An analysis of World-Wide Web Proxy Cache performance and its application to the modelling and simulation of network traffic.

David Barnes and Neil Smith
The Computing Laboratory,
The University
Canterbury
Kent CT2 7NF
England.
Email: N.G.Smith@ukc.ac.uk

Abstract

Previous studies of World-Wide Web traffic patterns have, quite necessarily, been limited in their scope. Without exception they have had to choose to represent either a client's [1] or a server's [2,3] point of view, or solicit user responses in interactive surveys [4,5]. In all of these cases the methodology is able to provide neither a complete nor an impartial picture of the activities being carried out on the Web. The unique position of the HENSA Unix proxy cache, sitting between large numbers of clients and servers, means that over the past two years we have been able to monitor and analyse several hundreds of millions of transactions on the World-Wide Web. In this paper we present our findings as far as traffic trends are concerned. These trends show both the growth of the Web and typical characteristics of a large volume of Web traffic.

The configuration of the HENSA cache has changed significantly over the two years of its existence, and one of the purposes of this study is to better understand the demands placed upon large proxy caches, in an effort to simulate and model more effective cache configurations.

1 Introduction

The results presented in this paper are of a preliminary nature and we are in the process of performing further, more detailed, analyses on the enormous quantity of data which has been collected during the HENSA proxy cache project. However, we believe that the dissemination of this information now will be useful in influencing future development decisions on the World-Wide Web.

All the information that we present here has been generated from two years worth of data collected in the logs of the HENSA Unix proxy cache which serves the '.ac.uk' community. Currently this service is delivering more than 1.1 million documents each day. This typically accounts for more than 15GB of data delivered to clients and the generation of more than 160MB of raw log information each day.

We have divided this paper into a number of sections, each of which examines a particular aspect of the development of traffic on the Web. The first of these looks at the growth in the number of servers publishing information, this is followed by an examination of the number of individual documents on the Web, and an analysis of their size and media types. We go on to look at the growth in the client population, and the trends in the use of different protocols on the Web.

Our ultimate aim in performing these analyses is to build a reliable model of traffic flow on the Web. We hope that this model will then allow us to parameterise and simulate this flow. We will go on to use these simulation models to test the performance of proxy caches in different configurations. We are striving for configurations that give efficient use of bandwidth with acceptable document refresh times.

We conclude with a discussion of the experience that we have gained so far through this work, and with a look forward to further work that needs to be carried out.

2 History

The HENSA Unix Public Web Cache [6] has existed in one form or another since November 1993. A Web cache aims to save bandwidth by acting as a proxy for a large number of Web browsers. Each page requested by a client is kept in a disk based cache on the proxy. As the number of browsers making use of the proxy increases so does the chance that the cache will have a copy of a page requested as a result of a previous request. In the simplest case this page can be served straight from the cache, with no international connection required at all.

The HENSA Web Cache was installed, experimentally, in reaction to the poor international connectivity seen in the United Kingdom, and the resulting poor performance of the Web. At the time, the UK had a single 2Mbps link to the United States. The 150Mbps national backbone allowed great demand to be placed on this link. The situation is now a little better, and the UK's current international links include 4Mbps to the US, 4MBps to Europe and 2Mbps to Scandinavia.

The significant bandwidth savings and latency reduction that the service can offer has meant that demand has always exceeded the service's ability to meet it. Recently, however, the service has been accepted as the UK National Web Cache and received equipment sufficient to meet the demand (at least for a short time to come).

3 Number of servers

When analysing traffic flow and determining the effectiveness that the caching of Web documents may have, it is important to know how the Web is growing over time. As the population of Web servers grows it is reasonable to assume that the client population will be accessing a larger and larger number of servers. This thinner spread of access may result in a cache of fixed size gradually becoming less and less effective.

The first area studied was the number of distinct Web servers that the cache visited as a result of user requests. Simple counts of servers were generated over periods of a week with the results being shown in Figure 1. A seasonal trend, very similar to that seen later for client accesses, is immediately apparent, with the number of servers visited remaining stable over the summer months and climbing sharply with the start of the academic year.

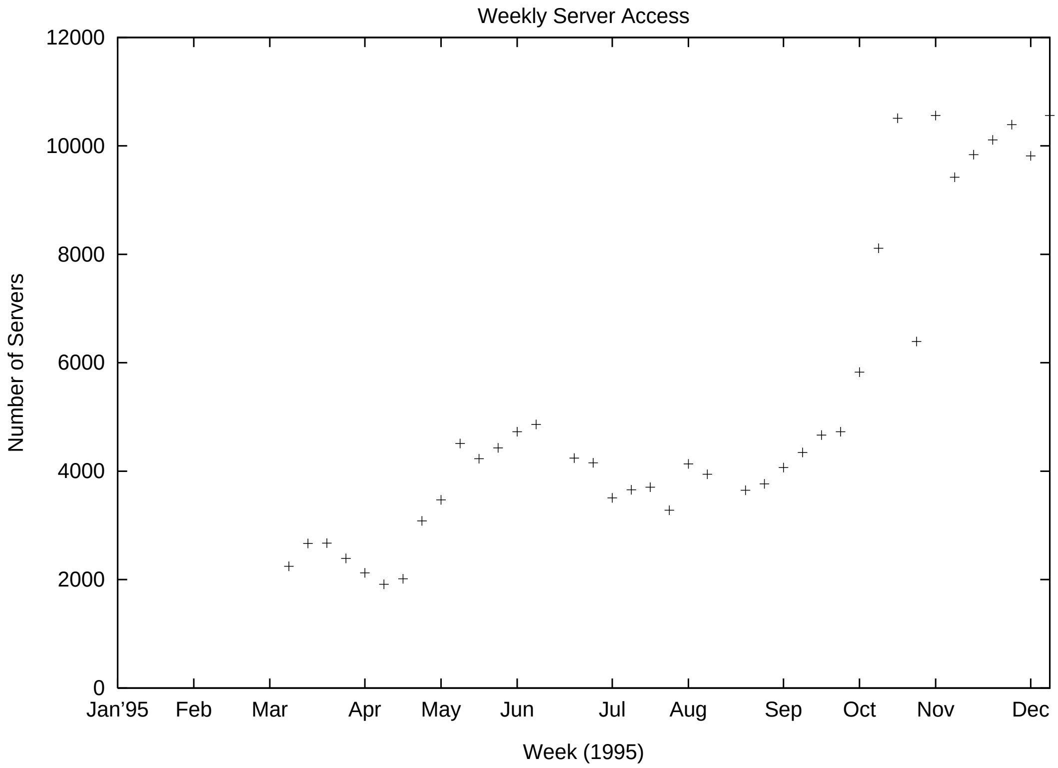

Figure 1: Number of Unique Servers per Week

It is unfortunate that our client population follows this seasonal trend, and that from year to year this population grows with the increasing popularity of our service. The combination of these effects makes it difficult, if not impossible, to differentiate growth based on more information becoming available from growth based on the simple fact that a larger client population is likely to cover a wider server population.

This problem is demonstrated by the fact that between January 1995 and January 1996 the number of requests served by the HENSA cache has increased fourteen-fold, from 78,000 to 1,100,000 per day. This compares to a growth in the number of servers seen in the order of five-fold. How much of the apparent growth in server population can be attributed to an increased number of clients, and how much is due to a real increase in the number of servers is difficult to say.

While the inability to resolve this problem is worrying, with more data, giving us more information about annual trends, we should still be able to gauge the number of active servers. After all, as far as a cache is concerned, there is no difference between a server that doesn't exist, and a server that exists, but is never accessed.

Figure 2 shows a count of servers visited as a result of requests by clients, but in this case a cumulative count of unique servers is kept. This means that once visited, a server is always counted towards the total. An expiry of 31 days has been placed on this count. This means that a server that has ceased to serve will be excluded from the count after that period of inactivity.

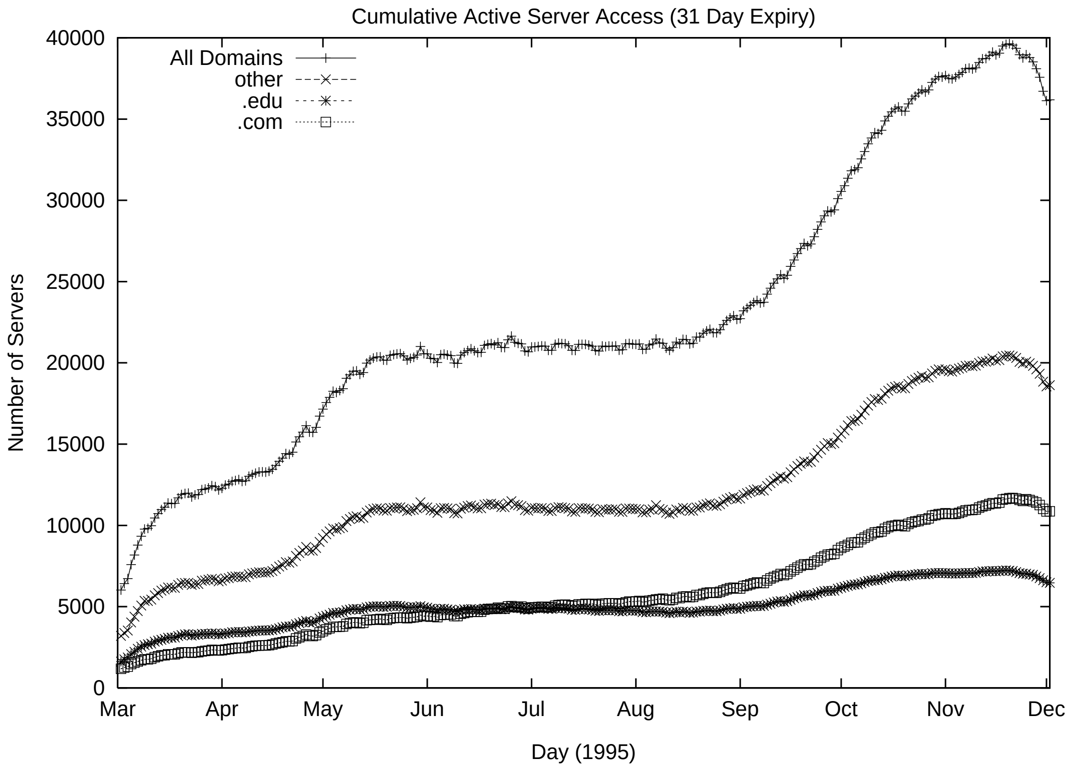

Figure 2: Cumulative Number of Servers per Day

This method of counting allows us to more accurately determine the number servers in existence, as compared to the number of servers currently attracting user attention. Looking at both figures, we could suggest that four times as many servers exist, as are visited in any particular week.

Additionally, on this figure the counts have been broken down by domain. `.edu`, `.com`, and the rest of the world, are represented as separate plots. The pattern overall is, again, similar to the seasonal trends that we expect. However, by looking at the relative speeds of growth, we can see the `.com` domain expanding, overtaking `.edu` in June, and accounting for more than 25 per cent of all sites visited by the end of 1995.

The size and rapid growth of the `.com` domain presents problems at all levels. Technically, the naming system designed to be efficient and scalable is becoming very unbalanced. Practically, the `.com` domain is geographically location independent (although, at the moment, the vast majority of `.com` sites are located in the US) and as far as caches are concerned this makes it difficult to develop global strategies based on a site's name. Along with the document location, this is all the information a cache has when a client requests a particular URL.

4 Server objects

An alternative method to consider in trying to measure the growth of the Web, along with the effect that it may have on caching, is based on the number of distinct documents seen, rather than the number of sites visited.

In Figure 3 we have collated the number of documents seen at seven of our most consistently popular sites. The documents counted are only those that can be cached. This excludes most documents generated on the fly (for example, the result of a CGI script run on the server or database search) which are very poor candidates for caching and so are not of interest in this study.

Site	Apr	May	Jun	Aug	Oct	Nov
Best	2581	3905	4498	7627	7644	7021
CNN					6478	7375
Infoseek	190	225	294	363	588	715
Netscape	1713	1841	1580	1854	3329	3373
Telegraph	502	852	977	683	1118	1249
Webcrawler	49	83	80	99	335	407
Yahoo	3493	5060	5702	8675	21828	19138

Figure 3: Site growth (Number of Static Documents Cached), 1995

The figures show that some sites are very bandwidth efficient when used with a cache, while others are more bandwidth wasteful. Comparing the three directory services in the top seven (Infoseek, Webcrawler and Yahoo) we can instantly see how these services achieve their goal (resource discovery on the Web).

Infoseek and Webcrawler are search engines that respond to a user query. The result of this query cannot be cached, as the contents of the search database could be different the next time that search is requested. In any case, these searches are based on keywords and, with a few exceptions, these keywords tend to be different for each request (even if the results are the same). Yahoo, on the other hand, is a hierarchical directory made up from a large number of static pages. These pages can be cached.

When modelling the caching process, and estimating the bandwidth savings made, it is important to bear in mind the different strategies used by these very popular services. Caching is an effective way to ensure that users of Yahoo make efficient use of bandwidth. The only way to make services such as Infoseek and Webcrawler more efficient is to distribute the databases that they rely on so that queries and responses do not have to cross congested links.

A further examination of the figures gives us another, more concrete, opportunity to gauge the growth of the Web. Services such as the search engine based directories and Netscape, have existed for a long time and are virtually independent of the nature of the client user. Given a certain minimum number of clients, it is reasonable to assume that all pages on such general servers will be visited in a relatively short period of time. As a result it is feasible to suggest that the growth perceived through the cache's log of transactions is a true representation of the growth of the site. Given a larger collection of such servers it should be possible to gain some idea of how this growth maps to the Web as a whole. This is work that will be followed up at HENSA Unix over the coming months.

5 Size of server objects and media mix

The Web has paved a large part of the way for global multimedia communication. This is one of the greatest contributing factors to the success of the Web and also to the bandwidth crisis experienced by the Internet community as a whole.

In order to ascertain how the composition of Web traffic has changed we have considered the four key media types: text, including of course HTML, still images, moving images and audio files. As far as caching, and efficient use of the networks are concerned this consideration of the different media types is important. Their characteristics, in terms of the network and cache overheads they impose, are distinct and different. Broadly speaking, large video and audio files fall into one of two categories, either frequently changing (for example, a weather report) or never changing (a movie clip). This allows us to consider these media types differently from HTML and still images within the cache. A large infrequently changing video or audio file can be kept in the cache for far longer without being refreshed. Heuristically, there is is far less chance that it will change spontaneously. This gives very good caching performance.

In a similar fashion, when comparing still images and HTML it is probably safe to say that HTML has a much greater likelihood of changing than a still image. In the case where the contents of a still image change, the name of the file is likely to change to reflect this fact and there is no chance of the cache serving the incorrect image. In the case where the contents of an HTML file change, the filename is likely to remain the same and so the cache has to check regularly to ensure that changes on the remote server are reflected in the cache.

A significant trend in the growth of one or more of these media types would allow us to apply these heuristic rules in order to determine future traffic composition. With our other considerations this would enable us to predict the future effectiveness of a Web cache.

Figure 4 shows how the four key media types are represented by the distribution of accesses. Text and still images account for all but a tiny fraction of transactions, but the size of audio and video data means that we should seriously consider their effect.

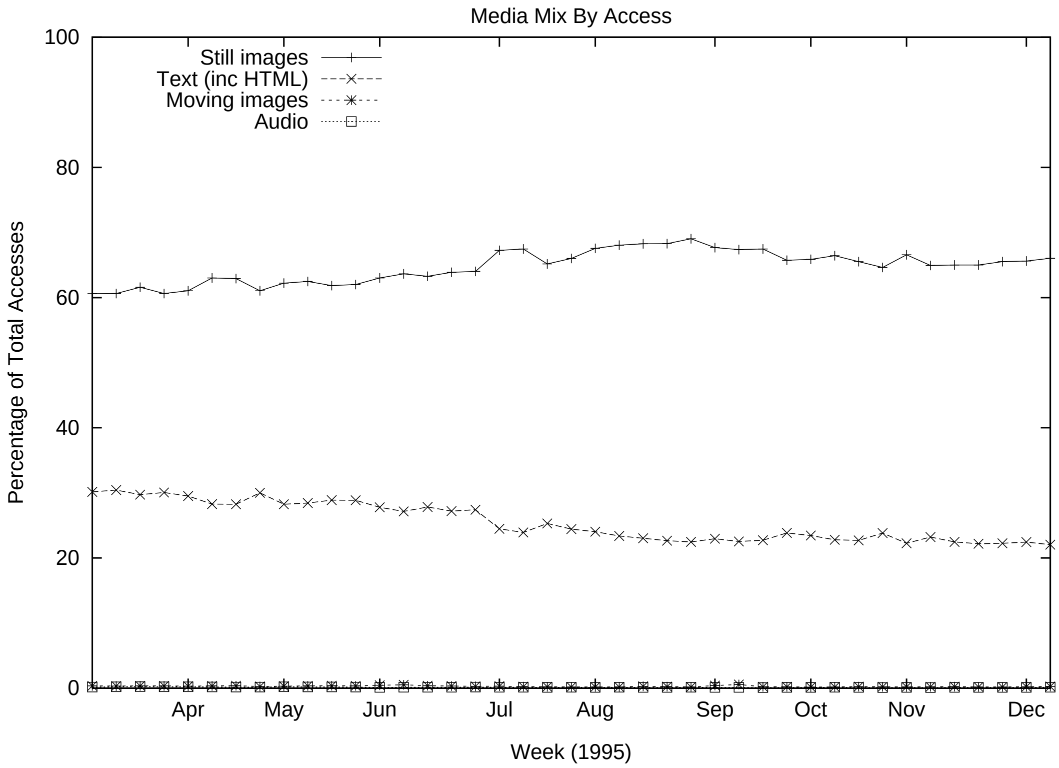

Figure 4: Media Proportions by Access

Accesses represent overheads in the number of connections made to remote servers, and we can see from this graph that video and audio material account for virtually nothing. With the promise of multimedia, we might expect to see a growing use of these two data types. It seems that the situation, at least as far as the UK is concerned, is stagnant. This is probably due to the restrictions that the limited bandwidth places on the retrieval of these files (the chance of successfully retrieving any file larger than about 100 kilobytes between the hours of 9:00 and 21:00 is very small). The apparent growth in the use of still images, with a reduction in textual material is promising as far as caches are concerned (although what it means for the information content of the Web is another matter). One possible explanation could be an increasing use of small images on text pages, as page design becomes more sophisticated, and advertising becomes a popular means of generating supporting revenue.

Figure 5 shows these media types in terms of the number of bytes that they account for. Even a casual glance at this graph shows that, other than apparently random fluctuations, the proportions of the four key media types have not changed over the past year (and further back, but not represented on the graph). There are no evident trends for any of the media types, either upwards or downwards.

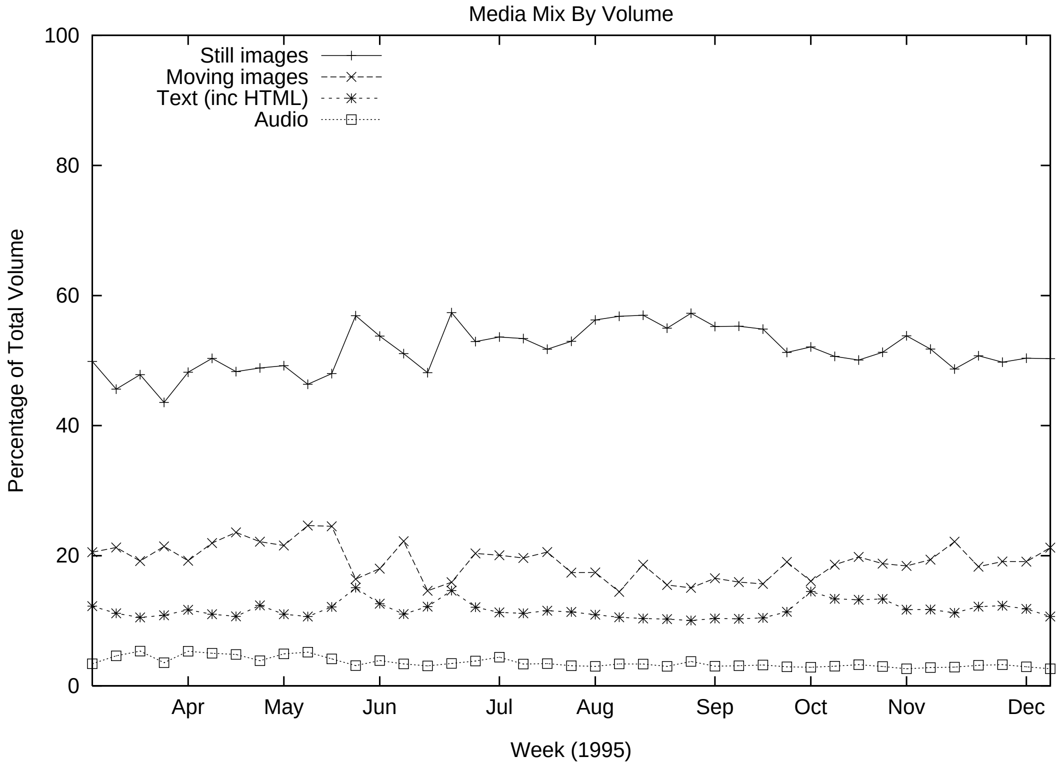

Figure 5: Media Proportions by Volume

6 Number of clients

By themselves Web servers make little use of bandwidth. To have a complete picture of network traffic and make suggestions for cache implementation it is essential to consider the growth trends at the other end of the transactions; the clients. As with server growth we have difficulty differentiating between an increase in the popularity of our service (that is people who were using the Web before, but have now switched to using HENSA Unix) and the true growth in the number of Web users.

Figure 6 shows how the observed client population has changed. The same seasonal trend is apparent, although it is easier to justify a large increase in the number of users from the beginning of the academic year as more institutions reequip over the vacation and then expose their machines to the Internet at the end of it.

In this case the analysis is additionally complicated by the fact that the cache can only log the name of the host from which each request comes. It is difficult to determine whether this host is a multi-user machine, a dedicated workstation or perhaps another proxy cache representing many thousands of users. However, by measuring the number of connections we see from a specific machine and the number of other hosts sharing that domain we can make some attempt to determine the distribution of these three types of system.

Compare, for example, the access patterns of Edinburgh and Leeds Universities. Over the period 10-19th October 1995, Edinburgh accounted for 463,000 requests from 850 distinct hosts. At the other extreme, Leeds accounted for 188,700 requests from 32 hosts. The greatest number of requests from a single host at Edinburgh was 7,500 while from Leeds it was 182,000. It is clear that Edinburgh has lots of individual machines making use of the cache, while Leeds have a local proxy that itself then makes use of the national facility. In this case, Edinburgh is a much more attractive source of data when trying to establish client growth trends as we believe that we can isolate, if not individuals, then at least small groups of users. This is further work that HENSA Unix will be carrying out.

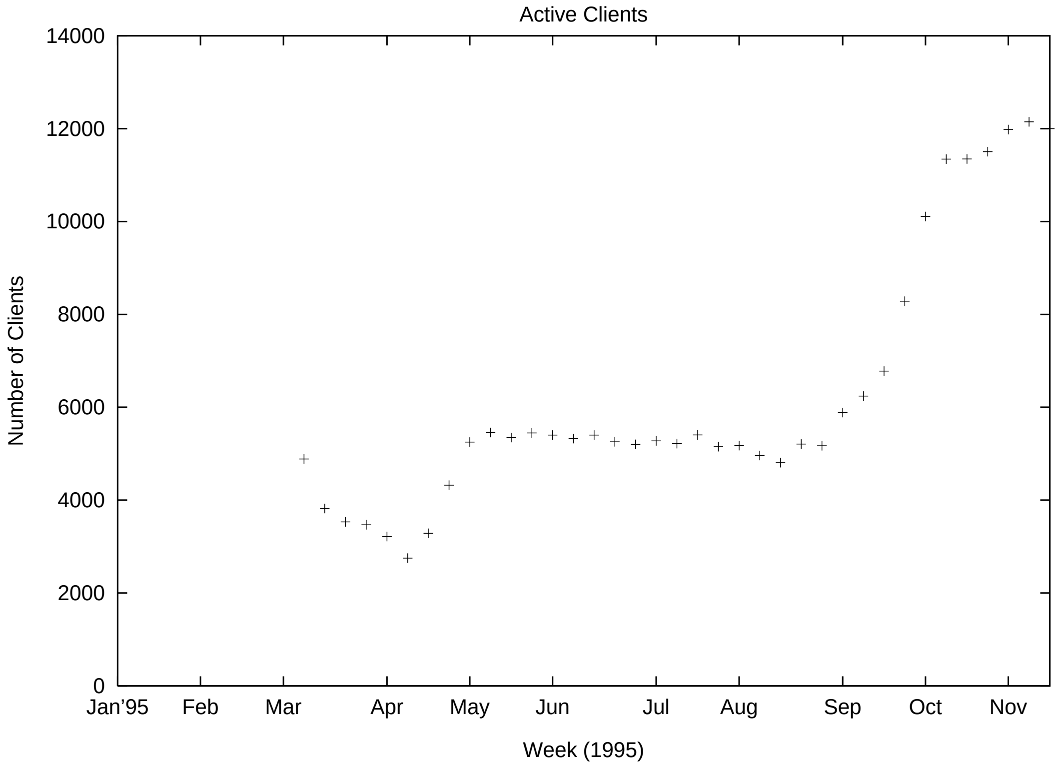

Figure 6: Number of Unique Clients per Week

7 Use of protocols

Some of the Web's success has surely been due to the fact that it encompasses the technologies that preceded it. While this makes standardisation and software development somewhat slower it also makes the transition from one of those older technologies that much easier. It is interesting to see how those older technologies are still being used.

The fact that HENSA Unix can proxy and cache for FTP as well as gopher is well advertised, but perhaps not appreciated by all, and it is certainly true to say that these older protocols are not as sophisticated in the ways in which they can interact with the proxy. For this reason we have not made a quantitative comparison of the connections attributed to each protocol, but merely a comparison of the proportions of those connections and the quantities of data transferred by each. This comparison can be seen in Figure 7. Over the period represented by this data, Accesses have risen from 0.6 million in October 1994 to 22 million in October 1995, whilst data volume has increased from 8 Gigabytes to 300 Gigabytes.

		10/94	11/94	12/94	1/95	4/95	6/95	8/95	9/95	10/95
	http	95	94	92	93	93	95	95	96	97
Access	gopher	3	3	2	3	2	1	1	1	<1
	ftp	<1	2	4	2	3	3	2	2	1
	http	92	87	80	83	91	85	88	89	92
Volume	gopher	3	3	3	4	1	1	1	1	<1
	ftp	3	8	15	11	6	12	10	9	6

Figure 7: Protocol Use. Percentage of traffic by access and volume.

While gopher has all but been abandoned, FTP is still a well used protocol. By comparing the data we can see that FTP, while accounting for a relatively small number of connections, still represents a significant proportion in terms of data transfer. This suggests that the crude mechanisms used by existing caching servers when dealing with FTP requests should be improved to ensure more effective bandwidth use. This would surely result in significant improvements as the most widespread use of FTP is to transfer large quantities of very infrequently changing data, i.e. software packages.

8 Modelling and simulation

Despite inconclusive results that will require more analysis, we have taken a first step towards modelling and simulation. Based on the logged transactions seen by the HENSA Unix cache we have simulated a simple proxy cache under different cache size constraints.

This simulation is based on the logged transactions from March to June 1995. Over this period of time the cache served 23 million requests accounting for 351 gigabytes of data transferred to clients. It should be clear that a cache serving a smaller client population will probably not achieve such good results.

The results from this simulation, seen in Figure 8, are of general interest to anyone wishing to deploy a Web cache as they show the likely savings that can be expected dependent upon the investment in disk space.

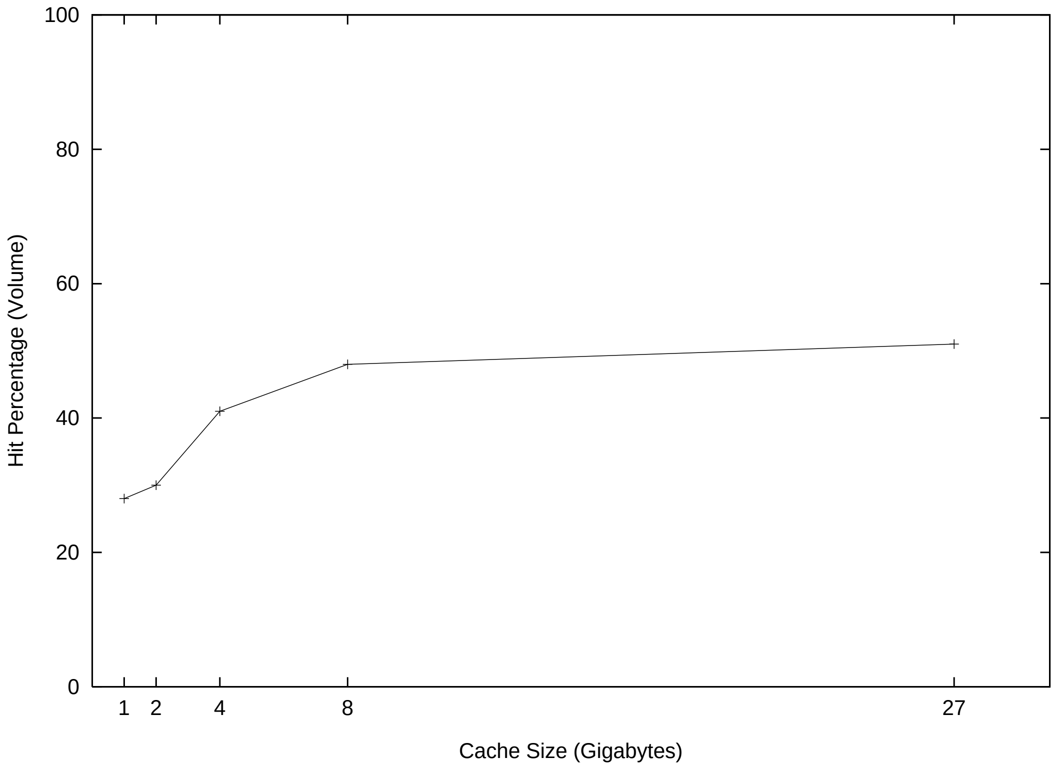

Figure 8: Simulated Hit Rate on Unix HENSA Proxy Traffic

With further work we hope that our models will allow us to simulate the behaviour of such a cache at any stage of the Web's development.

9 Conclusion

The aim of this study has been to use the information contained within the log files of a large and popular proxy cache in an attempt to chart trends in the use of the Web. We hope to use this information to model these trends and build an accurate simulation of the Web's use and growth. With this simulation we should then be able to analyse proposed schemes to combat the current bandwidth crisis.

The volumes of data involved in this study, and ever changing nature of the systems being studied have made it difficult to draw firm conclusions. However, some trends are clear, and this work has led to a number of interesting avenues that we shall be following up in the near future. We would be pleased to hear from other groups with additional material that would serve to enhance our model.

10 References

[1] Lara D. Catledge, James E. Pitkow, Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems* 27 (1995) 1065-1073

[2] Helen Fuell, M.D. Tedd, Experimental Analysis of JANET Web Servers. *Poster Proceedings Third International World-Wide Web Conference* (April 1995) 68-70

[3] James E. Pitkow, Krishna A. Bharat, WebViz: A Tool for World Wide Web Access Log Analysis. *Advance Proceedings First International World-Wide Web Conference* (May 1994) 271-277

[4] James E. Pitkow, Mimi Recker, Results from the First World Wide Web User Survey. *Advance Proceedings First International World-Wide Web Conference* (May 1994) 283-294

[5] James E. Pitkow, Margaret M. Recker, Using the Web as a survey tool: results from the second WWW user survey. *Computer Networks and ISDN Systems* 27 (1995) 809-822

[6] Neil G. Smith, What can Archives offer the World-Wide Web? *Advance Proceedings First International World-Wide Web Conference*, eds. R. Cailliau, O. Nierstrasz, M. Ruggier (May 1994) 101-111