Proactive Caching of DNS Records: Addressing a Performance Bottleneck

Edith Cohen
AT&T Labs–Research
Florham Park, NJ 07932 USA
edith@research.att.com

Haim Kaplan
Tel-Aviv University
Tel Aviv, Israel
haimk@math.tau.ac.il

Abstract

The resolution of a host name to an IP-address is a necessary predecessor to connection establishment and HTTP exchanges. Nonetheless, DNS resolutions often involve multiple remote name-servers and prolong Web response times. To alleviate this problem name servers and Web browsers cache query results. Name-servers currently incorporate passive *cache management where records are brought into the cache only as a result of clients' requests and are used for the TTL duration (a TTL value is provided with each record). We propose and evaluate different enhancements to passive caching that reduce the fraction of HTTP connection establishments that are delayed by long DNS resolutions. (A)* Renewal policies *refresh selected expired cached entries by issuing unsolicited queries. Trace-based simulations using Web proxy logs demonstrated that a significant fraction of cache misses can be eliminated with a moderate overhead. (B)* Simultaneous-validation *(SV) transparently uses expired records. A DNS query is issued if the respective cached entry is no longer fresh, but concurrently, the expired entry is used to connect to the Web server and fetch the requested content. The content is served only if the expired records used turn out to be in agreement with the query response.*

1 Introduction

The resolution of a host name to an IP-address is a necessary predecessor to communication between Internet hosts. In particular, it is required for connection establishment and HTTP exchanges with a Web server. DNS (Domain Name System) [21, 22] is in essence a distributed database that answers queries on mapping between names and addresses. Name-servers belong to a hierarchy where typically servers responsible for large domains delegate other name-servers to be in charge of subdomains. DNS was designed prior to the onset of the Web application, but fortunately, its design allowed it to scale and accommodate the explosive growth of the Internet. The perhaps unavoidable downside of this design is that resolving a DNS query often involves communication with at least one remote name-server, and may require following a delegation/referral chain of several remote name-servers. Furthermore, Since name-servers are different hosts than the HTTP servers that are contacted subsequently, DNS resolutions create additional potential point(s) of failure. Resolutions typically use UDP exchanges and use timeout and retransmission, which adds a delay on the order of seconds, in the event of packet loss.[1] The overall impact of DNS resolutions on user-perceived latency stems from both additional RTTs to remote servers and sensitivity to long timeouts.

Caching of query results at local name-servers decreases both overhead and user-perceived latency and is instrumental for performance. A caching mechanism for DNS was specified in RFCs [21, 22], and is integrated in BIND [1], the most popular name-server software. Name-servers resolve client queries about hostnames inside and outside their authoritative zones. Outside queries are resolved through communication with other name-servers, following referral chains to an authoritative name-server. Name-servers cache the results of queries sent to other name-servers. (They also keep and periodically refresh the zones they are authoritative for.) Each such piece of information (e.g., CNAME, IP-address, or an authoritative name-server) is provided with a TTL (Time To Live) value, and name-servers may use entries they are not authoritative for only until the TTL expires.

Caching is effective for requests constituting cache hits. Although "cache misses" (resolutions necessitating exchanges with remote servers) precede only a fraction of HTTP connections, their durations tend to be unpredictable and heavy-tailed. Studies show that when long Web waits

[1] In contrast, if packet loss occurs during a TCP connection, the timeout value is set adaptively according to previous history of RTT duration, which is typically considerably shorter than the default value.

do occur, DNS resolutions are a significant cause [6, 14]. Such occasional unexpected long delays significantly impact the consistency of service quality, which is often measured by extremes. High variance in "connecting" time was not critical for applications such as email, telnet, and FTP, that dominated the Internet when DNS specifications emerged, but is detrimental to Web browsing. Unfortunately, it seems that the impact of DNS resolutions is inherent in the current architecture. As bandwidth increases and content transmission time decreases, Web service speed would be increasingly dominated by RTTs, and the relative contribution of DNS resolutions would only increase.

Caching in BIND currently works in a passive manner: information for which a name-server is not authoritative for is obtained only as a consequence of a client query and is cached until either the TTL expires, or the name-server process dies. Here we propose and evaluate enhancements to basic passive caching aimed at reducing user-perceived latency due to DNS query time. Our enhancements fall within the framework of the current DNS architecture and can be locally deployed.

To put in context our proposed enhancement to DNS caching, we contrast it with the more well-studied subject of Web content caching. Content caching is integrated in popular Web browsers and widely deployed at proxy servers [15, 16]. Issues such as replacement policies, coherency and validation mechanisms, and understanding access patterns were extensively studied (e.g. [23, 24, 5, 4]). Content is typically cached beyond its freshness lifetime, but expired entries are validated with the origin before being served. HTTP provides mechanisms for client-driven validation and freshness control [13]. Proposed approaches to reduce user-perceived latency incurred on validation requests included validating objects prior to predicted requests [18, 9] and server-driven validations [20, 19]. Other studies proposed transferring stale cached data to a client while the data validity is being verified [12] or while the modified portion (the "delta") is being computed [2]. These approaches are of interest to us here since the issues of validation latency and DNS latency are conceptually related.

Although considerable research targeted validation latency for content caching, it seems that no analogous proposed enhancements were made so far for passive caching of DNS records. DNS caching differs in some basic aspects from content-caching: entries have considerably smaller sizes, storage space is ample with respect to the amount of data, and response sizes are small. Yet, there are also basic resemblances: Query time (cost of a cache miss) is significant and hence cache hit-rate is crucial for reducing perceived-latency; freshness typically expires well before the object is modified; and request sequences (to hostnames for DNS caching and to URLs for content caching) exhibit reference locality and characteristic frequencies.

Proactive DNS caching integrate automatically-generated "preemptive" queries that update the cache. Proactive caching approaches attempt to balance the number of *eliminated cache misses* and *overhead* of additional DNS queries issued to remote name-servers. *Renewal policies*, proposed and studied here, are a natural class of proactive caching algorithms.

Renewal Policies Renewal policies refresh cached entries upon their expiration time by issuing a new query. The different policies vary by when an entry is renewed. We consider several natural policies, based on reference locality (analogous to the cache replacement algorithm LRU), access-frequency (analogous to LFU), and an adaptive per-hostname policy (analogous to policies studied in [8, 5, 17]). These analogies are made based on properties of the request sequence that are exploited but the underlying cost/benefit measures are different. We experimentally evaluated and compared performance of the different polices using a heterogeneous set of proxy logs, consistently obtaining significant increase in hit-rate at reasonable overhead costs.

Renewal vs. Preresolving *Preresolving* (prefetching DNS queries) is a technique related to renewal. Preresolving was proposed in [6] as a low-overhead alternative to the prefetching of documents. Preresolving applies prediction schemes (e.g., by analyzing hyperlinks or tracking access patterns) to decide which hostnames (Web servers) to preresolve. The work [6] demonstrated a potential for considerable reduction in user-perceived latency when preresolving hostnames returned on search-engines responses. The tradeoff of latency and overhead also measures the performance of preresolving algorithms. The basic difference between renewal policies and preresolving is in their deployment: Preresolving utilizes predictions made based on per-user access patterns and currently-viewed hyperlinks. This information is available at the user's browser or proxy server, and therefore, preresolving queries would most naturally be initiated there and be viewed at the local name-server as regular client queries. Renewal policies, on the other hand, aggregate per-record patterns from DNS query sequences. Hence, they can be incorporated within the name-server cache and be transparent to its clients. Preresolving and renewal also differ in their impact on traffic: first, like document prefetching, preresolving is more likely to generate bursts [11]. Secondly, prediction-based preresolves are more likely to include loaded root servers whereas renewals are often directed only to lightly loaded servers lower in the DNS hierarchy.

Simultaneous Validation DNS TTL-based freshness control poses an inherent conflict for a domain administrator assigning TTL values. Smaller TTL values increase user-perceived latency and name-server load, and makes the site more likely to be inaccessible when the name server is down. Large TTL values, on the other hand, constitute long-term commitments. If the name-to-address translation changes, many users would look at the cached no-longer-valid IP address, and would be unable to reach the host until the TTL expires and a new DNS query is issued. In practice, TTL values are set conservatively: our measurements indicated that periods between changes are considerably longer than respective TTL values. This observation was our underlying motivation for introducing *simultaneous-validation* (SV). Under SV, when a client issues a request to a host (Web server) and a cached expired resolution is available, the proxy/browser issues the HTTP request(s) using the expired address while *simultaneously* issuing a DNS query to resolve the hostname. For transparency and consistency, fetched contents are held and displayed only if the stale address entry is validated by the DNS query results. SV reduces latency since DNS query and communication with the host are performed concurrently rather than sequentially. Our evaluation reveals that the mapping of names to IP-addresses is fairly static, and consequently, estimated SV success rate is over 98%. The SV approach is fundamentally different from preresolving or renewal policies. SV does not impose overhead of additional DNS queries. Deploying SV, however, necessitates caching of expired DNS records and support by both the DNS cache and the browser or proxy server.

Overview Section 2 provides background material on the domain name system and some relevant statistics. The bulk of our contribution is contained in Sections 3–5. In Section 3 and 4 we present and evaluate several renewal policies. In Section 5 we introduce and evaluate simultaneous-validation. We conclude in Section 6 and outline future research issues.

2 The Domain Name System

The domain name system (DNS) is a distributed database for name to address mapping of Internet hosts and mail servers (email addresses)[21, 22]. Prior to DNS, name-to-address mapping was performed centrally via a single file HOSTS.TXT maintained by the Network Information Center (NIC). This file was periodically *ftp*ed by the hosts in the network.

In graph-theoretic terms, the domain name space is a tree structure, where each node has a label. The *domain* or *domain name* of a node is the list of the labels on the path from the node to the root of the tree, separated with dots. If a node x is a descendant of y we say that x is in the domain of y or is a *subdomain* of y. Each node has information associated with it represented by *resource records* (RRs). This set of RRs comprises the domain database. It is a distributed database. The database is divided up into *zones*, which are distributed among *name-servers*. The number of name-servers contacted to resolve a particular name depends on the contents of the cache in our local name-server and in servers we communicate with through the resolution process. The maximum possible such number equals the number of subdomain delegations on the path from the root to the node representing our queried name in the domain-name tree, and, if name-servers are not located in subzones, additional queries to find addresses of authoritative name servers for the delegated subdomains. Our measurements indicate that the number of delegations on the path to a typical host is often not very large. As an upper bound we measured the length of the path from the root of the domain name tree to the nearest delegated ancestor of the name. Histograms for the length of this path for about 35K servers sampled from the NLANR logs and 9K from the AT&T research proxy log (see Table 1) show that the length of this path for about 80% of hostnames in the AT&T log and 90% of the hostnames in the NLANR log is two. This statistics implies that a typical resolution will query the root server which is also authoritative for the `.com` `.org` and `.net` domains. The root server will return a referral to the name-server authoritative for the name. If the returned name-server lies outside the root-server's zone, we would have to resolve it separately. Finally we query the authoritative name-server itself. The reason for longer path names for the AT&T log hosts is that hosts in the newer NLANR log contain higher fraction of vanity names, whereas the AT&T list contains a higher fraction of foreign domain names (outside the US) and names inside universities, which are typically delegated further to individual departments.

Associated with each RR is a time-to-live (TTL) parameter that states how long it can be cached before it should be discarded. About 25% of the hosts have TTL no greater than an hour and about 90% of them have TTL values no greater than a day.

3 Renewal Policies

Name-servers receive and resolve DNS queries and cache and reuse records for the time period specified in their TTL value.[2] A client query that can be answered from the local cache is labeled *cache hit*. Otherwise, resolving the client query involves issuing queries to remote name server(s) and the client query constitutes a *cache miss*. Under *passive caching*, currently practiced by BIND, DNS queries are issued by the name-server only as a result of a cache miss. In contrast, *proactive caching* uses

[2] We limit the discussion in this section to data not associated with the local zone.

"unsolicited" queries (*automatic-queries*) in order to increase hit-rate on client queries. On average, more than one automatic-query is performed in order to avoid one cache miss. Our premise, however, is that user-time is valuable and thus a cache miss is "costlier" than a respective automatic query.

3.1 Cost Model

We associate costs with both automatic-queries and cache misses. The cost of an automatic-query attempts to price the overhead imposed on name-servers and the network. Cache misses cost correspond to the associated user-perceived latency. Our *basic cost model* counts the number of automatic queries vs. the number of cache misses. *Refined cost measures* account for varying query-complexity (number of different name-servers contacted) for measuring automatic-query cost and for elapsed query times to measure miss cost. We use the basic cost model in Section 4.1 and address the refined cost measures in Sections 4.2 and 4.3. For each proposed policy we consider the tradeoff that correspond to different ratios between *miss cost* and *automatic-queries cost*. Overhead and latency costs are dependent, as significant increase in traffic and name servers would result in increased latency. We chose, however, to separate the two since DNS traffic even if increased by a small constant factor would still constitute a small fraction of the overall traffic and should not strongly affect user response time. We also expect that even if this dependence is factored in the relative performance of different policies would stay the same.

This model assumes that storage space is plentiful and records do not have to be evicted before they expire. This is consistent with our data since only about 200K distinct servers were seen by the 3 large NLANR caches combined (see Table 1) over a period of two weeks. The AT&T proxy trace included only 13K distinct servers. The combined size of all associated RRs can easily fit on a small part of a hard disk (or even in memory of a dedicated PC). This assumption of abundant storage is also implicit in (the UNIX version of) BIND, that allows its cache to grow until maximum process-size is exceeded and the application is killed [1].

The *passive policy* corresponds to a single point on the performance curve. It performs a single query for each miss. The following property is established by a simple induction argument on the request sequence.

Lemma 1 *The passive policy performs the minimum possible number of DNS queries needed to serve all client requests.*

The optimal proactive caching policy is the omniscient OPT that uses knowledge of the future: An automatic-query is issued by the name-server *just before* it receives each client-query that would otherwise be a miss. OPT incurs no overhead (issues the minimum number of queries needed to serve the sequence of client-requests[3]) and suffers no cache misses. In the following, we restrict our attention to proactive caching policies that are only allowed to extend freshness of cached items by renewing items as they expire. These policies exclude predictive renewal of long-expired items or prefetching of new ones. We label such policies *Renewal policies*.

3.2 Policies

All the policies that we considered are *renewal policies* which are defined as follows. A *Renewal* of a cached entry is performing a new resolution upon its expiration, and updating the cached copy and extending expiration time accordingly. A *Renewal policy* associates with each cached item a *renewal credit*, which is an integer stating the remaining number of renewals. The credit may be updated (increased) when the cached item is used. When the item expires (or about to expire) and has a positive credit, it is renewed and the credit is decremented.

We describe our renewal policies by specifying for each how it assigns renewal credits. Our policy-design principles were to only use information available locally at the name-server and to be at least as simple to implement as popular cache-replacement policies. We named our renewal policies after analogous cache-replacement policies, where analogies were made based on the property of the request sequence exploited. Policies are parameterized and cost-benefit tradeoffs are obtained by sweeping the parameter value.

- R-FIFO(r): A fixed number of renewals, r, is associated with each item upon a cache miss. The renewal credit is assigned at the point of entry into the cache and is not increased by subsequent hits. This is analogous to the cache replacement policy FIFO (First-In-First-Out) that upon a miss evicts the item with the earliest entry-time into the cache.

- R-LRU(r): Each item is renewed for r times *passed the time of the most-recent cache-hit involving the item*. Hence, when a cache-hit involves an item, its renewal credit is set to r. This is analogous to LRU (Least Recently Used) that evicts the item with the least-recent cache hit.

- R-LRU(ϵ, r) is similar to R-LRU(r), but the renewal credit is reset to r only as a result of hits that occur after an ϵ-fraction of the "current" TTL interval had passed.

[3] assuming that under passive, requests do not occur at the exact end of a TTL interval.

Our motivation for considering R-LRU(ϵ, r) is that plain R-LRU always grants at least 1 renewal when the miss is followed by at least one consecutive hit. R-LRU(ϵ, r) avoids an extra renewal if there is no "deeper" evidence of need such as later hits stemming from requests initiated at a different session.

- R-LFU(r): The first request to the item *in each TTL interval* increases the renewal credit by r. R-LFU(r) is analogous to LFU (Least Frequently Used) that evicts the cached item with smallest number of hits. Both LFU and R-LFU give additional value to an item for every cache hit.

- R-ADAPTIVE: is an enhancement of R-LRU where different credit $r(h)$, is associated with each host h. To find the mapping $r(h)$, R-ADAPTIVE collects per-host statistics from "learning data." Each host h that had sufficiently many misses under R-LRU with $r = 1$ on the "learning data," gets a value $r(h) \geq 1$. A generic value $r(h) \equiv r \geq 1$ is used for hosts for which there was no sufficient data. The mapping of hosts to values $r(h)$ is performed using similar techniques as the ones used in [8, 7] and analyzed in [5]. The implementation of R-ADAPTIVE is considerably more involved than R-LRU, R-FIFO, or R-LFU and requires generating and maintaining the mapping $r(h)$.

- R-OPT: is the optimal renewal policy. The performance curve of R-OPT gives for any number of renewals, the minimum possible number of cache misses. R-OPT relies on knowing the future. Given the request sequence as an (offline) input, the optimal tradeoff-curve can be computed using dynamic programming [4]. For our experiments we implemented R-OPT(r), a simpler greedy approximation of R-OPT. The greedy approximation performs a single pass on the input and grants renewals only if the gap between the expiration and the next request is less than r TTLs. The greedy approximation can be viewed as analogous to Belady's cache replacement algorithm [3]. The approximation collapses to the optimal solution if renewals can be granted for fractional TTL values.

  Note that R-OPT differs from the optimal proactive caching policy outlined above since it follows the more restrictive framework of renewal policies. As such, it provides a more realistic tighter performance upper-bounds.

[4]Consider each hostname separately. Let $r_1, \ldots, r_n$ be the sequence of requests to resolve one hostname. Let $c(i, j)$ be the "best way" to cover $r_1, \ldots, r_j$ with $i \leq j$ misses, that is among the coverings that minimize the number of queries, we consider the one that extends as much time beyond r_j as possible. Note that $c(i, j)$ is not defined for all values. The maximum i for which $c(i, k)$ is defined corresponds to the passive policy. The range of values $c(i, k)$ (minimum number of queries for covering the complete sequence with i misses) corresponds to a tradeoff curve of queries vs. misses. For the optimal policy we actually focus on additional queries vs. reduction in misses with respect to passive. For each hostname, the quantities $c(i, j)$ can be computed using dynamic programming from $c(i', j')$ such that $(j', i') < (j, i)$ in lexicographic order. To obtain a globally optimal solution, we use a threshold value V to select the number of renewals used for each hostname in an "equal way." The "fractional knapsack" strategy outlined in [5] is applicable.

label	requests (thousands)	hosts (thousands)	time-period
LJ	4103	63	May 18–June 5/99
UC	10837	91	May 18–June 5/99
PA	6886	104	May 18–June 5/99
AT&T	489	10.5	Nov 8–19/96

Table 1. Proxy Logs

Varying renewal and miss costs can be naturally integrated into our policies definitions. Our initial evaluation uses the basic model and we subsequently adjust by incorporate projected query times.

4 Performance Evaluation of Policies

Our data included logs from 3 of the large NLANR Web caches (downloaded from the NLANR site [16]) and a proxy log from the AT&T Research proxy. The NLANR caches are high-volume, with large rate of requests and clients that include many proxy caches, whereas the AT&T proxy log reflects the activity of about 460 individual users (IP-addresses). Properties of the different logs are provided in Table 1. We considered only requests made to hosts logged by name that we were able to resolve successfully. This was the vast majority for the NLANR logs, but included only 10.5K out of 13K servers for the older AT&T proxy log. The NLANR logs were split into two parts in order to obtain a learning and test data for evaluating the R-ADAPTIVE policy. Learning part of each log included the first 8 days, and the second part the remaining 11 days.

The proxy logs include information about (scrambled) user identity[5], Web server (host), requested resource (URL), and cache performance (whether the request was a local hit, obtained from another NLANR cache, or obtained from the Web server). The logs, however, do not include any information on DNS queries and results. DNS data is crucial for our performance evaluation and was projected from independent measurements.

Obtaining DNS Data We extracted the list of distinct hosts, and resolved them using `DIG` [1]. Thus, we obtained information on associated IP addresses, aliases (canonical names), immediate name-servers, and DNS query times,

[5]For the NLANR caches a different scrambling was used every day.

along with respective TTL values. We issued several queries for each host name, with varying time intervals between them, in order to estimate rate-of-change of name-to-address mappings, query times, and TTL values. We used these TTL values to emulate the local name-server cache and its performance under various policies. The name-to-address rate-of-change was used for evaluating the performance of SV (see Section 5)

Methodology When projecting on the activity of the local name-server, we assumed in our evaluation that it exclusively handles DNS queries associated with all requests in the proxy log. The number of resolutions, however, is considerably smaller than the number of logged HTTP requests. With persistent HTTP connections (incorporated in HTTP/1.1 [13]), requests for embedded contents and requests issued shortly-after a previous requests to the same host re-use an existing TCP connection, and implicitly, re-use DNS resolution results. Even though some RR have very small TTL values (even 0), (persistent) TCP connections to a host are not aware of expiration of the address RR, and are not terminated when the latter expires. To simulate name-servers co-located at a proxy, we assumed that no resolutions are performed on requests occurring less than 60 seconds after a previous request to the same host. Interestingly, most Web browsers impose larger minimum TTL values when caching name-to-address entries, since actual TTL values are disregarded altogether and an LRU based fixed size cache or a fixed timeout value are used instead. We performed sensitivity analysis by varying minimum effective TTL between 0 and 15 minutes: The DNS cache miss-rate decrease by about 20% (since many projected misses are due to small TTLs), but our results did not qualitatively change (relative improvements of the policies remain the same).

4.1 Performance in the basic cost model

For each policy we measure the tradeoff between the relative decrease in cache misses and the relative increase in DNS queries (issued by the name-server). As the natural baseline we use the passive policy, whose performance on the various logs is listed in Table 2. The table provides, for each log, the total number of DNS misses, the number of DNS misses divided by the number of HTTP requests and the fraction of DNS misses that are associated with the first HTTP request to a hostname. The number of queries issued by the passive policy is equal to the number of DNS misses.

Table 2 shows that the fraction of HTTP requests that incur a DNS cache miss at the local name server is about 6.5%-8.7%. A better but less explicit metric is to look at the fraction of "Web pages" affected, since with HTTP/1.1, fetching of a group of objects while reusing open connections (such as objects located on the same page and served by the same hostname) is relatively efficient (TCP connection establishment and server queueing delays are incurred only once). On average "Web pages" contain 10–20 embedded images, and thus, we expect 0.65-1.7 DNS misses per page.

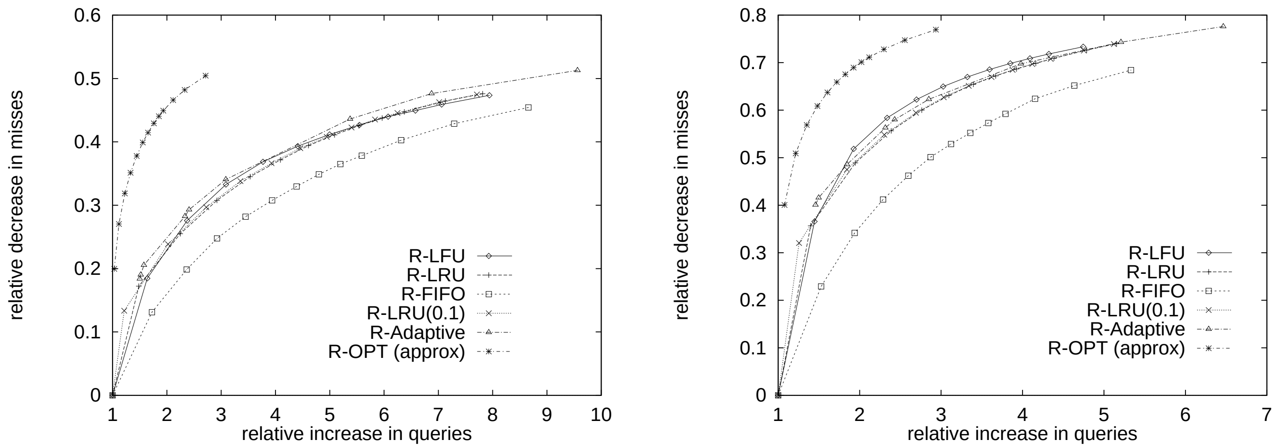
log	DNS misses (% HTTP requests)	FS misses (% DNS misses)
LJ	337964 (8.2 %)	18.7%
UC	941150 (8.7 %)	9.7%
PA	592278 (8.6%)	17.6%
AT&T	31732 (6.5%)	33.0%

Table 2. Performance of passive caching

We refer to DNS misses that are associated with previously-seen hosts as *PS misses* and to misses incurred on the first appearance of a hostname as *FS misses*. The number of PS misses constitutes an upper bound on the performance of renewal policies, since only such misses can be eliminated. It is meaningful to also consider performance in terms of PS misses, since it is robust to varying log duration. The fraction of PS misses increases and converges with log duration (e.g., 150K distinct host seen on the first 8 days of the 3 NLANR caches whereas the following 6 days included only 50K additional hosts).

Figure 1 plots the performance of the various policies on the AT&T proxy log and the second part of the UC log. The relative performance of the different policies was consistent across all 4 logs and the two parts of each log. R-FIFO consistently under-performed other policies. The performance of R-LRU, R-LFU, R-LRU(ϵ =0.1) and R-ADAPTIVE was comparable. R-LRU(ϵ =0.1) and R-ADAPTIVE provided only minor performance improvements over plain R-LRU. The perhaps smaller-than-expected advantage of R-ADAPTIVE is explained by a large number of hosts with a small number of "misses" on the learning data. Thus, R-ADAPTIVE associated the generic renewal credit amount with these hosts. These results also imply that the added complexity of implementing R-ADAPTIVE is not justified by the performance gain.

Figure 2 plots the performance of R-LFU on all 4 logs. The tradeoffs obtained by R-LRU closely follow R-LFU. The convexity of the curves suggest that cache misses were more likely to occur on popular and more recently-requested objects. Better tradeoffs were obtained for the larger logs. This is due to both the fraction of PS misses that increases with log period and to larger number of clients. When the fraction of PS misses is factored out (by measuring reduction in PS misses instead of reduction in the total number of misses), the tradeoffs obtained for different logs move closer together. Across the 3 NLANR logs the tradeoffs completely converged, but the AT&T log that has a considerably-smaller client-base still exhibits worse trade-



AT&T log UC log (part 2)

Figure 1. Performance of renewal policies

offs. On the NLANR logs, R-LRU and R-LFU eliminate about 60% of PS-misses with query overhead of 2 and eliminate 80% of PS-misses with query overhead of 5. The respective reductions for the AT&T log are 36% and 63%.

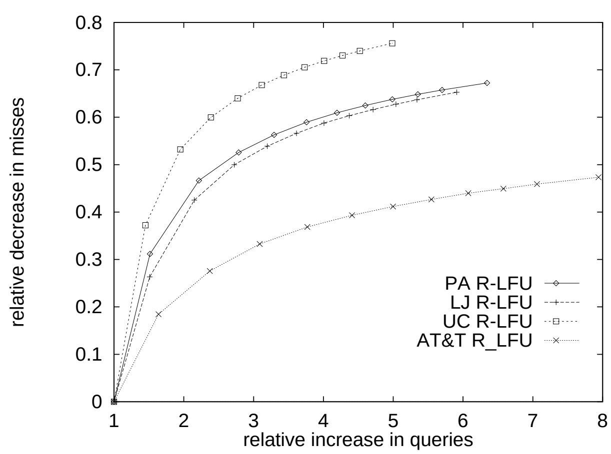


Figure 2. Performance of R-LFU on the 4 logs

4.2 Renewal costs

Renewal costs is measured by the overhead inflicted on name-servers throughout the distributed DNS system. It is hard to capture precisely, but can be estimated by the number of different subqueries issued in the recursive resolution process, or by considering the number of queries issued to the loaded root DNS servers (see Section 2). Root DNS servers typically impose minimum TTL values for top-level domains they are authoritative for, and generally, TTL values tend to decrease when going down the hierarchy. Hence, although the vast majority of from-scratch resolutions involve at least 2 subqueries, the first issued to a root DNS server (see Section 2), renewals are more likely to involve only a subquery to a lower-level server. We considered TTL values for hosts and their immediate name-servers and noticed that TTL values for both immediate name-servers and their address records are larger or equal to the host TTL for 89% of hostnames in the AT&T log and for 87% of hostnames in the first week of the LJ log. This suggests performing renewals *just before* the expiration of the host address record, by directly contacting (the generally still available) immediate name-server. Such *pre-expiration* renewals are resolved by a single subquery and guarantee an authoritative response with the maximum TTL value. This subquery often renews the name-server record itself, allowing for continued pre-expiration benefits in subsequent renewals. Thus renewals, and in particular pre-expiration renewals, are more likely to involve only lower-level lightly-loaded name servers.

4.3 Integrating query times

Query times associated with cache misses may vary substantially and depend on the hostname and cache content (see Section 2). Thus, when evaluating performance, it is important to relate the reduction in the number of misses with the reduction in the number of misses *with long query times*. Since respective DNS query times were not provided with the traces, we project query times using separate measurements. The cache is primed by previous requests to the same hostname and hostnames sharing higher-level domains, hence the elapsed time since a previous request should be accounted for when projecting resolutions times. We issued DNS queries to hostnames in the AT&T log with varying time intervals between them. We associated them with DNS misses in our simulation according to elapsed time since previous cache-miss involving the same hostname. Figure 3 shows the number of DNS misses with (projected) query times exceeding $x \geq 500$ms. The uppermost curve corresponds to misses incurred by the passive policy. There were 31.7K misses in total (see Table 2), and the figure shows that about 1.6K misses (5%) had query times exceeding 3 seconds. The lower-most curve aggregates across all FS misses. This is the baseline curve, since renewal policies (and SV) apply only to PS misses. The curve for FS misses is halfway of the passive-policy curve throughout query times, showing that FS misses account for about 1/2 of the misses with long query times (over 500ms or more). Recall that FS misses account for about 1/3 of the total number of misses of the passive policy on the AT&T log (see Table 2). This gap arises because FS misses are incurred on hostnames that were not previously resolved, and hence, are more likely to incur longer query times. The figure also shows the latency-distribution for misses incurred under R-LFU with different values of r and respective renewal overhead. The number of long query times does not decrease proportionally to the total number of misses. This is expected, since renewal policies target misses occurring sooner after a previous resolution. The figures also show, however, that the number of long query times still significantly decreases as the total number of misses decreases. R-LFU with $r = 2$, for example, performs 140% more queries than passive and eliminates 40% of the long query times incurred on PS misses. R-LRU (not shown) exhibits comparable performance. Last, the figure shows query times under SV, which is discussed in Section 5.

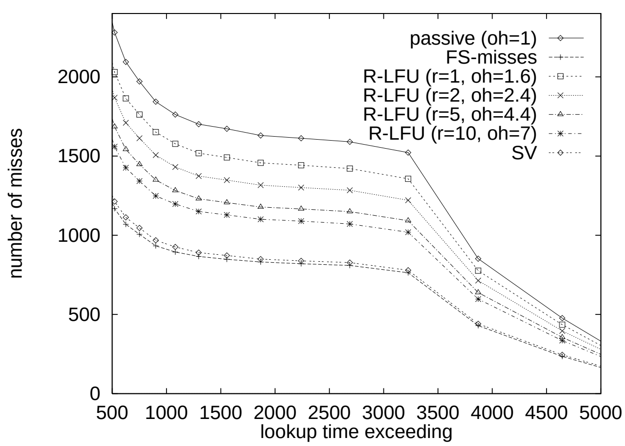


Figure 3. Number of misses with long query times

Implementation Most effectively, renewal policies can be integrated in the name-server software. The implementation can then use the cache content and query history and benefit from performing pre-expiration renewals. Alternative implementation is as a process external to the name-server that issues queries to the name-server. Incorporating a renewal policy in the name-server software requires both (1) maintaining and updating "remaining number of renewals" and (2) scheduling renewals. The policies considered here update the number of remaining renewals only when the item is fetched, is involved in a cache hit, or is being renewed. Thus, updates happen only when a RR is accessed and if the remaining number of renewals is stored together with each record, the update overhead is minimal. The scheduling and issuing of automatic-queries can be performed through periodic scans of the cache. Periodic scans for removing expired items are incorporated in BIND 8. Different implementation that allows for frequent renewals is via a *priority queue* [10]. Scheduling renewals in a queue allows for prompt and quick identification of currently-due items while avoiding frequent complete scans of the data. When TTL lengths allow for sufficient flexibility, it may be desirable to perform renewals at off-peak times.

5 Simultaneous Validation

Simultaneous validation (SV) reduces total document-fetching time by *concurrently* performing the hostname resolutions and subsequent parts of the process (TCP connection establishment and HTTP request-response). It is potentially effective when a fresh entry for the hostname is not available at the local cache, but the cache contains an expired entry obtained from a previous resolution or from elsewhere. SV needs to be supported at the entity initiating TCP connections to Web servers, typically, browsers or proxy servers. With SV deployed, expired DNS records are not discarded, but kept cached for SV use. When a client issues an HTTP request, SV proceeds as follows:

1. If a fresh address record is available, or if no (fresh or expired) address record is available, the process proceeds as if SV is not deployed.

2. Otherwise, if an expired address record is locally available (from browser cache or from a local SV-supporting name-server cache), then concurrently:
   (a) the expired address record is used to attempt a TCP connection establishment. If a connection was established, HTTP request(s) are issued and responses are stored locally.
   (b) The resolver performs a DNS query of the host name. The query response (and subquery responses) update and replaces respective cached records.

SV may continue in several manners, depending on the desired approach. A *conservative* implementation guarantees that the deployment of SV is completely transparent to the user's view (other than reduced latency). Contents fetched in the first step are cached and transferred to the browser (if deployment is at proxy) or displayed to the user (for Web browser deployment) only if the IP-address is validated by the results of the simultaneous DNS query. The latency experienced by the user under a conservative implementation is the maximum (rather than the sum) of DNS query time and the remaining portion of the process serving the request.

Under an *aggressive* implementation of SV, fetched contents are displayed before DNS query response is received. An aggressive implementation is beneficial when DNS query times extend longer than the rest of the process or when relevant name-servers are temporarily unavailable. In these cases, perceived-latency is reduced by DNS-query time. The main disadvantage of an aggressive implementation is compromising transparency. When query response does not validate the expired address, it may result in serving the user with stale or unintended contents. If validation failed, the aggressive implementation has the following options: 1) consider the displayed results as valid; 2) validate the fetched content [13] by contacting the host using a valid address; 3) fetch the content from a valid address and redisplay if there is discrepancy.

Rate of Change of IP-addresses We estimated the rate-of-change of hostname to IP-address mappings. We extracted the list of hostnames from the logs and resolved these names repeatedly with time intervals varying from 10 minutes to 10 days. To estimate change intervals, we took the first IP-address in each resolution of a host-name and checked whether it is present in the list of addresses provided in each of subsequent resolutions. When the address was not included in the resolution results, we considered it as an address change. Our measurements showed a low rate-of-change for the groups of hosts in the NLANR and AT&T logs, where only 2%-3% of hosts had any address changes. A small subset of hosts, including many image and advertising servers, exhibited very frequent changes. About 0.25-0.5% changed daily and fewer than 0.1% changed hourly.

Performance evaluation We evaluated the performance of SV using trace-based simulations with the data and general methodology as outlined in Section 3. Our simulation assumed underlying passive caching at the DNS cache. That is, that cached records are obtained only as a consequence of serving client queries. We used the estimated rate-of-change measurements to associate address changes with DNS misses.

When implemented over a passive cache, SV is applicable only to PS misses (see Table 2 for the fraction of DNS misses that are PS-misses). Hence, we measure the performance of SV by considering all PS misses and measuring the fraction of such misses that were *SV hits*, that is SV was successful (the address did not change since the time of the previous resolution). Performance results for SV on the 4 logs are provided in Table 3 showing success rate of 98%-99%. Note that our evaluation correspond to a conservative implementation of SV, and an aggressive implementation may yield an even higher success rate. Figure 3 shows the corresponding projected query times for SV misses (including FS misses). As we can see, SV is also highly effective on PS misses with long query times. The gap between the SV curve and the FS-misses curve in Figure 3 represents the 2% of PS-misses that SV does not eliminate.

log	SV success rate (% PS misses)
LJ	98.8%
PA	99.1%
UC	99.1%
AT&T	97.9%

Table 3. Performance of conservative SV on various logs: percentage of SV hits out of PS-misses

Implementation Simultaneous-validation requires a two-fold resolution of host names: locally-available expired address records are returned quickly and a resolution is initiated in the standard manner to obtain current addresses. SV can be supported by stand-alone implementation at browsers or proxy-servers. Implementation at proxy-servers benefits from aggregation across multiple users, with more up-to-date and extensive caches.

Stand-alone implementations at browsers or proxy benefits from modifying only a single entity. Nevertheless, browsers or proxies can still benefit from SV support by the local name-server. Name-servers are the natural provider of two-fold resolutions: they already handle DNS requests from multiple users and cache DNS records. To support SV, they need to (i) cache, instead of discard, expired contents[6] and (ii) provide protocol support for two-fold resolutions. Protocol support for SV can be facilitated through new types of queries through which a client indicated interest in possibly expired records. The resolver will use such a query to indicate to its name-server that it is willing to accept stale cache entries. When a stale entry is returned, the name-server should follow up with a valid response.

Coherence Support for a Stale Database A future application of SV, that extends its applicability to cover FS misses, is to provide coherence to an extensive hostname database (e.g., obtained by extracting lists of hostnames from multiple sources). Since the database entries are not guaranteed to be fresh, the extent and reliability of refreshes can be limited. The effectiveness of such a database can be further increased if other attributes are tracked such as frequency of change of IP-addresses and usage of Round-Robin DNS. It can then focus on hostnames with stable

[6]Currently BIND 8 periodically discards expired records.

addresses which are more likely to be SV hits and also support round-robin rotation of mirrors. It is interesting to contrast this with the historic HOSTS.TXT approach that predated DNS (see Section 2). HOSTS.TXT made the complete name-to-address database locally availability, and thus, DNS lookup times were minimal. This approach, however, was not scalable because of growing file size and a coherence mechanism which did not allow for distributed control. In retrospect, storage is less of a concern but coherence remains crucial. The combination of SV and the current DNS infrastructure addresses coherence.

6 Discussion

Latency incurred on DNS misses is inherent in the hierarchical/distributed nature of DNS and is often dominated by RTTs to multiple destinations. As such, query time is not considerably shortened when bandwidth is increased. Nonetheless, reducing perceived-latency due to query times is crucial for improving the experience of web users. We view enhancement to current passive caching of DNS data as a necessary step in the ultimate goal of reducing Web latency. To this end, we proposed and demonstrated the effectiveness of two orthogonal approaches: renewal policies and simultaneous-validations. Renewal policies are effective for hostnames where the typical time interval between requests is within some small factor of the TTL. SV is effective when frequency of change of hostname to address mapping is lower than the frequency of requests. They also differ in their place of implementation: Renewal policies can be implemented only inside the name-server whereas SV needs to also be supported by the entity issuing HTTP requests. These two different solutions suggest a combined approach.

Future work could benefit from simultaneous tracing of DNS data and HTTP requests. Such data (query times, cached records and TTLs, subqueries issued) can be used for evaluation of renewal policies at the name-server or RR level. Another use could be to evaluate the potential benefits of "cooperative DNS caching" where local name-servers exchange non-authoritative cached information.

References

[1] P. Albitz and C. Liu. *DNS and BIND*. O'Reilly, Cambridge, MA, 3 edition, 1998.

[2] G. Banga, F. Douglis, and M. Rabinovich. Optimistic deltas for WWW latency reduction. In *Proceedings of the USENIX Annual Technical Conference*. USENIX Association, 1997.

[3] L. A. Belady. A study of replacement algorithms for virtual storage computers. *IBM systems journal*, 5:78–101, 1966.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *Proceedings of the IEEE INFOCOM'99 Conference*, 1999.

[5] E. Cohen and H. Kaplan. Exploiting regularities in Web traffic patterns for cache replacement. In *Proc. 31st Annual ACM Symposium on Theory of Computing*. ACM, 1999.

[6] E. Cohen and H. Kaplan. Prefetching the means for document transfer: A new approach for reducing web latency. In *Proceedings of the IEEE INFOCOM'00 Conference*, 2000.

[7] E. Cohen, H. Kaplan, and J. D. Oldham. Policies for managing TCP connections under persistent HTTP. In *Proceedings of the World Wide Web-8 Conference*, 1999.

[8] E. Cohen, B. Krishnamurthy, and J. Rexford. Evaluating server-assisted cache replacement in the Web. In *Proceedings of the 6th European Symposium on Algorithms*, pages 307–319. Springer-Verlag, Lecture Notes in Computer Science Vol. 1461, Aug. 1998.

[9] E. Cohen, B. Krishnamurthy, and J. Rexford. Improving end-to-end performance of the Web using server volumes and proxy filters. In *Proceedings of the ACM SIGCOMM'98 Conference*, Sept. 1998.

[10] T. Cormen, C. Leiserson, and R. Rivest. *Introduction to algorithms*. McGraw-Hill Book Co., New York, 1990.

[11] M. Crovella and P. Barford. The network effects of prefetching. In *Proceedings of the IEEE INFOCOM Conference*, 1998.

[12] A. Dingle and T. Partl. Web cache coherence. In *Proceedings of the Fifth International World Wide Web Conference*, May 1996.

[13] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, and T. Leach, P. Berners-Lee. Hypertext Transfer Protocol — HTTP/1.1. RFC 2616, ISI, June 1999.

[14] Internet quality of service assesment. *ftp.telcordia.com/pub/huitema/stats/quality_today.html*.

[15] Inktomi Traffic Server. `http://www.inktomi.com`.

[16] A Distributed Testbed for National Information Provisioning. `http://www.ircache.net`.

[17] S. Keshav, C. Lund, S. Phillips, N. Reingold, and H. Saran. An empirical evaluation of virtual circuit holding time policies in IP-over-ATM networks. *J. on selected areas in communication*, 13, 1995.

[18] B. Krishnamurthy and C. E. Wills. Study of piggyback cache validation for proxy caches in the world wide web. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, Monterey, California, December 1997.

[19] D. Li and D. R. Cheriton. Scalable web caching of frequently updated objects using reliable multicast. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, 1999.

[20] C. Liu and P. Cao. Maintaining strong cache consistency in the world-wide web. In *Proceedings of the 17th IEEE International Conference on Distributed Computing Systems*, May 1997.

[21] P. Mockapetris. Domain names – concepts and facilities. RFC 1034, ISI, Nov. 1987.

[22] P. Mockapetris. Domain names – implementation and specification. RFC 1035, ISI, Nov. 1987.

[23] J. C. Mogul. Hinted caching in the web. In *Proceedings of the 1996 SIGOPS European Workshop*, 1996.

[24] S. Williams, M. Abrams, C. R. Standbridge, G. Abdulla, and E. A. Fox. Removal policies in network caches for world-wide web documents. In *Proceedings of the ACM SIGCOMM Conference*, pages 293–305, August 1996.