Understanding and Reducing Web Page Latency

Kevin Curran and Noel Broderick

Abstract - ***Studies have shown that surfers spend a lot of time impatiently waiting for Web pages to emerge on screen and HCI guidelines indicate ten seconds as the maximum response time before users lose interest. This paper presents research into the observed usage of Web images and the effect on page retrieval times. The prevalent factor that affects how quickly a Web site performs is the type of Web hosting environment that the site is deployed in. Web users are faced with sliding scale of delays in reality, with no one Web page taking the same time to load on two separate occasions. It is arguable that the magnitude and variance of network delay between a client and server are generally proportional to the distance spanned, assuming that all other influencing factors remain constant. Web can tweak their content to reduce the loading time of their sites.***

Index Terms: ***Web page latency, Performance measurements, Image compression***

1. INTRODUCTION

Web access is the most popular service and statistics from NUA speculated that there are an overwhelming 605 million users online [1]. Evidently, this immense demand placed on the World Wide Web infrastructure has led to reduced bandwidth which is a primary contributor to client latency. Studies have shown that Web users spend a lot of time impatiently waiting for Web pages to emerge on screen. Slow page retrieval times is the most widely reported problems and users prefer Web pages to be presented on a computer screen as quick as one can turn over a new leaf of a book. HCI guidelines indicate ten seconds as the maximum response time before users lose interest [2]. Such delays impact the sites success and are expensive in terms of lost business opportunity or users productivity [3]. Since the Mosaic browser was introduced in 1993 with its ability to display images [4], there was a proliferation in the number of Web pages using a combination of text and image-heavy design.

Manuscript received December 5, 2004.

Kevin Curran is a lecturer with the school of Computing and Intelligent Systems, at the University of Ulster, UK. He can be contacted at kj.curran@ulster.ac.uk.

Noel Broderick is an MSc student with the school of Computing and Intelligent Systems, University of Ulster, UK.

The use of images contributes heavily to slow loading sites [5] and is straining the capacity of the Internet further as the current network infrastructure that also supports bandwidth intensive applications such as email, video conferencing and online gaming, will not cope with the stress of end users increasing demand for more bandwidth. Any technique that saves on bandwidth and make browsing more pleasurable should be explored.

This paper presents research into the observed usage of Web images as it is within the control of Web developers and will yield the most value. Findings of the study will provide a better understanding and help to devise a strategy on what can one do to eliminate or at least reduce potentially harmful effects of very slow page retrieval times. The top level domains (TLD), i.e. home pages, of forty seven academic Web sites were chosen. They span across five different countries: UK, Ireland, Canada, USA and Australia. They have many Web users and typically they all hit the same home page. There will be potential students checking out the course prospectus. Perhaps commercial users are looking for consultancy. The five countries under scrutiny have their own private multi-gigabit data communication network reserved specifically for research and education use and are linked to international peer networks.

The first study examines network latency by visiting the Web sites for the first time. The efficiency of cache mechanism in reducing the client latency was also assessed. Latency measurements was obtained from two different sources, from a workstation at the author's institution and from a Web performance monitoring service provided by TraceRT [6]. The second study surveys the academics' Web sites and account for variations in page retrieval times particularly to images which was the interest of this study.

The final study explores image compressions and assesses how efficiently Web developers are optimising images for their Web sites.

Another topic awaiting exploration was to trim the file size of images while retaining visual fidelity.

The effects of reduced image size (in bytes) have on page retrieval times had been looked into.

2. NETWORK DELAY COMPONENTS

A modern server uses Path Maximum Transmission Unit Discovery (PMTUD) heuristics to determine the Maximum Segment Size (MSS) which is the safe packet size that can be transmitted. This technique was adopted to address the poor performance and communication failures associated with oversized packets which are fragmented at routers with small MTU [7]. Today, the PMTUD concept is imperfect as it uses the Internet Control Message Protocol (ICMP) which some network administrators view as a threat and block them all, disabling PMTUD, usually without realising it [8]. This led to increased packets overheads due to retransmissions and eventually connection time-outs. Lahey [9], suggested a workaround where after several time-outs, the server network should be reconfigured to accept an altered ICMP packet with the 'Do Not Fragment' bit disabled. Consequently the PMTUD feature is bypassed, but detection can take several seconds each time, and these delays result in a significant, hidden degradation of network performance.

2.1 Transmission Control Protocol's (TCP) Flow control

The flow control mechanism of TCP uses slow start and congestion avoidance algorithms as a mechanism to control the data transmission rate. This helps to reduce packets loss caused by congested routers. However, lost packets can be recovered using TCP's retransmission feature, but this incurs added delivery time. The aggressive behaviour of multimedia applications involving audio and video, in which developers employ UDP compounds the problem of congestion. UDP are not TCP friendly and they do not respond to packet drops which typically hint congestions. This aggressive behaviour degrades and even shuts out TCP packets such as Hyper Text Transfer Protocol (HTTP) and prevents them from obtaining their fair share of their bandwidth when they battle for bandwidth over a congested link. Lee et al. [10] examined the use of TCP tunnels at core routers to isolate different types of traffic from one another. Benefits include reduced TCP's retransmission per connected resulting in packets being processed using the same amount of memory resources. This concept is not used extensively on the current Internet infrastructure.

2.2 Domain Name Servers (DNS) Lookup

DNS is responsible for translating domain names into an equivalent IP address needed by the Internet's TCP. The latency between DNS request and response is a random variable as the DNS lookup system uses the client's cache file, the hierarchical nature of the domain name and a set of DNS operating at multiple sites to cooperatively solve the mapping problem. A survey from Men and Mice [11] showed that 68% of the DNS for commercial sites (e.g. .COM zones) has some configuration errors, thus making them vulnerable to security breach and denial of service. The often can be misconfigured. An intelligent DNS management system was recently developed by Liu et al. [12] which offers administrators support in DNS system configuration, problem diagnosis and tutoring,

2.3 Protocol

The network delay for Web page loading is dominated by the current version of the HTTP/1.1 standard. It is an application level protocol for transfer of Web contents between clients and servers. Due to increasing Internet traffic, HTTP/1.1 makes inefficient use of the network and suffers from high latencies for reasons such as TCP's three-way handshakes for opening a connection which adds extra round trip time delay and multiple parallel TCP streams which do not share the same congestion avoidance state. Spreitzer et al. [13], have composed a prototype for HTTP 'next generation' which should address these latency issues.

2.4 Cache mechanism

Caching mechanisms can exist on a client's local disk, network servers or at Internet Service Provider locations. Its rationale is to assuage congestion, reduce bandwidth consumption, improve retrieval times by temporary storing Web objects closer to the clients and reduce the burden on the site server as it handles fewer requests. Caching is often deliberately defeated as not all Web contents are cacheable. A modern day Web page contains both dynamic and static contents. Dynamic items are non-cacheable and typically they contain interactive and changeable items that provide a far richer experience for users, but they are not happy to wait for them [14]. Cached components characteristically contain items that do not change, i.e. they are static. An intelligent cache engine has emerged recently that serves dynamic elements of Web page and reduces the latency time by 90% [15]. It works by estimating future client's behaviour at a site based on pass and present access patterns. The downside with caching is that if the user does not use the cached items, then congestion may have been caused needlessly.

2.5 File size of embedded objects

Recommendations that were made to improve Web page designs have positive impact to page retrieval times as well as usability. The adoption of Cascaded Style Sheets (CSS) [16] and more compact image representations, Portable Network Graphics (PNG) [17], have added value of reducing the file size and speeding up page downloads without sacrificing graphics design [14]. PNG was designed to be successor to the popular GIF files, but it was not until the late 1997 when browser wars came to an end as many old browsers finally caught up and are able to read PNG formats. Another Web image format is JPEG which uses lossy compression.

3. EVALUATION

Forty seven TLD sites belonging to worldwide universities were selected for this study. This was comprised of twelve UK sites, ten USA sites, nine Canadian sites, eight Irish sites and eight Australian sites. Universities selected for this research are linked to the NRENs [18], [19], [20], [21]. Information about the network topology for the five NRENs was gathered and checks were made to ensure that there were no intermittent brief outages or reported performance issues.

For this study, the response time was obtained from two sources. From the author's institution, the response time was the period it took for the requested Web page to be fully presented on the browser window. This measured the performance of Web pages delivered within the international NRENs infrastructure via the UK's JANET network. The test was conducted using Netscape Navigator 6.2 on Windows 2000 Professional, with a 10Mbit/s link to JANET. The experimental method was to request a Web page via specially prepared bookmarks. The download timer in Navigator gave the loading time measurements. Two types of request were used:

1. *First time retrieval*: equivalent to a browser visiting a site for the first time. In Navigator the memory and file cache was cleared.

2. *Cache request*: equivalent to revisiting a site online. The static contents were already available in the client's local cache. This meant that static items are displayed on screen more quickly the next time the page is visited and any dynamic items had to be retrieved from the server.

To account for network idiosyncrasies, latency measurement was collected three times and the measured mean was used. The second measurement source was provided by TraceRT. This service was used to measure the speediness of Web sites as seen from six measurement points (called agents) around the world. The agents are commercial sites and operate outside the NRENs infrastructure. To account for changing server loads and different time zones, the response time investigations was repeated at approximately the same time in the morning, afternoon and evening (in British Summer Time) for seven consecutive days. Long latency link may have a major influence on the total response time for serving a set of Web page objects from the server to the client. The location of site servers was gathered using a diagnostic tool from NeoTrace Express [22]. To account for variations in retrieval times, statistics on the quantity and size of objects that a Web page contains was collated. The number of embedded objects gave an indication of how many server requests must be made and the file size implicates how quickly the heterogeneous network could present them.

Using GIF files that were extracted from the sample sites, the author tried to re-express these images into compact PNG formats. This was done by the means of a batch image conversion tool, ReaConverter Pro v3.4 [23]. Where possible an attempt was made to create transparent PNGs so that images could rely on the background colour of the site's home page. With Adobe Photoshop 8.0, all JPEG images were optimised for the Web using sixty as the quality factor. The optimised JPEG images were compared with the original to see if Web developers have used appropriate compressions. Next, the author trimmed the file size of optimised JPEG images by saving them, unchanged, as JPEG in Microsoft Paint v5.1. The file size for the new PNG, optimised JPEG and trimmed JPEG images was recorded.

3.1 Thresholds of interest

Web users tend to be sensitive to variations in loading delays and for this study there are two natural thresholds of interests: that of insignificant and that of pain. Delays that are less than the threshold of insignificant are not minded by the user. Delays that is greater than the threshold of pain result in users abandoning the system. Delays that fall between these thresholds normally results in a minor complaint from the user. Absolute values for these natural thresholds are not known as patience varies from user to user. In this study, values for insignificant and 'pain' threshold were taken to be three seconds and eight seconds, respectively.

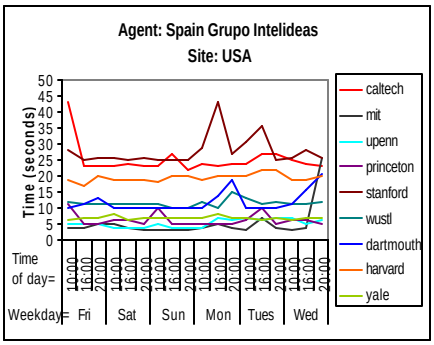

Figure 2: Speed of USA's Web sites as seen by an agent in Spain.

4. RESULTS

The first time retrieval and cache validation tests for Web pages downloaded via the JANET infrastructure is shown in Figure 1. The author found no obvious relationship with the file sizes or image counts to justify the response times. Long latency link could possibly explain the high retrieval times as experienced from two USA's site servers, 'Caltech' and 'Stanford', as they are the only site servers in the west coast of USA. Other USA sites are located in the east. Figure 2 confirms the speed (in seconds) of nine USA's Web sites as observed by a commercial agent in Spain.

While the Internet was behaving in a fluctuating manner it can be seen that five USA's sites would have missed out on possible business opportunities as they were in the pain zone. The other four sites fell inside the whinge sector. USA lags behind Spain by five to eight hours, therefore a safe assumption was made whereby on Monday afternoon (Spanish time) the condition of site servers became heavy as web users in the USA went online after the weekend break. Comparing the latency results shown in Figure 1 and 2, pages requested within the NREN infrastructure was presented much quicker. To seek out additional reasons behind the sliding scale of delays as seen in Figure 1 and 2, a packet sniffer [24] was used to count the number of packets involved for the transmission of images. When a sniffer was applied during individual requests for images from sample site servers, some interesting effects were noticed. Based on visual inspection of Figure 3, it has been noticed that no two images of equal size (in bytes), from five Australian's sites, arrived at the author's workstation with the same number of packets. As the file size of images increased the packet counts increased exponentially and without doubt so did the image loading time. Similar behaviours have been observed for sample sites in USA, Canada, UK and Ireland. Dissimilar PMTUD schemes used by site servers, server loads, congestion levels or fragmentation of oversized packets may have attributed to varying packet counts.

The outcome of the first time retrieval tests (from Figure 1) showed that five sites were in the pain sector, ten sites fell inside the whinge region and twenty eight sites in the insignificant zone. With the cache mechanism in place, the author noticed that for the five sites that were in the pain sector, one went to the insignificant zone while three moved to the whinge region. The last one stayed, but its response time improved by ten seconds. Three cache misses took place, but the user would not have cared or noticed because the average added time was 236ms and the affected sites did not shift from the insignificant zone. The cache system was very effective in reducing the retrieval times and made Web browsing more pleasurable. While cache misses augment page retrieval times the author carried out a survey to obtain the frequency of these misses.

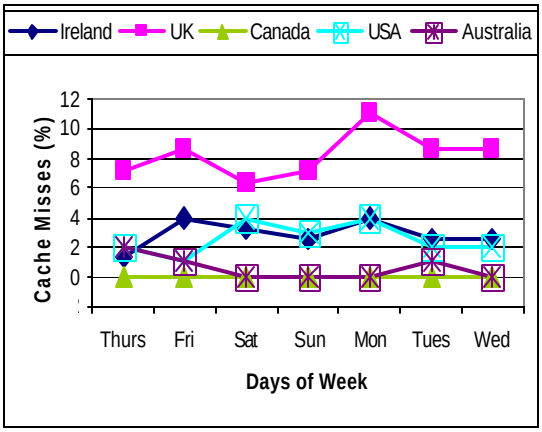

Figure 4: cache misses within the international NREN infrastructures.

The results are depicted in Figure 4. While there were no cache misses for Canada, UK had the highest percentage of misses due to a large number of sites containing dynamic items which had to be fetch from the site servers when the page was reloaded. It has been noticed that new items may increase the overall page size and this impacts how quickly the Internet could present the page's contents. The author took a peek at Navigator's temporary cache folder and noticed that only image files are readily cached than text files.

Forty seven test sites held a total of static 904 GIF images. Of these only 476 were successfully converted to PNG format. GIF images that were successfully converted totalled 1,048,299 bytes, while PNG equivalent resulted in a total 919,298 bytes, saving 129,001 bytes. The savings are modest because many of the images are very small. PNG conversion did not perform well on very low depth images in the sub-900 bytes category. It is thought that checksums and related data were added which made the file bigger. It is clear from Figure 5 that some sites in

the USA did not Web-optimised their JPEG images. By optimising the JPEG images and trimming the file size, the author was able to compress 452,250 bytes of original JPEG images down to 194,150 bytes. This represents a saving of 57.1%. Similar behaviour was also observed for sites in Australia, Canada, UK and Ireland. It is thought that by saving the optimised JPEG images in Microsoft Paint, it removed supporting bytes used by Photoshop. Based at the author's institution there are two public Web sites, each containing thirteen images. The total page size for Site A is 83.3KB and is composed of six GIFs and seven original JPEG images. Site B totalled at 51.9KB includes matching images as in Site A, but only compressed versions are used. The effects of reduced file size to loading times as seen by six foreign agents are depicted in Figure 6. It is evident that by making attempts to reduce file size, it will reduce user visible latency.

5. CONCLUSION

The prevalent factor that affects how quickly a Web site performs is the type of Web hosting environment that the site is deployed in. Web users are faced with sliding scale of delays in reality, with no one Web page taking the same time to load on two separate occasions. It is the number of application packets, not bytes, and the number of simultaneous users of the part of the Internet involved in the connection that determines the Web page latency and satisfaction levels. If Web developers take the time to tweak different file sections then the loading time of their Web sites will fall dramatically. While it is highly documented that PNG is a more compact image representation, they are not suited on low depth images in the sub-900 bytes group. Of the 904 GIF images sampled, 48% of them fell in the sub-900 group, but they do not graphically capture the meaning of the page. To achieve the graphical and functional goals of web sites within the technological limitations of the Internet infrastructure, the author wish to research the possibility of developing a web-authoring tool that will trim the file size of images autonomously. One technique as was adopted in this paper is to convert GIF images above 1KB to PNG formats. In addition, using known limitations of the human eye, the tool can further optimise the JPEG images to acceptable quality levels. An option would be available for developers to override these features. One drawback with these techniques is that they will stress memory resources as they will contain both original and compressed images.

REFERENCES

[1] http://www.nua.ie/surveys/how_many_online/ (September, 2002)

[2] Selvidge, P.R., Chaparro, B.S. and Bender, G.T. (2002). The world wide wait: effects of delays on user performance. *International Journal of Industrial Ergonomics.* 29:15-20

[3] Saiedian M.Z.H. and Naeem, M. (2001). Understanding and Reducing Web Delays. *IEEE Computer Society Press* 34(12): 30-37

[4] Fraser, Q. (2001). The Life and Times of the First Web Cam. *Communications of the ACM*, July, Vol.44, No.7, 25-26

[5] Zhi, J. (2001). Web Page Design and Download Time. *CMG Journal*, Spring, Issue 102

[6] http://www.tracert.com (July, 2004)

[7] Kent, C.A. and Mogul, J.C. (1987). Fragmentation considered harmful. *Digital Western Research Laboratory.* Research report 87/3, December. Taken from: http://research.compaq.com/wrl/techreports/abstracts/87.3.html (December, 1987)

[8] Knowles, S. (1993). IESG Advice from Experience with Path MTU Discovery. *RFC1435*. March. Available at: http://www.faqs.org/ftp/rfc/pdf/rfc1435.txt.pdf (March, 1993)

[9] Lahey, K. (2000) TCP Problems with Path MTU Discovery. *RFC2923*. Available at: http://www.faqs.org/ftp/rfc/pdf/rfc2923.txt.pdf (September, 2000)

[10] Lee, B., Balan, R., Jacob, L., Seah, W. and Ananda, A. (2002). Avoiding congestion collapse on the Internet using TCP tunnels. *Computer Networks*, Vol.39, No.2, 207-219

http://www.w3.org/TR/REC-CSS1 (January, 2004)

[11] http://www.menandmice.com/6000/61_recent_survey.html (February, 2003)

[12] Liu, C.L., Tseng, S.S. and Chen, C.S. (2004). Design and Implementation of an intelligent DNS management system. *Expert Systems with Applications* , 27:2, 223-236

[13] Spreitzer, M. and Janssen, B. (2000). HTTP 'Next Generation'. *Computer Networks* , 33:593–607

[14] Nielsen, H.F., Gettys, J., Baird-Smith, A., Prud'hommeaux, H., Lie, H. and Lilley, C. (1997). Network performance effects of HTTP/1.1, CSS1, and PNG. *Computer Communication Review*, 27: 4.

[15] Govatos, G. (2001). Accelerating dynamic Web site performance and scalability. *Chutney Technologies* , *Inc*., Available at: www.caching.com/pdf/Preloader_final.pdf (January, 2001)

[16] Lie, H. and Bos, B. (1996). Cascading Style Sheets, level 1. *W3C Recommendation, World Wide Web Consortium*, 17th Dec 1996, revised 11th Jan 1999. Available at:

[17] http://www.libpng.org/pub/png/ (May, 2004)

18] http://www.ja.net (July, 2004)

[19] http://abilene.internet2.edu/ (July, 2004)

[20] http://www.canarie.ca/canet4/ (July, 2004)

[21] http://www.heanet.ie (July, 2004)

[22] http://www.networkingfiles.com/PingFinger/ (April, 2004)

[23] http://www.convertzone.com/net/cz-ReaConverter %20 Pro-6-7.htm (July, 2004)

[24] http://www.etherdetect.com (July, 2004)

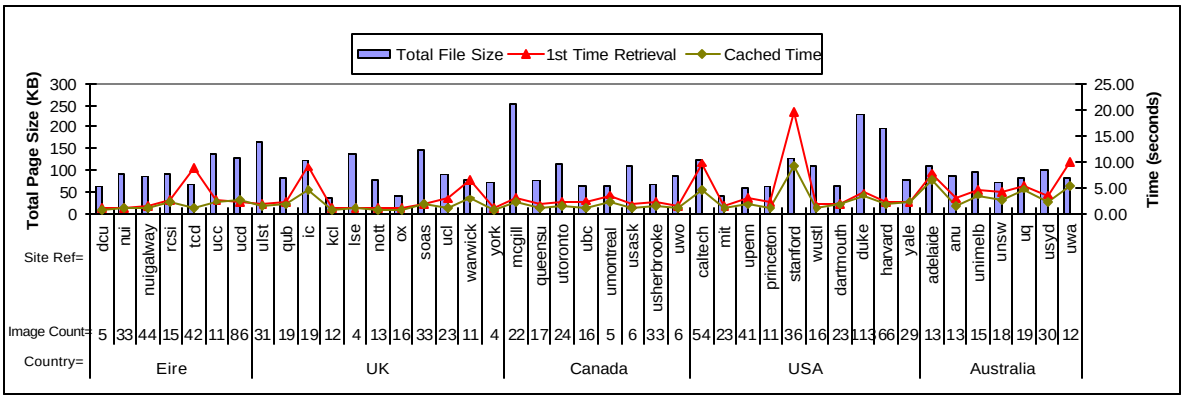

Figure 1: First time retrieval and cache validation response time results (for pages requested within the NRENs). Including image counts with histograms showing the total page sizes (in KB).

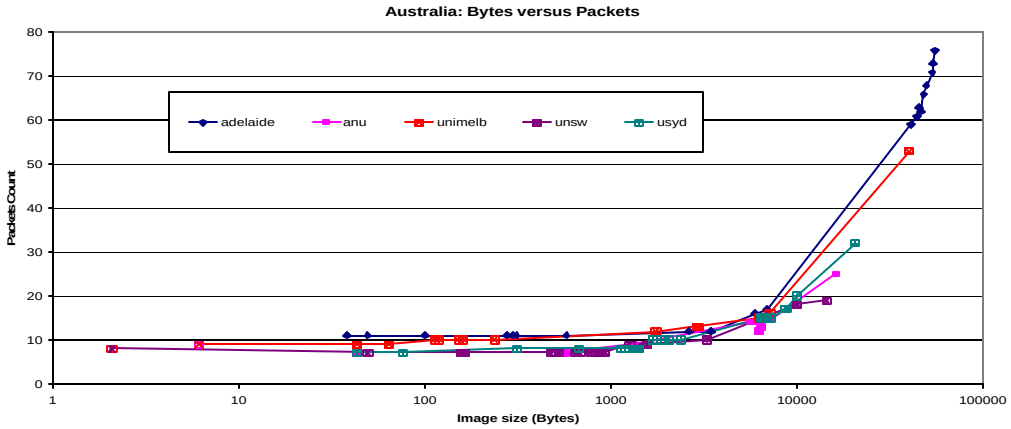

Figure 3: Australia: Image Size (in Bytes) versus Packet count from five site server

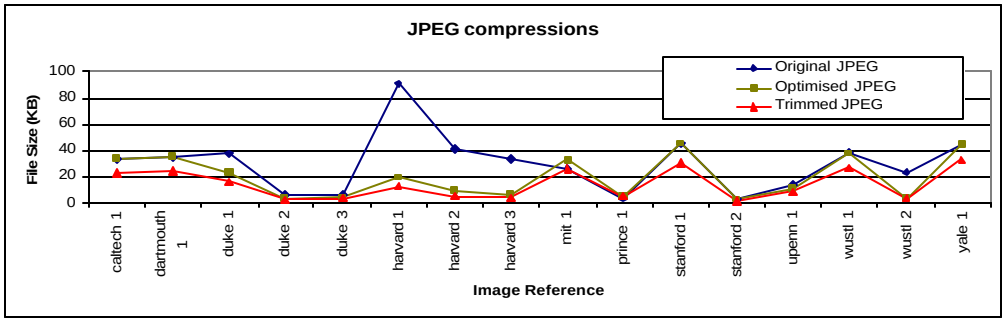

Figure 5: Effects of compression to JPEG images held by universities in the USA.

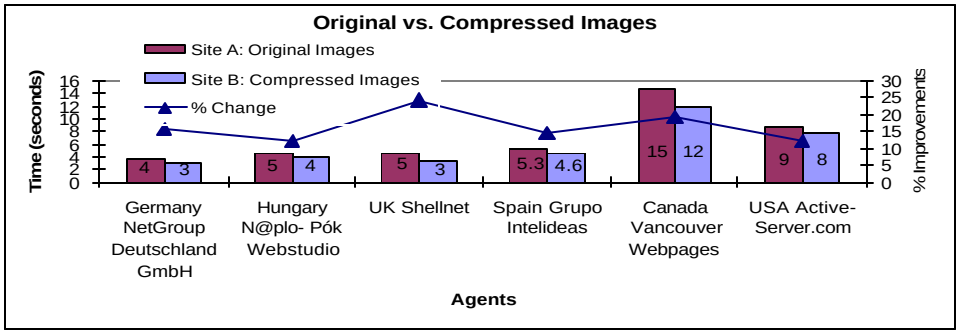

Figure 6: Effects of image compression to page loading times as seen by six agents.