# “Living in Barcelona” Li-BCN Workload 2010

Josep Ll. Berral, Ricard Gavaldà, Jordi Torres
Universitat Politècnica de Catalunya and Barcelona Supercomputing Center
Jordi Girona 31, 08034 Barcelona, Spain
{berral,torres}@ac.upc.edu, gavalda@lsi.upc.edu

January 14, 2011

**Abstract**

Nowadays lots of Internet users are clients of web hosting companies, willing to offer their web services, store their content, or just publish their web sites on the network. This has made the hosting companies to use big data-centers or just the Cloud, in order to serve a web server, domain names, disk space and bandwidth to this great demand.

In hosting companies, customers are often big companies or just private users or small business wanting to offer a web service or publish a website. Here we present and detail workloads from a set of different real web sites, of different owners and with different kind of content or offered web services. Some of them are personal or professional web-log sites, also small eCommerce sites, file storage/support sites, and information panel sites.

The presented workload brings pieces of loads, that compared with “killer applications” like Facebook or Google can be considered small, but represent loads and behaviors of the average individual web sites. Also, once studied the data provided by the workload, to reproduce or recreate it, it can be properly scaled in number of users or visit values depending on the levels of stress or the target of the experiments where this data is used.

## 1 Introduction

One of the most usual Internet business hold on data-centers are the hosting services, offering platforms as a service, infrastructures as a service, and services as a service. Companies and individual customers pay for using physical or virtual machines in order to run their jobs, using a pre-installed system and apply their web services, or just using pre-installed services in order to use them somewhat customized.

Here we present a workload obtained from some real individual web sites, owned by different persons and holding different services and contents. These pieces of the workload are today on-line offering content, services and information, as their respective owners use their contracted web space or contracted virtual machines for personal or professional usage. Some of the web sites are personal blogs used to distribute opinion, multimedia content or academic content, others are professional blogs used to share opinions or business information, also others are entertainment web sites such a message boards or news boards. Here is also included websites supporting other websites hosting files and content.

The workload described here, instead of being not from popular overcrowded sites, is an example of the load generated by small and medium companies or individual users. Instead of generating each site lots of visits and requests, each one generates a moderated amount, but summing all that customers with their loads and requirements, some high performance computing platform is required from the data-center hosting them. Also, as an advantage, each web site from the workload has different characteristics, so it can be used for statistical purposes and studies, and more.

The following report exposes ans is structured as follows: Section 2 brings a description of the service where the web sites are hosted, and also the description of the architecture where the data-center relies. Section 3 exposes the statistical details of each web site and relevant data. Section 4 exposes some usage that can be applied over the presented workload. And finally, Section 5 explains some improvements and future work to be done with and for the given workload.

## 2 Service Description

Current web service hosting companies hold in their clouds several clients, each one with the possibility of running their jobs or web services. Each client disposes of one or several web servers like Apache, Tomcat, etc. [1, 2] with the possibility of enabling application service modules like CGI, PHP, Perl or others [6, 7]. Also, each client disposes of a database server, usually MySQL, PostgreSLQ or Oracle [3, 4, 5], with its corresponding web server connection module. Each client is able to deploy over its hosting space his web site, web applications and web services, open to the Internet given a specific URL or domain name.

In order to assure security and the minimal interference between the hosted clients, each one is granted a Virtual Machine for its own. The isolation through virtualization brings to the client the idea of having exclusively all the system resources, while the VM manager distributes the physical machine resources in order to satisfy the demand of the virtual machines, so the client does not notice the effects of virtualization.

One virtual machine can contain one or more web sites or web services, but always under one web server of a kind. I.e. a virtual machine can be equipped with an Apache web server for web pages, a Tomcat for web applications, an Apache PHP module for PHP scripts, and a MySQL database server. This virtual machine can contain web sites of diverse kind, like static web sites or dynamic content web sites using PHP scripts and DB queries, and storage sites, containing images and multimedia resources supporting other websites. Also, each web site can be referred with one and more domain names, or be all grouped in one domain and separated by different URLs.

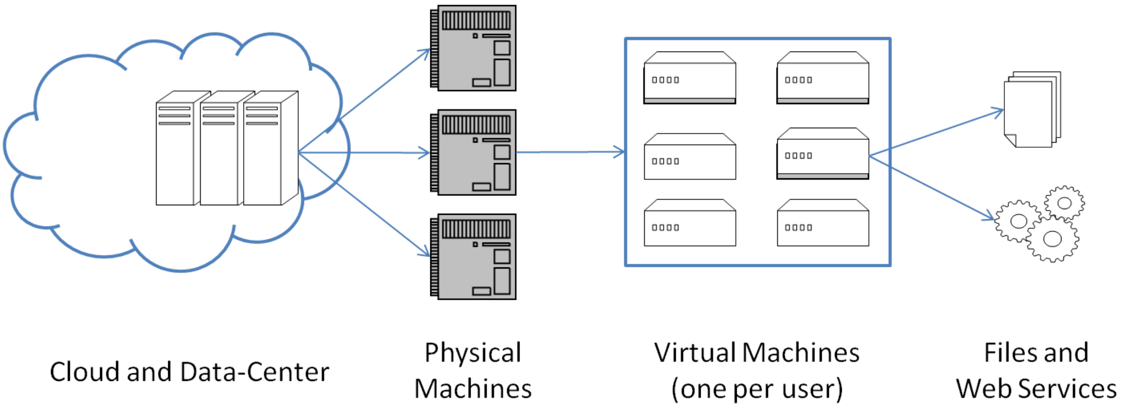

Figure 1: Hosting - Virtual Machines - Web Sites schema

## 3 Workload Description

The workload described in this document correspond to the extraction and analysis of behaviors, visits and received load of different real web sites property of individual customers of a current real hosting company. The names of web sites, companies and users involved in this work have been anonymized according to the request of the owners of the web sites. Although that, a detailed description of the load is provided, so tests with this workload can be reused and also reproduced.

The different web sites providing the information for the workload are from different hosting customers, disposing each an Apache v2.X web server, support modules for CGI, PHP, Perl and Python, and also a MySQL 5.5 DB. The kind of web sites offered are the following detailed ones:

**PHP-based Web-log**: A web-log (blog) web site for personal or professional reasons. These kind of sites are built using some popular web content applications like WordPress, Drupal or Joomla [8, 9, 10]. For the used websites, WordPress is used, a PHP application script using a MySQL database. As this kind of web sites are property of individual persons, most of the requests are just for browsing the site or download content available. The WordPress script can contain modules for catching content and also save DB queries. On the used websites there are 4 blogs: one is a personal blog with also a webcomic blog (containing some more multimedia content), a second one a professional blog, a third one a student blog with academic material and files, and a fourth one a news blog with text content and image files. The web-log scripts are also “search engine friendly”, so they are prepared to indicate to search engines the update periodicity and allow their bots to crawl the content.

**Imageboard Web Site**: A message board web site attaching images to each message. These kind of sites have become quite popular since the apparition of the ”2chan” Japanese imageboard [12]. The web application script of these sites is the popular among them Kusaba Script [11], a script made in PERL that is in charge of display the board and its images, and also post new comments or contributions to the board. This script uses a MySQL database in order to store the content to be displayed and also captcha mechanisms for preventing the intrusion of SPAMbots. As seen from the kind of clients received by these imageboards, the major part of requests from clients are ”GET” requests. The great majority of clients just browse the web site and view the images (passive users), and are few the ones who post new images (active users) in relation with the passive ones. For that reason the script generates, for each post, a cache web page in order to save DB queries.

**Information Web site**: A web site with a single page as a front page, directly displaying information and links to the customers. The used web site is a script page written in PHP displaying information from a database (always the same query), and also a XML file containing RSS feed content for announcing the page updates. The page users just browse the front page or query for the XML RSS feed file, so the load per user is extremely light, and also part of the clients instead of browsing the site just read the updates only from the RSS feed.

**eCommerce Web App**: A small electronic commerce web site. There are several application scripts for building eCommerces, like the one used here OsCommerce [13]. The OsCommerce application script uses PHP and MySQL to run. Like the previous web sites, customers of the eCommerce just browse or “shop”. On our eCommerce web site, minor products are sold, so cannot be compared with great commercial web sites, but we are interested just in the web application more than the business held here.

**File Support Web Site**: A web site containing images and content to be used by partner web sites. These websites usually are storages for images, banners or files used in other web sites, so the load for them is basically bandwidth and disk usage. It also contains some PHP scripts, but great part of the load comes from file transfer.

Tables 1 and 2, and Figure 2 show the load received by each web site, and its kind of load. The data from this workload comprehends 19 consecutive days, starting in Wednesday.

| Description | Web Server | Modules | reqs/day | fail reqs/day | Transf. Bytes/day |
|---|---|---|---|---|---|
| ImageBoard | Apache v2 | Perl, MySQL | 9822 | 278 | 112.60 MBytes |
| PHP Web-log (1) | Apache v2 | PHP, MySQL | 568 | 12 | 14.98 MBytes |
| PHP Web-log (2) | Apache v2 | PHP, MySQL | 79 4 | 2 | 6.95 MBytes |
| PHP Web-log (3) | Apache v2 | PHP, MySQL | 337 | - | - |
| PHP Web-log (4) | Apache v2 | PHP, MySQL | 1,788 | 68 | 83.90 MBytes |
| eCommerce App | Apache v2 | PHP, MySQL | 210 | 178 | 10.00 MBytes |
| File Support Site | Apache v2 | None | 9391 | 36 | 656.38 MBytes |
| Information Web | Apache v2 | PHP, MySQL | 788 | 15 | 13.93 MBytes |

| Description | GET reqs | POST reqs | other reqs | Main Content |
|---|---|---|---|---|
| ImageBoard | 99.89 | 0.03 | 0.08 | Images |
| PHP Web-log (1) | 96.79 | 3.13 | 0.08 | Text+Images |
| PHP Web-log (2) | 98.15 | 1.12 | 0.73 | Text+Files |
| PHP Web-log (3) | 92.99 | 6.98 | 0.03 | Text+Files |
| PHP Web-log (4) | 99.15 | 0.79 | 0.06 | Text+Images |
| eCommerce App | 99.27 | 0.70 | 0.03 | Text+Images |
| File Support Site | 98.10 | 1.83 | 0.07 | Files+Images |
| Information Web | 99.81 | 0.14 | 0.05 | Text |

Table 1: Average web site load details

| | Imageboard | | PHP Web-log (1) | | File Support Site | | Information Site | |
|---|---|---|---|---|---|---|---|---|
| File extension | #Reqs | %bytes | #Reqs | %bytes | #Reqs | %bytes | #Reqs | %bytes |
| .jpg [JPEG graphics] | 148621 | 71.892 | 588 | 8.125 | 138163 | 91.908 | 747 | 10.05 |
| .html [HTML web file] | 7479 | 1.180 | 154 | 0.036 | 121 | 0.001 | - | - |
| .png [PNG graphics] | 7344 | 1.206 | 788 | 20.459 | 6891 | 0.260 | 8495 | 71.195 |
| .pl [Perl scripts] | 225 | 0.004 | 1 | 0.002 | - | - | - | - |
| .css [Cascading Style Sheets] | 5076 | 0.736 | 474 | 0.580 | 557 | 0.008 | 373 | 0.364 |
| .js [JavaScript code] | 4759 | 0.491 | 8 | 0.036 | 396 | 0.002 | - | - |
| .ico | 3125 | 0.046 | 145 | 0.147 | 2697 | 0.748 | 371 | 0.035 |
| . [directories] | 2492 | 0.473 | 2453 | 4.472 | 2121 | 0.036 | 803 | 5.592 |
| .txt [Plain text] | 378 | 0.002 | 407 | 0.048 | 1047 | 0.001 | 216 | 0.032 |
| .php [PHP] | 6300 | 0.047 | 341 | 0.077 | 891 | 0.011 | 136 | 0.035 |
| .gif [GIF graphics] | 86 | 1.286 | 135 | 0.057 | 5822 | 2.206 | 1019 | 1.862 |
| .xml | 3 | 0.052 | 13 | 0.028 | 15 | 0.045 | 2774 | 10.815 |
| . [others] | 7475 | 22.575 | 6160 | 65.92 | 26905 | 4.771 | 17 | 0.001 |

| | PHP Web-log (2) | | PHP Web-log (3) | | PHP Web-log (4) | | eCommerce | |
|---|---|---|---|---|---|---|---|---|
| File extension | #Reqs | %bytes | #Reqs | %bytes | #Reqs | %bytes | #Reqs | %bytes |
| .jpg [JPEG graphics] | 1307 | 48.829 | 309 | 23.585 | 10653 | 92.755 | 71 | 11.377 |
| .html [HTML web file] | - | - | 855 | 1.484 | 1 | 2.060 | 4 | 0.002 |
| .png [PNG graphics] | 3138 | 1.420 | 3 | 0.007 | 198 | 0.004 | 6 | 0.058 |
| .pl [Perl scripts] | - | - | - | - | - | - | - | - |
| .css [Cascading Style Sheets] | 1834 | 3.340 | 734 | 3.767 | 2096 | 0.594 | 201 | 0.523 |
| .js [JavaScript code] | 102 | 0.247 | - | - | 3 | 0.002 | - | - |
| .ico | 71 | 0.017 | 102 | 0.059 | 4925 | 0.336 | 11 | 0.004 |
| . [directories] | 4168 | 27.958 | 1166 | 12.836 | 939 | 0.589 | 579 | 17.068 |
| .txt [Plain text] | 662 | 0.095 | 321 | 0.251 | 696 | 0.014 | 413 | 0.182 |
| .php [PHP] | 153 | 0.133 | 741 | 2.074 | 253 | 0.005 | 1248 | 14.144 |
| .gif [GIF graphics] | 1086 | 1.300 | 479 | 1.539 | 811 | 0.031 | 84 | 0.076 |
| .xml | 36 | 0.022 | 5 | 8.218 | 12 | 0.001 | 1 | 0.336 |
| . [others] | 2580 | 16.633 | 2158 | 46.174 | 14913 | 5.664 | 1718 | 56.224 |

Table 2: File transfer details for each site

# 4 Usage of the Workload

These workloads will be used in our next works towards stress tests over different kinds of machines and data-centers, the observation of behaviors from the same machines and data-centers, and also studies of the information from the workload itself that can be obtained for modeling and statistical usages.

## 4.1 Stress tests over machines

From this described data, one thing to notice is that usually a high performance machine is able to hold not only an average user website but several. As the volume of visits is perfectly supported by a dedicated high performance machine, in order to perform stress tests using this workload it must be scaled in users or scaled in load. Stress tests done with a Intel Xeon QuadCore @ 3Ghz, have shown that multiplying the number of visits of the imageboard web site by a factor of 1500 would bring the hosting physical machine to maximum usage, counting that the web server (Apache v2) is running inside a virtual machine (VirtualBox v3.1) and having the database (MySQL v5.5) in the same VM. Usually flash-crowds or DDoS attacks towards web sites reach and far surpass this factor, so it is reasonable to apply it in order to bring a HPC machine to a stress status.

Obviously, scaling clients would only scale the load for punctual instants, so an instant with no load will scale nothing. For this, the workload can be used scaled and interpolated, not only scaling the users, but making these virtual users or new added virtual users to query asynchronously, filling the time gaps between queries on the real workload. If the recreation uses only statistic parameters (like requests per second, bytes per second, etc.), and also the target is to stress a HPC machine like the described above, this interpolation can be emulated by multiplying the requests by another

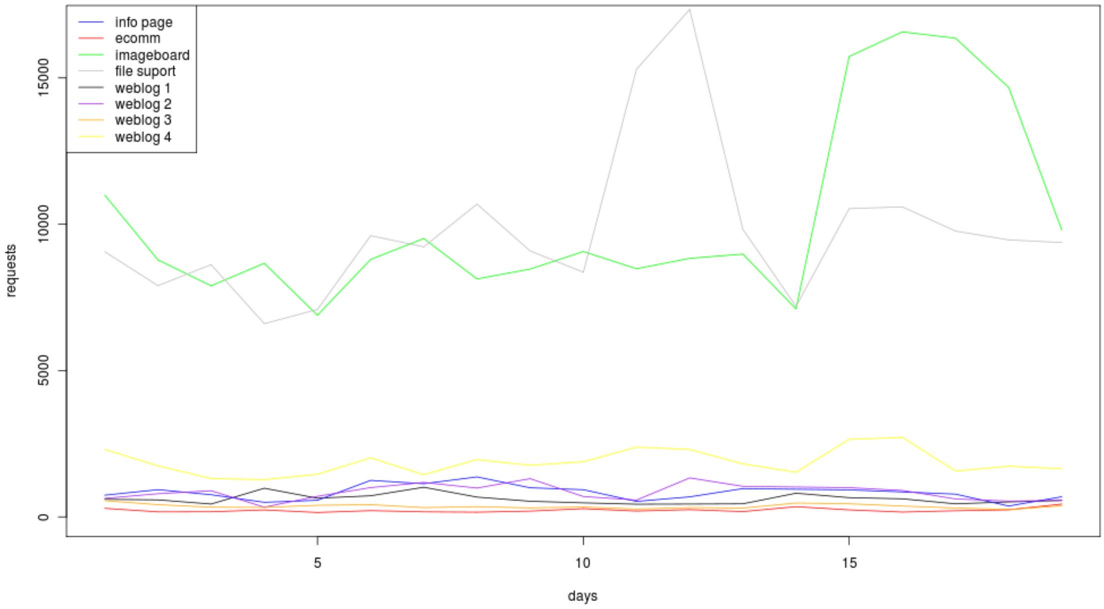

Figure 2: Requests per day of each web site

factor (60 to 600), in order to keep the stress during the workload ”inactivity period”. This factor depends on the selected web site, as some of them receives users (several consecutive requests) each minute or each 10 minutes. Of course, it can be scaled far away this factors in order to saturate the machine whether it is the objective of the workload usage.

These factors and stress tests have been done by recreating the workload using snapshots of the entire web site and database, and the visits registered in the corresponding log-files have been reproduced by attacking the web site while scaling the number of clients attacking, and also separating each client in time by some seconds in order to generate a continuous load on the server-side. This brings new apache logs of the same workload but recreated and scaled.

### 4.2 Data mining usages

Other usage for the workload is found in the field of Data mining. As being a real workload it provides the natural behavior of users visiting web-logs, information web-sites, and also it shows the behavior of pages where the users accede indirectly. From here on, studies on user behavior and web navigation can be performed for statistical purposes. Also, machine learning and other techniques can take profit from the collected workload, as the data is extremely diverse. Next works to do using this workload are mainly focused to data mining and knowledge discovery.

## 5 Next Improvements

Next improvements over this workload will be add new web site logs of different kind not seen here, and try to select new web applications not only from web sites but other web services, preferentially with more “high performance load” per request, more database load per request, and also trying to get specific logs for the database load.

Also, depending on the results that will be obtained and the requirements that we find missing in this workload while performing recreations or data ming, new versions will be made trying to add the necessary information.

# References

[1] The Apache HTTP Server Project, *http://httpd.apache.org/*, as seen in January 2011.

[2] The Apache Tomcat Project, *http://tomcat.apache.org/index.html*, as seen in January 2011.

[3] The MySQL DataBase Project, *http://dev.mysql.com/doc/*, as seen in January 2011.

[4] PostgreSQL, *http://www.postgresql.org/*, as seen in January 2011.

[5] Oracle, *http://www.oracle.com/us/products/database/index.html*, as seen in January 2011.

[6] PHP.net, *http://www.php.net/*, as seen in January 2011.

[7] The PERL programming language, *http://www.perl.org/*, as seen in January 2011.

[8] Wordpress.org, *http://wordpress.org/about/*, as seen in January 2011.

[9] Joomla CMS, *http://www.joomla.org/about-joomla.html*, as seen in January 2011.

[10] Drupal, *http://drupal.org/about*, as seen in January 2011.

[11] ImageBoard Kusaba Script, *http://kusabax.cultnet.net/wiki/*, as seen in January 2011.

[12] MessageBoard 2chan, *http://www.2chan.net/*, as seen in January 2011.

[13] Online Shop eCommerce, *http://www.oscommerce.com/*, as seen in January 2011.