Impact of Inter-Update Time Distributions on Wireless Data Access *

Wei-Ru Lai
Dept. of Electrical Engineering
Yuan Ze University
wrlai@saturn.yzu.edu.tw

November 19, 2001

Abstract

Modern mobile networks support wireless data applications. An application running on the wireless handheld device may repeatedly access a data entry received from the application server. If the data entry is not sensitive to time, then the customer may access the data stored in the cache of the wireless handheld device instead of querying the application server, and the expensive wireless transmission overhead is reduced. If the data entry is sensitive to time, then the current data entry should be provided from the application server. Some time-sensitive wireless applications can tolerate certain degree of inaccuracy. For this type of applications, we can set an expiration period t to predict when the data entry will be updated. During period t, the data entry in the handheld device is used. When t expires, the next data access results in a query to the mobile network. In this case, the application is *weakly consistent* where the wireless handheld device may occasionally access stale data. A mechanism is required to predict when a data entry expires. In Apache and Squid, a time-to-live (TTL) interval t is defined for the data entry stored in the wireless handheld device. We propose an analytic model to provide lower bound performance for the TTL prediction mechanism. Based on our model, we show how the mean, the variance, and the skewness of the inter-update time distribution affect the accuracy of TTL interval prediction.

*This research was sponsored in part by the National Science Council under contract NSC 89-2213-E-009-203 and the Lee and MTI Center for Networking Research, NCTU.

Keywords: wireless data, weakly consistency, time-to-live

1 Introduction

Modern mobile networks support wireless data applications. An example is Wireless Application Protocol (WAP) [7, 4]. In this environment, a mobile customer may use a wireless handheld device (e.g., a wireless PDA) to access data services from the application server through the mobile network. An application running on the wireless handheld device may repeatedly access a data entry received from the application server. If the data entry is not sensitive to time, then the customer may access the data stored in the cache of the wireless handheld device instead of querying the application server, and the expensive wireless transmission overhead is reduced. If the data entry is sensitive to time, then the current data entry should be provided from the application server. Some time-sensitive wireless applications can tolerate certain degree of inaccuracy (e.g., most web accesses and location dependent information in wireless applications). For this type of applications, we can set an expiration period t to predict when the data entry will be updated. During period t, the data entry in the handheld device is used. When t expires, the next data access results in a query to the mobile network. In this case, the application is *weakly consistent* where the wireless handheld device may occasionally access stale data. A mechanism is required to predict when a data entry expires. In Apache [1] and Squid [6], a

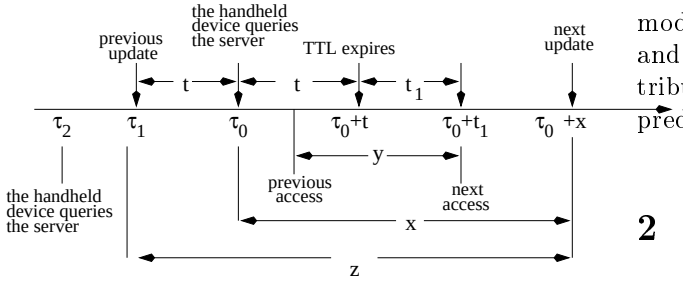

Figure 1: The Timing Diagram

time-to-live (TTL) interval t is defined for the data entry stored in the wireless handheld device. Consider the timing diagram in Figure 1.

In this figure, the consecutive updates to the data entry occur at time τ_1 and $\tau_0 + x$. When the server is queried for the data entry at time τ_0, the TTL interval t for the data entry is set as

$$t = c_f(\tau_0 - \tau_1) \tag{1}$$

where c_f is a system-defined fudge factor (in Figure 1, $c_f = 1$). Note that the TTL prediction mechanism is typically exercised with cache replacement such as LRU (least recently used) and LFU (least frequently used) [2] in a proxy cache for WWW accesses. Since the storage of a handheld device is limited, the wireless application may determine that no cache replacement algorithm is exercised for frequently accessed data (or they are likely to be replaced by infrequently accessed data). That is, when a wireless handheld device runs a particular application, some data used by this application are considered as "frequently accessed", and will always be kept in the handheld device until they expire. This is especially true for some location dependent services provided by the mobile operators. The customer may also enable a data entry as "frequently accessed", and the handheld device will not exercise cache replacement for this data entry until the frequently accessed indication is disabled.

We investigate the performance of the TTL-based algorithm for frequently accessed wireless data with weak consistency. The mechanism considered in this paper is the one used in Apache and Squid. We propose an analytic model to provide lower bound performance for the TTL prediction mechanism. Based on our model, we show how the mean, the variance, and the skewness of the inter-update time distribution affect the accuracy of TTL interval prediction.

2 Input Parameters and Output Measures

Define a *cycle* as the interval between two consecutive queries from the wireless handheld device to the server. For example, in Figure 1, the accesses at time τ_2 and τ_0 result in two consecutive queries to the server, and thus $[\tau_2, \tau_0)$ is a cycle. During the cycle time, the handheld device returns the cached copy to all accesses to the data entry. Let random variable K_1 represent the number of the non-stale accesses (to cache or the server). And let random variable K represent the total number of accesses (including stale and non-stale accesses) in a cycle. We derive the following primary output measures:

- The expected number $E[K_1]$ of *non-stale* accesses in a cycle: For a non-stale access, when the access occurs, the data entry in the cache is the same as that in the server. Note that the non-stale accesses include the one that results in the query to the server at the beginning of a cycle (for the cycle $[\tau_2, \tau_0)$ in Figure 1, this query occurs at τ_2).
- The expected number $E[K]$ of accesses in a cycle: This number includes the stale and the non-stale accesses in the cache plus the access resulting in a query from the handheld device to the server.

It is clear that the handheld device communicates with the server for every $E[K]$ accesses. Based on $E[K_1]$ and $E[K]$, we can investigate the accuracy of TTL interval prediction through the *staleness ratio* p_s, which is the probability that the handheld device returns a stale data entries for an access. That is,

$$p_s = \frac{E[K] - E[K_1]}{E[K]} \tag{2}$$

It is clear that the smaller the p_s value, the better the TTL prediction.

3 Bound Analysis for TTL Prediction

Assume that the data accesses are a Poisson stream with rate λ. Let random variable Y represent the inter-access time, then Y is exponentially distributed with mean $1/\lambda$, and

$$E[Y] = \frac{1}{\lambda} \tag{3}$$

Let random variable Z represent the inter-update time, which has a general cumulative distribution function $F(z)$, density function $f(z)$, Laplace transform $f^*(s)$, mean $E[Z] = 1/\mu$, variance σ^2, and skewness ω. Note that the variance is a measure of spread, i.e., if the values of a random variable tend to be far from their mean. The skewness is the third moment about the mean, which is a measure of asymmetry. Suppose that Z is a non-lattice random variable and $E[Z^2] < \infty$, then the residual life X of Z has the cumulative distribution function $R(x)$, density function $r(x)$, Laplace transform $r^*(s)$, mean $E[X] = 1/\bar{\mu}$ and variance $\bar{\sigma}^2$, where from [5]

$$\begin{aligned}
r(x) &= \mu\Big[1 - F(x)\Big] \qquad (4)\\
r^*(s) &= \left(\frac{\mu}{s}\right)\Big[1 - f^*(s)\Big]\\
E[X] &= \frac{1}{\bar{\mu}} = \frac{E[Z^2]}{2E[Z]} = \frac{\sigma^2\mu}{2} + \frac{1}{2\mu}\\
E[X^2] &= \frac{E[Z^3]}{3E[Z]} = \frac{\omega\mu}{3} + \sigma^2 + \frac{1}{3\mu^2}\\
\bar{\sigma}^2 &= E[X^2] - (E[X])^2\\
&= \frac{\omega\mu}{3} + \sigma^2 + \frac{1}{3\mu^2} - \left(\frac{\sigma^2\mu}{2} + \frac{1}{2\mu}\right)^2 \qquad (5)
\end{aligned}$$

Let $X_1, X_2, ..., X_n$ be n independent variates, each with cumulative distribution function $R(x)$, density function $r(x)$, mean $1/\bar{\mu}$ and variance $\bar{\sigma}^2$. From [3],

$$\begin{aligned}
E\left[\min_{1\le i\le n} X_i\right] &\ge \frac{1}{\bar{\mu}} - \frac{(n-1)\bar{\sigma}}{\sqrt{2n-1}}\\
&\text{and}\\
E\left[\max_{1\le i\le n} X_i\right] &\le \frac{1}{\bar{\mu}} + \frac{(n-1)\bar{\sigma}}{\sqrt{2n-1}} \qquad (6)
\end{aligned}$$

The inequality (6) is used later.

We consider the bound for p_s when $c_f = 1$. Suppose that a query to the server occurs at time τ_0 as shown in Figure 1. Let random variable T be the TTL interval computed in Apache [1]. For $c_f = 1$, random variable T is the reverse residual life of Z. In Figure 1, $T = t$. From the reversibility property of residual life, T has the same distribution as X (the residual lift of Z), and

$$E[T] = \frac{1}{\bar{\mu}} \tag{7}$$

During a cycle the non-stale accesses occur in the period $T_{min} = E[\min(T, X)]$. From (6),

$$E[T_{min}] \ge \frac{1}{\bar{\mu}} - \frac{\bar{\sigma}}{\sqrt{3}} \tag{8}$$

Since the accesses are a Poisson stream, from [5], (3) and (8), the expected number $E[K_1]$ of non-stale accesses in a cycle is

$$\begin{aligned}
E[K_1] &= 1 + \frac{E[T_{min}]}{E[Y]} \qquad (9)\\
&\ge 1 + \left(\frac{\lambda}{\bar{\mu}}\right) - \frac{\bar{\sigma}\lambda}{\sqrt{3}} \qquad (10)
\end{aligned}$$

Similarly, from [5] the expected values of the total number of accesses in a cycle is

$$E[K] = \frac{E[T]}{E[Y]} + 1 = \frac{\lambda + \bar{\mu}}{\bar{\mu}} \tag{11}$$

From (10) and (11), we have

$$p_s = \frac{E[K] - E[K_1]}{E[K]} \le \frac{\bar{\sigma}\lambda\bar{\mu}}{\sqrt{3}(\lambda + \bar{\mu})} = p_s^+ \tag{12}$$

An alternative approach to derive the upper bound p_s^+ is the following. Let

$$W = \begin{cases} 0, & \text{if } T < X \\ T - X, & \text{if } T \ge X \end{cases}$$

Random variable W represents the period where the data accesses are stale. Since T and X have the same distribution, $\Pr[T > X] = 1/2$. From (8), we have

$$\begin{aligned}
E[W] &= \Pr[T > X](E[\max(T, X)] - E[\min(T, X)])\\
&\le \left(\frac{1}{2}\right)\left(\frac{2\bar{\sigma}}{\sqrt{3}}\right)\\
&= \frac{\bar{\sigma}}{\sqrt{3}} \qquad (13)
\end{aligned}$$

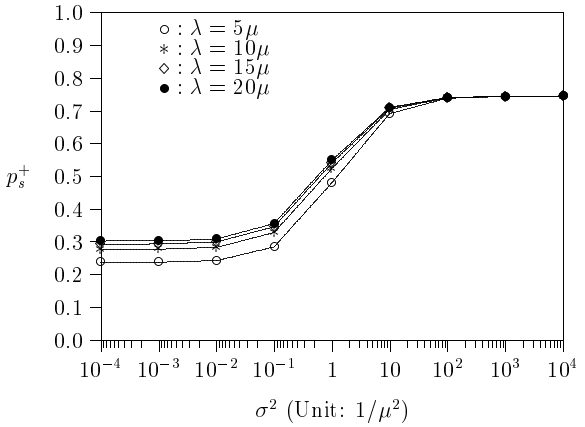

Figure 2: The Measure p_s^+

From (2), (11) and (13),

$$p_s = \frac{E[W]/E[Y]}{E[K]} \leq \frac{\bar{\sigma}\lambda\bar{\mu}}{\sqrt{3}(\lambda+\bar{\mu})} = p_s^+$$

which yields the same bound as (12).

4 Discussion

For the illustration purpose, we consider the inter-update interval Z that has a gamma distribution with mean $1/\mu$ and variance σ^2. That is

$$\begin{aligned} f^*(s) &= (1+\sigma^2\mu s)^{\frac{-1}{(\sigma\mu)^2}} \\ E[Z] &= -\left.\frac{df^*(s)}{ds}\right|_{s=0} = \frac{1}{\mu} \\ E[Z^2] &= (-1)^2\left.\frac{d^2f^*(s)}{ds^2}\right|_{s=0} = \frac{1}{\mu^2}+\sigma^2 \\ E[Z^3] &= (-1)^3\left.\frac{d^3f^*(s)}{ds^3}\right|_{s=0} = \mu+\frac{3\sigma^2}{\mu}+2\mu\sigma^4 \end{aligned} \quad (14)$$

Substituting (14) into (5), we obtain $\bar{\mu}$ and $\bar{\sigma}^2$, which are used in (12) to evaluate p_s^+.

Figure 2 plots p_s^+ as a function of λ (normalized by μ) and σ^2 (normalized by $1/\mu^2$). The figure indicates, for example, that if the variance σ^2 is less than $0.1/\mu^2$, at most two stale accesses are expected for every ten data accesses.

If Z has an exponential distribution, then $F(z) = e^{-\mu z}$. From (4), $r(x) = \mu e^{-\mu x}$ and

$$\begin{aligned} E[T_{min}] &= \int_{t=0}^{\infty} t\binom{2}{1} r(t) \int_{x=t}^{\infty} r(x)dxdt \\ &= \int_{t=0}^{\infty} t(2\mu)e^{-2\mu t}dt = \frac{1}{2\mu} \end{aligned}$$

From (2) and (9),

$$p_s = \frac{\lambda}{2(\lambda+\mu)} \quad (15)$$

From (15) and (12), for the exponential inter-update interval case, the errors between p_s and p_s^+ are about 15% for various λ values.

In summary, this paper derives a simple equation that can easily evaluate the upper bound of staleness of data access. The input parameters are the mean, the variance and the skewness of the inter-update intervals, which can be easily obtained from measurement.

References

[1] Apache 1.3. HTTP Server Document; http://www.apache.org, 2000.

[2] Cao, P., and Irani, S. Cost-Aware WWW Proxy Caching Algorithms. *Proc. Usenix Symp. Internet Technologies and Systems*, 1997.

[3] H.A. David. *Order Statistics*. Wiley And Sons, 2nd edition, 1981.

[4] Lin, Y.-B., and Chlamtac, I. *Mobile and Wireless Network Protocols and Services*. John Wiley & Sons, 2000.

[5] Ross, S.M. *Stochastic Processes*. John Wiley & Sons, 1996.

[6] Squid 2.3. Internet Object Cache Document; http://squid.nlanr.net/Squid, 2000.

[7] WAP Forum. Wireless Application Protocol Architecture Specification. Technical report, WAP Forum, 1998.