

# Enabling Network Caching of Dynamic Web Objects

Pankaj K. Garg<sup>1</sup>, Kave Eshghi<sup>1</sup>, Thomas Gschwind<sup>2</sup>, Boudewijn Haverkort<sup>3</sup>,  
and Katinka Wolter<sup>4</sup>

<sup>1</sup> HP Labs, Palo Alto, USA

{garg | keshghi}@hpl.hp.com

<sup>2</sup> Technische Universität Wien, Austria

tom@infosys.tuwien.ac.at

<sup>3</sup> RWTH Aachen, Department of Computer Science, Germany

haverkort@lvs.informatik.rwth-aachen.de

<sup>4</sup> Technische Universität Berlin, Germany

katinka@cs.tu-berlin.de

**Abstract.** The World Wide Web is an important infrastructure for enabling modern information-rich applications. Businesses can lose value due to lack of timely employee communication, poor employee coordination, or poor brand image with slow or unresponsive web applications. In this paper, we analyze the responsiveness of an Intranet web application, i.e., an application within the corporate firewalls. Using a new Web monitoring tool called WebMon, we found, contrary to our initial expectations, substantial variations in the responsiveness for different users of the Intranet Web application. As in the Internet, traditional caching approaches could improve the responsiveness of the Intranet web-application, as far as static objects are concerned. We provide a solution to enable network caching of *dynamic* web objects, which ordinarily would not be cached by clients and proxies. Overall, our solution significantly improved the performance of the web application and reduced the variance in the response times by three orders of magnitude. Our cache enabling architecture can be used in other web applications.

## 1 Introduction

Since its inception in the early 1990's, the World Wide Web has evolved into being the most widespread infrastructure for Internet-based information systems. As with traditional information systems, the response time (measured starting from the user's initiation of a request and ending when the complete response is returned to the user) for individual interactions becomes critical in providing the right Quality of Experience (QoE) for a web user. A fast, responsive application gives the impression of a well-designed and well-managed business and web service. A slow, un-responsive application, on the other hand, can suggest poor quality of business and may send users seeking alternative web sites or services [1].

Unlike traditional information systems, however, not all components required to accomplish a web transaction are necessarily under the control of the organization running the web application. Typically, end-users access a web application with a web browser, through a collection of proxies or caches, which ultimately send the request to a back-end web server farm with application servers and databases. Several networks participate in the intervening layers. Such *organizational heterogeneity*, in addition to the *component heterogeneity*, complicates the designs for responsiveness of web applications.

Caching of web objects has emerged as key enabler for achieving fast, responsive web applications [9]. The main idea is to maintain copies of web objects close to the end-user's browser and serve requests from caches to improve latency and reduce effects of queueing in network elements along the path from the browser to the server. Web browsers maintain cached copies of frequently requested objects. Similarly, large organizations or Internet Service Providers (ISPs) maintain caches of web objects frequently requested and shared by their user community.

To fully utilize the benefits on the Internet or Intranet, effective caching requires web objects available as **static** HTML pages. This means that the HTML page for that web object was not generated when a user requested the object, but was available as a stored HTML file. This is important because the intervening caches, e.g., the browser's cache, will normally not cache **dynamic** content but it will cache static content. Approaches for enabling caching of dynamic content through HTTP headers, e.g., with the **Expires** header, although a step in the right direction, are not universally and uniformly honored by intervening caches.

For performance reasons, web site administrators are well advised to turn dynamic content into static content whenever necessary and possible, to avoid the above problems.

In this paper we provide a framework for: (1) monitoring web applications performance to determine the need for a higher static-to-dynamic web-object ratio, and (2) an architecture to turn dynamic content into static content based on existing file systems and dependency tracking tools, e.g., using the well-known **make** facility [4]. Our approach relies on the fact that some dynamic objects do not change that often, and the data required to construct them is available in files rather than in a database.

We use WebMon [6] to monitor a web application's performance, which determine the requirements for enabling network caching of dynamic content. WebMon monitors the end-user perceived response time for web transactions and reports to a central server. At the same time, WebMon correlates client perceived response times with the server response time for each transaction. Statistical analysis of such client and server response times will help in understanding the requirements with respect to caching. Once a caching solution has been implemented, WebMon can again be used to quantify the benefits of caching.

When we turn dynamic content into static content, the user's request does not usually come all the way to the web server, but may be served by an intermediate cache. Although good for performance reasons, this is a drawback if the web server's administrator would have liked to use that *server hit* to count accesses

to the web server. WebMon has a component that monitors every access to a given URL, even though the corresponding request may not come all the way to the web server. With a minor modification, this mechanism also serves for providing an accurate hit count.

In this paper we describe our framework for enabling caching of dynamic objects. We first describe the WebMon tool in Section 2. We describe the analysis of a sample application's response time in Section 3. In Section 4 we describe an architecture for converting most of the dynamic objects of the sample application into static objects. We monitored the altered application and report the results in Section 5. We conclude in Section 6 with a summary and guidelines for future work. An extended version of this paper appears as [5].

## 2 WebMon: Web Transaction Monitoring

WebMon is a performance monitoring tool for World Wide Web transactions. Typically, a web transaction starts at a Web browser and flows through several *components*: the Internet, a web server, an application server, and finally a database. WebMon monitors and correlates the response times at the browser, web server, and the application server [6].

WebMon relies on a sensor-collector architecture. Sensors are typically shared libraries or script components that intercept the actual processing of a transaction request. Each sensor is logically composed of two parts: the *start* part and the *end* part. The start part of a sensor performs data correlation, whereas the end part forwards monitored data to a collector that gathers and further correlates the data. Since the browser is typically behind a firewall, we use a surrogate sensor web server to forward the browser's data to the collector. This web server can be the same as the original web server that is serving the web pages, or a different one. The collector makes the data available in a file, database, or a measurement server, for further processing.

Browser WebMon sensors are built using JavaScript code sent with each instrumented page. The instrumented page instructs the browser to inject the start and end parts of sensors as the event handlers for the appropriate events for the browser.

The start part of the browser sensor generates a request identifier, and passes the identifier along with its request to the web server. The end part of the browser sensor sends its measurement to the WebMon collector.

The start part of the web server sensor extracts the identifier from the request, and passes the identifier along to the application server. The end part of the web server WebMon sensor sends its measurement along with the request to the collector. Similarly, the application server sensor processes the request and sends its measurement data to the collector. The collector correlates the data received by the individual components on the basis of the unique identifier associated with each transaction. In the prototype we use a single collector, although multiple servers tailored for different deployment architectures are feasible.

### 3 Response Time Analysis

WebMon enables transactional analysis of web applications from a variety of perspectives. In this paper we are mostly interested in the *response time* perceived by the clients for any **transaction** with the application. A transaction *starts* when the user clicks on a URL, or types a URL in the browser window. The transaction *ends* when the user’s browser has finished displaying the complete page associated with that URL.

An application usually offers multiple transactions to its users. In addition to an overall transactional view, WebMon enables transactional segregation that can be used to identify poor performing transactions. Similarly, users of applications come from a variety of network *client nodes*. Accordingly, we analyze WebMon data to cluster and identify user locations that are poorly performing. Such views of the data determine opportunities for improving the web application.

**Table 1.** Sample mean, median, variance and squared coefficient of variation of the client and server response times.

	mean (ms)	median (ms)	variance ( $ms^2$ )	$CV^2$
<b>clientTime</b>	24 902.9	1993	$1.16413 \cdot 10^{11}$	187.7
<b>serverTime</b>	917.413	721	$3.86337 \cdot 10^5$	0.46

In Table 1, we list some statistics for the client and server response times of our sample application, Corporate Source [3]. Corporate Source is an *Intranet* application that enables sharing of source code among HP employees. Corporate Source is a two-tiered web application with a browser accessing web pages using the Apache web server, CGI scripts, and Server Side Include pages.

With respect to the `clientTime`, we observe that its mean value is very large, in fact, much larger than its median. This indicates that there are measurements that are very large, so large, that the `clientTime` distribution is skewed to the right. This is also confirmed by the extremely high variance and squared coefficient of variation. We attempted to fit a Pareto distribution to these measurements and found that a reasonable fit could only be obtained when the five largest observations are removed, thus implying that these “outliers” have an extreme impact on the observed mean. In the near future we will consider a fit to a suitably chosen hyperexponential distribution using the EM algorithm, cf. [8].

For the `serverTime`, we observe a much better behavior in the sense that the mean, which is much lower, is closer to the median, and the variance is very moderate as well. The squared coefficient of variation is only 0.5, indicating that the `serverTime` is rather deterministic.

By ordering the response times on client IP numbers, we found a dependence of the response times on client IP address. Combining the statistical data from

table 1 and the fact that the location of a client is a factor for the perceived performance, we conclude that the network, i.e., the Intranet, is responsible for the large variations and large absolute values of the user-perceived performance. The serverTime is always relatively small, and rather deterministic, so it cannot be the cause of the bad performance perceived at the clients.

**Remark.** At first instance, we were surprised to see the Intranet performance as bottleneck; we had expected the Intranet to show less variable and lower response times. In the case study at hand, however, it seems that the Intranet is behaving more or less like the ordinary Internet. An explanation for this might be the fact that HP is a global company with offices all over the world, and hence, the Intranet is virtually as widespread as the Internet.

In order to improve the user-perceived performance, we therefore propose to move the objects requested closer to the clients, in order to avoid, as much as possible, the use of the Intranet. This can be achieved by increasing the fraction of web-objects that are cached at the clients. Since many of the objects in the study at hand are dynamic objects, we have to devise a means to enable caching of such dynamic objects. This will be discussed in the next section.

## 4 Enabling Network Caching of Dynamic Objects

To enable caching of some dynamic web objects from the Corporate Source application, we have to re-engineer the application. We describe this process in Section 4.1. In such re-engineering, determining the appropriate *Cache Update Policies* is quite important. We discuss cache update policies in Section 4.2. Implementation issues are then discussed in Section 4.3, where the issue of counting accesses is given special treatment in Section 4.4.

### 4.1 Re-engineering “Corporate Source”

The Corporate Source application uses two main file-based data sources: (1) the meta-data source (`/usr/local/hpcs`), and (2) the source code, with version history (`/usr/local/cvsroot`).

The application utilizes the two main dynamic page generation capabilities of the Apache web server: (1) server-side includes (`.shtml` pages), and (2) Common Gateway Interface (CGI) scripts. Out of the nine transactions for the Corporate Source application, four are server-side include pages, and five are CGI scripts. Hence, for the entire application, **none** of the pages are cacheable! This fact explains, at least to a certain extent, the measured response times as shown previously: all requests have to be served over the Intranet, over which many other HP network applications run as well.

To improve the response times for this application, we must transform the service from one that is providing non-cacheable web pages to one that is providing cacheable content. The main change is that instead of directly supplying the pages from the server-side includes or CGI, the web server

supplies a set of HTML pages. The HTML pages, in turn, are periodically updated from the CGI and server-side include pages. Hence, for the web browser, *the web server behaves like a cache rather than a server generating dynamic pages*. Periodically, on requests from the Cache Generator, it behaves like a web server generating HTML pages from dynamic web objects. This is similar to the idea of a *reverse-proxy*, as used in several web servers, e.g., see <http://developer.netscape.com/docs/manuals/proxy/adminmt/revpxy.htm>. The effect, however, is quite different, since we *enable the caching of the web pages by intervening network entities*, whereas “reverse-proxy’ing” does not achieve this. Note that for our approach to operate correctly, all original HTML links to dynamic objects have to be converted to links to their static counterparts.

## 4.2 Cache Update Policies

The re-engineering process involves creating a cache of the dynamically generated web pages, and updating these cached pages as needed. For this purpose, it is useful to classify web pages as follows:

- **on-demand** pages are generated on demand; they are dynamic pages;
- **on-change** pages are generated when any underlying content for the page changes;
- **periodic** pages are generated periodically.

In the case of on-demand pages, the information given to the user is always up-to-date, but the disadvantage is possibly poor response time for the end-user. For on-change pages, the users may obtain up-to-date pages, but the machinery required to monitor all possible changes is complex and may be expensive for the web server, e.g., see Challenger et al. [2]. For pages that are updated on a periodic basis, the users may sometimes receive stale information, but the machinery and effort required by the web server are not expensive. Hence, we decided to implement the periodic update policy.

To see how good the periodic update policy is, we consider the following analytical model. Define: the object request rate ( $\lambda$ ), the object change rate ( $\tau$ ), and the cached object update rate ( $\gamma$ ). We *assume* that both the process of object requests and the process of object changes are Poisson processes. Furthermore, due to the periodic update policy with rate  $\gamma$ , the inter-update time is a constant, denoted  $D = 1/\gamma$ .

We can now view the time-line as partitioned by updates, all exactly  $D$  time units apart. Consider the “arrival” of an object request (as part of a Poisson process). The time that has passed since the last periodic update is a random variable  $Y$ , with distribution

$$F_Y(y) = \frac{1 - F_X(y)}{E[X]},$$

with  $F_X$  the inter-update time distribution and  $E[X] = D$  the mean inter-update time [7, page 102].  $F_X(x)$  is a unit step-function at  $x = D$ . Accordingly,  $Y$  is

uniformly distributed on  $[0, D]$ . This is intuitively appealing; the page change occurs truly randomly in an update interval.

Similarly, we can compute the distribution of  $Z$ , the time until the most recent page change. Since the page changes form a Poisson process as well, the time between page changes is a negative exponentially distributed random variable, which is, by definition, memoryless. Hence,  $Z$  has a negative exponential distribution with rate  $\tau$  as well.

The probability that the request for the page will result in an up-to-date copy, then equals the probability  $\Pr\{Y < Z\}$ , that is, the probability that the last page update took place more recently than the last page change. Conditioning on the time  $Y$  until the last update instance, we obtain:

$$\begin{aligned} \Pr\{\text{“okay”}\} &= \Pr\{Y < Z\} = \int_0^D \Pr\{Y < Z|Y = t\} \Pr\{Y = t\} dt \\ &= \int_0^D \frac{e^{-\tau t}}{D} dt = \frac{1}{\tau D} (1 - e^{-\tau D}). \end{aligned} \tag{1}$$

One can easily show the following properties for this probability, which we for convenience write as  $p(x)$ , with  $x = \tau D \geq 0$ . The following limits are obeyed by  $p(x)$ :  $\lim_{x \rightarrow 0} p(x) = 1$  and  $\lim_{x \rightarrow \infty} p(x) = 0$ . Furthermore, the derivative  $p'(x) < 0$  for  $x > 0$ , so that  $p(x) \in [0, 1]$ , and it decreases monotonously from 1 to 0, for  $x \geq 0$ .

As an example, consider the case where the update rate  $\gamma = 10$  per month (every third day an update;  $D = 0.1$ ) and the change rate  $\tau = 1$  (per month). The probability that a request will result in the right (most recent) copy of the page then equals  $p(0.1) = 10(1 - e^{-0.1}) = 0.95$ . If we set  $\gamma = 30$  (daily updates;  $D = \frac{1}{30}$ ) we obtain success probability 0.984. Notice that the current model does not have  $\lambda$  as one of its parameters. The probability  $p(x)$  holds for every request that arrives as part of a Poisson process.

### 4.3 Implementation Considerations

To implement the periodic updates, we re-engineered the web application to use static HTML pages instead of server-side includes and CGI scripts. We first created a sub-directory “Cache” that contains all the static pages offered by the web server. A Unix “cron” job periodically accesses the CGI and server-side include web pages and generates the static pages for the web server. In the following, we use an example to illustrate the method of converting dynamic pages to static pages.

One of the transactions provided by the Corporate Source application is to describe the details of a given software project, using the meta-data stored about the project. This meta-data is stored as an XML file under the directory:

```
META-DATA/<project-id>/<project-id>.xml
```

To show the meta-data, a CGI PERL script parses the XML file and generates the HTML code for the user’s browser to display. The `<project-id>` is a

parameter to the script, through the fat URL method. Hence, the URL to ask information for the project “hpcs” would be:

```
http://src.hpl.hp.com/cgi-bin/sdata.cgi?software=hpcs
```

To develop an effective cache for such web pages, we need a script that converts the XML files into the required HTML files. This script should run only when data in the META-DATA directory has changed. Also, this script should run for all projects submitted to Corporate Source, even projects that may not have been submitted when the script was first written, i.e., we cannot hard-code the projects in the script.

An elegant solution for such a script can be developed using the Unix `make` utility [4]. `make` is a utility that takes a specification (usually called `makefile`) of dependencies among source and derived files, and when executed with the right specification (re-)produces all derived files for which the sources have changed. The user specifies the tools used for transforming the source to the derived files. For our case, we know the source files as the XML files that contain the meta-data. The derived files are the HTML files to be generated. We can use the existing web server and its CGI script as the tools to transform the XML to HTML.

#### 4.4 Counting Accesses

As we will show in the next section, the redesign of the Corporate Source application does significantly improve the user-perceived performance. It does, however, create one problem: in the new architecture, some of the user requests may not make it anymore to the actual web server, so that we do not know anymore how many users have viewed a particular software project, or how many have downloaded it. With the dynamic pages it was easy to obtain this information from the web server logs. This seems to be a general problem when a dynamic page is changed to a static, cacheable page.

To overcome this problem, we again used the WebMon tool [6]. In order to send browser response times back to a measurement data collector, WebMon introduces an extra HTTP request back to a web server that forwards the measured data to the measurement collector. This additional web server can be another web server, on a different network, in case performance of the original web server is an issue. In our case, this additional WebMon request can be used to log accesses to the server, and hence give us the count of dynamic web page access.

To achieve this, we must modify the original WebMon client side instrumentation to include the URL of the previous request in the data collection request.

## 5 Analysis of the Data with Cached Pages

We adapted the Corporate Source application as described above, and subsequently monitored it for about two weeks.

From the new data set of 521 observations, we computed the sample mean, median, variance and squared coefficient of variation, as given in Table 2. Note that the serverTime could not be monitored for the cached pages, simply because request to these pages do not make it to the server. The serverTime has been observed for only 166 of the 521 entries of the data set.

The serverTime has an extremely small coefficient of variation, which means that there is hardly any variability observed. An Erlang- $k$  distribution with  $k \approx 5$  could be used to describe the data. We also observe that the serverTime has decreased (the mean serverTime has decreased to 69% of the former value, the median to 74%), most probably due to the decreased load on the server.

**Table 2.** New sample mean, median, variance and squared coefficient of variation for serverTime and clientTime

	mean (ms)	median (ms)	variance ( $ms^2$ )	$CV^2$
<b>client</b>	4524.3486	1121	$5.209 \cdot 10^8$	25.4
<b>client-15</b>	1081.2	1549.061	$5.3033 \cdot 10^6$	2.21
<b>server</b>	638.988	539	81219.2	0.2

The mean clientTime is reduced to only 20% of its former value, however, the variance of the clientTime is reduced by three orders of magnitude (which corresponds to a reduction of roughly 99.5%). The coefficient of variation is still large, but considerably smaller than before. The same is true for the variance. Also observe that the mean-to-median ratio (previously equal to 12.5) has reduced to only 4.0; the clientTime distribution is considerably less skewed now.

Interestingly, if we drop the 15 largest observations for the clientTime, the variance is reduced by another two orders of magnitude and end up with a “normal” coefficient of variation. This shows that a few extremely long response times cause these unusual statistics; the majority of the observations is much more moderate. More advanced statistical methods are required to filter these outliers.

## 6 Conclusions

In this paper, we have proposed and evaluated a performance enhancement of an Intranet web application, resulting, for the case study at hand, in a mean response time reduction of 80% and in a response time variance reduction of three orders of magnitude. The reengineering of the web application was motivated by the insights gained with the novel tool, WebMon, that can monitor end-user perceived response times for web applications and correlate them with server response times. WebMon helped us in determining the performance bottleneck at hand, i.e., the Intranet. In the subsequent reengineering process of the web application studied, we took measures to reduce the load on the Intranet. In

particular, we proposed a method which transforms dynamically generated web pages (that cannot be cached) into static web pages. Thus, this transformation enables caching of web objects at the client, and therefore reduced network traffic and load.

Our approach relies only on software tools that are widely available (like the Unix `make` facility), so that it can be used by other web application developers for improving the responsiveness of their web applications.

Although useful for a large number of web applications with dynamic objects, our approach is not directly suitable for web applications with frequently changing objects, such as commonly found in personalized e-commerce shopping web applications and news suppliers. In the future, we will address the performance issues for such web applications as well.

## References

1. N. Bhatti, A. Bouch, and A. Kuchinsky. Integrating User-Perceived Quality into Web Server Design. Technical Report HPL-2000-3, Hewlett-Packard Labs, Palo Alto, CA., January 2000.
2. J. Challenger, P. Dantzig, and A. Iyengar. A Scalable System for Consistently Caching Dynamic Web Data. In *Proceedings of the 18th Annual Joint Conference of the IEEE Computer and Communications Societies*, New York, New York, 1999.
3. J. Dinkelacker and P. Garg. Corporate Source: Applying Open Source Concepts to a Corporate Environment: (Position Paper). In *Proceedings of the 1<sup>st</sup> ICSE International Workshop on Open Source Software Engineering*, Toronto, Canada., May 2001.
4. S. Feldman. Make - A Program to Maintain Programs. *Software Practice and Experience*, 9(3):255–265, March 1979.
5. P. K. Garg, K. Eshghi, T. Gschwind, B. Haverkort, and K. Wolter. Enabling Network Caching of Dynamic Web Objects. Technical report, HP Labs, 1501 Page Mill Road, Palo Alto, Ca 94304, 2002. To Appear.
6. T. Gschwind, K. Eshghi, P. Garg, and K. Wurster. Web Transaction Monitoring. Technical Report HPL-2001-62, Hewlett-Packard Laboratories, Palo Alto, Ca, April 2001.
7. B. R. Haverkort. *Performance of Computer Communication Systems*. John Wiley & Sons, 1998.
8. R. El Abdouni Khayari, R. Sadre, and B.R. Haverkort. Fitting world-wide web request traces with the EM-Algorithm. In R. van der Mei and F. Huebner-Szabo de Bucs, editors, *Proceedings of SPIE: Internet Performance and Control of Network Systems II*, volume 4523, pages 211–220, Denver, USA, August 2001.
9. D. Wessels. *Web Caching*. O'Reilly & Associates, Inc., 101 Morris Street, Sebastopol, CA 95472, June 2001.