Fi, $N = 14$ should be selected without compression and $N = 13$ with compression. In these selections, the extra network resources are $E[n^*] \leq 12.5$ for 3G without compression,

$E[n^*] \leq 9$ for 3G with compression, $E[n^*] \leq 6.5$ for Wi-Fi without compression, and $E[n^*] \leq 6$ for Wi-Fi with compression.

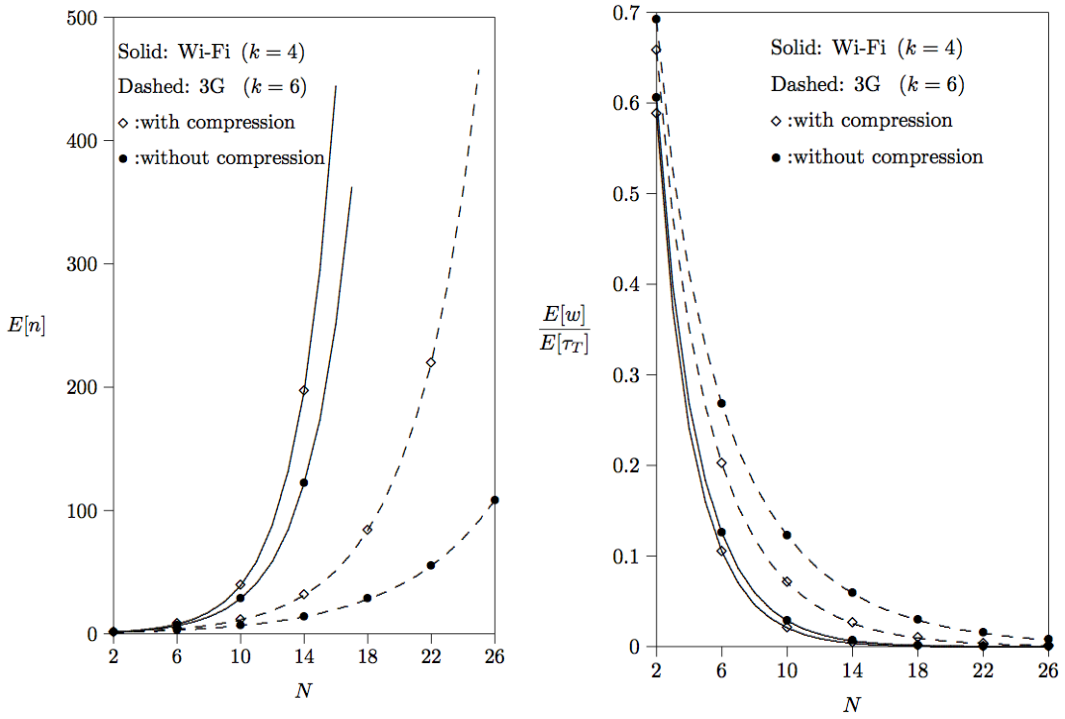

Fig. 4. Effects of N (V_V=10 $E[\tau_V]^2$)

Fig. 3 and Fig. 4 indicate when V_V increases from $0.1E[\tau_V]^2$ to $10E[\tau_V]^2$, to achieve the same user experience (e.g., $E[n] > 100$ and $E[w]/E[\tau_T] < 0.1$), the selected N should be increased from, for example, 2 to 19 for 3G with compression.

Let α be the improvement on $E[n]$ due to image compression. Similarly, let β be the improvement on $E[w]$ when the images are compressed. Fig. 5 indicates that as V_V increases, the improvements (i.e., α and β) decrease. Fig. 6 indicates that as N increases, the improvements increase. The improvements offered by compression are more significant for 3G than those for Wi-Fi. For example, when $N = 26$ and

V_V=10 $E[\tau_V]^2$, the α improvement for 3G is 1.31 times that of Wi-Fi and the β improvement for 3G is 2.91 times that of Wi-Fi. In summary, if the mobile album service is offered under the 3G environment, then prefetching with compression is highly recommended.

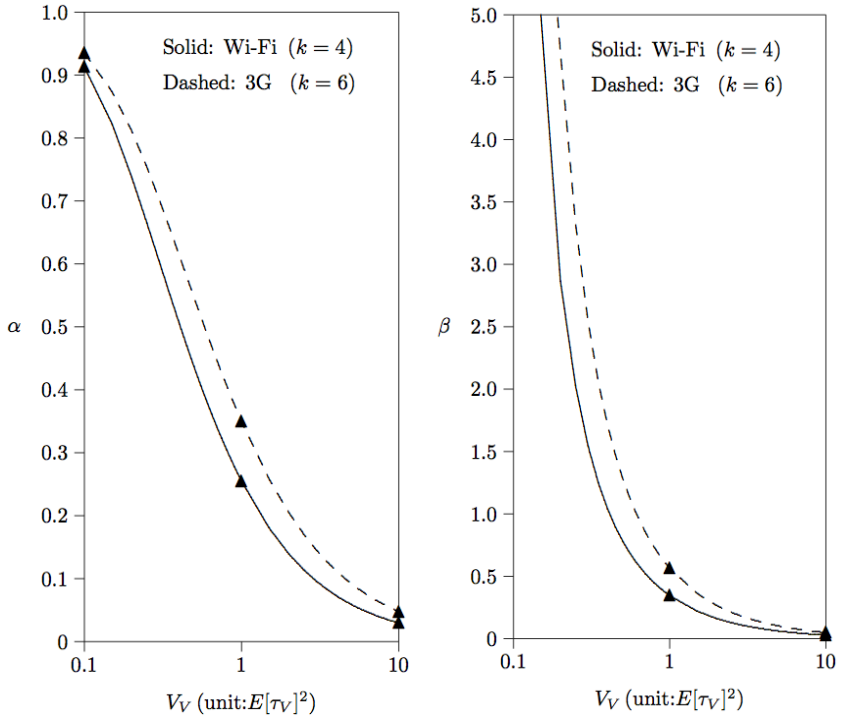

Fig. 5. Effects of V_V and compression ($N = 2$)

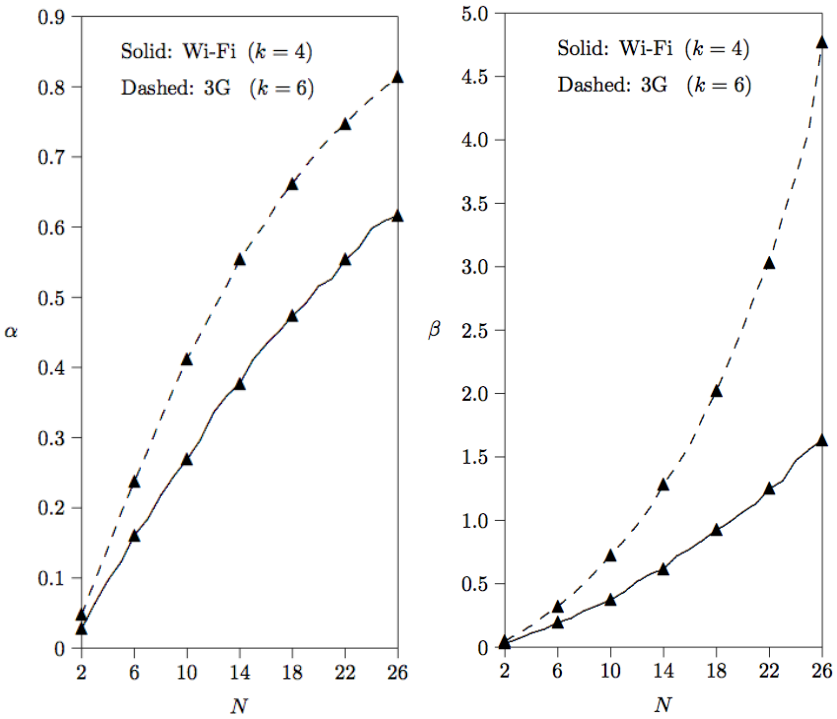

Fig. 6. Effects of N and compression (V_V=10 $E[\tau_V]^2$)

6. Conclusions

This paper proposed a prefetching mechanism that enhances user experience on accessing mobile web albums. The output measures are the expected number of the images $E[n]$ that the user can continue to access without waiting, and the expected waiting time $E[w]$ that the user has to wait for the arrival of the next image. The buffer size N of the user equipment (UE) affects the performance of prefetching. The larger the N value, the better the user experience (a larger $E[n]$ and a smaller $E[w]$). However, a large N value means that many images will be perfected. If they are not viewed by the user, the network resources for transmitting these images are wasted. The number $E[n^*]$ of wasted transmitted images is $E[n^*] \leq \frac{N-1}{2}$.

We proposed analytic and simulation models to select the smallest N (the optimal N value) so that the expected user experience can be achieved. Our study indicated that $E[n]$ and $E[w]$ are significantly affected by the variance V_V of the viewing time distribution. For example, when $V_V = 10\ E[\tau_V]^2$, if the user experience goals are $E[n] > 100$ and $E[w]/E[\tau_T] < 0.1$, then $N = 19$ should be selected for 3G with compression, and $N = 14$ for Wi-Fi with compression. On the other hand, to achieve the same $E[n]$ and $E[w]$ performances when $V_V \leq 0.04E[\tau_V]^2$, $N = 2$ is enough. Further investigation indicates that image compression significantly improves the performance in the 3G environment. As a final remark, the prefetching mechanism of CloudPocket can be easily implemented to improve the user experience, and this mobile web album service is an award winner in a nation-wide competition of Taiwan in 2012. In the future, we will explore user experience factors other than image transmission delay and viewing time to accommodate more complicated user behavior of web album; for example, visual attention based image browsing [19].

Acknowledgement

This work was supported in part by NSC 100-2221-E-009-070 and 101-2221-E -009-032, Academia Sinica AS-102-TP-A06, Chunghwa Telecom, IBM, Arcadyan Technology Corporation, the ITRI/NCTU JRC Research Project, the ICL/ITRI Project,

Nokia Siemens Networks, Department of Industrial Technology (DoIT) Academic Technology Development Program 101-EC-17-A-03-S1-193, and the MoE ATU plan.

Reference

[1] HTC, "HTC One," 2013. [Online]. Available: http://www.htc.com/tw/smart phones/htc-one/.

[2] Samsung, "GALAXY S III," 2012. [Online]. Available: http://www.samsung. com/global/galaxys3/.

[3] ITRI, "CloudPocket," 2012. [Online]. Available: http://developers.buddysquare. com/Case_Studies_cloudpocket.php. (in chinese)

[4] W. Li, "Autopager Chrome," 2012. [Online]. Available: http://chrome.google. com/webstore/detail/autopager-chrome/.

[5] J. Hung, "jQuery plugin," 2012. [Online]. Available: http://github.com/jayhung/ prefetch.

[6] J. F. Kurose and K. W. Ross, *Computer Networking: A Top Down Approach Featuring the Internet*. Addison-Wesley, Inc., 2007.

[7] M. Connolly, "ZipArchive," 2010. [Online]. Available: http://github.com/ mattconnolly/ZipArchive.

[8] Y. Fang and I. Chlamtac, "Teletraffic Analysis and Mobility Modeling for PCS Networks," *IEEE Transactions on Commun.*, vol.47, no.7, pp. 1062-1072, Jul. 1999.

[9] S.-R. Yang, "Dynamic Power Saving Mechanism for 3G UMTS System," *ACM/Springer Mobile Networks and Applications*, vol. 12, no. 1, pp. 5-14, Nov. 2007.

[10] H.-L. Fu, P. Lin, Y. Fang, and T.-Y. Wang. "Tradeoff between Energy

Efficiency and Report Validity for Mobile Sensor Networks." to Appear in *ACM Transactions on Sensor Networks*.

[11] K. Thomas, and Z.-G. Gao, "Some divisibility properties of Catalan numbers, " *Mathematical Gazette*, 95:96–102, 2011

[12] 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.

[13] Y.-B. Lin, "Eliminating Tromboning Mobile Call Setup for International Roaming Users," *IEEE Trans. Wireless Commun.,* 8(1): 320-325, 2009.

[14] W.-E.Chen, Y.-B. Lin, and R.-H. Liou, "A Weakly Consistent Scheme for IMS Presence Service," *IEEE Trans. Wireless Commun*., vol. 8, pp. 3815-3821, Jul. 2009.

[15] H.-N.Hung, Y.-B. Lin, and C.-L. Luo, "Deriving the Distributions for the Numbers of Short Message Arrivals," to appear in *Wireless Communications and Mobile Computing Journal.* (pdf file available in http://onlinelibrary.wiley.com/doi/10.1002/wcm.2194/abstract)

[16] Y.-C. Sung, and Y.-B. Lin, "IPsec-based VoIP Performance in the WLAN Environment," *IEEE Internet Computing*, 12(6): 77-82, 2009.

[17] Y.-B. Lin, et. al. "Performance Measurements of TD-LTE, WiMAX and 3G Systems," to appear in *IEEE Wireless Communications*. (pdf file available in http://liny.csie.nctu.edu.tw/document/tdp.pdf)

[18] Fu. Lin, "Improve the Quality of Wireless Internet Services by Using Agent," *IEEE International Conference on Computer Commun*., pp. 1-4, 2006.

[19] Fan, Xin, et al. "Visual Attention Based Image Browsing on Mobile Devices," *IEEE International Conference on Multimedia and Expo*, 2003.