Analyzing the Browse Patterns of Mobile Clients

Atul Adya, Paramvir Bahl, Lili Qiu
Microsoft Research, Redmond, Washington
{*adya, bahl, liliq@microsoft.com*}

Abstract—
We study the dynamics of a large popular commercial Web site designed specifically for users who access it via their cell-phones and PDAs. Unlike most previous Web studies that have analyzed accesses seen by proxies and servers from clients connected via the wired network, we focus primarily on client accesses made over wireless channels and made for downloading content on small devices for offline browsing. We carry out user-behavior analysis as users authenticate themselves before accessing and then every access is logged with a unique user identifier. Using browser traces gathered over a period of 12 days, we perform detailed content analysis, document popularity analysis and server load analysis. We answer questions like what sorts of content wireless users are most interested in, when and how much load they put on the servers, and how much time they spend on the channel while accessing the Web wirelessly. We discuss the implications of our findings for techniques such as query caching, server scheduling, channel use and TCP optimization.

I. Introduction

Over the last decade the cellular phone industry and the World Wide Web have experienced a phenomenal growth as people around the world have embraced these technologies at a remarkable rate. Understandably, there is much excitement around the impending deployment of third (3G) and fourth generation (4G) wireless cellular networks at the core of which are cellular phones and Personal Digital Assistants (PDAs) capable of accessing the Web wirelessly from anywhere at anytime. Today, most major wireless service providers in the United States, Europe, and Japan offer wireless Internet services and many Internet companies provide content that has been adapted to suit the limited display, bandwidth, memory, and processing power of small devices. It is therefore reasonable to assume that wireless Internet is real and that it is growing rapidly. To help sustain this growth over the next several years, a critical issue that needs to be examined is the performance of the "wireless Web" servers and the popularity of the content provided by Internet service companies to users accessing the Web wirelessly via small devices.

In this paper we study and analyze user behavior and its effect on the Web server that is designed specifically for cell-phone (wireless) users who browse the Web in real-time and for offline users who download content on to their PDAs or small mobile devices for later (offline) browsing. Our analysis is important from the perspective of content providers, wireless ISPs, and Web site managers as we answer questions such as: what type of content users are most interested in, what can be said about document popularity, how long do users stay on the wireless channel, how loaded are the Web servers and if catching can help in improving the performances of these specialized Web servers.

Some of our findings are: (i) As reported in some previous wired line Web trace studies the majority of client requests for wireless clients are also concentrated on a small number of documents. Unlike most previous studies, the popularity distribution of these documents do not follow Zipf-like distribution. (ii) More than 60% of the browse pages accessed at the Web server are due to offline PDA users (or automated program) and less than 7% are due to wireless clients. (iii) Most of the replies to wireless clients (real users) are less than 3 KBytes, and to offline clients (robots) are 6 KBytes.

Some of the implication of our results are as follows: (i) It is possible to cluster users according to their session times, the number of sessions they start up in a given time period, the number of bytes they typically download, and the link (wireless or wired) over which they browse data. With this information it is possible to create service strategies to manage expectations. (ii) The high concentration of requests to documents in the browser log implies that caching the results of popular queries would be very effective in reducing the wireless Web server load. (iii) The fact that offline PDA users generate significantly more requests than wireless users suggests that system designers have to be careful in ensuring that wireless clients are given a higher priority over offline PDA users.

The rest of this paper is organized as follows. In Section II, we describe the data we analyzed In Sections VI, III, IV and V we present detailed analysis of the browser logs focusing on wireless and offline users. In Section VII, we survey previous work. In Section VIII, we summarize our key results, discuss the broader implications of our findings and conclude with a brief discussion of ongoing and future work.

II. Data Gathering and Categorizing

The data we analyze in this paper is the Web browsing logs of requests that come from wireless clients, automated programs that download information into PDAs for offline access, and desktop clients. The logs are for a 12 day period between August 15, 2000 and August 26, 2000 and there are 33,005,944 entries. The Web site provides different types of content such as news (sports, local, national), weather, stock quotes, mail, yellow pages, travel reservations, entertainment information, etc. All browsing accesses are logged individually by each server along with a unique user id; we use this id to perform our user-based analysis. In all, we analyze 58,432 wireless users and 50,968 offline users.

The total size of the raw data was over 15 GB. To manipulate this data efficiently, we consolidated the logs and bulk-copied them into a commercial database system. The database allows us to run queries efficiently over the large dataset.

We categorized an access as being desktop, wireless or offline based on the browser type stored in the log entry corresponding to that access. For example, entries with browser type "Mozilla Windows", "Avantgo", "UP.Browser" are categorized as desktop, offline and wireless, respectively.

III. Document Popularity

Several studies, such as [1], [2], [3], [4], [11], have found that Web accesses follow Zipf-like distribution, that is, the number of requests to the i^{th} most popular object is proportional to $\frac{1}{i^{\alpha}}$. The estimates of α range from 0.5 to 1 in the Web proxy logs [6], [9], [4], and range from 1 to 2 in the Web server logs [2], [11]. It is very interesting to examine if the wireless Web exhibits similar property.

In Figure 1, we plot the number of requests to URLs versus their popularity ranking on log-log scale for the 8/15 trace. As we can see, there are three distinct linear regions in the graph: (i) URLs 1 - 25, (ii) URLs 25 - 134, (iii) URLs 134 - 290. This implies that the requests do not follow Zipf-like distribution when considering all the documents. We compute the slope of each regions using least square fitting, and find that the slopes for regions (i), (ii) and (iii) are -0.7914, -2.2997, and -11.9796, respectively. Similar performance is observed for the other days' logs.

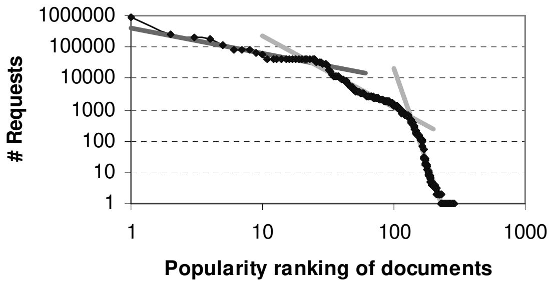

Fig. 1. Frequency of document accesses versus document ranking in log-log scale.

It is interesting to note that the popularity patterns do not match those of regular Web sites but they do bear resemblance to the popularity distributions observed in the 1998 World Cup Web site [1]. Those results exhibit a similar discontinuity of three linear regions in the popularity distribution curve. Though the boundaries of the three regions and their slopes in the World Cup logs are different from ours, they share the same characteristics, i.e., the first region is relatively flat, and the third region has a sharp drop. A possible reason for such discontinuity and deviation from Zipf-like distribution in both logs is that there are a small number of unique files served by these Web sites; Zipf-like distribution tend to exhibit in a relatively large data set.

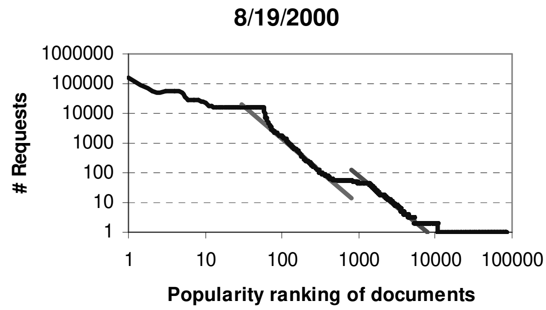

Fig. 2. Frequency of document accesses versus document ranking in log-log scale, where a document is defined as a unique URL and input parameter combination.

Since most Web pages in our dataset are dynamically generated, we now look at the distribution of requests to documents by taking the input parameters into account. For ease of discussion, in the remaining of this section, a document is referred to as a unique URL and parameter combination. We plot the number of requests to documents, and find after ignoring the top 100 samples, the number of requests decreases almost linearly with the number of documents. However for some days, the curve fits well with one straight line, and for other days, the curve fits well with two separate lines with similar slopes but different shifts; Figure 2 shows such a curve for 8/19.

Figure 3 shows the cumulative distribution of the re-

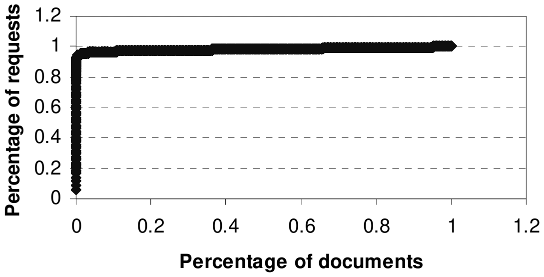

Fig. 3. CDF of requests to documents.

quests to documents for the 8/15 log; the logs for other days are similar. As we can see, majority of the requests are concentrated on a very small number of documents. In particular, 0.1% - 0.5% URL and parameter combinations (i.e. about 121 - 442 unique combinations) account for 90% requests. This implies that very small amount of memory is needed to cache popular query results and achieve a significant reduction in the Web server load. For example, unlike regular Web sites, the memory requirements for highly popular documents are considerably low (less than 2 MB from the above numbers). Wireless Web server designers can even replicate these documents at every front door server (if the data changes infrequently). Note that the real benefit of such optimizations is not to reduce the download time of a wireless client (which is constrained by its wireless link) but to reduce the load on the server CPU and disks, thereby rendering the server more scalable.

IV. User Behavior Analysis

Classifying users according to their access patterns is useful for personalization, targeted advertising, prioritizing, and capacity planning. We now present user-based analysis for the browser logs by taking advantage of the unique user identifier that is logged with every Web access. In [7], Kelly describes an alternative technique that measures Web client access patterns by modifying a proxy to intercept and serve all responses to clients marked as uncacheable.

A. Distribution of Wireless User Sessions

In order to understand user behavior, we would like to know how long a wireless user stays on the channel as he browses the Web. We use the notion of a *session* to model such a sequence of interactions initiated by a user on his micro-browser (i.e., browser on a cell-phone or a PDA).

A wireless service provider can utilize information about the number of sessions and length of sessions for effective pricing, capacity planning, and for providing service differentiation between users with different usage characteristics. A Web server administrator can classify users as short or long-session users and utlize this information for better load balancing (e.g., a uniform mix of long-session users and short-session users may result in fewer resource bottlenecks) and prefetching/caching strategies (e.g., cache or prefetch time-insensitive information for long-session users).

Since it was infeasible for us to instrument client micro-browsers for demarcating when a user sessions starts and ends, we approximate session times using a hueristic: if a user is idle for a “sufficiently long” time (called the *session-inactivity period*), we say that the session has ended. We now discuss our heuristic to determine the session-inactivity period.

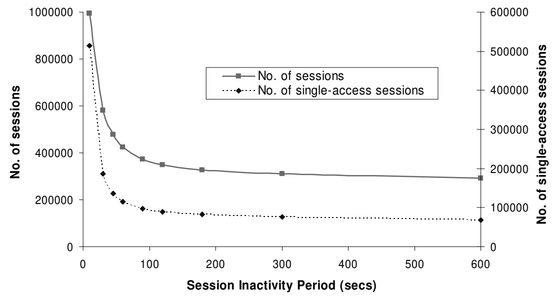

Fig. 4. Determining the session-inactivity period based on single-access sessions and number of sessions

When a user starts browsing, he accesses the home page (by default) and then traverses the Web site. Given the structure of the Web site under study, we expect few sessions in which only one page is accessed (i.e., the user connects but does not browse). Thus, if we choose a very small session-inactivity period, many single-session accesses will be incorrectly counted as part of separate sessions. As the session-inactivity period increases, these accesses are (correctly) classified as part of larger sessions and the number of single-hit sessions will decrease. The appropriate session-inactivity period is at the knee point where the change in its value does not produce a significant change in the number of single-hit sessions.

Figure 4 shows the number of single-hit sessions versus the session inactivity period. As we can see, the knee point is between 30 to 45 seconds, and we use it as the session-inactivity period. Note that even though our analysis is based on correctly classifying single-hit sessions, it impacts the classification of multi-hit sessions as well.

We can verify our chosen period in another manner. As the session-inactivity period is increased (starting at a small value), smaller sessions may merge into larger ones, thereby decreasing the total number of sessions (until all sessions are merged into one big session). There is a point at which the rate of decrease of sessions becomes relatively steady/low. Again, this is the region where the real session-inactivity period lies. In Figure 4, there is a relatively steep drop in the number of sessions from 10 to 30 seconds and then a smoother curve from 30 seconds onwards. Thus, this analysis confirms the fact that the session-inactivity period is approximately 30 to 45 seconds. This value is different from the one reported in [8] (where a value of 90 seconds was used based on the dynamic IP address timeout policy).

Using 30 seconds as our session-inactivity period, we now analyze user sessions in detail. For each user, we determined the total session time, the largest session time and the number of sessions initiated during this period. We then classified users according to these three parameters, e.g., how many users have 1 session, 2 sessions, and so on. Figure 5 shows that most users browse the wireless Web for short periods of time, e.g., for 95% of the users, the largest session time is less than 3 minutes (not shown). Similarly, the total browse time for 95% of the users is less than 7 minutes for the entire trace period. Moreover, 95% of the users initiated less than 35 sessions during the trace period and 98% of the users had less than 200 hits during the trace period.

There could be several explanations for this behavior. First, browsing the Web on cell-phones is cumbersome due to their small form factor. Second, unlike the traditional wired connections, browse time on the cellular network is not free (subscribers have to pay for airtime). Finally, wireless Web services are still in their infancy; in time, with the availability of better content, better display technology and with cheaper airtime, more users will stay on the channel for longer periods of time.

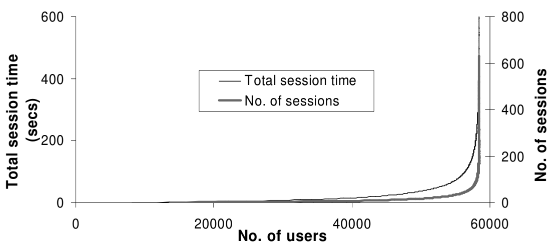

Fig. 5. Distribution of session times and number of sessions for wireless users during the trace period

B. User Byte Access Distribution

We also analyzed the number of bytes downloaded by each user and observed that 90% of wireless users fetched less than 100 KBytes during the trace period. As expected, offline users have a different distribution compared to wireless users: the former set of users access significantly more data (more than 95% of offline users downloaded 100 KBytes of data during the trace period). However, this result does not mean that offline users actually access all the data; it just reflects the amount of data downloaded by the "sync" program (to synchronize a PDA's data with a copy of the data on a PC) according to user's registered profile.

Currently, an offline user's profile is registered with the user's "sync" program. If the profile was registered with the server, it could be used effectively to quickly retrieve all the relevant pages for that user when the first request is received from the "sync" program (i.e., prefetch the data for that user). In fact, with this change, the sync program can send just one HTTP GET request to the server, thereby reducing the number of messages and delay.

V. System Load Analysis

We now present an analysis of the message reply distribution and the load experienced by the Web server during different times of the day.

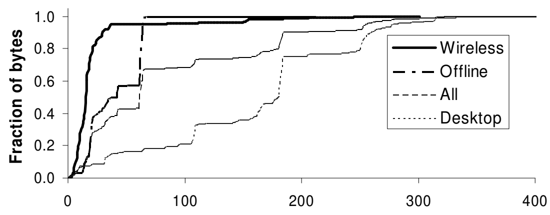

Fig. 6. CDF of total bytes sent vs. the size of the reply message

Figure 6 shows the CDF of the total bytes sent to users versus the reply size (CDF for the number of replies is similar and is not shown). Since the Web site has been designed with small pages, most of the reply sizes in the traces are small: 98% of wireless client replies are below 3 KBytes and 99% of offline user replies are below 6.3 KBytes. This indicates that the wireless Web server should be highly optimized in sending short replies. In particular, using TCP slow start procedure to probe available network bandwidth usually takes several roundtrips. This is very inefficient for sending small files. Previous proposals on optimizing TCP for short Web transfers, such as [10], [13], would be especially useful and yield even higher performance benefit for wireless documents since they are even smaller than most regular Web pages.

Replies for desktop clients are larger since most of the personalization and signup activity happens through desktop clients (and these pages are relatively bigger); a significant fraction of bytes is sent in larger replies (80% of the total bytes sent in replies have sizes 10 KBytes or more).

Figure 7 shows the amount of data sent by the server at different times of the day over the trace period (the graph for number of hits is similar). As expected, there are more hits on the Web site during the daytime than during the night; server administrators can utilize this time period for daily maintenance tasks. However, as more international users come online, we expect the graph to become flatter; in fact, such an effect can be observed in Figure 7: there is a long period during the day when the load on the server is not significantly below the peak load.

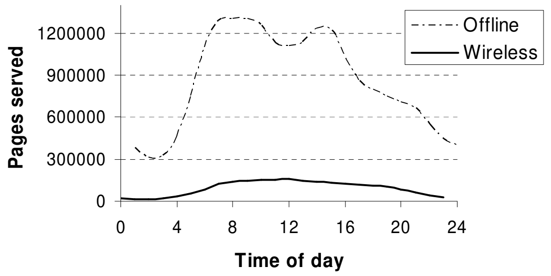

Fig. 7. Number of pages served by the Web site at different times of the day

We find that a large fraction of data (and requests) received by the server are from an automated "sync" program. The server can detect such requests and give them lower priority than wireless clients (where an end user is waiting for the page).

VI. Content Analysis

The study of content dynamics is important for the following four reasons: (1) It is important to Web servers, as it has implications on cache consistency control mechanisms and hence on the effectiveness of Web caching [11] strategies. (2) It is important to content providers, as it has implications on what content the provider should focus on providing. (3) It is important to cellular carriers and wireless ISPs since it affects airtime and hence revenues. Carriers like to have their clients stay on the network longer because that translates to higher airtime and hence more money. One way to keep users connected longer is to offer content that is useful to them. (4) Finally, content dynamics is important to Internet companies who are involved in providing downloads to small mobile devices.

With this motivation, we consider content in terms of categories such as entertainment, stocks, weather, news, travel, yellow pages etc. Doing analysis along these categories was convenient since it is both intuitively appealing and the Web site is structured in a similar manner.

Figure 8 shows the top three categories for wireless users, offline users, and desktop users. We note here that in the subsequent graphs, *mail* shows up as being low on popularity as well. The explanation for this is that currently, the wireless Web site does not handle mail browsing. It simply re-directs the client queries to the user's Internet mail service provider's Web site where the interfaces to the mail service have not yet been adapted to small devices. Our conjecture is that when the Web site is ready to handle mail accesses, the popularity of mail service will go up.

	Rank # 1	**Rank # 2**	**Rank # 3**
Wireless	Stock Quotes	News	Yellow Pages
Offline	Help	News	Stock Quotes
Desktop	Sign-ups	Mail	Sports

Fig. 8. Top three preferences for different kinds of users

Figure 9 shows the weekday and weekend activity for wireless users; the Y-axis shows the average number of bytes received and sent by the Web server in a day.

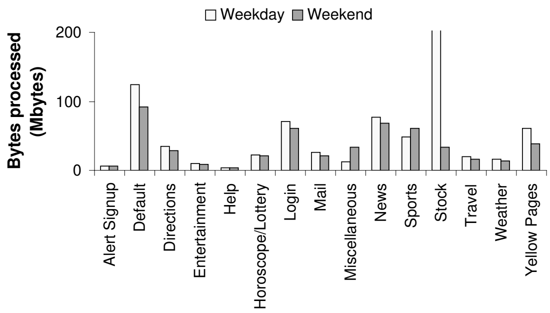

Fig. 9. Weekday versus weekend content analysis for Wireless users

Although not shown here, we observed that the browse patterns during a weekday and weekend are similar for both offline and wireless users but are different when compared to one other. Offline users "sync" up their devices more during the weekday than during weekends.

VII. Related Work

There are numerous studies on characterizing Web workload from the perspective of proxies [4], [12], browsers [5], [3], and servers [2], [11]. However most studies look at the workload of clients at wired networks, and there are very limited studies on the workload of clients at wireless networks. The study closest to ours is [8]. The authors analyze network trace data generated by a mobile browser application focusing on user behavior as determined by session characteristics. User behavior is in the form of bytes transferred and time spent on the wireless link. While the work is interesting, it is severely limited by the size of the data analyzed. Although the traces were collected over a period of seven months, only 80K entries were logged. It is unclear how well the inferences drawn from this study scale up to large commercial sites. In contrast, we analyzed 33 million entries generated over a period of 12 days using a large commercial site. Using this data, among other things, we show that the definition of what constitutes a sessions as described in [8] is not sufficient; thus, some of the results reported in that paper may not necessarily reflect reality. Also, our study is broader as we focus on user behavior, server load, content and document popularity analysis. Furthermore, we are able to carry out user-behavior analysis whereas the authors in [8] cannot since the only user information that is available to them are client IP addresses, which can change for the same user.

VIII. Conclusions

In this paper, we have presented the dynamics of a popular commercial Web server that serves wireless and offline PDA users. We believe our work is useful since we have analyzed the access patterns of a Web site that is designed primarily for small wireless and/or mobile devices; earlier research has focused on Web sites designed for desktop clients.

A. Summary of Key Results

Our main findings are:

1. The distribution of document popularity does not closely follow Zipf-like distribution. Majority of requests are concentrated on a small number of documents. In particular, we find that 0.1% – 0.5% URL and parameter pairs (i.e. about 121 - 442 pairs) account for 90% requests.
2. More than 60% of the (browse) pages accessed at the Web server are due to offline PDA users and less than 7% of the accesses are due to wireless clients.
3. Most of the replies to wireless clients are less than 3 KBytes. For offline clients, most of the replies are less than 6 KBytes (with a similar distribution as wireless users).
4. Users tend to have short sessions when interacting with the Web site: the largest session time for 95% of the users was less than 3 minutes. We empirically determined the session-activity threshold to be somewhere between 30 to 45 seconds.
5. Stock quotes, news, and yellow pages are the top categories accessed by wireless clients. For offline clients, help is the topmost category followed by news and stock quotes, though offline users may not access all the help pages downloaded automatically by the programmed agent.
6. The relative importance of different categories did not change between weekdays and weekends (except stock quotes and sports). However, the amount of data accessed over the weekend drops by 45%.

B. Performance Implications

Our trace analysis of browser logs has the following performance implications:

1. The high concentration of requests to documents in the browser log implies that caching the results of popular queries would be very effective in reducing the wireless Web server load.
2. The fact that offline PDA users generate significantly more requests than wireless users suggests that system designers have to be careful in making sure that wireless clients are given a higher priority over offline PDA users.
3. Most replies sent to wireless and offline users are very small, 3 - 6 KBytes, which suggests that a wireless Web server should highly optimize sending short replies to clients.
4. We propose a heuristic to determine the session-inactivity period. Our experimental results confirm that the heuristic is able to determine the session-inactivity period. This heuristic can be useful for wireless service providers to reclaim IP addresses. In particular, we find that the session-inactivity period is between 30 to 45 seconds, which suggests that we may reclaim IP addresses more quickly than 90 seconds used in WAP.

References

[1] M. Arlitt and T. Jin. Workload Characterization of the 1998 World Cup Web site. *IEEE Network*, May 2000.

[2] M. Arlitt and C. Williamson. Internet Web Servers: Workload Characterization and Performance Implications. In *IEEE/ACM Trans. on Networking*, 1997.

[3] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web Client Access Patterns. *World Wide Web Journal*, 1999.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web Caching and Zipf-like Distributions: Evidence and Implications. In *Proc. of INFOCOMM '99*, Mar 1999.

[5] C. Cunha, A. Bestavros, and M. Crovella. Characteristics of WWW Client-based Traces. *Boston Univ CS Dept Technical Report TR-95-010*, June 1995.

[6] S. Glassman. A Caching Relay for the World Wide Web. In *Proc. of 1st WWW Conference*, May 1994.

[7] T. Kelly. Thin-client web access patterns: Measurements for a cache-busting proxy. In *Proc. of the Sixth Web Caching and Content Delivery Workshop*, Jun 2001.

[8] T. Kunz, T. Barry, X. Zhou, J. Black, and H. Mahoney. WAP Traffic: Description and Comparison to WWW Traffic. *ACM Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Aug 2000.

[9] N. Nishikawa et al. Memory-based Architecture for Distributed WWW Caching Proxy. In *Proc. of 7th WWW Conference*, April 1998.

[10] V. Padmanabhan and R. Katz. TCP Fast Start: A Technique for Speeding up Web Transfers. In *Proceedings of IEEE Globecom'98*, Nov 1998.

[11] V. Padmanabhan and L. Qiu. The Content and Access Dynamics of a Busy Web Site: Findings and Implications. In *Proc of ACM SIGCOMM 2000*, Aug 2000.

[12] A. Wolman et al. On the Scale and Performance of Cooperative Web Proxy Caching. In *Proc. of SOSP*, Dec 1999.

[13] Y. Zhang, L. Qiu, and S. Keshav. Speeding up Short Data Transfers: Theory, Architectural Support, and Simulation Results. In *Proc. of NOSSDAV*, Jun 2000.