Squeezing More Bits Out of HTTP Caches

Jeffrey C. Mogul, Compaq Computer Corporation Western Research Laboratory

Abstract

Computer system designers often use caches to solve performance problems. Caching in the World Wide Web has been both the subject of extensive research and the basis of a large and growing industry. Traditional Web caches store HTTP responses, in anticipation of a subsequent reference to the URL of a cached response. Unfortunately, experience with real Web users shows that there are limits to the performance of this simple caching model, because many responses are useful only once. Researchers have proposed a variety of more complex ways in which HTTP caches can exploit locality in real reference streams. This article surveys several techniques, and reports the results of trace-based studies of a proposal based on automatic recognition of duplicated content.

HTTP accounts for most of the bytes flowing over the Internet backbone (up to 75 percent, in one study [1]). This bandwidth demand requires continued investment in link and switch capacity, and leads to congestion, which increases user-perceived latency. At the edges of the Internet, which are often bandwidth-constrained, every extra byte transferred adds incremental delay; this is a particular problem for home users, most of whom do not yet have a cost-effective means to increase bandwidth above 56 kb/s. And every round-trip through the Internet adds delay, often several hundred milliseconds.

Almost any computer system that suffers from latency or bandwidth problems can benefit from caching. The Web is no exception, and caching mechanisms have been part of HTTP almost since its inception. Caching is perhaps the one aspect of the Web most easily amenable to academic studies, and many research papers have been published. Web caching is also sufficiently useful to have led to the creation of a rapidly growing industry.

Caching works when a reference stream has locality. Temporal locality exists when an item is referenced more than once — a cache can store the item on the first reference, and then return it for subsequent references. Traditionally, Web caches have exploited temporal locality, with a URL as the granularity of reference. Such a cache stores a response to a request for a URL, and then a subsequent request for the same URL yields a cache hit.

Other caches in computer systems, such as CPU data caches, often reach hit rates approaching 100 percent, but numerous studies of actual Web reference streams report much lower hit rates, often 50 percent or less. Indeed, recent studies have shown intrinsic limits to the hit rates achievable with URL-granularity temporal locality: no matter how large the cache or user population, many references will never be cache hits.

This article surveys some of the techniques proposed to extend the effectiveness of HTTP caches, by exploiting other forms of locality or ensuring coherency requirements without having to defeat caching. These techniques include cooperative caching, prefetching, support for partial transfers, differential cache updates (also known as delta encoding), HTML macros. The article also surveys several methods that have been used to evaluate these techniques, and presents recent results suggesting that a proposal for automatic duplicate suppression can improve cache utility.

Reasons for HTTP Caching

Web caches have proved useful for three purposes:

- Latency reduction: A cache can often deliver content to a client faster than the *origin server* (HTTP's term for the original source of the content).
- Bandwidth conservation: Whenever a cache avoids the transmission of bytes to or from the origin server, this reduces bandwidth requirements on that portion of the network, resulting in cost savings and lower congestion. Reduced congestion can also improve overall latency.
- Disconnected operation: A cache can provide access to information when the origin server is unavailable, due to either network disconnection or server failure.

Subsequent sections of this article, describing various techniques for improving HTTP caching, will indicate how each approach affects the first two of these metrics.

Web caches do not yet support fully disconnected operation, because this would require anticipating all (or almost all) of a user's future requests. While several projects have demonstrated impressive results using caches to support disconnected or weakly connected distributed file system access [2, 3], these results may prove hard to transfer to the HTTP protocol. (Web caches designed for latency and bandwidth reduction can still, as a side effect, reduce the likelihood that a cachable resource is perceived as inaccessible.)

Cost and Benefit Synergies and Trade-offs

Sometimes the motivations for caching conflict: the ultimate user might prefer reduced latency, while a corporate accountant might prefer reduced bandwidth charges. A cache can reduce latency by avoiding network round-trips, avoiding the use of low-bandwidth links, or prefetching content. Prefetching imposes a trade-off: a successful prefetch pleases everybody (reducing latency for the user without moving unnecessary bytes), but most prefetching mechanisms often fetch unnecessary bytes as well, increasing the overall bandwidth requirements.

All caching mechanisms also impose a trade-off between the idealized benefits of caching, and the practical costs of building and operating a computer system. These costs include CPU cycles, RAM, disk storage, and other hardware costs; software costs, especially as implementation complexity increases; and operational costs, particularly when cache algorithms require expert intervention to maintain effectiveness. These costs are not limited to the caches themselves; many HTTP mechanisms intended to support caching impose additional costs on origin servers, and some impose added costs on clients.

Server and cache operators often resist spending more on hardware and software. However, hardware capacities are still increasing at exponential rates (for CPU speeds, RAM sizes, and disk capacity), while the speed of light (and hence round-trip time) remains constant. Bandwidth, unlike latency, is not intrinsically limited, although it is not universally improving (especially not on wireless links). Almost any caching mechanism that promises to eliminate round-trips will therefore become cost-effective at some time in the future, and caching mechanisms that improve bandwidth without harming latency should also pay off at some point. On the other hand, a caching system that adds much complexity or network communication in order to avoid CPU or storage costs is not likely to be useful in the long term.

We can encapsulate these considerations into four principles to guide the evaluation of HTTP caching proposals:

- Exploit Moore's law to avoid having to change the speed of light
- Avoid round-trips, unless the payoff is significant
- Use extra header bytes, sparingly, to gain efficiency
- Use hints to improve the payoff of speculative approaches

Later we summarize how the various proposals described in this article address these trade-offs.

In addition, any realistic proposal must take into account the difficulty of deploying protocol changes in an installed base that already includes tens of millions of users. A protocol that can be incrementally deployed may have significant practical advantages over a theoretically superior protocol that does not prove useful until significant numbers of systems are upgraded.

Alternatives to Caching

Caching is not the only way to avoid sending bytes over a network; content simplification and data compression can also pay off.

Content simplification works, to a point. For example, Web designers can use common sense to reduce page complexity, or special software tools to optimize image coding (e.g., [4]), but the desire for richer user experiences usually prevails over such pragmatism. New content formats, such as cascading style sheets [5], can increase coding efficiency without reducing expressiveness. But some content, such as medical images, broadcast-quality video, and executable software, cannot be simplified without loss of meaning. And other emerging formats, such as MPEG-1 layer 3 (MP3) audio [6], provide new bandwidth challenges.

Data compression directly targets the transmission of redundant bits within a single transfer. Existing general-purpose compression algorithms provide significant size reductions, typically reducing text file sizes by a factor of three or more. However, the bulk of Web-related bytes transferred come from content-types such as images, video, and audio, which are already heavily compressed using highly efficient content-specific compression algorithms; so the net benefit of applying general-purpose compression to all Web traffic would be relatively small [7]. The trend toward increased use of nontext media further reduces the potential for general-purpose compression in HTTP.

With the increased transmission of new data types (e.g., Java byte codes and other software), one can expect to see progress on new type-specific compression techniques. For example, general-purpose algorithms can be tuned for improved compression of binaries [8], and special-purpose compression algorithms based on parse trees can do even better [9]. Even so, compression has its limits, and data types such as program binaries inevitably gain complexity with time.

Other Proxy Functions

HTTP supports the use of caches directly integrated with the end client (e.g., a browser), closely associated with a particular server (*server accelerator*), or integrated with a proxy server. Many proxy systems serve other functions, such as security protection (i.e., as a firewall), censorship, and transcoding [10] (a non-end-to-end form of content simplification). These functions are not intrinsically related to caching, and are outside the scope of this article.

Proxy systems set up as server accelerators operate on very different reference streams than do other HTTP caches; therefore, server accelerators are also outside the scope of this article.

Methods for Evaluating HTTP Caching Mechanisms

Because caching exploits locality in a reference stream, we can evaluate the potential performance of various caching designs by obtaining a reference stream, and then either analyzing the stream or simulating the performance of a caching design on this stream.

The input reference stream can be either a real stream, captured by tracing the references of actual users, or a stream generated by a model of user activity. Several workload generators have been devised either for use in benchmarking proxy caches [11] or for more general uses [12].

Trace-based studies require more effort to capture and store the trace logs, but they avoid the simplifying assumptions made when parameterizing a workload generator (and, in any case, the parameters for a realistic workload generator must be based on some set of traces.) While a few studies have been able to capture the reference stream generated by instrumented browsers [13], almost all interesting traces have been captured by either logging at proxies, or capturing network packets and then reassembling the TCP streams.

Inaccuracies are possible even when using traces to drive simulations or performance measurements, because this still requires certain simplifying assumptions. In particular, a cache may alter the response time seen by users, who in turn might alter their behavior. One can test the modified system under a live load, but because of unpredictable variations in usage patterns, this might mean giving up the repeatability provided by traces or load generators. Consequently, it may be hard to discern whether a change in the system under test leads to a significant performance improvement.

One can bypass the repeatability problem when comparing two or more implementations by running them simultaneously, and randomly splitting the incoming reference stream among the systems under test. In principle, this should eliminate most of the variation in load, so any difference in performance can be attributed to the implementation differences. In practice, it might be difficult to achieve truly random load splitting, and also to split the load without affecting locality properties.

Limits of Simple Approaches to HTTP Caching

Most studies of Web reference streams and proxy caches report the *hit ratio* (HR) per resource and the *byte hit ratio* (BHR), weighted by the size of the response body (also sometimes called the *weighted hit ratio*, WHR). The bandwidth reduction ratio should be similar to the BHR, although the BHR does not account for protocol header overheads. Simulation studies have reported HRs ranging from 30–49 percent and BHRs ranging from 14–36 percent (assuming an infinite cache size) [14, 15]; the variations may be due to differences in user community, geography, or when the traces were obtained. Reports generated from in-service measurements of the National Laboratory for Applied Network Research (NLANR caches [16] show actual BHRs vary tremendously over short timescales.

It seems impossible to significantly increase HRs and BHRs above a (fuzzy) threshold, because of several intrinsic aspects of Web reference streams:

Uncachable Resources — Some responses cannot be cached; for example, stock quotes, query results, or e-commerce "shopping baskets." Other responses could be provided from a cache, save for the server site's desire to gather demographic information or advertising revenue.

Particular causes of cache misses include:

- **Coherency misses**: Any system where data items are frequently updated, caching is employed, and use of an out-of-date value could lead to erroneous behavior requires a mechanism to maintain cache *coherency*. Various coherency mechanisms have been developed for distributed systems, often using techniques such as callbacks or leases. HTTP, however, has no way to guarantee coherency for a resource except by disabling caching for that resource.
- **Nonce URLs**: Some sites implicitly defeat caching by generating a unique URL for each reference to a resource. For example, the same advertising banner may appear on millions of pages, but each time the banner appears with a different URL.

Zipf's Law — The Web is so large that many pages will never be referenced more than once in the reference stream seen by any one cache. Most cache hits come from a small set of resources, but many references are made to resources outside this set. Studies have shown that page re-reference frequencies follow a distribution similar to Zipf's law, in which the relative probability for a reference to the kth most popular page is a proportional to $1/k^a$, for some constant a [17–19]. This implies that in a large universe of pages, most of these pages are extremely unlikely to be referenced twice in the same reference stream. The first reference via a given caching proxy to any given resource is a *compulsory miss*, since the value cannot possibly be in the cache.

High Rate of Change — Many potentially cachable resources change fairly rapidly [20], which would lead to cache incoherence if responses for these resources were allowed to be cached. If the semantics of the resource does not require absolute coherency, HTTP allows the origin server to limit the cachable lifetime of a response, in the hope that it will expire before the resource is updated (effectively a form of lease). But this trades improved latency against potential incoherency.

Voelker *et al.* point out that most Web resources appear to change more rapidly than they are re-referenced, unless the reference stream comes from an enormous population [19].

Resource Size Distribution — Why is the byte hit ratio almost always lower than the simple hit ratio? Williams *et al.* [21] observed that most references in their traces were for small resources. Breslau *et al.* [17], working from several trace sets, showed that there is no strong correlation between resource size and access frequency, although the mean size of popular resources is smaller than the mean for unpopular ones. One possible explanation for these observations is that a small set of small resources accounts for most of the cache hits.

Practical Limits on HTTP Caches

It is tempting to evaluate an HTTP caching design by simulating an idealized implementation of the design. However, this runs afoul of several practical constraints on HTTP caches.

The most obvious one is cache capacity. Any real cache has finite storage, and once it fills up, it must evict old entries in order to store new ones, following some *replacement policy*. One simple policy is Least Recently Used (LRU), but various studies (e.g., [21, 22]) have proposed other policies that reduce the number of *capacity misses*, relative to LRU, in some cases. A replacement policy also imposes implementation costs (including meta-data storage, update, and lookup costs), and some policies might not be feasible to implement.

In current technologies, the working set of a large HTTP cache does not fit into an economically reasonable amount of RAM; some or all of the cached data must be stored on disk. Disk access adds latency: not only the disk's average access time, on the order of 10 ms, but also queuing delays if the disk is the bottleneck. These queuing delays can be quite large.

If we are using a cache to reduce latency, we cannot simply evaluate it on the basis of HR. Instead, we must estimate its weighted access latency, using

$$weighted\_access\_latency = HR * hit\_cost + (1 - HR) * miss\_cost$$

$$miss\_cost = retrieval\_latency + cache\_check\_latency$$

where *HR* is the mean hit ratio, *hit_cost* is the mean total cost of processing a cache hit, and *miss_cost* is the mean latency for a cache miss. The *miss_cost* is the sum of two terms: not just the *retrieval_latency* for retrieving the response from the origin server, but also the *cache_check_latency* to determine if the cache entry is present. (The *hit_cost* may also include *cache_check_latency*.)

If the hit ratio is quite high and the *cache_check_latency* not too excessive, the *weighted_access_latency* will be better than the *retrieval_latency* (the latency without caching), and the cache will pay off. However, observed hit ratios are usually well below 50 percent, and queuing delays at the disks of an overloaded proxy can be quite large. Therefore, the *hit_cost* may be significant, and one cannot simply assume that an

HTTP proxy cache will actually improve overall latency. In particular, the disk subsystem must be carefully designed to avoid imposing excessive latencies [23, 24].

The *cache_check_latency* depends mostly on whether the data structures used for the cache lookup can be held in RAM, or whether the lookup even requires a disk access. Also, if the data structure is stored on disk, lookups must compete with frequent updates for disk bandwidth. Unless the replacement policy has perfect future knowledge, it will store responses that will never be used for cache hits. Hence, the data structure might be updated almost as frequently as it is used for lookups.

With modern large-capacity disks, it might be reasonable to configure a cache with sufficient disk storage that it almost achieves the HR of an infinite cache. But it might not pay to increase the cache size beyond the point where the lookup data structure no longer fits in RAM, since capacity misses decline only slowly past a certain cache size.

Storage capacities of RAM and disk are still growing exponentially. These growth rates could increase the effectiveness of HTTP caches; but will hardware improve fast enough, or will Web reference rates grow even faster? One study (of a very small and atypical user population) reported that "potential for caching requested files in the network has declined" between 1995 and 1998 [13], but more extensive studies are warranted.

Given that RAM size is increasing exponentially, but disk access times increase very slowly, and most cache hits come from a relatively small subset of the cachable content, at some point it might become feasible to build RAM-only HTTP proxy caches. That is, if the replacement policy does not justify storing a response in a large RAM, it might not be worth the cost of storing the response on disk: the extremely low chance that it would be useful in the future might not offset the cost of the disk write or of maintaining relevant meta-data.

Cooperative Caching

Since the HR of a Web cache is a function of the user population, perhaps by combining the efforts of several *cooperative caches* [25], we could improve overall performance. When a proxy *P*1 receives a request but has no corresponding valid cache entry, *P*1 could send a request to one or more cooperating caches *P*2, *P*3, ..., *Pn*, asking if they have a valid entry for the resource. This pays off if the entry is indeed in one of the cooperating caches, and if these caches are close enough that the intercache transfer has a latency or bandwidth advantage over direct retrieval from the origin server.

One might expect cooperative caching to pay off for two reasons. First, the effective cache size is larger (reducing the probability of a capacity miss). Second, the cooperating caches see a combined reference stream from a larger set of users, which reduces the probability of a compulsory miss; a resource being referenced by a client of *P*1 might already have been referenced by a client of *P*2.

Cooperative caching does have costs. The intercache messages add more latency to a reference that ultimately misses in all the caches; these messages and their responses also increase bandwidth demands. If the cooperating caches are arranged in a hierarchy, as is common, the higher levels of the hierarchy may become bottlenecks (imposing queuing delays). Several recent proposals have greatly reduced these costs [15, 26] but they cannot entirely be eliminated.

Do the benefits of cooperative caching outweigh the costs? One recent study by Voelker *et al.* [19] suggests that cooperative caching is useful only in limited circumstances. The overall cache HR does increase with increasing population size, but as the population grows above a few thousand users, the additional benefits are very small. Voelker *et al.* state that the population size that can significantly benefit from cooperative caching "could easily be handled by a single proxy cache." For larger populations, the benefits of cooperative caching appear to be minimal, given current access patterns.

There might be administrative or geographical constraints that prevent the use of a single proxy for a set of small clusters of users, but these same constraints might also work against cooperative caching: separate administrations might have security concerns about shared caches, and between geographically distinct clusters the network latencies and bandwidth might make the intercache protocol infeasible.

In any case, a cooperative caching system serving a given set of users cannot provide a better HR than an idealized infinite cache. Studies have shown that even infinite caches could not provide especially good HRs, which suggests that we must look elsewhere for improved Web bandwidth and latency.

Squeezing More Performance Out of HTTP Caches

Given the apparent limits on the performance of simple caches, researchers and vendors have developed several techniques to extend the utility of Web caches. If the basic principle of simple "reuse" caching is to exploit repeated references to entire cached responses, the basic principle of these extended mechanisms is to exploit partial information present in caches.

There are at least three such kinds of partial information:

Clues About Future References — Because a cache sees an entire reference stream, it can use the information in the reference stream to make predictions about future references. This may allow the cache to more accurately prefetch data before it is actually referenced, potentially resulting in much lower latencies. It may also allow the cache to make better replacement decisions.

Filling In or Replacing Gaps — A cached response might hold some, but not all, of the bits required to satisfy a subsequent request for the same URL. If the cache entry contains missing information, the gap could be filled in using a partial transfer. Or the underlying resource might have changed, but in such a way that it would be more efficient to transfer the differences between the cache entry and the current resource instance, rather than retrieving an entire new response.

Alias Discovery — Frequently, the same content appears in the Web under more than one URL: there are multiple *aliases* for a given piece of data. If the cache can detect such aliasing, it might be able to avoid storing multiple copies, or retrieving data that it already has stored under another alias.

Techniques that exploit partial information typically impose a trade-off. For example, they may require additional computation or storage; within certain limits, this is usually worthwhile if it reduces network bandwidth or latency. Or a technique may improve latency at the expense of bandwidth, by either transferring a few additional HTTP header bytes in some cases, or transferring additional messages. Or one might be able to improve overall bandwidth by transferring a few additional bytes in carefully chosen cases.

The rest of this article describes various proposals to exploit partial information in order to increase the utility of HTTP caches. For each proposal, I will try to evaluate the trade-offs involved.

Prefetching and Replacement Hints

If a cache can predict a user's future references, and if there is spare bandwidth, the cache can prefetch the expected resources. When the prediction is correct and timely, this can increase the BHR and reduce user-perceived latency [27–29], but inevitably increases bandwidth requirements (because of false prefetches). False prefetches also consume cache space, and might displace useful cache entries. Prefetching forces a cache operator to choose between improving latency and minimizing bandwidth.

Successful prefetching therefore requires solution of two problems: making accurate predictions, and deciding whether sufficient extra bandwidth exists.

Prefetch prediction algorithms use the reference stream, usually on a per-user-session basis, in two ways. First, they observe the stream over relatively long periods in order to construct a model of the conditional probabilities of observing certain references given a set of previous references. Second, they use the recent behavior of the stream (sometimes just the most recent reference) as input to this model, which may then generate a prediction of some set of future references. For example, if the most recent reference is to http://ieee.org/index.html, the model might predict a reference to http://ieee.org/about/ with probability 0.2, and a reference to http://ieee.org/conferences/ with probability 0.6.

The first proposals for Web prefetch prediction employed first-order Markov models [30, 31]. Other model-based mechanisms, with better prediction performance or storage requirements, have since been described [32].

Regardless of the particular prediction algorithm, such a prefetching system cannot predict every reference: many references will never have appeared before in the reference stream, or involve parameters (as in a POST-based form) that are not visible to a proxy cache. Also, since a prefetch delivers data in advance of the actual reference, a response that cannot be cached (e.g., due to coherency requirements) cannot be prefetched. Kroeger *et al.* report a trace-based study demonstrating that, with infinite cache capacity and infinite bandwidth, at best one can get a 60 percent latency reduction from any prefetching mechanism that obeys reasonable causality constraints [28]. In other words, at least 40 percent of the latency in this trace was due to compulsory and coherency misses. Any real prefetching proxy would be further constrained by both cache size and bandwidth.

Alternatively, a proxy can parse the HTML responses it forwards, extract the link URLs from these, and then prefetch the link targets [33, 34]. This approach, however, has limits: it cannot on its own distinguish between likely and unlikely prefetch targets, and it depends on parsing HTML documents.

Prefetching makes sense only if sufficient unused bandwidth exists; otherwise, the extra transfers for unnecessary prefetches would cause network congestion. In a few special cases, one can readily determine that extra bandwidth is available (e.g., when a home user's dialup modem is idle), but in the general case this is a difficult problem, with little available research. Maltzahn *et al.* were able to show that a diurnal variation in demand-fetch bandwidth requirements leaves capacity, during the early morning, for some prefetching, but their work made several simplifying assumptions [35].

Predictions about future references can also be useful in making replacement decisions. Caches typically make such predictions by applying an algorithm, such as LRU, to their own reference streams (for HTTP caches, better algorithms than LRU have been developed [22]). However, because of the Zipf's law distribution of references, a cache will seldom have sufficiently precise information about most of its entries. A server could supply hints, in the HTTP headers, about the appropriate replacement strategy for a response [36, 37]. Hint information could include both the expected frequency of future references, and the expected cost of a future retrieval.

Partial Transfers, Delta Encoding, and Macros

An HTTP transfer can terminate in midstream because of a network error, because the user clicks the stop button, or because the user clicks on a link before the entire page is loaded. In HTTP/1.0, the result of a partial transfer is not worth caching, but in HTTP/1.1, a cache can fill in the missing data using a *range retrieval request* [28]. The ability to retrieve ranges of a response makes partial cache entries useful: although the proxy cannot avoid an additional HTTP request/response round-trip, it can avoid transferring bytes that have already been sent over the network. Unfortunately, there are no published statistics for the prevalence or size distribution of partial transfers, but anecdotal evidence suggests they are not uncommon.

The use of range retrievals in HTTP/1.1 imposes a minimal trade-off, limited to the transmission of one or two extra request header lines. Since the requesting cache knows that it has partial content and wants the full response, the only uncertainty is whether the server is able to generate a partial response.

Partial content as the result of a truncated transfer represents a special case of a more general situation: a cache entry already contains some arbitrary subset of the bits of the desired response. Proposals intended to exploit these bits include *delta encoding*, *cache-based compaction*, and *HTML macros*.

The motivation behind delta encoding [39, 40] is the observation that, although many frequently referenced Web resources change too frequently to allow for useful caching, the changes are often minor. Instead of sending the entire current instance of a resource, the server sends just the difference (or *delta*) between the cached and current instances. With careful encoding [41], these differences can be quite small. The potential savings vary by content-type, because much Web content, such as images and continuous media, tend to change more radically than HTML text, but one trace-based study [7] showed potential overall bandwidth savings of 8.5 percent and latency savings of 5.6 percent, and significantly greater improvements in text content.

The use of delta encoding in HTTP requires relatively little protocol overhead and no extra round-trips, but it does require the origin server to either store previous instances of a changing resource, or be able to dynamically recreate these instances. It also requires a modest amount of computation to create and decode the deltas.

Woo and Chan proposed cache-based compaction [42], which they describe as a further generalization of delta encoding, combined with the use of dictionary-based compression. Dictionary-based compression algorithms, such as Lempel-Ziv-Welch [43], use a dictionary to map between input symbols and compressed codes. In order to decompress the output, the receiver must have the dictionary, which means that it is normally transmitted (in some form) as part of the compressed file.

Whereas delta encoding transmits the difference between the current instance of a resource and one older cached instance of (usually) the same resource, cache-based compaction uses a larger set of cached older instances, possibly from a variety of different resources. This set of older instances functions as a large compression dictionary; the new instance is compressed using this effective dictionary, which does not itself have to be transmitted (because both cache and server are using the same set of older instances).

Cache-based compaction should improve on the performance of delta encoding (because it includes delta encoding as a special case). However, the protocol is more complex, because the mechanism must choose an appropriate subset of cache entries (it would be infeasible for the cache to inform the server of its entire state), and because the cache and server need to agree on which set of older instances is employed. So far, the proposal has not been extensively evaluated, but it appears to have advantages mostly for low-bandwidth links.

The approaches described previously in this section operate independent of the syntax and semantics of the data being transferred (although delta encoding algorithms for images may require some specialization). They function by decomposing responses at the bit or byte level into currently cached and need-to-be-transferred components. One can also do this decomposition at a higher level. Douglis *et al.* [44] describe an *HTML macro* mechanism in which a set of similar HTML pages is decomposed into a constant component (akin to a macro body) and a variable component (akin to macro arguments). In many cases, the variable component can be quite small; this means that once the constant component is in a cache, references to similar pages require fetching only the small variable component, at a significant cost savings over transferring a monolithic response.

The main drawback to the HTML macro approach is that it requires direct involvement by the designer (or software) when generating the Web pages, including some careful attention to the decomposition of a set of similar pages. It might also require some additional language-level standardization, although this perhaps could be obviated through the use of Java-based macros.

Alias Discovery and Automatic Duplicate Suppression

The techniques for extending cache utility discussed above do not exploit one possible mechanism for exploiting existing cache entries: if two distinct resources generate exactly identical responses, a cache entry for one of them could be used to provide a cache hit for a request referencing the other.[1] This technique is called *duplicate suppression*, and in principle could avoid a lot of data transfer and related latency.

In practice, its utility depends on:
- A simple and efficient way to detect exact duplication
- A simple and efficient HTTP protocol extension to carry the necessary meta-information
- A sufficient rate of duplication to justify deploying the protocol extension

The first requirement is met by the use of a digest (or checksum) algorithm, such as MD5 [45], for which it is difficult to generate identical digest values ("collisions") for two different inputs. However, there is some suspicion that MD5 is not collision-proof [46], and if an attacker could generate a carefully chosen collision, this would create a security hole. Other more secure digest algorithms, such as SHA-1 [47] or RIPEMD-160 [48], could increase security but at the cost of increased HTTP header sizes and computational costs.

Several protocols have been proposed to address the second requirement. The first such design, the Distribution and Replication Protocol (DRP) [49], proposed creating a special Universal Resource Name (URN) out of the MD5 or SHA digest for a resource. A later refinement of this proposal retains the traditional HTTP URL mechanism for naming resources, and transmits the digest in a new HTTP header field [50]. We present a simplified description of this proposal below; it involves sending hint information from servers to clients, and other hint information from clients to caches, in the hope that the caches can avoid requesting duplicated responses directly from the servers.

Finally, we need to evaluate whether the rate of duplication is high enough that it justifies the extra protocol overhead, computational overhead, and implementation complexity of the proposal. We summarize the results of a trace-based study that addresses this question.

Related Work

Several other techniques also use digests to exploit exact data equality. If duplication is indeed common, a caching proxy could end up storing multiple copies of many response bodies. By computing a digest of every cachable response body and maintaining an index keyed by the digest results, the proxy can detect this situation and arrange to store only one copy of the duplicated body. Inktomi's Traffic Server [51] uses this technique.

Given that duplication of HTTP bodies is common, one might also expect to see exact duplication at the packet level. Santos and Weatherall [52] describe a router-based technique which detects when a packet body is a duplicate, and sends its digest value instead of the entire body. This mechanism requires the sending router to have fairly accurate information about the state of the receiving router's packet cache. Santos and Weatherall report bandwidth savings of about 20 percent, with relatively little overhead. They also report an HTTP-specific packet duplication ratio of about 26 percent.

Proposed Duplicate Suppression Protocol

This section presents a simplified duplicate suppression extension to HTTP in order to make the rest of the article more concrete. Complete specifications for several protocols are available [49, 50].

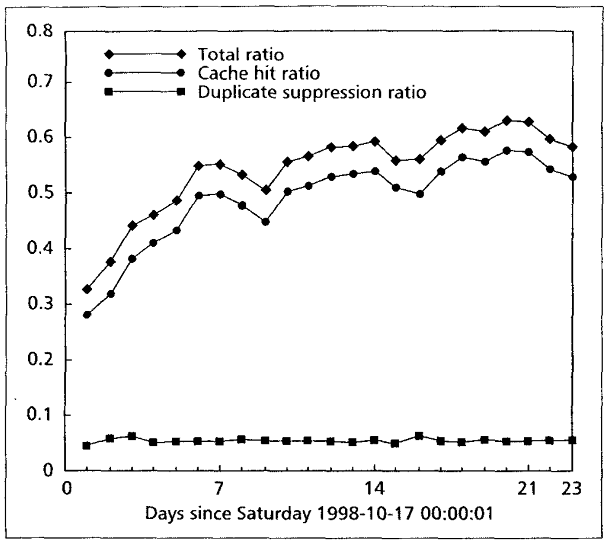
Although users occasionally load Web pages by typing a URL, in most cases an HTTP transfer is initiated when the browser software follows a link. Except in the relatively infrequent case where a link leads to another server, the source of the linkage information is also the source of the linked-to resource; that is, the same server often controls both the link information and its target.

This allows the server to provide meta-information about the link target as part of the linkage information. (For example, HTML supports the HEIGHT and WIDTH attributes of an IMG tag, allowing the browser to reserve screen space for an image before loading it.) To support duplicate suppression, the server would include in this meta-information an MD5 (or similar) digest of the link target.

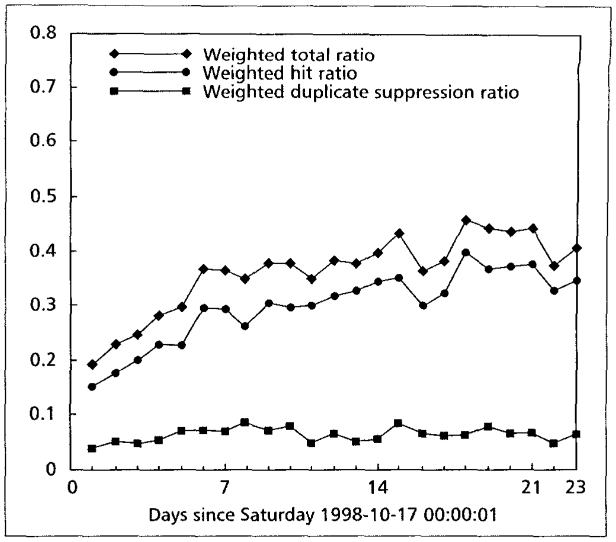
As an alternative to using a new HTML attribute, the digest value could be transmitted in a structured type. DRP introduces a new "index" content-type to provide meta-information for a consistent set of link targets [49]; similarly, WEBDAV [53] specifies a similar "collection" resource, whose state consists of a list of member URLs and an extensible set of properties.

Assume then that a client, about to make a request for URL *U*, has a server-supplied hint that the proper response

[1] *Cache-based compaction can do this in principle, but only if the two resources have similar URLs; this is often not true in practice.*



■ Figure 1. *Daily ratios.*



■ Figure 2. *Daily ratios, weighted by response size in bytes.*

has an MD5 digest value *D*. The client can check its cache not only for an existing entry for *U*, but also for an existing entry with a digest value of *D*. Either cache entry should therefore be a satisfactory substitute for getting a response from the actual server.

If the client's local cache does not contain the target, it could send its request via a proxy cache that might. This request would be, in essence, "please send me either a response for URL *U*, or a response with MD5 digest value *D*." If the proxy cache has a response cached under either key, it can return the cache hit rather than forward the request to the server. Thus, once the client knows the proper MD5 digest value, it can use both its own cache and a proxy's cache to find a duplicate with the same digest, rather than waiting for a response from the actual server.

The complete specification of a duplicate suppression protocol would require attention to a number of other issues, such as whether HTTP header information for a cached response (e.g., authentication information) can properly be associated with a duplicate suppression response for a different resource, and whether the cache entries discovered by duplicate suppression hints are timely with respect to the requested URL.

Potential Benefits of Duplicate Suppression

Duplicate suppression can be evaluated by simulating its effect on a reference stream trace, such as one taken from a proxy server; the trace would have to include the digest value of every response body. However, without actual deployment of a duplicate suppression protocol, one cannot know how often clients would actually receive server-supplied digest hints. For this reason, a trace-based simulation of current references can provide only an upper bound on the potential improvements.

I obtained a 23-day trace from a noncaching proxy at Compaq Computer Corporation. The raw trace includes 29 million entries, although only about 19 million are useful for this study. The trace was then fed to a simulator, which modeled an infinite cache; this avoids any capacity misses, and thus reduces the apparent advantage of duplicate suppression. The simulation also models a "perfect coherency" cache, since the MD5 digests allow the simulator to know for sure whether a coherency miss would be necessary or not.

Space permits only a brief summary of the simulation results; more detail is available in [54].

The infinite perfect-coherency cache simulation, without duplicate suppression, yielded an HR of 50.4 percent and a BHR of 32.2 percent. Duplicate suppression "hits" always take the place of cache misses, not cache hits, so their benefit always adds to that of simple caching. Had the duplicate suppression mechanism been applied to every eligible request, it would have avoided an additional 5.4 percent of the retrievals in the trace. Weighted by bytes transferred, duplicate suppression would have saved an additional 6.2 percent of the server response bytes.

Figure 1 shows the unweighted cache HR and duplicate suppression ratios, sampled at 24-hr

Technique	Latency	Bandwidth	Coherency	CPU	RAM	Applicability
Prefetching	++	– –	+/–	–	–	General
Replacement hints	+	+?	=	=	+	General
Partial transfers	+	++	=	=	–	General
Delta encoding	+	+	=	–	– –	Mostly HTML, text
Cache-based compaction	+?	+?	=	–	=?	Mostly HTML, text
HTML macros	–	+	=	=	=	HTML only
Duplicate suppression	+	+	=	–	– –	Image, stream, and program objects; maybe HTML and text
++: Much improved	+: Somewhat improved	=: Not significantly affected				
+/–: Depends on circumstances	–: Somewhat worse	– –: Much worse	?: Needs more study			

■ Table 1. *A comparison of HTTP caching techniques.*

intervals. The curves trend upward because the simulations start with an empty ("cold") cache. The figure also shows the unweighted "total" ratio (including both cache hits and duplicate suppression "hits"). Figure 2 shows the corresponding byte-weighted ratios.

The weighted cache HR is much lower than the unweighted cache HR, but the weighted duplicate suppression ratio is higher than the unweighted duplicate suppression ratio. Therefore, the net bandwidth improvement due to duplicate suppression is more significant than one would expect from the unweighted ratios.

The unweighted duplication ratio never gets much above 6 percent on a daily basis, which implies that it might not be worth the extra protocol overhead. This statistic is calculated over all responses in the trace, but one might expect that some sets of resources are far more likely than others to be subject to duplication. If so, one could limit the protocol overhead of duplicate suppression to these contexts, and concentrate the effort where the benefits justify the costs. For example, "audio/midi" content (relatively rare in this trace) yields a duplication ratio of 15 percent and a weighted ratio of 15.5 percent. Other content-types with relatively high ratios include Java byte codes, other program binaries, GIF and JPEG images, and some video formats.

Almost 84 percent of the URLs in the trace were never involved in duplication; 16 percent were duplicated exactly once, and 0.15 percent were duplicated exactly twice. In other words, very few URLs are duplicated more than once, but some highly duplicated URLs account for most of the duplication. In fact, half of all duplicate responses come from URLs that give rise to at least 406 different duplicate responses (i.e., the median of the nonzero duplication counts is 406). Also, most duplication is limited to a very small subset of the server hosts in the trace.

These results and others suggest that one could predict, based on content-type, server host, or other factors, whether a response is likely to experience duplication. This would allow the server to avoid sending digest hints except in cases where they would probably pay off.

Summary

This article has surveyed a number of techniques for better exploiting the bits in HTTP caches. How do these techniques compare? Table 1 summarizes the trade-offs for each technique on five metrics: latency, bandwidth requirements, cache coherency, CPU time (on servers, proxies, and clients), and proxy RAM requirements. The last column lists the contexts in which each approach appears to be most applicable. The table entries are guesses, at best, pending more extensive studies.

Note that most of the entries in Table 1 do not affect cache coherency. Several techniques to improve HTTP cache coherency have been proposed [55, 56]. Because current methods for avoiding incoherency often disable caching, these new techniques may enable improvements on other metrics. For example, if more cache entries are known to be coherent, this should provide more opportunities for delta encoding and partial transfers.

Conclusions

HTTP caching remains a fertile area for both research and development of commercial products and services. Web caching researchers initially focused on replacement policies and cooperative caching, but these lines of research may be reaching diminishing returns. The wide variety of recent proposals to improve the effectiveness of Web caches suggests that the space of possible solutions has not yet been fully explored. Even for the techniques described in this article, we still lack sufficient understanding of their utility, and how best to implement and deploy them.

Acknowledgments

I would like to thank Mark Manasse and Andrei Broder for their help in explaining to me various aspects of cryptography and digital signatures. I would also like to thank Carey Williamson for prodding me to write this article, and the anonymous reviewers for their helpful comments.

References

[1] K. Thompson, G. J. Miller, and R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics," *IEEE Network*, vol. 11, no. 6, Nov./Dec. 1997, pp. 10–23.
[2] J. J. Kistler and M. Satyanarayanan, "Disconnected Operation in the Coda File System," *Proc. 13th ACM Symp. Op. Sys. Principles*, Oct. 1991, pp. 213–25.
[3] L. B. Mummert, M. R. Ebling, and M. Satyanarayanan, "Exploiting Weak Connectivity for Mobile File Access," *Proc. 15th Symp. Op. Sys. Principles*, Copper Mountain, CO, Dec. 1995, pp. 143–55.
[4] Equilibrium, DeBabelizer Product Information page, http://www.equilibrium.com/ProductInfo/ProdInfo.html
[5] H. Lie and B. Bos, "Cascading Style Sheets, level 1," Rec. REC-CSS1, W3C, Dec. 1996, http://www.w3.org/pub/WWW/TR/REC-CSS1
[6] ISO/IEC, 11172-3, "Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to About 1.5 Mbit/s — Part 3: Audio," 1993.
[7] J. Mogul *et al.*, "Potential Benefits of Delta Encoding and Data Compression for HTTP," *Proc. SIGCOMM '97*, Cannes, France, Sept. 1997, pp. 181–94.
[8] T. L. Yu, "Data Compression for PC Software Distribution," *Software — Practice and Experience*, vol. 26, no. 11, 1996, pp. 1181–95.
[9] J. Ernst *et al.*, "Code Compression," *Proc. ACM SIGPLAN '97 Conf. Prog. Lang. Design and Implementation*, Las Vegas, NV, June 1997, pp. 358–65.
[10] A. Fox *et al.*, "Adapting to Network and Client Variation via On-Demand Dynamic Transcoding," *Proc. ASPLOS VII*, Cambridge, MA, Oct. 1996, pp. 160–70.
[11] J. Almeida and P. Cao, "Measuring Proxy Performance with the Wisconsin Proxy Benchmark," *Proc. 3rd Int'l. WWW Caching Wksp.*, Manchester, England, June, 1998.
[12] P. Barford and M. Crovella, "Generating Representative Web Workloads for Network and Server Performance Evaluation," *Proc. Joint Int'l. Conf. Meas. and Modeling of Comp. Sys. (SIGMETRICS '98/PERFORMANCE '98)*, Madison, WI, June, 1998, pp. 151–60.
[13] P. Barford *et al.*, "Changes in Web Client Access Patterns: Characteristics and Caching Implications," *World Wide Web*, Special Issue on Characterization and Performance Evaluation, vol. 2, no. 1, Jan. 1999, pp. 15–28.
[14] B. Duska, D. Marwood, and M. Feeley, "The Measured Access Characteristics of World-Wide-Web Proxy Caches," *Proc. USENIX Symp. Internet Tech. and Sys.*, Monterey, CA, Dec., 1997, pp. 23–35.
[15] L. Fan *et al.*, "Summary Cache: A Scalable Wide-Area Web Cache Sharing Protocol," *Proc. SIGCOMM '98*, Vancouver, BC, Sept. 1998, pp. 254–65.
[16] NLANR Hierarchical Caching System Usage Statistics, http://www.ircache.net/Cache/Statistics
[17] L. Breslau *et al.*, "Web Caching and Zipf-like Distributions: Evidence and Implications," *Proc. INFOCOM '99*, New York, NY, Mar. 1999, pp. 126–34.
[18] NLANR, Cache Popularity Index, http://www.ircache.net/Cache/Statistics/Popularity-Index
[19] G. Voelker *et al.*, "On the Scale and Performance of Cooperative Web Proxy Caching," *Proc. 17th SOSP*, Kiawah Island, SC, Dec. 1999, pp. 16–31.
[20] F. Douglis *et al.*, "Rate of Change and Other Metrics: A Live Study of the World Wide Web," *Proc. USENIX Symp. Internet Tech. and Sys.*, Monterey, CA, Dec. 1997, pp. 147–58.
[21] S. Williams *et al.*, "Removal Policies in Network Caches for World-Wide Web Documents," *Proc. SIGCOMM '96*, Stanford, CA, Aug. 1996, pp. 293–305.
[22] P. Cao and S. Irani, "Cost-Aware WWW Proxy Caching Algorithms," *Proc. USENIX Symp. Internet Tech. and Sys.*, Monterey, CA, Dec. 1997, pp. 193–206.
[23] C. Maltzahn, K. J. Richardson, and D. Grunwald, "Reducing the Disk I/O of Web Proxy Server Caches," *Proc. 1999 USENIX Annual Tech. Conf.*, Monterey, CA, June 1999, pp. 225–38.
[24] E. P. Markatos *et al.*, "Secondary Storage Management for Web Proxies," *Proc. 2nd USENIX Symp. Internet Tech. and Sys.*, Boulder, CO, Oct. 1999, pp. 93–104.
[25] A. Chankhunthod *et al.*, "A Hierarchical Internet Object Cache," *Proc. 1996 USENIX Tech. Conf.*, San Diego, CA, Jan. 1996, pp. 153–63.

[26] R. Tewari *et al.*, Design Considerations for Distributed Caching on the Internet," *Proc. 19th IEEE Int'l. Conf. Dist. Comp. Sys.*, Austin, TX, May 1999, pp. 273–84.

[27] A. Bestavros and C. Cunha, "A Prefetching Protocol Using Client Speculation for the WWW," Tech. rep. TR-95-011, Boston Univ., CS Dept., Boston, MA, Apr. 1995.

[28] T. Kroeger, D. Long, and J. Mogul, "Exploring the Bounds of Web Latency Reduction from Caching and Prefetching," *Proc. USENIX Symp. Internet Tech. and Sys.*, Monterey, CA, Dec. 1997, pp. 13–22.

[29] T. Loon and V. Bharghavan, "Alleviating the Latency and Bandwidth Problems in WWW Browsing," *Proc. USENIX Symp. Internet Tech. and Sys.*, Monterey, CA, Dec. 1997, pp. 219–30.

[30] A. Bestavros, "Using Speculation to Reduce Server Load and Service Time on the WWW," *Proc. 4th Int'l. Conf. Info. and Knowledge Mgmt*, Baltimore, MD, Nov. 1995.

[31] V. Padmanabhan and J. Mogul, "Using Predictive Prefetching to Improve World Wide Web Latency," *Comp. Commun. Rev.*, vol. 26, no. 3, 1996, pp. 22–36.

[32] J. Pitkow and P. Pirolli, "Mining Longest Repeated Subsequences to Predict World Wide Web Surfing," *Proc. 2nd USENIX Symp. Internet Tech. and Sys.*, Boulder, CO, Oct. 1999, pp. 139–50.

[33] CacheFlow, Inc. Active Web Caching Technology, http://www.cacheflow.com/technology/wp/activecaching.html, 1999.

[34] K. Chinen and S. Yamaguchi, "An Interactive Prefetching Proxy for Improvement of WWW Latency," *Proc 7th Annual Conf. Internet Soc.*, Kuala Lumpur, June, 1997.

[35] C. Maltzahn *et al.*, "On Bandwidth Smoothing," *Proc. 4th Int'l. Web Caching Wksp.*, San Diego, CA, Mar. 1999.

[36] E. Cohen, B. Krishnamurthy, and J. Rexford, "Evaluating Server-Assisted Cache Replacement in the Web," *Proc. Euro. Symp. Algorithms*, Venice, Italy, August, 1998, pp. 307–19.

[37] J. Mogul, "Hinted Caching in the Web," *Proc. 7th ACM SIGOPS Euro. Wksp.*, Connemara, Ireland, Sept. 1996, pp. 103–8.

[38] R. Fielding *et al.*, "Hypertext Transfer Protocol — HTTP/1.1," RFC 2616, HTTP Working Group, June 1999.

[39] G. Banga, F. Douglis, and M. Rabinovich, "Optimistic Deltas for WWW Latency Reduction," *Proc. 1997 USENIX Tech. Conf.*, Anaheim, CA, Jan. 1997, pp. 289–303.

[40] B. Housel and D. Lindquist, "WebExpress: A System for Optimizing Web Browsing in a Wireless Environment," *Proc. 2nd Annual Int'l. Conf. Mobile Comp. and Networking*, Rye, New York, Nov. 1996, pp. 108–16.

[41] J. Hunt *et al.*, "An Empirical Study of Delta Algorithms," *IEEE Soft. Config. and Maint. Wksp.*, Berlin, Germany, Mar. 1996.

[42] M. C. Chan and T. Woo, "Cache-based Compaction: A New Technique for Optimizing Web Transfer," *Proc. IEEE INFOCOM '99*, New York, NY, Mar. 1999, pp. 117–25.

[43] T. Welch, "A Technique for High Performance Data Compression," *IEEE Comp.*, vol. 17, no. 6, 1984, pp. 8–19.

[44] F. Douglis, A. Haro, and M. Rabinovich, "HPP: HTML Macro-Preprocessing to Support Dynamic Document Caching," *Proc. USENIX Symp. Internet Tech. and Sys.*, Monterey, CA, Dec. 1997, pp. 83–94.

[45] R. Rivest, "The MD5 Message-Digest Algorithm," RFC 1321, Network Working Group, Apr. 1992.

[46] M. Robshaw, "On Recent Results for MD2, MD4 and MD5," RSA Labs bulletin, vol. 4, no. 12, Nov. 12, 1996, pp. 1–6.

[47] National Institute of Standards and Technology, Secure Hash Standard, FIPS Pub. 180-1, U.S. Dept. of Commerce, Apr. 1995, http://csrc.nist.gov/fips/fip180-1.txt

[48] B. Preneel, A. Bosselaers, and H. Dobbertin, "The Cryptographic Hash Function RIPEMD-160," *CryptoBytes*, vol. 3, no. 2, Autumn 1997, pp. 9–14.

[49] A. van Hoff *et al.*, "The HTTP Distribution and Replication Protocol," Technical Report NOTE-DRP, W3C, Aug. 1997, http://www.w3.org/TR/NOTE-drp-19970825.html

[50] J. Mogul and A. van Hoff, "Duplicate Suppression in http," Internet-Draft draft-mogul-http-dupsup-00, IETF, Apr. 1998; work in progress.

[51] Inktomi Corporation, "Inktomi Traffic Server Product Info," http://www.inktomi.com/products/network/traffic/product.html, 1999.

[52] J. Santos and D. Wetherall, "Increasing Effective Link Bandwidth by Suppressing Replicated Data," *Proc. USENIX 1998 Annual Tech. Conf.*, New Orleans, LA, June 1998, pp. 213–24.

[53] Y. Goland *et al.*, "HTTP Extensions for Distributed Authoring — WEBDAV," RFC 2518, IETF, Feb. 1999.

[54] J. Mogul, "A Trace-Based Analysis of Duplicate Suppression in http," Research rep. 99/2, Compaq Comp. Corp. Western Research Lab., Nov. 1999, http://www.research.digital.com/wrl/techreports/abstracts/99.2.html

[55] P. Cao, J. Zhang, and K. Beach, "Active Cache: Caching Dynamic Contents on the Web," *Proc. Middleware '98*, Lake District, England, Sept. 1998, pp. 373–88.

[56] B. Krishnamurthy and C. E. Wills, "Study of Piggyback Cache Validation for Proxy Caches in the World Wide Web," *Proc. USENIX Symp. Internet Tech. and Sys.*, Monterey, CA, Dec. 1997, pp. 1–12.

Biography

JEFFREY C. MOGUL (mogul@pa.dec.com) received an S.B. from the Massachusetts Institute of Technology in 1979, an M.S. from Stanford University in 1980, and his Ph.D. from the Stanford University Computer Science Department in 1986. He has been an active participant in the Internet community, and is the author or co-author of several Internet standards; most recently, he has contributed extensively to the HTTP/1.1 specification. Since 1986, he has been a researcher at the Compaq (formerly Digital) Western Research Laboratory, working on network and operating systems issues for high-performance computer systems, and on improving performance of the Internet and the World Wide Web. He is a member of ACM, Sigma Xi, and CPSR, and was Program Committee Chair for the Winter 1994 USENIX Technical Conference, and for the IEEE TCOS Sixth Workshop on Hot Topics in Operating Systems.