Performance Models of Web Caching and Prefetching for Wireless Internet Access[*]

Hisashi Kobayashi Shun-Zheng Yu

hisashi@ee.princeton.edu syu@ee.princeton.edu

Dept. of Electrical Eng., Princeton University, Princeton, NJ 08544

ABSTRACT There is a fair amount of study reported on the performance of Web caching and prefetching for wired Internet access. In a related matter, statistical characteristics of Web traffic have been recently studied by a number of researchers. In a wireless network, the movement of mobile users presents a scenario of access to the Internet that is substantially different from the wired network. Requests for content issued by a mobile user depend on its mobile state (e.g., location, velocity and direction). Therefore, we need to consider the probability with which mobile users access Web pages. The performance of caching and prefetching in the proxy gateways that connect the mobile users to the Internet may be highly predictable.

In this paper, we first present a brief survey of these statistical properties that have been reported in the literature. Then we construct a new analytical/numerical model that characterizes mobile user behavior in a general state-space using a semi-Markov process representation. Based on the mobility model and the resultant request model, we analyze the content access patterns and then obtain estimates of the total average latency, hit ratio, cache capacity and bandwidth resources required for the wireless and wired network. Finally, we charaterize dynamic behavior of the aggregate request rate and the aggregate traffic rate.

Keywords: Wireless Internet, Cache, Prefetch, Performance

1. Introduction

In a wireless communications network, the movement of mobile users presents significant technical challenges for us to provide efficient wireless access to the Internet. For a given individual mobile user, his or her location, velocity and direction will vary in time. It is therefore imperative to take into account dynamic mobile behavior in evaluating performance of Web caching and prefetching performed by proxy gateways and in allocating resources to traffic at the wireless and wired network interface.

In this paper we introduce a new integrated model of mobility and traffic that differs from existing work [1][2][3]. The model allows us to exploit recent results in queuing and loss network theory [4] and to characterize the macroscopic mobility and traffic behavior in the wireless network.

We apply this new model to the important problems of caching and prefetching performance evaluation and efficient resource allocation in wireless networks. First, we show how the mobility information obtained from our model can be used in constructing request patterns that mobile users access documents. Second, we show how request information obtained from the model can be used to evaluate the performance of caching and prefetching, including the total average access latency, hit ratio, cache capacity and transmission rate in wireless and wired networks. Third, by exploiting recent results in diffusion approximation [5], we obtain the dynamic behavior of aggregate request stream and aggregate response traffic.

The remainder of the paper is organized as follows. Section 2 surveys relevant results that have been reported in the literature. Section 3 develops our integrated model of user mobility and requirements in the wireless network. Section 4 discusses request patterns with which mobile users access documents, based on the mobility and requirement model. Section 5

[*] This work has been supported, in part, by grants from the NJCWT (New Jersey Center for Wireless Telecommunications), and NTT DoCoMo, Inc., Japan.

discusses the total average access latency, hit ratio, cache capacity and transmission rate in wireless and wired networks. Section 6 presents the simulation results. Finally, Section 7 concludes the paper.

2. Properties of Web Traffic

There are many factors that can affect the performance of Web proxy caching and prefetching. Statistical properties of Web traffic over the wired Internet have been reported in a number of papers. They serve as useful references in evaluating the performance of Web caching and prefetching for the wireless Internet access.

Highly skewed pattern in document access:

Document popularity: The popularity of Web documents is highly skewed [6] - [13]. The distribution of document requests generally follows a Zipf's law -like distribution where the probability of requests for the i th most popular document is proportional to $1/i^{\alpha}$, with α typically taking on some value less than unity [6][14][15].

Concentration of references: Most accesses concentrate on a small number of "hot" documents [6][16]. In a typical situation 30 percent of the documents may account for 60-80 percent of the requests.

One-time referencing: Typically 15-30 percent of the documents accessed in a log are accessed only once in the log, regardless of the duration of the access log studied [17].

Temporal locality: This refers to the property that a recently referenced file is likely to be referenced again in the near future [7][18][19][20]. Temporal locality can be measured by the Least Recently Used (LRU) stack-depth analysis. When a file is referenced, its current depth in the stack is recorded, and then the file is moved to the top of the stack. A high degree of temporal locality will give a small average stack depth recorded relative to the maximum stack depth. Conversely, a low degree of temporal locality will lead to a large mean stack depth.

Document types: HTML and image documents account for close to 95 percent of the total requests and image documents are consistently the most frequently requested document type (68-78 percent), followed by HTML documents (about 20 percent) [7] [8][9] [17].

Limits of Caching:

For an infinite sized cache, the hit-ratio for a web proxy grows as a logarithmic function of the client population of the proxy and of the number of requests seen by the proxy [19]-[22]. The hit-ratio of a web cache grows in a log-like fashion as a function of the cache size [8] [13] [16] [18] [19] [21] [23] [24]. The factors that affect the hit ratio include uncachable dynamic content (e.g., stock quotes, advertising banner with nonce URLs), "consistent misses" due to frequent update of documents, and "compulsory misses" due to one-time referencing.

High rate of change: Many potentially cacheable Web resources appear to change more rapidly than they are re-requested, unless the request stream comes from a very large population [15][25]. This would lead to "cache inconsistency" if responses for these resources were allowed to be cached. The correlation between a document's access frequency and its average amount of modifications per request is generally quite low [6].

Request process:

Interpage request time: The author of [28] verified that the aggregate Web page request process can be approximated by a Poisson process if bandwidth availability is high and page download time low. This approximation is also fairly good, even when the mean page download time is not negligible.

Heavy-tailed document size distribution:

Document size distribution: The document size distribution is heavy-tailed [8]-[11] [17] [26]. A distribution is considered heavy-tailed if $P[X>x] \sim x^{-\alpha}, \quad x \to \infty, 0<\alpha<2$. This means that the variable X is distributed over a very wide range and that the larger values of x, even if less probable, may still account for a non-negligible portion of the traffic. The document size distribution has been found to be *lognormal* [12][29]. That is, after applying a logarithmic transformation to the data, the data appears to be normally distributed.

3. Mobility Model

Abstract Mobility State Space

We define the state of a mobile user in terms of a vector (x_1, …, x_n), where the i th component, x_i, represents a value from a finite *attribute* space A_i. The attribute spaces represent properties of the mobile user such as location, moving direction, speed, etc. The set of possible states for a mobile user is an n -dimensional vector space given by

$$S = A_1 \times \cdots \times A_n, \qquad (1)$$

where × denotes the Cartesian product. The abstract space S can be made as rich as desired by including the appropriate attributes as components in the state vector. The dynamic motion of a user, as defined by its time-varying attribute values, can then be described by its trajectory in this space.

We enumerate all possible states in S and label them as 1, …, M such that the state space S can more simply be represented as follows:

$$S = \{\ 1, \ldots, M\ \}. \qquad (2)$$

We introduce two *inactive* states in addition to the set of *active* states S: the *source* state 0 and the *destination* state d. A user enters the system by assuming the state 0. A user exits the system by assuming the state d. Thus, the user can assume states in the augmented state-space $S' = S \cup \{0, d\}$.

No transitions occur from states $j \in S$ to the source state, i.e., $a_{j0} = 0$. From any such state j, the user next assumes the destination state d with probability a_{jd}. No transitions are allowed from the destination state. Hence, the state d is considered to be the *absorbing* state of the Markov chain. Further, no transitions occur from state 0 to state d, i.e., $a_{0d} = 0$. The state transitions of a user are characterized by a Markov chain with transition probability matrix:

$$\boldsymbol{A}' = \begin{array}{c} d \\ 0 \\ 1 \\ : \\ : \\ M \end{array} \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & a_{01} & a_{02} & \ldots & a_{0,M} \\ a_{1d} & 0 & a_{11} & a_{12} & \ldots & a_{1,M} \\ a_{2d} & 0 & a_{21} & a_{22} & \ldots & a_{2,M} \\ : & : & : & : & : & : \\ a_{M,d} & 0 & a_{M,1} & a_{M,2} & \ldots & a_{M,M} \end{bmatrix} \qquad (3)$$

We note, however, that transitions among the states is limited and we may assume that from a given state transitions can occur to on the order of ten neighboring states. Such considerations imply that the transition probability matrix will be highly sparse in practical applications.

We assume the dwell time of a user in state $m \in S$ to be generally distributed with mean d_m. Hence, the state process of a user is a semi-Markov chain. The transition probability matrix and the state duration distributions can be estimated by means of a parameter estimation algorithm discussed in [32].

The aggregate behavior of the system of mobile users can be represented by the vector process

$$\boldsymbol{N}(t) = (\ N_1(t), ..., N_M(t)\ ), \qquad (4)$$

where $N_m(t)$ represents the number of mobile users in state m at time t.

Observe that the above system is equivalent to an open queuing network with M infinite-server stations corresponding to the states in S. Clearly, the source and destination stations of the queuing network correspond to 0 and d, respectively. Results from the theory of queuing and loss networks [4] show that the steady state distribution of $\boldsymbol{N}(t)$ is insensitive to the distributions of the dwell times at each station.

From

$$e_m = a_{0m} + \sum_{n \in S} e_n a_{nm}\ , \qquad m \in S \qquad (5)$$

we get the value e_m, which can be interpreted as the average number of visits that a user makes to state m during its sojourn in the system starting from the source state 0 until reaching the destination state d.

Our proposed abstract mobility state space model differs from other proposed mobility models (cf. [30][31]) in that it leads to a simple parametric representation of the mobile behavior that can be related to a general queuing network with multi-class users in which each service center is infinite server (IS) with multiple types. This representation allows us to capitalize on recent results in queuing and loss network theory [4] which show that the steady-state distribution is surprisingly robust to all state time distributions and state transition behaviors. This result in turn implies that to obtain the state distribution of mobile users, we need only have two sets of parameters: the mean dwell time, d_m, in state m and the expected number of visits, e_m, the user makes to state m in its lifetime per user class. Thus, only 2 M pieces of numeric data per user class provide sufficient statistics of the user mobility, as far as the steady-state distribution and related performance measures are concerned.

Departure rate from each state

Let N_m denote the expected number of users in state m in equilibrium (m=0, 1, ..., M). The mean departure rate from state m is given by

$$\lambda_m = N_m/d_m = \lambda_0\, e_m, \qquad m=1, 2, \ldots, M, \qquad (6)$$

where λ_0 is the total rate at which mobile users transit from inactive state 0 to an active state, i.e., the total rate of entry to the system, and d_m is the mean dwell time in state m.

4. Access to Document

Request rate for each document

We can augment the above mobility model by introducing state-dependent information. Let j =0, 1, ..., J represent a set of user requirements, where 0 specially represents no requirement. We shall suppose that a mobile user entering state m will require content of type j from the network with probability $c_m(j)$, which satisfies:

$$\sum_{j=0}^{J} c_m(j) = 1\,, \qquad m=1, \ldots, M. \qquad (7)$$

The request rate for content j is then given by

$$R_j = \sum_{m=1}^{M} \lambda_m c_m(j) = \sum_{m=1}^{M} \frac{c_m(j)}{d_m} N_m\,, \qquad j=1, \ldots, J. \qquad (8)$$

The instantaneous request rate for content j can be defined by

$$R_j(t) = \sum_{m=1}^{M} \frac{c_m(j)}{d_m} N_m(t)\,, \qquad j=1, \ldots, J. \qquad (9)$$

If the dwell time distributions of the user states are assumed exponential, $\boldsymbol{N}(t)$ is a Markov process, hence the request rate process $R_j(t)$ defined here can be viewed as a Markov modulated rate process (MMRP) as studied in [5]. If we allow the dwell times to have general distributions, $R_j(t)$ becomes what we may term as a semi-Markov modulated rate process. Let $\boldsymbol{X}(t)$ be an M-dimensional diffusion process that approximates the M-dimensional semi-Markov process $\boldsymbol{N}(t)$. Under a set of reasonable assumptions [5], $\boldsymbol{X}(t)$ can be expressed as an M-dimensional Ornstein-Uhlenbeck (O-U) process. Hence, the process $R_j(t)$ can be approximated by a Gaussian process

$$\tilde{R}_j(t) = \sum_{m=1}^{M} \frac{c_m(j)}{d_m} \cdot X_m(t) \qquad j=1, \ldots, J. \qquad (10)$$

Aggregate request rate

Let R denote the total request rate for the set of J contents.

$$R = \sum_{j=1}^{J} R_j = \sum_{m=1}^{M} \lambda_m \sum_{j=1}^{J} c_m(j) = \sum_{m=1}^{M} \frac{1-c_m(0)}{d_m} \cdot N_m\,. \qquad (11)$$

Define the instantaneous total request rate by

$$R(t) = \sum_{m=1}^{M} \frac{1-c_m(0)}{d_m} \cdot N_m(t)\,. \qquad (12)$$

Then, the process $R(t)$ can be similarly approximated by a Gaussian process

$$\tilde{R}(t) = \sum_{m=1}^{M} \frac{1-c_m(0)}{d_m} \cdot X_m(t)\,, \qquad (13)$$

Access probability for each document

Define the ratio

$$\gamma_j = R_j / R \qquad (14)$$

represent the probability of request to content j, j =0, 1, 2, …, J.

Average hit rate for each document

Let q_j be the probability that there is at least one request to content j during the content's lifetime μ_j. Suppose we observe k update intervals for content j. Then there will be on the average kq_j such intervals in which at least one request is made to content j. The first request in a given update interval will miss a fresh copy of the content and must fetch it from the origin server. The consequent requests in the interval use the cached copy. Thus, the average hit rate for this document is given by

$$h_j = 1 - \frac{kq_j}{k\mu_j R_j} = 1 - \frac{q_j}{R_j \mu_j}, \quad j=1, 2, \ldots, J, \qquad (15)$$

where $k\mu_j R_j$ is the total number of requests to content j during the interval $k\mu_j$.

5. Performance of Caching and Prefetching

The total access latency

Now we derive general expressions for the average latency and hit ratio of a generic cache and prefetch scheme. Due to finite capacity of cache, any prefetch scheme cannot cache all the documents. Suppose that r documents are prefetched to the cache. By "prefetch" we mean the action that a proxy Web server takes by automatically caching and updating any of the selected r documents as soon as it has expired, and this action is not driven by the client requests. Therefore, the average latency L for a prefetch scheme is given by

$$L = \sum_{j=r+1}^{J} \gamma_j \left[h_j T_{j,c} + (1-h_j) T_{j,s} \right] + \sum_{j=1}^{r} \gamma_j T_{j,c}$$

$$= \sum_{j=1}^{J} \gamma_j \left[h_j T_{j,c} + (1-h_j) T_{j,s} \right] - \sum_{j=1}^{r} \gamma_j (1-h_j)(T_{j,s} - T_{j,c}), \qquad (16)$$

where γ_j is the access probability to content j given by (14), h_j is the hit rate for content j given by (15), $T_{j,c}$ is the time required to fetch content j from the cache to the mobile user, whereas $T_{j,s}$ is the time required to fetch content j from its origin server via the cache to the mobile user.

As is seen from (16), the average latency consists of two terms: the first term is solely determined by the document request rates $\{R_j\}$, update intervals $\{\mu_j\}$ and the response times $\{T_{j,c}, T_{j,s}\}$, and represents the latency of a "no-prefetch" cache scheme, i.e., the conventional caching; the second term of (16) is the reduction in latency that a prefetch scheme can provide. This reduction is determined by the number r and the selection of r documents to be prefetched, thus depends on the specific prefetch scheme to be adopted.

The total hit ratio

The total hit ratio is given by

$$H = \left(\sum_{j=r+1}^{J} R_j h_j + \sum_{j=1}^{r} R_j \right) \Big/ R = \sum_{j=1}^{J} \gamma_j h_j + \sum_{j=1}^{r} \gamma_j (1-h_j), \qquad (17)$$

where R_j is the request rate to document j, R is the total request rate, h_j is the hit rate, and γ_j is the access probability given by (14).

Obviously, the first term is independent of a specific choice of prefetch scheme, which is the total hit ratio of a "no-prefetch" cache scheme (i.e., the conventional caching); the second term is the increment of the hit ratio compared with the "no-prefetch" cache scheme.

The total cache capacity required

Now we proceed to derive expressions for the required cache capacity and bandwidth. Instead of considering a specific cache scheme, we generally assume that once a cached content expires, it is deleted from the cache. In other words, missing a "fresh" copy of document j in the cache means that the cache currently does not store any bytes of the document. Thus, the probability that a document is stored in the cache is just the hit probability of the document. For prefetching, we assume that the selected r documents are prefetched into the cache. Whenever any of the prefetched r contents expires, a fresh copy will be immediately fetched from its origin server. That is, the selected r documents are stored in the cache permanently until changing the prefetching selection. Therefore, the required cache capacity C is:

$$C = \sum_{j=r+1}^{J} z_j h_j + \sum_{j=1}^{r} z_j = \sum_{j=1}^{J} z_j h_j + \sum_{j=1}^{r} z_j (1-h_j), \qquad (18)$$

where z_j is the size of document j. Therefore, the prefetch scheme needs an extra cache capacity of $\sum_{j=1}^{r} z_j (1-h_j)$ compared with the caching only scheme. Thus, the improvements in the average latency L and the hit ratio H are obtained in exchange for the increase in cache capacity.

The total transmission rate required for wired network

The total transmission rate required for transmitting documents from the origin servers to the cache over the wired network should be

$$B_{wired} = \sum_{j=r+1}^{J} R_j (1-h_j) z_j + \sum_{j=1}^{r} z_j / \mu_j$$

$$= \sum_{j=1}^{J} q_j \frac{z_j}{\mu_j} + \sum_{j=1}^{r} (1-q_j) \frac{z_j}{\mu_j} \,, \qquad (19)$$

where we used Eq. (15). Therefore, a prefetch scheme needs an extra transmission bandwidth of $\sum_{j=1}^{r} (1-q_j) \frac{z_j}{\mu_j}$ over the bandwidth required by a cache only scheme. Thus, the improvement in the average latency and the hit ratio is achieved at the expense of increased bandwidth usage.

Aggregate traffic over the wireless network

The average transmission rate required for transmitting documents to the mobile users over the wireless network should be:

$$B_{wireless} = \sum_{j=1}^{J} R_j z_j \qquad (20)$$

The aggregate wireless transmission rate is

$$B_{wireless}(t) = \sum_{j=1}^{J} R_j(t) z_j = \sum_{m=1}^{M} \frac{\sum_{j=1}^{J} c_m(j) z_j}{d_m} \cdot N_m(t) \qquad (21)$$

where $R_j(t)$ was defined by (8) and can be approximated by a Gaussian process. Therefore, the process $B_{wireless}(t)$ can be approximated by a Gaussian process of the form:

$$\tilde{B}_{wireless}(t) = \sum_{m=1}^{M} \frac{\sum_{j=1}^{J} c_m(j) z_j}{d_m} \cdot X_m(t) \qquad (22)$$

where $X_m(t)$ is a diffusion process to approximate $N_\text{m}(t)$.

6. Simulation Results

In a serving area (about 1 km by 1 km), there are 128 street segments in a mesh layout. Each street segment is about 100 meters long. Assuming in each street segment, there are two walking states (in two directions), two driving states (in two directions) and one shopping state. There are in total 640 active states plus one inactive source state and one absorbing state. Each active state can transit to about ten other states in the neighboring street segments. The mean dwell time for walking state is about 3 minutes, that for a driving state is 16 seconds, and that for a shopping state, 12 minutes. There are 800 mobile users involved in the wireless Internet services. There are 20 categories associated with each street segment, each category has 20 contents. Therefore, there are total 51,200 contents for the serving area. The simulation results for request rate are shown in Figure 1 and Figure 2.

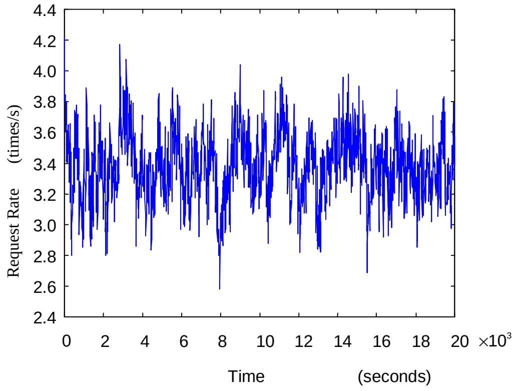

Figure 1 Request rate

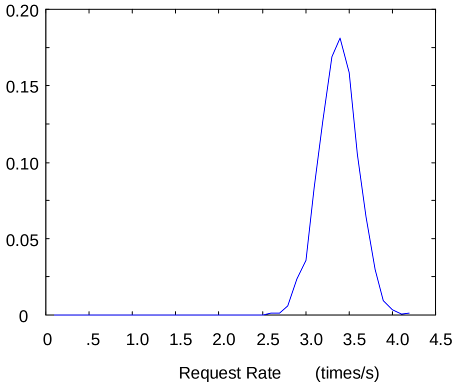

Figure 2 Request rate distribution

We assume that the content lifetimes $\{\mu_j\}$ are exponentially distributed with mean 24 hours. The average time $T_{j,c}$ required to request and transfer a content from the cache to the requesting mobile user is assumed 1s, and the average time $T_{j,s}$ required to request and transfer a content from the origin Web server to the requesting mobile user is 1.5s, for all j=1, 2, …, J.

For an infinite cache capacity, as the total number of active mobile users increases, the average latency L will decrease and the average hit ratio H will increase, because the more frequent access to the cache will increase the re-access probability to the cached documents, as shown in Figure 3 and Figure 4. We set the total number of prefetch contents as r = 200 and choose r documents as prefetched contents. From Figure 3 and Figure 4, we can see that the performance of latency and hit ratio can be significantly improved by performing prefetch scheme.

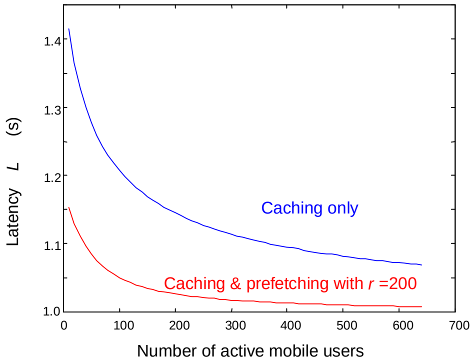

Figure 3 Average latency

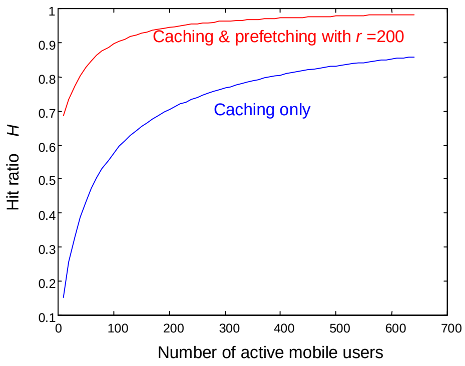

Figure 4 Average hit ratio

7. Conclusion

In this paper, we have proposed a new analytical/numerical model that characterizes mobile user behavior in a general state-space, based on a semi-Markov process representation. Based on the mobility model and the resultant request model, we obtained access patterns to documents and then obtained estimates of the total average latency, hit ratio, cache capacity and bandwidth resources required for the wireless and wired network. Finally, we obtained expressions for the dynamic behavior of the aggregate request rate and the aggregate traffic rate.

References

[1] P. C. Chen, "A cellular based mobile location tracking system," In *Proc. IEEE VTC'99*, pages 1979–1983, 1999.

[2] H. Maass, "Location-aware mobile applications based on directory services," *Mobile Networks and Applications*, (3):157–173, 1998.

[3] G. Y. Liu and G. Q. Maguire, "A predictive mobility management scheme for supporting wireless mobile computing," *Walkstation Project Technical Report, 1995-02-0.* (*available online: http://www.it.kth.se/labs/ccs/WS/papers*), 1995.

[4] H. Kobayashi and B. L. Mark, "Product-Form Loss Networks," In J. H. Dshalalow, editor, *Frontiers in Queueing: Models and Applications in Science and Engineering*, pages 147–195. CRC Press, 1997.

[5] Q. Ren and H. Kobayashi, "Diffusion process approximations of a statistical multiplexer with Markov modulated bursty traffic sources," *IEEE Jour. of Select. Areas in Commun.*, 16(5):679–691, 1998.

[6] L. Breslau, P. Cao, L. Fan, G. Phillips and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," *IEEE INFOCOM*, 1999, pp.126-134.

[7] M. Arlitt and C. Williamson, "Internet Web Servers: Workload Characterization and Performance Implications," *IEEE/ACM Trans. Net.*, vol. 5, no.5, Oct. 1997, pp.631-45.

[8] C. Cunha, A. Bestavros, and M. Crovella, "Characteristics of WWW Client-Based Traces," *Tech. Rep*. TR-95-010, Dept. of Comp. Sci., Boston Univ., Apr. 1995, available at ftp://cs-ftp.bu.edu/techreports/95-010-www-client-traces.ps.Z

[9] G. Abdulla *et al.*, "WWW Proxy Traffic Characterization with Application to Caching," *Tech. Rep*. TR-97-03, Comp. Sci. Dt., Virginia Tech., Mar. 1997, available at http://www.cs.vt.edu/~chitra/work.html

[10] H. Braun and K. Claffy, "Web Traffic Characterization: An Assessment of the Impact of Caching Documents from NCSA's Web Server," *Comp. Netorks and ISDN Sys*., vol. 28, no. 1 &2, Jan. 1995, pp.37-51.

[11] P. Barford and M. Crovella, "Generating Representative Web Workloads for Network and Server Performance Evaluation," *Proc. 1998 ACM SIGMETRICS Conf. Measurement and Modeling of Comp. Sys.*, Madison, WI, July

1998, pp.151-60, available http://cs-www.bu.edu/faculty/crovella/papers.html

[12] P. Barford et al., "Changes in Web Client Access Patterns," *World Wide Web J., Special Issue on Characterization and Performance Evaluation*, vol. 2, 1999, pp. 15-28, available at http://cs-www.bu.edu/faculty/crovella/paper-archive/traces98.ps

[13] V. Almeida et al., "Characterizing Reference Locality in the WWW," *Proc. 1996 Int'l. Conf. Parallel and Dist. Info. Sys.*, Miami Beach, FL, Dec. 1996, pp.92-103, available at http://www.cs.bu.edu/~best/reslpapers/pdis96.ps

[14] NLANR, Cache Popularity Index, http://www.ircache.net/Cache/Statistics/Popularity-Index

[15] G. Voelker et al., "On the Scale and Performance of Cooperative Web Proxy Caching," *Proc. 17th SOSP*, Kiawah Island, SC, Dec. 1999, pp. 16-31.

[16] S. Williams, M. Abrams, C. Standridge, G. Abdulla, and E. Fox, "Removal Policies in Network Caches for World-Wide Web documents," *Proc. SIGCOMM*'96, Stanford, CA, Aug. 1996, pp. 293-305.

[17] A. Mahanti, C. Williamson, and D. Eager, "Traffic Analysis of a Web Proxy Caching Hierarchy," *IEEE Network*, May/June, 2000, pp.16-23.

[18] L. Rizzo and L. Vicisano," Replacement Policies for a Proxy Cache" *Technical Report* RN/98/13, University College London, Dept. of Computer Science, UK, 1998, available: http://www.iet.unipi.it/~luigi/caching.ps.gz

[19] P. Cao and S. Irani, "Cost-aware WWW Proxy Caching Algorithms," *Proc. of the 1997 USENIX Symp. on Internet Tech. and Sys.*, pp. 193-206, Dec. 1997. available: http://www.cs.wisc.edu/~cao/publications.html

[20] T. M. Kroeger, J. Mogul, and C. Maltzahn, "Digital's Web Proxy Traces," ftp://ftp.digital.com/pub/DEC/traces/proxy/webtraces.html, August 1996

[21] S. Gribble and E. Brewer, "System Design Issues for Internet Middleware Services: Deductions from a Large Client Trace," *Proc. of the 1997 USENIX Symp. on Internet Tech. and Sys.*, Dec. 1997, available: http://HTTP.CS.Berkeley.EDU/~gribble/

[22] S. Gribble and E. Brewer, "UCB Home IP HTTP Traces," available: http://www.cs.berkeley.edu/~gribble/traces/index.html, June 1997

[23] S. Glassman, "A Caching Relary for the World Wide Web," *First International Conference on the World-Wide Web*, CERN, Geneva, Switzerland, May 1994, available: http://www1.cern.ch/WWW94/PrlimProcs.html

[24] B. M. Duska, D. Marwood, and M. J. Feeley, "The Measured Access Characteristics of World-Wide-Web Client Proxy Caches," *Proc. of the 1997 USENIX Symp. on Internet Tech. and Sys.*, Dec. 1997.

[25] F. Douglis et al., "Rate of Change and Other Metrics: A Live Study of the World Wide Web," *Proc. USENIX Symp. Internet Tech. and Sys.*, Monterey, CA, Dec. 1997, pp. 147-58.

[26] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, Dec. 1997, pp. 835-46.

[27] Deng, "Empirical Model of WWW Document Arrivals at Access Link," *Proc. IEEE ICC*, 1996, pp. 1797-1802

[28] M. Molina, P. Castelli, and G. Foddis, "Web Traffic Modeling Exploiting TCP Connections' Temporal Clustering through HTML-REDUCE," *IEEE Network*, May/June, 2000, pp.46-55.

[29] M. Arlitt, R. Friedrich, and T. Jin, "Workload Characterization of a Web Proxy in a Cable Modelm Environment," *ACM SIGMETRICS Perf. Eval. Rev.*, vol. 27, no.2, Aug. 1999, pp. 25-36.

[30] S. Tekinay, "Modeling and analysis of cellular networks with highly mobile heterogeneous sources," *Ph.D. dissertation*. School of Information Technology and Engineering, George Mason University, 1994.

[31] K. K. Leung, W. A. Massey, and W. Whitt, "Traffic models for wireless communication networks," *IEEE J. Select. Areas in Comm.*, 12(8):1353–1364, Oct. 1994.

[32] S.-Z. Yu and H. Kobayashi, "A Forward-Backward Algorithm for Hidden Semi-Markov Model and its Implementation," *submitted for publication*, Feb 2000.

[33] D. R. Cox, *Renewal Theory*. London, U.K.: Methuen, 1962.

[34] J. C. Mogul, "Squeezing More Bits Out of HTTP Caches," *IEEE Network*, May/June, 2000, pp.6-14.

[35] M. Arlitt and T. Jin, "A Workload Characterization Study of the 1998 World Cup Web Site," *IEEE Network*, May/June, 2000, pp.30-37.