Analysis of WWW Traffic in Cambodia and Ghana

Bowei Du
Computer Science Division
University of California
Berkeley, CA 94720-1776
bowei@cs.berkeley.edu

Michael Demmer
Computer Science Division
University of California
Berkeley, CA 94720-1776
demmer@cs.berkeley.edu

Eric Brewer[*]
Intel Research Berkeley
2150 Shattuck Ave
Berkeley, CA 94704
eric.a.brewer@intel.com

ABSTRACT

In this paper we present an analysis of HTTP traffic captured from Internet cafés and kiosks from two different developing countries – Cambodia and Ghana. This paper has two main contributions. The first contribution is a analysis of the characteristics of the web trace, including the distribution and classification of the web objects requested by the users. We outline notable features of the data set which effect the performance of the web for users in developing regions. Using the trace data, we also perform several simulation analyses of cache performance, including both traditional caching and more novel off-line caching proposals. The second contribution is a set of suggestions on mechanisms to improve the user experience of the web in these regions. These mechanisms include both applications of well-known research techniques as well as offering some less well-studied suggestions based on intermittent connectivity.

Categories and Subject Descriptors

C.2.2 [**Computer-Communications Networks**]: Network Protocols—*Applications*; C.4 [**Performance of Systems**]: Measurement Techniques; C.2.4 [**Computer-Communications Networks**]: Distributed Systems—*Client/server*

General Terms

Design, Measurement, Performance

Keywords

Caching, Cambodia, Classification, Delay tolerant networking, Developing regions, Dynamic content, Ghana, Hypertext Transfer Protocol, HTTP, Measurement, Performance analysis, Proxy, Redundant transfers, Trace, WWW, World Wide Web

1. INTRODUCTION

In this paper we present an analysis of HTTP traffic captured from shared-use Internet cafés and kiosks from two different developing regions, Cambodia and Ghana. Based on these traces, we present some general characteristics of the data. These characteristics include distributions of object types and sizes, the popularity of URLs, patterns of request times, and a classification of the content of the requested sites. For each axis of classification, we identify interesting and/or unexpected features in the data set.

We show from this traffic data that there are some mismatches between web applications developed in (and for) the industrialized world to the constraints faced by users in developing regions. For example, in one of the web traces, web based e-mail constitutes approximately 15% of the total downloaded bytes. Inherently, e-mail is an asynchronous application; it neither requires end-to-end connectivity nor is it sensitive to variance in data delivery time, yet the web access model introduces both of these constraints. In addition, accessing e-mail over the web incurs a significant bandwidth overhead because of embedded presentation elements in each page. As many developing regions are characterized by expensive, constrained, and intermittent access to bandwidth, web based e-mail is in many ways a poor fit for the needs of users in these environments, who may be better served by store-and-forward delivery models such as SMTP.

Using the results of our trace analysis, we offer our second main contribution, which are suggestions of methods and techniques to improve the performance and user experience of web access in developing regions. These suggestions consist both of applications of known techniques, as well as some more novel approaches.

We focus our suggestions around the context of bandwidth and connectivity constrained conditions. Bandwidth access is both expensive and limited in developing regions. As we show in Section 2.2, the cost per byte of connectivity in developing regions can be significantly more expensive than in industrialized world nations such as the United States. Given that the overall purchasing power of developing countries is already diminished when compared to that of the industrialized world, developing regions pay a significant price premium payed for network access.

Within a single country or region, we have found that degree of remoteness of a location and the cost of connectivity from the location are often directly related. In Cambodia, remote locations rely on expensive satellite or cellphone connections whereas Internet cafés closer to urban centers can

[*]Dr. Brewer is also a professor in the Computer Science Division at the University of California, Berkeley.

This material is based upon work supported by the National Science Foundation under Grant No. 0326582.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

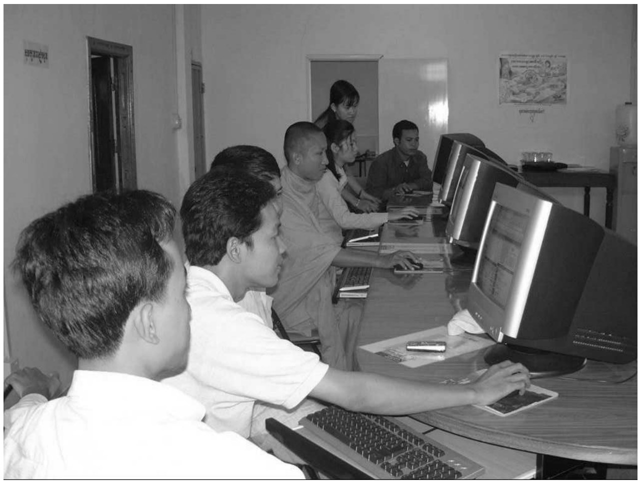


Figure 1: Users at a CIC center in Cambodia

use dialup or DSL. Furthermore, the user demand for this bandwidth is also increased because a single connection is often multiplexed among a large user population in a computing center. Figure 1 shows a typical shared usage situation in a CIC center, where several users use a common (slow) network link.

Furthermore, most of the web is designed with the assumption of continuous, relatively low-latency, end-to-end connectivity from client to server. Yet we note two factors that make asynchronous and/or disconnected web applications interesting for developing regions. First, the costs of connectivity can vary significantly with time. For example, one Cambodian ISPs charges between one half and one third their normal rates during nights and weekends [7]. This fee structure motivates the development of mechanisms for time shifting bandwidth usage. Second, there has been recent research interest within the delay tolerant networking community in providing access via physical transport [4, 11, 22]. This kind of network connectivity has the potential deliver a significant amount of bandwidth at low cost, but is only applicable for systems that can tolerate asynchronous operation.

The rest of the paper proceeds as follows: We first explore some related work on web traffic analysis and cache analysis as background and present motivations for our work in this subject. We then present the main findings from our traffic analysis, including highlighted features from the data sets. Finally, we proceed to present our suggestions for improving the performance, including some simulations to explore the efficacy of intermittent connectivity, then offer our conclusions.

2. BACKGROUND AND MOTIVATIONS

Although the area of web traffic analysis has a rich history, to our knowledge this paper is the first in-depth analysis of the traffic patterns of information centers in developing regions. To provide background and some context for this analysis, we briefly review some of the prior work on web traffic and cache analysis, then outline our motivations and goals for this paper and explain the methodology we used for gathering and analyzing the trace data.

2.1 Related Work

Early efforts to gather web trace data were undertaken at Digital Equipment Corporation [14] using a proxy log, and at Boston University [10] using instrumented client browsers. In the latter work, the authors suggest that the correlation of web requests to their popularity follows a power law distribution. Crovella and Bestavros [9] continue this analysis to demonstrate that WWW traffic exhibits evidence of a self-similar traffic model.

Gribble [12] presented an analysis of a large web trace gathered from at-home student and faculty access at UC Berkeley. The main results from this work include confirmation of an expected diurnal cycle, small time-scale burstiness of traffic load, and that long request latency requires a large number of simultaneous connections. One interesting aspect of these analyses is is that the primary mode of Internet access at the time was through dialup modems at speeds around 28.8 kbps [8], speeds comparable to the dialup or VSAT access that is prevalent in developing regions. Although the content of the web has certainly changed significantly since the prior studies, the results cited in these early traces present an interesting avenue for comparison due to the similar bandwidth access rates.

Using these (and many other) traces and simulations, a rich wealth of research has been undertaken into web caching. Pierre [19] compiled a bibliography of over 300 citations into various aspects of web caching analysis.

One particular avenue of investigation related to web caching that continues to attract research interest is the area of delta encoding [15, 17, 21]. This technique relies on cooperation between the cache and the server to transfer only the changed bytes instead of whole objects, while still maintaining compatibility with the existing protocols.

Another area of research considers the issues of aliasing or rotating documents in the web [13, 16]. In this case, the problem is not that files may have been slightly modified since the time when they were cached, but rather that a notable degree of aliasing is present in the web. This aliasing implies that standard caches that use the URL as a key into the cache index cannot find duplicate documents.

Another technique that combines the effects of both delta encoding as well as aliasing is work on value-based web caching [20]. In this approach, cooperation between the proxy cache and the client means that the URLs are effectively ignored for the purposes of caching; indices are made based on the content of the document that are then used to reduce bandwidth requirements.

2.2 Motivations

Despite the wealth of work cited above, it seems that the area of web traffic analysis has gone slightly out of vogue in the research community, as relatively little work has been done recently in this area. As such, in this work we do not claim to compare rigorously the traffic analysis patterns of the developing world to those in the industrialized world, as we have few sources of current research to use as the basis of such a comparison.

Instead, our goals in this work are two-fold: the first being to gain a high-level understanding of the traffic patterns in developing regions. We suggest that the traces we gathered are likely to be fairly representative, and hope to gain insight into the needs of users in developing regions, and how effectively those needs are being met.

Connectivity	Installation	Monthly	Hourly
United States			
Dialup 56 Kbps	-	$10	-
DSL 2 Mbps	-	$25	-
Cambodia			
GPRS 9.6kbps	-	$250	
Dialup 33.6Kbps	-	$10	$1-$1.67
DSL 128 Kbps	$10	$99	
DSL 512 Kbps	$10	$400	
VSAT 256 Kbps	$3,250-$3,550	$770	
Ghana			
Dialup	-	-	$1.30
Canopy 64kbps	$950	$299	-
Canopy 128kbps	$950	$499	-
Canopy 256kbps	$950	$799	-
Canopy 512kbps	$950	$1399	-
VSAT KU-band	$3,500-$15K	$650	-
VSAT C-band	$3,500-$15K	$1,800	-

Table 1: Cost of connectivity in various countries (with rates quoted in US dollars)

The second goal of our analysis is to consider some optimizations and techniques that could improve the user experience in these regions. To this end, we suggest applications of known techniques, as well as presenting some more novel ones. We focus this analysis on the question of how (if at all) the fact that the primary client base for the web is in the industrialized world contributes to a design mismatch with the needs and situations in the developing world.

As can be seen from Table 1, bandwidth in the countries under consideration is significantly more expensive than similar data rates in the United States. Also, the cost of bandwidth is in many cases not directly correlated to the amount of bandwidth provided. The difference in cost is often due to the fact that limited infrastructure restricts connectivity options to costly, low bandwidth options such as cellphone modem (GPRS).

We note that there is generally more diversity in the pricing and variety of connectivity plans in developing regions, in contrast to the nearly ubiquitous flat rate plans offered in the United States. A usage based rate plan allows a system to make an economic analysis of the value of transmitting data, versus operating with potentially missing or stale information. With a flat monthly rate, no economic incentive exists for adding such features to a system.

Finally, in many developing regions, a network access link is often shared among a group of users for economic reasons, in an model like that of the CIC centers or an Internet Café. A group of users will therefore be competing for the resources of that link, thus wasted or unnecessary transmission can cause delays for the whole user group.

The combination of expensive costs, low data rates, and high demand means that there is a high per-byte value on network transmissions in developing regions. Hence any improvements on the network bandwidth have a greater effect on the user benefit.

Field	Value
time	2005-06-01 00:00:50
uri	http://www.kapook.com/index01-2.gif
request size	318 bytes
response size	598 bytes
mime type	image/gif
http status	200
client ip	*ignored*
cache info	0, Inet

Table 2: Example log entry data

2.3 Traffic Details

The web traffic analyzed in this paper was collected from proxy server logs of two different user groups, the Community Information Center project in Cambodia and an Internet café in Ghana. Unless explicitly noted, statistics cited in the text refer to aggregated values for both sets of data.

Although both sets of logs were collected from Internet access points with relatively high bandwidth, we view these traces as representative of the patterns of real user demands in developing regions, and we posit that they are relevant to inform the design of Internet access infrastructure. While more constrained bandwidth would likely have some effect on web traffic patterns, these traces remain useful as an approximation of user access patterns.

Cambodia CICs

The Community Information Centers (CIC) in Cambodia are Internet enabled computing centers established by the Asia Foundation [2]. CIC centers were created in all twenty two provinces. Our study focuses on six of the centers, each of which is connected to the Internet via a 64 - 128 kbps link to a central service provider. All ISPs in Cambodia connect to the Internet by VSAT.

For a six month period from May to September of 2005, the centers reported a total of 15,729 visitors, corresponding to an average of 85 users per day. We gathered detailed web logs over a period from June to September of 2005 from an IIS proxy server deployed in Phnom Penh. All client machines were configured to use this proxy server, and direct Internet connectivity was blocked by a firewall.

The logs contain ∼12.6 million entries, with ∼3.3 million unique URLs from 63,944 unique hosts. The proxy reported downloading a total of ∼110 GB from the Internet.

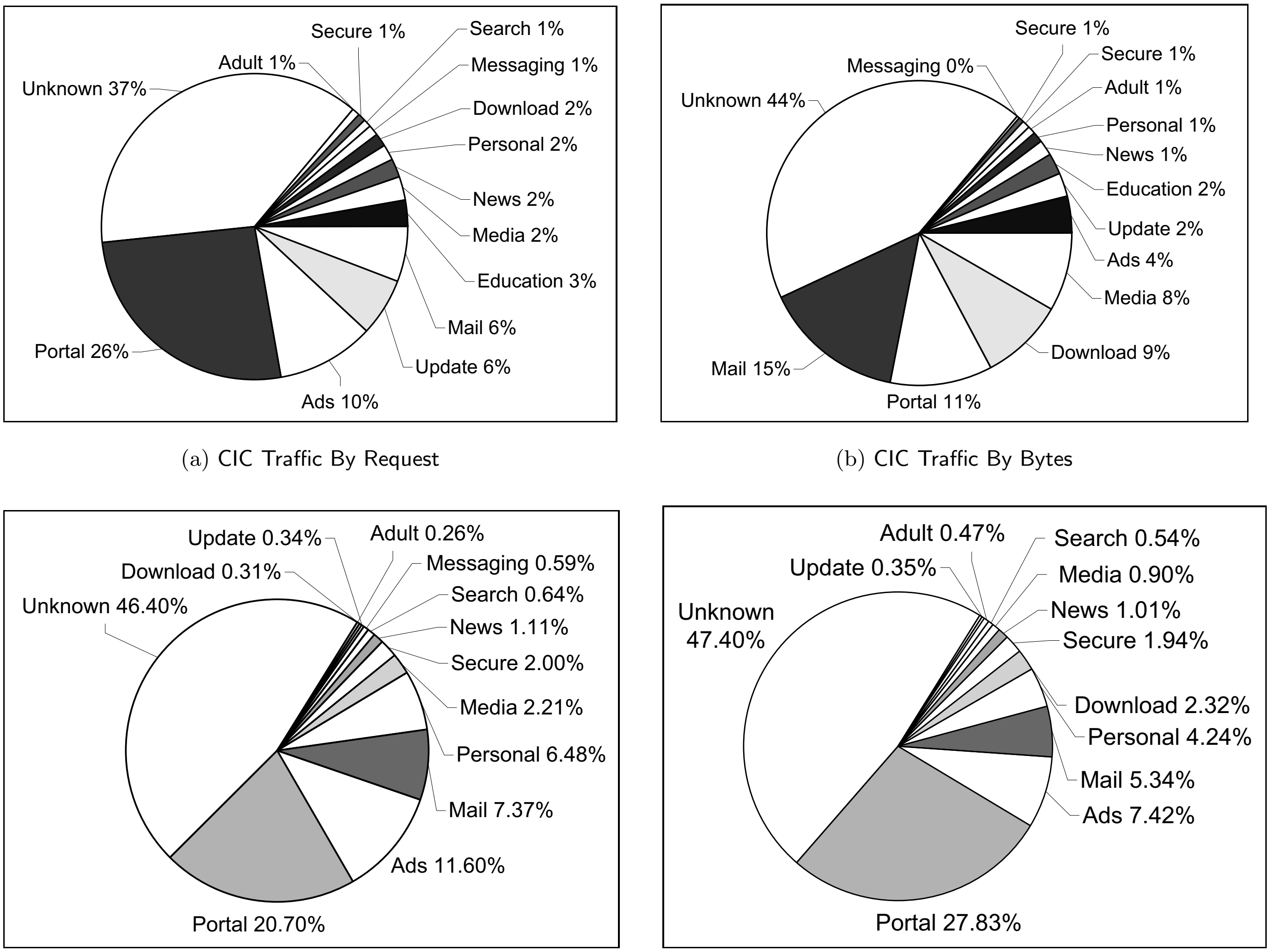
Ghana Internet Café

The Ghanaian trace was taken from an IIS proxy log at a Busy Internet [3] café in Accra. The proxy was physically colocated in the café and was connected to the Internet by VSAT as well. Roughly 100 users visit the café each day. We gathered proxy logs from the café for the period of May - June, 2004.

The logs contain ∼13.7 million entries, with ∼3.9 million unique URLs from 79,840 unique hosts. The proxy reported downloading ∼106 GB from the Internet.

Log format

Table 2 lists data from a sample entry from the log data we collected. The request size and response size fields are the sizes of the HTTP request made by the client and the



(a) CIC Traffic By Request

(b) CIC Traffic By Bytes

(c) Ghana Traffic By Request

(d) Ghana Traffic By Bytes

Figure 2: Traffic classifications: *Portal* – general information sites like Yahoo!; *Ads* – advertisement image sites; *Mail* – web-based email sites; *Personal* – dating and online community sites; *Download* – large binary files; *Secure* – all https traffic; *Media* – online audio and video files; *Education* – university sites and online course providers; *Messaging* – chat sites; *Update* – software update traffic; *News, Search, and Adult* – self-explanatory; *Unknown* – unclassified URLs

response sent by the server, respectively. The cache info field is a value from the proxy cache that indicates whether or the request was served by the cache. Other fields in the log format are self-explanatory. To avoid potential privacy issues, and because all access from the CIC centers passes through Network Address Translation (NAT), we ignored the client IP address in our analysis. Unfortunately, the MIME type and cache status information was not recorded in the Ghana logs.

3. TRAFFIC DATA

In this section we present an analysis of interesting characteristics of the observed traffic in these two centers.

3.1 Classification

To help gain further understanding of the traces, we examined the URLs from the traces and classified them by hand into a set of broad categories. The results of our classifications are shown in Figure 2. Note that we break down the results for both countries into ratios based on URL accesses and total bytes transferred.

Unknown

The first and most obvious result in these figures is the amount of data that we did not classify or escape classification. The URL accesses in our traces follow a well-known pattern, evidenced by a relatively small number of popular URLs and a very long tail of other URLs. After spending hours classifying the top URLs, the diminishing returns gained from continuing were too laborious for us to continue.

Portals

One of the largest category comprises sites we classified as portals. The main contributors to this class are well-

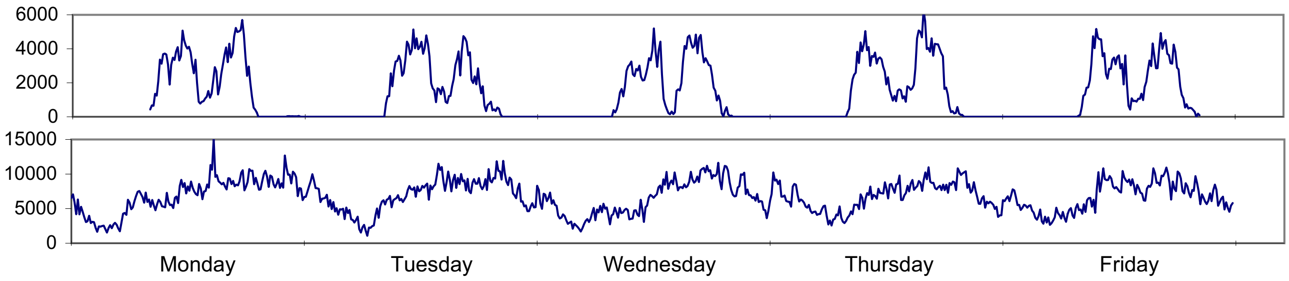


Figure 3: Web usage over the period of a week as a function of time. Requests totals were aggregated in ten minute buckets. The graphs shown are for Cambodia (top) and Ghana (below).

MIME Type	Request %	Bytes %
images	52.70%	20.81%
html/plaintext	27.67%	37.59%
other	8.85%	3.26%
javascript	4.35%	1.96%
css stylesheet	3.47%	1.15%
binary/exe	1.48%	17.09%
shockwave/flash	1.05%	3.02%
video	0.11%	2.22%
audio	0.19%	8.15%
compressed	0.05%	3.63%
pdf	0.05%	0.55%
msoffice	0.04%	0.58%

Table 3: Breakdown of CIC traffic by MIME Type

known sites like Yahoo! and MSN, but we also found some smaller, more niche-focused portals. The high dominance of US-based traffic in this class suggests that despite the language barriers, high-quality local information portals are not very prevalent in developing regions. As such, local-language versions of popular US portals would likely be well-received.

Advertising

Another large category is that of ads. We used some well-known pattern matching lists culled from the Firefox Filter Set G extension [18] as well as other obvious ad sites to determine which URLs are likely to be ads. As we discuss further in Section 4, the target population in developing regions are not likely to be customers of the advertised services in many cases, so this data is essentially wasted bandwidth.

Search

Another point to note is that the search category represents a large number of URL requests, but a relatively small amount of total bytes transferred. This distinction results from the fact that we logged only those URLs directly related to the search function in that category, excluding any images or other presentation that augment the search results. Thus while a single fetch to a complex portal UI might require tens of URL fetches to render, a simple search interface like Google only requires a few.

Center-Specific Traffic

Several distinctions can be drawn between the Cambodian and the Ghanaian data sets. One such distinction is in the amount of educational content, which represents 2-3% of the CIC traffic but negligible in the café traffic. Alternatively personals sites (e.g. online dating) comprised over 6% of the Ghanaian traffic but only 2% in the CIC centers.

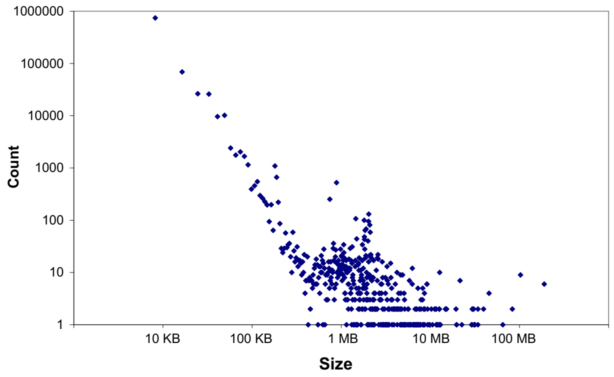


Figure 4: Logarithmic Plot of File Size Distribution from the CIC crawl data.

We posit that one reason for this distinction stems from the goals of the centers themselves. The emphasis on education that is part of the foundation of the CIC centers may have predisposed the users in those centers to furthering education. Similarly, the social atmosphere of the Ghanaian café environment may have influenced the high degree of dating related browsing.

By examining the relative ratios of request frequency and total bytes transferred, a few other features can be noted. For example, while portal traffic comprises a relatively even percentage of the requests from the two data sets (26% CIC and 20% Ghana), the percentage of bytes transferred is much higher in the the Ghanaian set (11% and 27%); indeed, the average object size from a portal site in Ghana is 10,260 bytes versus 4619 bytes in the CICs. Perhaps this discrepancy is due to richer portal content in Ghana, or simply a different site design for the popular portals.

3.2 Traffic Sizes and Requests

Table 3 summarizes the MIME type results from the CIC traffic analysis, and Figure 4 is a logarithmic graph of the size distribution of the URLs fetched within the a one week period of the CIC traffic, aggregated into 8 KB buckets.

Large Binaries

A significant fraction of the overall download data was made up of large binary files. These files constitute the tail

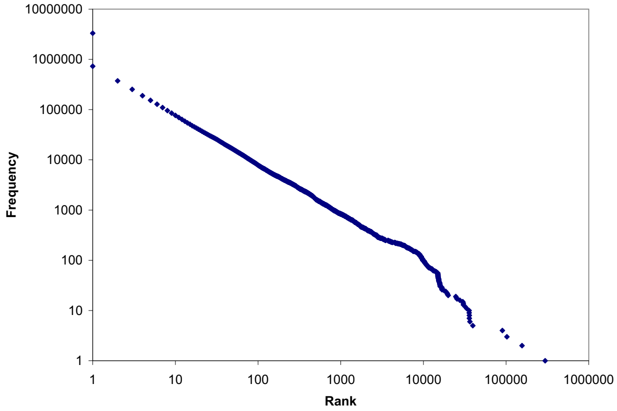


Figure 5: Logarithmic plot of CIC URL frequencies by rank, showing an expected Zipf distribution with the exception of the quickbooks updater URL.

(above 10 MB) of the file size distribution and were in some cases hundreds of megabytes. Although 99.8% of the objects are less than a megabyte in size, the remaining 0.2% of the web objects formed 58.9% of the total bytes downloaded.

In most cases, these large downloads were electronic game demos or movies. One interesting note is that the connections in question are shared-use links with a capacity in the tens of kilobits, so the users of these centers must be willing to wait several hours for each download to complete. Because of the shared nature of the Internet centers, users trying to download large file must physically stay in the center during the entirety of the download. As we suggest in Section 4, several optimizations including chunking and time-shifting can improve the efficiency of these downloads.

Media

As is the case in the industrialized world, the web experience in developing countries shows a rich multimedia flavor. We see not only presentation objects such as flash but a significant presence of audio and video files. From the Cambodian logs, we found several highly popular sites that distributed local Cambodian music. The effect of the multimedia content can be seen in the size distribution plot as a large number of files in the 1 MB - 5 MB range. The relatively high prevalence of media and game files was rather surprising to us, as we expected the use of the centers to be more focused on information access.

Images

One final note when examining the types of the objects is the slightly non-intuitive results for image files, which comprise over 52% of the URI requests but only about 20% of the bandwidth. We would have expected image files to be relatively large with respect to other objects like HTML and plain text. However investigating further, we found a large number of small image files (e.g. 1 pixel by 1 pixel) that are used as part of the layout of a site, or to referenced with dynamically-generated, non-cacheable URLs and used to record hits while still allowing larger objects to be cached. Hence the average size of an image object is ∼3,400 bytes as opposed to ∼11,900 bytes for plain text or HTML objects.

URL frequencies

Figure 5 shows the plot of URL request frequency from the CIC centers on a logarithmic scale relative to the rank of that URL. This plot is generally in-line with other studies on web traffic patterns, as it shows a Zipf-like distribution. The data from Ghana shows a similar pattern, hence we omit the plot here.

One notable exception from the CIC data is the highest-rank URL, which is out of line with respect to the general trend. It turns out that this URL is `http://qb13bgpatchsp.quickbooks.com/ud/38541` – an auto-updater URL that was requested 296,679 times over the course of our log period – a frequency of over 3,000 accesses per day! It is possible that this anomaly is due to a misconfiguration or simply an highly aggressive update mechanism. In either case, the frequency was a clear surprise.

Traffic Requests over Time

Figure 3 is a graph of the number of requests at each location center, aggregated into ten minute intervals. The operational hours of the two regions are clearly evident – the Cambodian Internet centers are only powered on during the work day (with an offline period during lunch), but the Ghanaian Internet café was always operational with high demand at midday and in the early evenings.

3.3 Caching

As mentioned above, all our trace data is derived from IIS proxy caches. To better understand the efficacy of the caches, we decided to run a portion of the traffic logged through a caching simulator to determine the sensitivity to various parameters.

For this analysis, we selected a random week from the Cambodia CIC log trace and then ran a crawling process to download all the URLs from the one week long trace[1]. We crawled 273,140 unique URLs (all but 40,145 successfully) from 904,708 log entries. From this crawled data, we extracted the downloaded object sizes as well as the values from the `Cache-Control`, `Pragma`, and `Expiration` HTTP headers, values that were not present from our original trace data.

We then ran a simple caching simulator to ascertain the effects of various parameters such as the cache size and object expiration policies on the hit rate. Figure 6 plots the distribution of cache performance characteristics for a variety of cache sizes. In addition to the totals for *hits* and *misses*, we also include the number of *expired* objects – requested objects that were in cache but where the cached copy had exceeded its expiration time, *nocache* objects – those explicitly tagged by the server as uncacheable, and finally *errors* – objects we did not retrieve in our crawl.

As can be seen from the simulation results, setting the cache size to a only a few hundred megabytes approaches the maximum hit rate possible for this data set. Furthermore, as the hit rate in our simulations is very good – approaching 50% in the limit, this experiment suggests that the use of

[1] For this crawl we used the wget utility, set to save all HTTP headers for cacheability information, and with a timeout set five minutes per URL to speed up the crawl. Any unfetchable objects were marked as errors, including genuine errors like mistyped URLs, objects that no longer exist, and servers that were unreachable during the crawl process. Although the content of the web has evolved since the time of the trace, and hence the objects downloaded by our crawl are not necessarily the same ones that were originally requested in the trace, we believe this change should not substantially affect our results.

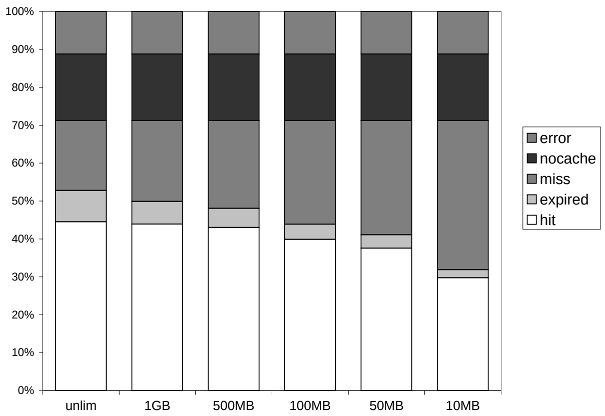


Figure 6: Simulation experiment to determine the effect of various cache sizes on a trace from the CIC centers in Cambodia. Each plot includes hits, misses, expired in-cache objects, uncacheable objects, and HTTP errors.

proxy caches should significantly save on bandwidth usage. Indeed, anecdotal reports from the CIC operators confirm that their observed performance improved with the deployment of the IIS proxy.

4. SUGGESTIONS

In this section, we present some mechanisms that we believe would help to improve performance for web access in developing regions. For all of these suggestions, we assume that bandwidth access will remain constrained by availability, reliability, and/or cost for some time; hence the bulk of these suggestions are aimed at mitigating the negative effects of a constrained bandwidth link.

The access model depicted in Figure 7 is a simple architecture involving shared-use proxies near the clients, as well as one (or more) entities at a well-connected center. Many of the suggestions below take advantage of this model by leveraging cooperation between the client-side proxy and the server in a data center to limit the traffic over the constrained link that connects them.

Although this model is not novel, we believe it to be appropriate and applicable to developing regions for several reasons. First of all, a sizable majority of users in developing regions do not own their own computers, but rather use shared facilities like the CIC centers or the Busy Internet café. Thus the computing costs of a shared-proxy in a center can be amortized for several users' benefit. Furthermore, the high price of bandwidth relative to the declining cost of storage and computation mean that in some cases, caches can pay for themselves. A similar line of reasoning holds for the operation of a data center in a well-connected area (possibly in a different part of the world).

4.1 Web based e-mail

We found from the traces that web based e-mail comprised of a sizable fraction of the HTTP traffic. We tracked the bytes and HTTP requests generated for the top three web e-mail providers – Yahoo!, Hotmail and Google. This traffic comprised 6.6% of total URL requests and 11.88% of the total bytes transferred. Note that these traffic measurements include only bytes from the HTML containing the e-mail and rendering message and does not include embedded objects such as advertisement images in the HTML page.

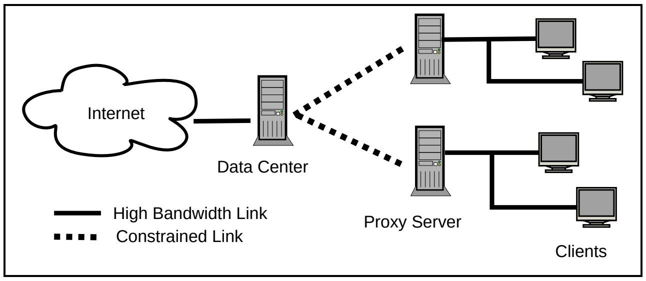


Figure 7: Infrastructure architecture for access improvement, involving a "smart" proxy near by to the clients and a cooperating entity at a high-bandwidth connection to the Internet.

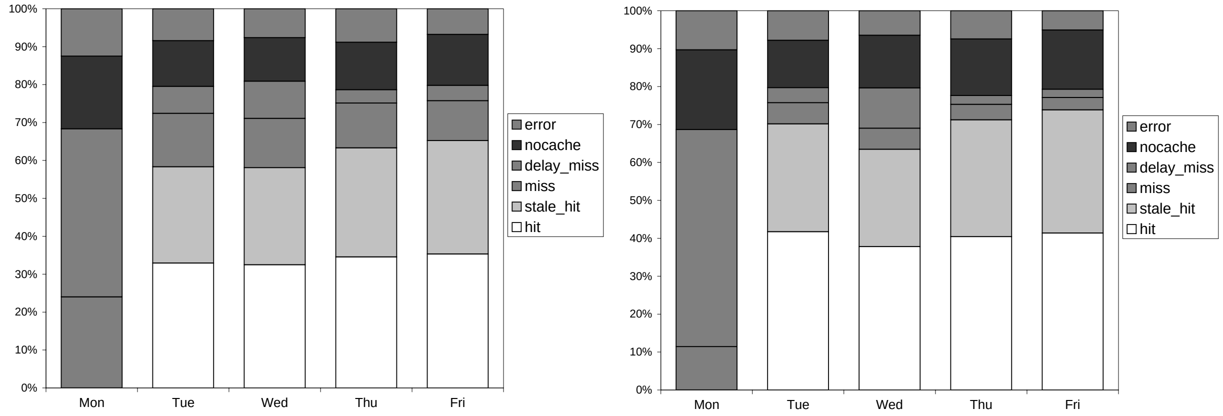
To estimate the amount of traffic wasted by the webmail protocol, we compared the size of an HTML page containing a message from the top e-mail provider in our traces, Yahoo! mail [1], with the size of an RFC822 mail message itself. We found that extra formatting and script data averaged 52 KB per mail message. In the CIC trace data, there were 70,026 individual URL requests to read mail from Yahoo! mail, resulting in a wastage of approximately 3.4 gigabytes of bandwidth over the trace period.

Other client access methods, such as IMAP, are much more efficient in terms of client-server bandwidth than the web. More widespread deployment of these services would help address this bandwidth overhead. However, to address this inefficiency while still maintaining the ease of use and familiar nature of the web mail interface, one could add functionality to the proxy cache to take care of generating the layout-related parts of a page, such that the only data that needs to be fetched from the server is the actual e-mail message itself.

4.2 Convert polling to push

Many software programs installed on the computers in the kiosks have auto-update checks that poll a central server for version information and update patches. We found that update traffic was 5.6% of the total web requests in the CIC centers, and was the most frequently requested URL (see Section 3.2). Almost every response returned from the auto update checks was identical, which we suspect reflects the fact that software versions do not change anywhere near the rate of polling. Yet for programs such as virus scanners or other security systems, it is critical for the software to be patched soon after an update is available, hence it is an inadequate solution to simply reduce the rate of update checking.

With a cooperative proxy model, a server in a well-connected data center could do the frequent polling on behalf of a set of remote centers with constrained bandwidth. When updates are found, they would be proactively pushed out to the client-side proxy cache. Any update check requests from an application could then be intercepted and immediately answered by the client-side proxy without traversing the bandwidth-constrained link. In this manner, the client application will still be patched as soon as an update is available, but the constrained link is not used for frequent redundant update checks.



(a) all traffic

(b) non-dynamic traffic

Figure 8: Intermittent web connectivity experiment on a one week trace of URL data from the CIC centers. The first group considers all URLs from the week; in the second group we removed URLs from known dynamic content. The plots show the daily breakdown of cache traffic including uncacheable data, hits, misses, and "disconnected misses", i.e. cases where URL was accessed more than once between the connectivity periods.

4.3 Irrelevant advertising

We found that advertising URLs comprised 10.1% and 11.6% of the requests and 3.6% and 7.5% of the total bytes transferred for the CICs and Ghana, respectively. Most (if not all) of the targeted advertising content is irrelevant for web users in developing countries, as many of the offered goods an services are unavailable for purchase in those countries. Thus proxy-based filters that would automatically remove these ads from the traffic are unlikely to impact the revenue to the advertisers, yet the filters could have a significant impact on bandwidth reduction.

Alternatively, these ads could be replaced with local advertisements, served locally from a proxy cache in each center. This would provide the center operators and other local entities with a potential additional revenue stream, plus could be an effective channel for users to be informed of new products.

4.4 Availability vs. Freshness

When designing the protocols for HTTP caching, a strong emphasis was placed on maintaining identical access semantics when using a cache to vanilla access. Indeed, this tradeoff makes logical sense for the common situations for web access – namely reliable, relatively high-bandwidth access.

One question that we raise is whether a different set of assumptions could motivate modifications to the caching protocols. For instance, the only mechanisms to classify the cacheability of objects are provided to the content publisher in the form of Cache-Control and Expires tags. Yet we posit that there are situations in which some of this control should be delegated to the client-side as well.

For instance, consider a news site in which articles change frequently throughout the day. Therefore, to achieve the semantics whereby the presence or absence of a cache does not affect the site, the cache server must set a short expiration time for the pages, or in the extreme, mark them as uncacheable. However, in a highly bandwidth-constrained environment, a user might be willing to tolerate a news report that is stale by a matter of hours or even a day, if that report could be accessed quickly and without paying the bandwidth costs to go to the server. Given that the user is the ultimate consumer of the requested information, it seems logical that the user could (and should) have insight as to the granularity at which the information is useful.

To effect this change on a protocol level, one could add analogous headers to `Cache-Control` and `Expiration` to the http request protocol, to inform proxy caches that the user is willing to accept stale data. Caches would of course need to be augmented with functionality to continue to store old data past its expiration time. Finally, a user interface modelled after to the various popup and advertisement blockers could be integrated into the web browsers, to allow users to express their preferences for individual sites.

4.5 Offline Caching

Expanding on this idea of caching for availability, we consider whether a offline or disconnected mode of web caching could still provide a reasonable user experience. The motivation for this analysis stems from the fact that in many cases, connectivity is less expensive and more available at off-peak hours such as at night. In addition, intermittent transports such as those envisioned by Delay Tolerant Networking [11], First Mile Solutions [4] and the TEK project [22] could provide an economical intermittent connectivity option. WWWofffle [5] is an example of a web proxy which provides offline browsing of its cache contents.

For this experiment, we took a one week trace of the CIC center traffic and ran it through our web caching simulator to model a case where connectivity is only available once per evening. For simplicity, we ignore the effects of cache sizes and assume that all URLs requested in a given day are fetched that night and available in the cache the following day.

The results from our experiments are shown in Figure 8. In these graphs, we plot the percentages of *hit* and *miss* objects, *error* objects that were unfetchable, *nocache* objects that are explicitly uncacheable, and two other values – *stale_hit* and *delay_miss*. The percentage marked *stale_hit* refers to objects that would be served from the cache but have exceeded their expiration times, i.e. an old version of the site. Finally, *delay_miss* counts the requests for which the same URL was fetched at least once before during the same day, but because the system was disconnected, the object is not available.

These results suggest that an offline web cache may be able to offer a tolerable user experience for users in developing regions. Even with no other optimizations, the combination *hit* and *stale_hit* objects is over 50% of all URLs during the week; after we remove known dynamic content sites (ads, redirection servers, etc) from the experiment, then around 70% of the total requests could be served by an offline cache. The other techniques suggested in this section may improve this ratio even more.

4.6 HTTP Chunking and Time Shifting

An issue that can arise when downloading large files is that if the connection breaks or becomes too congested, the download is cancelled. This means that a user must start over from the beginning to fetch the file. The use of HTTP chunking headers would allow the already downloaded data to be used, and could significantly help in situations where the link is severely constrained. Though some browsers do this already, a proxy cache could transparently add these headers to requests opportunistically for the client, maintaining the partially downloaded results in the cache. Also, a proxy cache could combine chunked requests from multiple clients, again avoiding downloading the same content multiple times.

Additionally, large downloads could be time-shifted so they occur at night when the network link is less congested. In this model, a download could be initiated during the day, then queued by the proxy cache to be processed overnight, and would be available to the client in the morning. Given that users are already waiting hours for large downloads, the one-day delay would likely be tolerable at least for very large objects. This would also keep the daytime usage of the network free for interactive browsing.

4.7 Other Caching Techniques

As mentioned in Section 2.1, there is a wealth of research literature on various approaches to web caching, including delta encoding, value-based caching, various forms of cooperative caching, and a host of other variants. We suspect that these techniques would be beneficial in a developing regions context, though may require some modifications or tuning. In particular, techniques that expect low-latency (if not high-bandwidth) connectivity between cooperating endpoints may not function well when round trip times exceed hundreds of milliseconds (or in the more extreme case, when the round trip is being taken on foot). Therefore, we believe that continued study and application of some of these other caching techniques to a developing regions context is a fruitful area for further research.

4.8 Presentation vs. Content

Several systems such as Loband [6] and the TEK project [22] apply transcoding and content modification techniques to reduce bandwidth usage. Yet as the complexity of web pages increases through the use of Javascript and dynamic HTML these techniques become more and more challenging.

Figure 9: A cellphone-based GPRS modem providing 9.6kbps connectivity to a CIC center in Cambodia.

Obviously, paring down the complexity of web sites would make them more amenable to low bandwidth situations or caching. For sites that want to target developing regions but still want a rich presentation layer, a separation of the content from the presentation component can improve the caching performance. However, some of the more complex techniques such as Asynchronous Javascript and XML (AJAX) and other RPC-based protocols subvert the normal web caching protocols, so other approaches should be investigated.

In general, content providers interested in targeting developing regions should treat the information density of the transmitted content as the highest priority, separating out any redundant or static data into easily cacheable units.

4.9 Compression

Compression is a well-known technique that could be more ubiquitously used to reduce the bandwidth consumption. A simple experiment where we applied LZ compression (with gzip) to all of the crawled data resulted in a 34.5% reduction in the data size. As this data set includes a significant amount of already-compressed data (e.g. JPG images), more selective application of compression techniques to the HTML or plain text pages would likely achieve an even greater compression ratio. Still, a 1/3 reduction in the traffic with no data loss is certainly attractive.

5. CONCLUSION

In this paper we analyzed the traffic patterns of Internet centers in two developing countries. From the traffic patterns, we noticed several interesting aspect of the developing regions traffic. Users were willing to use the shared low bandwidth link to download large media files, even if the download time was significant. Also, we found that some software updaters have aggressive poll behavior which can lead to Internet connection fees during user idle times. In addition, from the trace we noticed some current web applications are ill suited for low bandwidth environments. Web e-mail, one of the most popular applications in the trace, requires approximately 52 KB of non-content data to be sent with each e-mail.

Many of the problems of a low bandwidth environment have been examined in the past in the context of caching and compression. We evaluated the effect of standard caching and compression on the data set. We also considered the effect of supporting disconnected web browsing. We found that fraction of the traffic can be viewed in a disconnected fashion, however the existence of a long tail of never before seen URLs may detrimentally affect the user experience.

Almost all of the techniques we have suggested for improving the WWW experience for users in developing regions involves an advanced proxy at either end of the bandwidth challenged link which cooperate to reduce bandwidth usage and mask intermittent connectivity.

6. ACKNOWLEDGMENTS

Thanks to Pauline Tweedie, R.J. Honicky, the Asia Foundation, and Busy Internet for for providing us with access to their proxy logs.

7. REFERENCES

[1] http://mail.yahoo.com. Website.

[2] http://www.asiafoundation.com. Website.

[3] http://www.busyinternet.com. Website.

[4] http://www.firstmilesolutions.com. Website.

[5] http://www.gedanken.demon.co.uk/wwwoffle/. Website.

[6] http://www.loband.com. Website.

[7] http://www.online.com.kh. Website.

[8] Graphics, Visualization, and Usability Center, 6th WWW User Survey, 1996. `http://www.cc.gatech.edu/gvu/user_surveys/survey-10-1996/`.

[9] Mark E. Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE /ACM Transactions on Networking*, 5(6):835–846, 1997.

[10] Carlos Cunha, Azer Bestavros, and Mark Crovella. Characteristics of World Wide Web Client-based Traces. Technical Report BUCS-TR-1995-010, Boston University, CS Dept, April 1995.

[11] Kevin Fall. A delay-tolerant network architecture for challenged internets. In *SIGCOMM*, pages 27–34, 2003.

[12] Steven Gribble. System design issues for internet middleware services: Deductions from a large client trace. *Masters Thesis, UC Berkeley*, 1995.

[13] T. Kelly and J. Mogul. Aliasing on the world wide web: Prevalence and performance implications. In *WWW*, 2002.

[14] Tom Kroeger, Jeff Mogul, and Carlos Maltzahn. Digital's web proxy traces. `ftp://ftp.digital.com/pub/DEC/traces/proxy`.

[15] J. Mogul, B. Krishnamurthy, F. Douglis, A. Feldmann, Y. Goland, and A. van Hoff. Delta encoding in http. RFC 3229, Dec 2001.

[16] Jeffrey C. Mogul, Yee Man Chan, and Terence Kelly. Design, implementation, and evaluation of duplicate transfer detection in http. In *NSDI*, 2004.

[17] Jeffrey C. Mogul, Fred Douglis, Anja Feldmann, and Balachander Krishnamurthy. Potential benefits of delta encoding and data compression for HTTP. In *SIGCOMM*, pages 181–194, 1997.

[18] Pierceive. Filter set g. http://www.pierceive.com/filtersetg/.

[19] Guillaume Pierre. A web caching bibliography. `http://www-sor.inria.fr/projects/relais/biblio/`.

[20] Sean Rhea, Kevin Liang, and Eric Brewer. Value-based web caching. In *WWW*, 2003.

[21] Anubhav Savant and Torsten Suel. Server-friendly delta compression for efficient web access. In *8th International Web Content Caching and Distribution Workshop*, 2003.

[22] William Thies, Janelle Prevost, Tazeen Mahtab, Genevieve Cuevas, Saad Shakhshir, Alexandro Artola, Binh Vo, Yuliya Litvak, Sheldon Chan, Sid Henderson, Mark Halsey, Libby Levison, and Saman Amarasinghe. Searching the world wide web in low-connectivity communities. In *WWW (Global Community Track)*, May 2002.