Analysis Tool for Web Hosting Service Providers

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

Abstract

As the popularity of the Web grows, an increasing number of businesses are wishing to seize the potential market opportunities that it offers. The shared Web hosting service uses the possibility to create a set of virtual servers on the same physical server. Each virtual server is set-up to write its own access log. Such 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.

Web Hosting Analysis Tool (WHAT) aims to provide 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.

Key words: *shared Web hosting, virtual servers, physical server, server access logs, traffic characteristics, system's resource requirements.*

1 Introduction

Demand for Web hosting and e-commerce services continues to grow at a rapid pace. Few companies, however, have the resources, money or 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%. More than two-thirds of all corporate web sites are now hosted (outsourced), according to Forrester Research Inc.

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's 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. This data is useful for business sites to recognize who their customers are and what documents or products get most attention, as well as geographical distribution of their customers and some other business related observations.

Such implementation and set-up, however, 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.

*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

†Work has been done while J.James 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: jobin@cs.ucr.edu

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.

There are several web log analysis tools freely available (Analog [1],Webalizer [5], WebTrends [6] to name just a few). They give detailed data analysis useful for business sites to understand their customers and customers interests. However, these tools lack 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.

Web Hosting Analysis Tool (WHAT) aims to provide a Web hosting service profile and characterize the system's usage specifics and trends.

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 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's version exists 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 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 traffic access patterns - the information which is of special interest to system administrators and service providers.

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 89 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.

We plan to extend **WHAT** with additional functionality for capacity planning and provisioning goals.

We will illustrate the use of tool with analysis of HP Web Hosting Service. To keep the anonymity of the hosted sites and their businesses, we substituted their names with numbers: site 1, site 2, etc.

3 Service Characterization

The study [3] 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 sites' 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 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.

In our analysis below for HP Web hosting service, we assumed that the sites are served via a web cluster with four nodes, i.e. total web server's capacity is 400% both for memory requirements and for load to be distributed. For example, if the access rate for a site is 95%, then, it means, that this site contributes 95% traffic of 400% total traffic for all the sites (served via 4 servers).

WHAT provides the absolute numbers too (i.e. in MBytes). We dont show them here to keep the example and explanations simpler.

Tables below show ten sites with the largest working sets and ten sites with the largest access rate for April, May, June, and July correspondingly. Additionally, we include these sites' access rates (or working sets) for understanding the sites' profile and their impact on total service composition and overall traffic.

April: Web Hosting Service has 71 hosted sites

Site	Largest WS%	AR%	Site	WS%	Largest AR%
62	213.9	40.2	57	56.8	95.6
57	56.8	95.6	20	2.7	46.7
17	14.3	3.43	62	213.9	40.2
42	12.2	10.0	67	2.4	34.2
60	12.1	9.8	51	2.7	28.3
48	10.4	7.1	10	4.5	20.2
41	10.3	12.2	13	5.8	19.5
13	5.8	19.5	50	1.7	14.9
34	5.4	1.5	41	10.3	12.2
47	4.8	2.0	21	2.5	11.4

As we can see, the site 62 accounts for largest working set (213% of total working set for the whole service. Note, that the service is provided by four servers - 400% of total memory to be used), i.e. more than a half of total memory requirement comes from this site. However, the site 62' access rate requirement (the load provided on a system by a traffic which comes to this site) is much lower and accounts only for 40.2% of the total traffic.

This data shows, that there are sites, like site 20, which have a very small working set (2.7% of total) which "attract" as much as 40.2% of the total load on a system. Such sites have small number of extremally "hot" files.

May: Web Hosting Service has 74 hosted sites

Site	Largest WS%	AR%	Site	WS%	Largest AR%
62	135.8	28.4	10	37.7	50.7
57	68.7	47.3	57	68.7	47.3
10	37.7	50.7	20	7.6	43.8
60	19.0	10.7	67	3.2	28.8
31	13.1	22.9	62	135.8	28.4
42	13.0	13.7	51	2.9	23.7
34	9.8	2.7	31	13.1	22.9
48	7.7	6.1	13	7.1	17.7
20	7.6	43.8	21	2.8	14.8
47	7.1	2.0	50	2.0	14.4

Data for May shows that the service' aggregate profile changes: some "old" sites account for less memory and load requirements, while few new sites start to require more of the system resources.

June: Web Hosting Service has 84 hosted sites

Site	Largest WS%	AR%	Site	WS%	Largest AR%
62	136.4	35.4	57	74.5	50.6
57	74.5	50.6	20	4.6	42.7
10	18.6	41.0	10	18.6	41.0
60	14.1	10.4	62	136.4	35.4
13	12.9	25.6	67	2.9	32.0
42	11.3	10.6	51	2.9	27.6
31	11.0	5.2	13	12.9	25.6
48	9.7	7.2	21	2.9	16.4
17	8.7	2.2	1	0.4	15.2
34	7.2	1.4	42	11.3	10.6

Data for June shows further trends in changing proportion of some sites' traffic: the memory and load requirements for sites 10 and 20 continue their steady growth while some other sites disappear from the "leaders" (for example, site 31 accounting for about 23% of traffic during May does not "show up" among the largest access rate sites. At the bottom of the list, we can see site 1 with very small working set (0.4%) but accounting for 15% of load on a server.

July: Web Hosting Service has 89 hosted sites

Site	Largest WS%	AR%	Site	WS%	Largest AR%
10	64.2	43.6	20	8.1	46.1
57	49.4	12.6	10	64.2	43.6
5	38.7	6.3	13	12.4	34.5
34	28.4	5.0	1	0.7	34.1
60	19.6	16.6	21	3.1	17.1
62	15.0	1.9	67	3.4	16.9
78	14.0	5.0	60	19.6	16.6
42	13.9	12.8	37	5.9	14.5
48	12.4	12.0	50	1.4	14.2
31	12.4	2.1	42	13.9	12.8

Data for July shows a clear change of "leaders": sites 10 and 20 become the largest sites with respect to working set and access rate requirements correspondingly. Site 1 with still very small working set (0.7%) accounts now for 34% of the load on a server! Site 62' "contribution" diminishes (compare this site data for April). In general, the whole service profile becomes more balanced: there are no sites with excessively large working sets (memory requirements) or access rates (load on a system).

This data provides a high level characterization of 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.

The sites profiles accumulated on a daily (weekly) basis allow to derive growth trends for those sites. "Combined trends" help to evaluate and, more important, to predict the overall "aggregate" service growth, and do capacity panning and scaling of the underlying infrastructure accordingly.

4 Traffic Characterization

WHAT provides analysis of combined traffic to all sites.

It reports the number of successful requests (code 200), *conditional_get* requests (code 304) and errors (the rest of the codes).

The status code of 200 means that the request was successfully completed by the server. The status code of 304 relates to the documents cached somewhere in the Internet (or by proxy caches) which send a "request-validation" whether the document was *"modified since"* the last requested time, no data bytes need to be transferred in this case. The percentage of *conditional_get* requests often indicates the "reuse" factor for the documents on a server. The rest of the codes specify "unsuccessful" requests, which web server was not able to satisfy (typically, a reason why the response was unsuccessful is given).

The following Table summarize the results for April, May, June and July.

Month	Total Number of Requests	Successful (code 200)	Cond_Get (code 304)	Errors (other codes)
April	1,674,215	51.2%	44.4%	4.4%
May	1,695,774	53.0%	43.6%	3.4%
June	1,805,762	53.1%	43.4%	3.5%
July	1,315,685	60.3%	35.3%	4.4%

Interestingly enough, the traffic characteristics for April, May and June are somewhat similar. July shows clear change in a service profile. Most of additional, new hosted sites do not have much traffic, since they are not yet well known. Some of the old sites show diminishing traffic as well. July exhibits less requests and less of those requests are *conditional_get* (code 304) requests. The possible explanation is that most of new sites have small customer population yet, and most of the requests have to be initially fetched from the web server.

WHAT provides statistics for the 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 requests in a style which is similar to SpecWeb96 [4] - the industry standard benchmark for measuring web server performance.

Month	Average Response Size	Average Response Size for 30/60/90%
April	22.2 KB	0.8 KB / 1.6 KB / 4.5 KB
May	21.5 KB	0.8 KB / 1.7 KB / 4.6 KB
June	22.8 KB	0.8 KB / 1.6 KB / 4.6 KB
July	18.4 KB	0.6 KB / 1.3 KB / 4.2 KB

Average response size for 30/60/90% of all (200 code) requests are surprisingly consistent for all four months. The decrease of the average response

size for the service indicates a shorter "tail" of very large rearely accessed files in the workload.

WHAT reports a percentage of the files requested only a few times - the files requested less than 2/6/10 times.

Month	Files Requested less than 2/6/11 times (as % of all (200 code) Requests)
April	34.4% / 66.0% / 74.6%
May	33.9% / 60.8% / 71.4%
June	36.9% / 62.2% / 69.2%
July	50.3% / 71.0% / 76.6%

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 disk. Additionally, service will not benefit from having these files in RAM. Hence, the "aggregate service (sites)" working set less "onetimers" working set helps to evaluate the needed RAM size for practically optimal configuration.

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).

Month	Total Access Rate	Total Working Set
April	14,859.8 MB	994.2 MB
May	14,658.1 MB	878.4 MB
June	13,909.2 MB	884.9 MB
July	8,713 MB	711.6 MB

WHAT provides the combined size of "onetimers". High percentage of "onetimers" and small memory size could cause bad site performance.

Month	Working Set of "Onetimers" (in MBytes)	Working Set of "Onetimers" (as % of Total WS)
April	370.0 MB	37.2%
May	374.5 MB	42.6%
June	311.2 MB	35.2%
July	298.3 MB	41.9%

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 (200 code) requests.

Month	Working Set for 97/98/99% of all (200 code) Requests (in MBytes)
April	242.7 MB / 362.1 MB / 556.3 MB
May	249.2 MB / 296.3 MB / 419.9 MB
June	196.1 MB / 304.8 MB / 475.1 MB
July	155.1 MB / 276.1 MB / 487.9 MB

The smaller numbers for 97/98/99% of the working set indicate higher traffic locality: this means that the most frequent requests target a smaller set of documents.

Month	Working Set for 97/98/99% of all (200 code) Requests (as % of Total WS)
April	24.4% / 36.4% / 56.0%
May	28.4% / 33.7% / 47.8%
June	22.2% / 34.4% / 53.7%
July	21.8% / 38.8% / 68.6%

As we can see, 97% of the requests target the documents which account only for 21.8%–28.4% of the total fileset.

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. The amount of traffic per day experienced by *www.hp.com* web site was significantly higher (order of magnitude) than the Web hosting service (described above) per month. So, our analysis was done for a single, usual (rather quiet) day in September, 1999. [1]

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.

[1] We do not report the number of total requests to hp.com, as well as the amount of bytes transferred by the site, because it is a business sensitive data.

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

WHAT identified 145 first level directories. To keep the anonymity of the sub-directories and their businesses, we substituted their names with numbers: dir 1, dir 2, etc.

In our analysis below for *www.hp.com* web site, we assumed that this site is served via a web cluster with four nodes, i.e. total web server's capacity is 400% both for memory requirements and for load to be distributed. For example, if the access rate for a sub-directory is 131%, then, it means, that this site contributes 131% traffic of 400% total traffic for all the sites (served via 4 servers).

Table below shows ten sub-directories with largest working sets and ten sub-directories with largest access rate. Additionally, we include these sub-directories access rates (or working sets) data for understanding the *www.hp.com* site' profile and sub-directories impact in total site composition and overall traffic.

Site	Largest WS%	AR%	Site	WS%	Largest AR%
33	174.3	131.8	33	174.3	131.8
138	25.5	4.6	56	1.2	84.1
30	19.0	16.4	122	0.1	19.2
80	18.4	8.7	30	19.0	16.4
6	10.4	3.6	78	6.3	11.2
20	9.6	0.5	39	6.2	11.0
50	8.7	2.1	80	18.4	8.7
103	8.2	6.3	28	2.6	8.5
54	7.	2.3	18	1.6	6.6
120	7.0	3.08	103	8.2	6.3

The analysis shows that sub-directories' profiles on this site are quite different: there is one sub-directory (dir 33) accounting for almost 1/2 of total memory usage and 1/3 of the load (traffic) on a cluster (174.3% and 1131.8% from total of 400% for working set and access rate correspondingly). As another extreme, the dir 138 accounts for 84.1% of load for very small working set (1.2%) consisting of few extremally "hot" pages.

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.

We used **WHAT** to build one-day traffic profile for *www.hp.com* web site. The following Table summarize 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 indicates the "reuse" factor for the documents on a server. These are the documents cached somewhere in the Internet by proxy caches.

Successful (code 200) Responses	Cond_Get (code 304) Responses	Errors (other) codes)
73.4%	23.7%	2.8%

WHAT provides the statistics for an average response-file-size (averaged across all successful requests with 200 code).

Average Response Size(200 code) (in KBytes)	Average Response Size for 30/60/90% of 200 code Requests
4.7 KB	0.6 KB / 0.8 KB / 1.4 KB

Overall average response size and the average response size for 30/60/90% of all the requests on *www.hp.com* web site are much smaller than for HP web hosting service which was analyzed in previous Sections, this indicates more thorough design of the site.

WHAT reports a percentage of the files requested only a few times - the files requested less than 2/6/11 times.

Files Requested less than 2/6/11 times (as % of all (200 code) Requests)
30.7% / 63.8% / 74.7%

This is an important characterization of traffic which gives indication of document reuse and memory (RAM) efficiency for the analyzed workload, as well as helps to identify which web server configuration is most efficient (to optimize performance/cost ratio). Most likely "onetimers" are the requests served from the disk (or evicted from the RAM after some period).

Memory requirements for *www.hp.com* web site is summarized in the following Table.

Total WS (in MBytes)	WS of "Onetimers" (in MBytes)	WS of "Onetimers" (as % of Total WS)
3,001.6 MB	891.2 MB	19.7%

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.

WS for 97/98/99% of all (200 code) Requests
258.4 MB / 527.3 MB / 1031.6 MB

The smaller numbers for 97/98/99% of the working set indicate higher traffic locality: this means that the most frequent requests target a smaller set of documents.

WS for 97/98/99% (as% of Total WS)
8.6% / 17.6% / 34.4%

The locality for *www.hp.com* web site is much higher than for HP web hosting service we analyzed in previous Sections. This indicates very efficient site design.

This completes our one-day analysis of *www.hp.com* web site.

WHAT's approach to the analysis of large single web sites helps 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

Understanding the nature of the web servers' workloads is crucial to properly designing and provisioning current and future web services. Web increasingly becomes a core element of business strategy. The number of public web sites grows exponentially, and the business users account for the majority of that growth.

With the rapid growth of the web traffic, most popular web sites need to scale up their server capacities. Issues of workload analysis, performance modeling and capacity planning become ever more critical.

There are several web log analysis tools freely available (Analog [1],Webalizer [5] 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.

We used **WHAT** as a part of load balancing solution **FLEX** [2] for a shared web hosting service implemented on a cluster of machines [2]. **FLEX** allocates hosted web sites to different machines in the cluster based on the sites' processing and memory requirements which are estimated using the site logs. **FLEX** is not based on static inflexible partitioning and can adapt to gradual changes in sites' traffic patterns.

8 Acknowledgements

We would like to thank Bill Kepner from Firehunter team for first encouraging and very useful discussions. The tool could not be possible without server access logs and help provided by Guy Mathews, Dan Schram, Wai Lam, and Len Weisberg. Their help is highly appreciated.

Shankar Ponnekanti (from Stanford University), who worked with us during the summer on related project, has contributed with discussions and valuable insights.

References

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

[2] L. Cherkasova: FLEX: Design and Management Strategy for Scalable Web Hosting Service. HP Laboratories Report No. HPL-1999-64R1,1999. http://www.hpl.hp.com/techreports/1999/HPL-1999-64R1.html

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

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

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

[6] WebTrends: http://webtrends.com/