# WHAT: Web Hosting Analysis Tool

Ludmila Cherkasova, Mohan DeSouza*
Hewlett-Packard Laboratories
1501 Page Mill Road,Palo Alto, CA 94303
e-mail: {cherkasova,mdesouza}@hpl.hp.com

## 1 Introduction

As the Web increasingly becomes a core element of business strategy, the task of hosting web content has become mission critical. Few companies, however, have the resources, money and expertise to build their web site entirely in-house. For this reason, many businesses choose to outsource their Web hosting to Internet Service Providers and some equipment vendors which, according to Forrester Research Inc., can slash costs by 80%.

Although the recent survey revealed over 1 million web servers on the Internet, the number of web sites exceeds this number by several times. The illusion of more web sites existing than actual web servers is created through the use of *virtual servers (hosts).*

The shared Web hosting service is based on this technique. The shared Web hosting market targets small and medium size businesses. The most common purpose of a shared hosting web site is marketing (in other words, it means that most of the documents are static). In this case, many different sites are hosted on the same hardware.

A Web hosting service uses the possibility to create a set of virtual servers on the same server. There are different alternatives to how this can be done. Unix web servers (Netscape and Apache) have the most flexibility in addressing the Web hosting problem. Multiple host (domain) names can be easily assigned to a single IP address. This creates the illusion that each host has its own web server when, in reality, multiple, "logical" hosts share one physical host.

Each virtual server is set-up to write its own access log. This is a very convenient configuration for the hosted sites (customers). The site access logs allow us to analyze incoming traffic to the site both quantitatively and qualitatively. Access logs provide invaluable information on both the most oftenly requested documents and the most active, frequent visitors of the site.

However, this implementation and set-up splits the "whole picture" of web server usage into multiple independent pieces, making it difficult for the service provider to understand and analyze the "aggregate" traffic characteristics.

The situation gets even more complex when a Web hosting infrastructure is based on a web server farm or cluster, used to create a scalable and highly available solution.

## 2 WHAT's Design Approach

Our goal was to develop a tool which characterizes an overall Web hosting service profile and system resource usage in both a quantitative and qualitative way. We have chosen to report information which

*Work has been done while M.DeSouza worked at Hewlett-Packard Labs during the summer of 1999. His current address is University of California,Department of Computer Science, Riverside, CA 92521; e-mail: mdesouza@cs.ucr.edu could be used by a Web Hosting Service Provider to evaluate the current solution and to improve and optimize the relevant components using overall service profile data.

**WHAT** performs an analysis which is entirely based on web server access logs collected from multiple sites hosted on a server (web server farm or cluster). The tool is written in Perl for the Common Log Format, which is the most popular default for web server access logs.

**WHAT** is aiming to provide:

- *service characterization* - a service profile, a comparative analysis of system resource usage by different customers (i.e. by hosted web sites);
- *traffic characterization* - a comprehensive analysis of overall workload with extraction of a few main parameters to characterize it;
- *system requirements characterization* - a related system resource usage analysis, especially memory requirements.

These characteristics provide an insight into the system's resource requirements and the traffic access patterns - the information which is of a special interest to system administrators and service providers.

# 3 Service Characterization

The study [MS97] asserts that the three primary issues that characterize a site are:

- site composition and growth;
- growth in traffic;
- user access patterns.

Our Web hosting site analysis supports this statement too. The monthly growth of the requests rates for different sites differ significantly. While the typical growth for most of the sites is exponential, it takes different times for different sites to double. Some of the sites experience decrease of the traffic rates and actually demonstrate negative growth. User access patterns differ significantly too. For example, some sites have a few, very popular documents or products. The accesses to such sites are heavily skewed: 2% of the documents account for 95% of the site's traffic. In order to design an efficient, high quality Web hosting solution, the specifics of access rates and users' access patterns should be taken into account. The traffic growth/decrease and the users' access patterns' changes should be monitored in order to provision for those changes well in time and in the most efficient way.

**WHAT**'s design and development was driven by the case study of HP Web Hosting Service provided to internal customers. We performed the analysis which covers a four-month period: from April,1999 to July,1999. Originally, in April, the service had 71 hosted sites. By the end of July, the service had 94 hosted web sites. During this period, **WHAT**'s analysis allowed us to monitor and analyze each particular site's traffic contribution to the overall traffic, and the evolution of the whole service by itself.

**WHAT** identifies all the different hosted web sites (from the given collection of web server access logs). For each hosted web site $i$, the tool builds a site profile by evaluating the following characteristics:

- $AR_i$ - the access rates to a customer's content (in *bytes* transferred during the observed period);
- $WS_i$ - the combined size of all the accessed files (in *bytes* during the observed period, so-called "working set");

- $FR_i$ - the table of all accessed files with their frequency (number of times a file was accessed during the observed period) and the files sizes.

We normalize both $AR_i$ and $WS_i$ with respect to $AR$ and $WS$ combined over all the sites in order to identify the percentage contribution of each particular site.

The access rate $AR_i$ gives an approximation of the load to a server provided by the traffic to the site $i$. The working set $WS_i$ characterizes the memory requirements by the site $i$.

These parameters provide a high level characterization of customers (hosted web sites) and their system resource requirements. This characterization is especially useful when it is time to scale the system. It can help to identify whether additional memory is going to be enough, or whether the service provider needs to add a new server. If a new server is added, often the content is going to be partitioned as well. The site profiles help to create a balanced partition with respect to a system's resources, avoiding the "bad" partitions where the "memory hungry" sites are left on one server and the "high load" sites are moved to a new server. **WHAT** provides valuable sites analysis to be used for capacity planning and balancing tasks.

# 4 Traffic Characterization

**WHAT** provides the analysis of the combined traffic to all the sites.

It reports the number of successful requests (code 200), *conditional_get* requests (code 304) and the errors (the rest of the codes). The percentage of *conditional_get* requests often indicates the "reuse" factor for the documents on a server. These are the documents cached somewhere in the Internet by proxy caches. The *conditional_get* request is sent to fetch the document only in case it was modified.

**WHAT** provides the statistics for an average response-file-size (averaged across all successful requests with 200 code). We also build a characterization of the file size distribution. For this purpose, we build a table of all accessed files with their sizes and access frequency information, ordered in increasing order by size. It allows us to build a file size distribution of the request in a style which is similar to SpecWeb96 [SpecWeb96] - the industry standard benchmark for measuring web server performance.

Thus, for example, in our case study (for April), the average request size was 22.7 KB. The average request size (for 30/60/90% of the requests) was 0.8 KB/1.6 KB/4.6 KB respectively. From the data above, one can conclude that a workload (in terms of file size distribution) has a long tail of rarely accessed very large files.

**WHAT** reports a percentage of the files requested only a few times - the files requested less than 2/6/10 times. In our case study, files requested less than 2/6/10 times account for 34.4%/66.0%/75.7% of all the requests. This is another important characterization of traffic which has a close connection to document reuse and gives indication of memory (RAM) efficiency for the analyzed workload. Most likely "onetimers" are the requests served from the disk. This data is helpful in understanding whether performance improvements can be achieved via optimization of the caching or replacement strategy.

Here the *traffic characterization* comes very close to *system requirements characterization.*

# 5 System Requirements Characterization

System requirements are characterized by the combined access rate and working set of all the hosted sites (during the observed period of time). The tool also provides the worst hour access rate analysis.

**WHAT** provides the combined size of "onetimers" (in our case study - this summed up to 388 MB from total of 1042.5 MB). High percentage of "onetimers" and small memory size could cause bad site performance.

In order to characterize the "locality" of overall traffic to the site, we build a table of all accessed files with their sizes and access frequency information, ordered in decreasing order by frequency. **WHAT** provides working sets for 97/98/99% of all requests.

In our case study, 97/98/99% of the requests required 262.6 MB/375.7 MB/600.4 MB of memory respectively. The smaller numbers for 97/98/99% of the working set indicate higher traffic locality: it means that the most frequent requests target smaller set of documents.

# 6 Large Single Sites Analysis

**WHAT** was designed for Web hosting service analysis needs. We realized, however, that its usage can be extended to provide the analysis of large single sites in a very useful way.

Our second case study was analysis of the *www.hp.com* web site.

**WHAT**'s functionality was extended to identify all the first-level directories. First-level directories give direct indication of web site composition. Often, the first-level directories represent different business units or reflect the company products, and the traffic analysis of these directories is of interest to these units.

After that, we performed an analysis similar to the Web hosting service analysis, where the first-level directories were treated as different web sites.

Such an approach to the analysis of large single web sites allows us to outline the site composition as well as determine the percentage of traffic going to the site's different parts. It allows to create accurate "sub-site" profiles in terms of memory usage and load on a system. Such analysis helps observance of the site evolution and the design of more efficient web sites.

# 7 Conclusion

There are several web log analysis tools freely available (Analog [Analog],Webalizer [Webalizer] to name just a few). They give detailed analysis of the most frequent accesses and the user population. This data is useful for business sites to recognize who their customers are and what documents or products get most attention.

However, these tools are lacking the information which is of interest to system administrators and service providers; the information which provides insight into the system's resource requirements and traffic access patterns.

When the site is a collection of different sites created through the use of *virtual servers (hosts)* a new analysis tool is required to understand the site's contributions to overall traffic, as well as the resource requirements imposed by each particular site. Such sites evolve in a special way: since the different sub-sites "live" their different lives. WHAT's analysis helps to observe site evolution and to provision for changes well in time and in the most efficient way.

# 8 References


[Analog] Analog: http://www.statslab.cam.ac.uk/˜ sret1/analog .

[MS97] S.Manley and M.Seltzer: Web Facts and Fantasy. Proceedings of USENIX Symposium on Internet Technologies and Systems, 1997, pp.125-133.

[SpecWeb96] The Workload for the SPECweb96 Benchmark.
URL http://www.specbench.org/osg/web96/workload.html

[Webalizer] Webalizer: http://www.mrunix.net/webalizer/