



LAN Based Web Caching Q&A - Part One

Moore's law states that the number of transistors on an integrated circuit and therefore performance will double every 18 months. However, it takes fewer than 18 months for the Internet to double in size. This means that the increased processing power of the latest networking hardware cannot keep up with the increased burden on Internet backbones, routers and servers brought about by the rapid growth of the Internet. Problems of congestion on the Internet will therefore not be solved by improved hardware technology and increased bandwidth alone. Higher bandwidth Internet connections such as cable modem or ADSL only improve delivery of content from the ISP to the end user. They will not solve problems caused by overburdened servers or congested Internet routers. LAN based web caching however, can help improve this situation both for the user of the caching server and for the Internet at large.

We are presenting this information in a Q&A (Questions and Answers) format that we hope will be useful. Although our knowledge of this subject stems from development of our own web caching server products, much of the information contained in this document will apply to caching servers from other vendors. Please [click here](#) if you wish to download free trial web caching software.

We welcome feedback and comments from any readers on the content of this document or its usefulness.

We are providing the best information available to us as at date of writing and intend to update it at frequent intervals as things change and/or more information becomes available. However we intend this Q&A as a guide only and recommend that users obtain specific information to determine applicability to their specific requirements. (This is another way of saying that we can't be held liable or responsible for the content).

The full Q&A is divided into two parts. Part one is general in nature and less technical, Part two deals with more technical matters.

This KnowledgeShare Document addresses the different types of LAN based web caching techniques in use today. Vicomsoft have gained significant experience in the area of web caching and would like to make this information available to those interested in this subject. For those who would like to study this subject in more detail useful links are listed at the end of this document.

Part One: Questions

1. What is Caching?
2. What are the main benefits and trade-offs of caching?
3. What different types of caching exist?
4. How does LAN based web caching work?
5. How is the freshness of content controlled?
6. What are the advantages of LAN based web caching?

Content of this page in its entirety is protected by US & UK Copyright
© 2002 Vicomsoft Ltd

Reproduction in electronic and written form is expressly forbidden without written permission.



7. What is the difference between a proxy server and a caching server?
8. What exactly does a web caching server store copies of?
9. Does LAN based caching improve DNS lookup response times?
10. Do I still need caching if I have a high bandwidth connection such as cable modem, ADSL or T1?
11. Is web caching useful on a small LAN, or only on large corporate networks?
12. Is there any real cost justification for a web caching server?
13. Don't caching servers encourage copyright violation?
14. Are there any limitations that I should be aware of?
15. What is the bottom line? What does Vicomssoft recommend?

Part One: Answers

1. What is Caching?

Caching is a technique that is used to temporarily store a copy of information that has been requested by a piece of software or hardware. Cached information is generally stored closer to the requester than the permanent information is. It is also generally stored on a physical device capable more rapid delivery than the originating device. For example, computers have both disk and memory caches. Frequently used information can be retrieved much more rapidly from the memory cache than from the computer's main memory. Frequently accessed data can be retrieved from a disk cache much more quickly than if they were repeatedly retrieved from disk. This is because the disk cache is in RAM, which is much faster to access than a hard disk by several orders of magnitude. Web browsers also cache web pages on disk and in memory. Retrieving web pages from either the disk or memory cache is much faster than returning to the original site and retrieving them over the Internet.

2. What are the main benefits and trade-offs of caching?

Caching in all of its various forms offers faster access to cached content. Further, when cached resources are copied from remote locations, caching reduces the use of wide area bandwidth. There is a risk however that in some cases cached content may not be up to date. There are means of dealing with this situation that will be examined below.

3. What different types of caching exist?

The focus of this Q&A will be on LAN based web caching. There are, however, a number of different types of caching in common use today:

Carrier-class caching servers on the Internet store temporary copies of highly popular Web content, for delivery to thousands of ISPs, and reducing congestion on the Internet backbone.

Local Area Network (LAN) based caches are used in corporations, small business, schools and universities to improve network efficiency.

Individual Web browsers use disk and memory caching to store local copies of web pages.

Hard disk drivers cache frequently accessed information to improve access times.

Computers have memory caching systems which are faster to access than regular RAM.

Content of this page in its entirety is protected by US & UK Copyright

© 2002 Vicomssoft Ltd

Reproduction in electronic and written form is expressly forbidden without written permission.

4. How does LAN based web caching work?

A software program known as a web caching server is located on the LAN. All requests for web content are directed at this server. When it receives a request, it first looks in its own disk cache to see if the content is present and has not expired. If so, it delivers the content to the requester. If not, it fetches the content from the source, stores a copy of it on the local disk and delivers it to the requester. Some high performance caching servers store and deliver the content concurrently.

5. How is the freshness of content controlled?

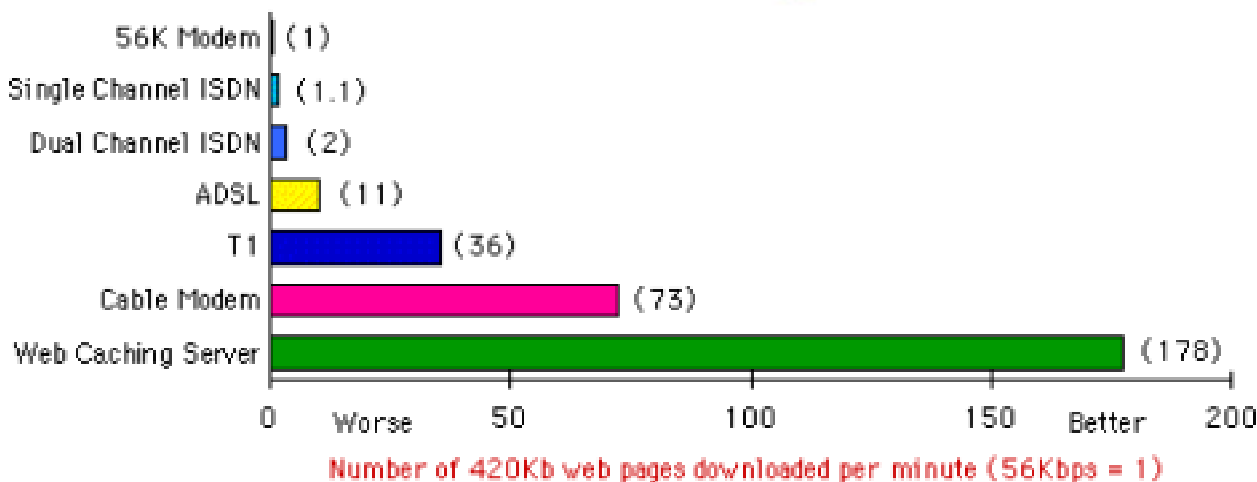
Web servers may deliver a caching directive with their web pages. The directive may instruct the web caching server not to cache the content. A frequently updated news page, or a page displaying stock quotes, for example, should not be cached at all. In this example the caching directive instructs the caching server whether or not to cache certain content, but not how long to keep it.

The caching directive may also contain additional information called "time-to-live". Some content can be cached indefinitely, while other content may only be valid for a month or an hour. A professional caching server will follow the originating server's caching directive when setting the time-to-live attribute of cached content. If the originating server instructs the caching server to store the content for a limited time, the caching server will stamp the content with an expiry date and time. Up to this date and time, the caching server will fulfil requests for the content from its own cache. After expiry, it will check to see if the cached content has been superseded before fulfilling the request.

The end result is that the user should always receive content identical to what would be delivered had the request been forwarded to the originating server.

6. What are the advantages of LAN based web caching?

The purpose of LAN based caching is to improve network efficiency by reducing the amount of traffic between the LAN and the Internet. The most obvious, and most frequently cited benefit is the shorter time required for the caching server to deliver cached content. Delivery time and therefore the end user experience are enhanced dramatically. For example, delivery of a 100KB web page from the originating server to the end user would take about 17 seconds over a 56Kbps modem, or 7.8 seconds over a dual channel Multilink PPP ISDN connection, assuming that there was no additional traffic congestion at the ISP or on the Internet backbone (this is not necessarily the case). The same page would take one second to deliver over a T1 connection, again assuming perfect Internet traffic conditions. However, this same page would be delivered from the caching server to the end user in about one tenth of a second *regardless of Internet traffic conditions*. This is the first and most obvious benefit of caching.



Some will point out that this is only a benefit if the same content is viewed a second time by a different user. If this does not occur, it may be argued, caching would be of no benefit. However, repeat visits to the same Web sites are more frequent than one may think. A recent test² representing a small business setup with a LAN comprising about 30 computers revealed that up to 70% of content delivered during any one hour period came from the cache, and as little as 30% came from the Internet. An independent laboratory study³ recently showed average response times reduced by 87% with the use of a caching server.

There are additional benefits as well. By delivering content from its own cache, the caching server reduces bandwidth use between the LAN and the Internet. This means that more bandwidth becomes available for users requesting fresh content directly from the Internet. These users experience improved response times even if they request content that is not stored in the cache.

Ironically, there are cases where a browser may display web content from the LAN based web cache faster than from its own disk cache. Because browsers are optimized for content delivered over the network, some may actually display a page delivered over the LAN more quickly than if the same page were read from their own computer's disk.

7. What is the difference between a proxy server and a caching server?

Most proxy servers are also caching servers, and many caching servers are also proxy servers. However, the term "proxy" and the term "caching" indicate two entirely different concepts.

A proxy is an agent that acts on behalf of another. In Internet technology this simply means that the proxy server is connected to the Internet on behalf of the end user's computer. The end user's computer is not in fact connected to the Internet at all, but rather to the proxy server. The distinction may appear trivial, but the implications are not. If the users on a LAN are not directly connected to the Internet, then they do not need public IP addresses. Further, they are isolated from intentional hostile attacks from the Internet, as the proxy server acts like a firewall. A whole LAN can use certain Internet services with only one official public address. The disadvantage is that all of the end user, or "client" computers must be individually and specifically configured to use the proxy server. If the proxy server is moved or renamed, all of the client computers must

Content of this page in its entirety is protected by US & UK Copyright
© 2002 Vicomsoft Ltd

Reproduction in electronic and written form is expressly forbidden without written permission.



be reconfigured. Fortunately, the use of a proxy server is not the only way to achieve the benefits of sharing an IP address or firewall protection. The Vicomssoft KnowledgeShare document on Network Address Translation addresses this subject in greater detail.

A cache on the other hand is a device that temporarily stores copies of information. Most proxy servers are also caching servers since this is a logical place to put a cache. However, a cache need not be a proxy, and there are products on the market that give end users the benefits of LAN based caching, Internet connection sharing and firewall protection without the inconveniences of proxy client configuration. If one is seeking the benefits of caching, as opposed to the benefits of proxy, it is important to remember not to restrict one's research to proxy servers. The results of such research may be disappointing.

8. What exactly does a web caching server store copies of?

This question is not as simplistic as one may think. A web caching server, unsurprisingly, stores copies of web pages. A web page however is made up of many elements. First, there is the HTML code that describes the page. If you want to know what HTML code looks like, go to a web page with your browser and choose "View Page Source" from the appropriate menu. This will open a text window with the HTML that describes the page. But this is only a very small part of what makes up a page. There may be many graphics, photos and illustrations that each may occupy kilobytes or megabytes of disk space. There may also be animation and Java scripts to instruct the browser to perform specific operations. Each one of these is a separate element and must be requested separately from the hosting server. Viewing the source code of a popular news site revealed 36 graphical elements from five different servers in addition to the server hosting the web page itself. For the technically minded, it may interest you to know that if that particular news page were cached, each of its individual elements would be stored separately as a disk file or as an object in a database.

9. Does LAN based caching improve DNS lookup response times?

Before a web browser can fetch a page, it must perform a DNS Lookup for each domain referenced on the web page. All of the content on a web page does not necessarily come from the same server. This is obviously the case for banner advertising that comes from an ad agency. In the above example the web browser would need to fetch information from six different servers. This would require six DNS lookups. In addition to the time it takes for a browser to fetch the page content, an additional delay is needed for each DNS lookup. Some advanced web caching servers will also cache DNS information. This effectively eliminates delays due to DNS lookups.

10. Do I still need caching if I have a high bandwidth connection such as cable modem, ADSL or T1?

It is a commonly held belief among users of modem Internet connections that by upgrading to a cable modem or ADSL connection, all of their bandwidth worries will disappear. Nothing could be further from the truth. Let us consider what happens when a user clicks on a hyperlink or types a URL in their browser:



1. The browser performs a DNS lookup which in and of itself can take several seconds.
2. The DNS lookup returns a numerical IP address that the browser requires in order to find the web page indicated by the hyperlink.
3. The browser then sends a request to its own gateway or router.
4. The gateway or router forwards the request to another router. This is repeated for as many routers as there are between the client and the server. When the request is forwarded, this is called a "hop". There can be up to ten or more hops in a connection. Internet congestion can cause the delay in a single hop to last for more than one second.
5. The server that is hosting the web page receives the request and responds by returning the object requested by the browser.
6. Step 4 is repeated in the opposite direction.
7. The browser receives the object, which is essentially a list of all of the other objects that make up the web page. It then goes through the list of items comprising the web page, and for each item (there may be 10, 20, 30 or more), steps 3 to 6 are repeated, including any delays between hops. Steps 1 and 2 may also be repeated for items to be retrieved from different servers.

This is a sample traceroute illustrating the number of hops required to get to a popular web site (www.ebay.com) from a user's browser.

Hop	Result	Min	Avg	Max	IP	Name
1	3/3	0.003	0.006	0.010	192.168.5.254	(none specified)
2	3/3	0.106	0.115	0.124	38.1.1.1	38-default-gw.psi.net
3	3/3	0.110	0.121	0.141	38.18.19.1	(none specified)
4	3/3	0.120	0.132	0.153	38.1.43.8	rc8.nw.us.psi.net
5	3/3	0.112	0.125	0.151	38.1.23.193	rc1.nw.us.psi.net
6	3/3	0.129	0.133	0.140	38.1.10.109	serial.portland.or.psi.net
7	3/3	0.132	0.149	0.159	209.1.169.25	ibr02-12-1-0.sntc01.exodus.net
8	3/3	0.127	0.142	0.150	216.33.147.65	dcr04-p0-0.sntc01.exodus.net
9	3/3	0.139	0.150	0.167	216.33.147.36	bbr02-g6-0.sntc01.exodus.net
10	3/3	0.131	0.157	0.195	209.185.249.142	bbr01-p5-0.sntc03.exodus.net
11	3/3	0.144	0.692	1.785	216.33.153.3	dcr03-g3-0.sntc03.exodus.net
12	3/3	0.145	0.155	0.171	209.185.84.70	(none specified)
13	3/3	0.226	0.272	0.302	216.32.179.198	(none specified)
14	3/3	0.361	0.366	0.371	10.1.2.3	(none specified)
15	3/3	0.233	0.266	0.300	216.32.120.113	tokay.ebay.com

As can be seen, to get from the user's browser to the eBay web, a request must travel across 15 hops. This took approximately three seconds to complete.



This eBay page consists of 25 images, in addition to the actual page. Each image has to follow the same (or similar) route as the original request for the home page shown in the traceroute.

This is the list of images to make up the page:

1. Image: http://pics.ebay.com/aw/pics/logo_home_tb.gif
2. Image: http://pics.ebay.com/aw/pics/h_category.gif
3. Image: <http://pics.ebay.com/aw/pics/new.gif>
4. Image: <http://pics.ebay.com/aw/pics/home/spacer.gif>
5. Image: http://pics.ebay.com/aw/pics/h_stats3.gif
6. Image: <http://pics.ebay.com/aw/pics/navbar/home-top.gif>
7. Image: http://pics.ebay.com/aw/pics/home/home_myebay_map_425.gif
8. Image: http://pics.ebay.com/aw/pics/yptc_mid.gif
9. Image: <http://pics.ebay.com/aw/pics/sellyouritem.gif>
10. Image: http://pics.ebay.com/aw/pics/news_chat.gif
11. Image: <http://pics.ebay.com/aw/pics/ww-homepage.gif>
12. Image: http://pics.ebay.com/aw/pics/p_register_tb.gif
13. Image: http://pics.ebay.com/aw/pics/h_feature2.gif
14. Image: <http://pics.ebay.com/aw/pics/more.gif>
15. Image: <http://pics.ebay.com/aw/pics/promo/ebayvisa1.gif>
16. Image: http://pics.ebay.com/aw/pics/home/welcome_widget.gif
17. Image: <http://pics.ebay.com/aw/pics/funstuff.gif>
18. Image: <http://pics.ebay.com/aw/pics/rosie1.gif>
19. Image: <http://pics.ebay.com/aw/pics/promo/Cover1.jpg>
20. Image: <http://pics.ebay.com/aw/pics/home/jpn-help.gif>
21. Image: <http://pics.ebay.com/aw/pics/88X31INF.gif>
22. Image: <http://pics.ebay.com/aw/pics/photonet.gif>
23. Image: <http://pics.ebay.com/aw/pics/usps8ac2.gif>
24. Image: <http://pics.ebay.com/aw/pics/bbb.gif>
25. Image: http://pics.ebay.com/aw/pics/truste_button.gif



Please note that this is a well designed page that makes moderate use of graphics. Many popular sites have twice this many elements.

Changing from a modem to a high bandwidth connection will only improve the time taken for the first hop or two. The remaining 13 hops will take exactly the same amount of time. Consider however, what happens when a user requests cached content from a LAN based caching server:

1. The browser performs a DNS lookup which in many cases is managed through DNS caching and takes a fraction of a second.
2. The browser sends the request to the caching server.
3. The caching server immediately returns the list of page items with no hops and no delays.
4. The browser goes through the list and requests all of the items from the caching server, which are in turn delivered immediately at LAN speeds.

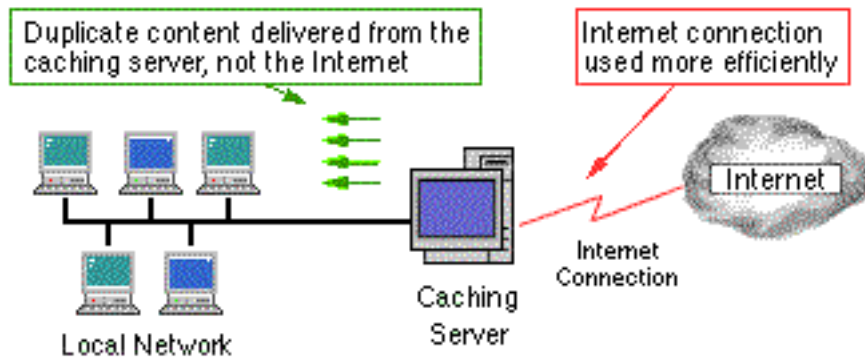
Many servers deliberately throttle back their throughput to prevent users from dominating their bandwidth throughput when viewing cumbersome graphics. This contributes to additional delays, and happens most frequently on sites with pages containing large elements. The larger the elements, and the more of them there are, the greater the benefit of caching. Thus, the performance gains experienced by users of web caching servers are potentially as substantial for T1 connections as they are for modem connections.

In our calculations we have not yet taken into account the distance travelled by the information. If the server is thousands of miles away and the page requested is made up of many components, the accumulated distance travelled by all of the requests and responses may add up to hundreds of thousands of miles. Signals travel at approximately 60% of the speed of light. Even with eutopic bandwidth and traffic conditions, it may take several seconds for content to travel such distances as opposed to near instantaneous LAN delivery.

Whatever the type of Internet connection, there are benefits to optimizing bandwidth usage, particularly as Internet content volumes increase.

11. Is web caching useful on a small LAN, or only on large corporate networks?

While the benefits of caching are more evident where a large number of users will potentially view the same content, studies indicate that even on LANs with as few as five users, performance gains can be impressive. This is particularly so in small focused workgroups where multiple users tend to use the same information sources.



12. Is there any real cost justification for a web caching server?

Time spent waiting for web pages to load is not used productively. If this waiting time is reduced, there are demonstrable cost savings. In addition, the bandwidth economy realized through the use of a caching server will in many cases allow a business to avoid upgrading to a higher bandwidth Internet connection. Cost savings achieved in this way can be considerable.

13. Don't caching servers encourage copyright violation?

Each time a user visits a web site, they are potentially viewing material that is protected by copyright. There is concern among certain providers of information, and among some politicians, that caching web content is tantamount to making illicit copies, and therefore a violation of copyright. It is not the purpose of this document to answer questions of a legal nature, yet users of caching servers have a legitimate concern in wishing to do the right thing. Caching servers are not intended to assist or encourage users to misuse or abuse copyrighted material. By complying with the caching directives that accompany HTML pages and the 'Conditions of Use' published by the Internet content providers, users can be sure that they are not infringing on copyrights in any way.

14. Are there any limitations that I should be aware of?

Some caching servers only cache web content. For many users this is not an issue, as the Web represents their primary if not unique use of the Internet. Some users however may require FTP or Gopher caching as well. This should be taken into account when choosing a caching server.

While web caching systems improve response times, they may also delay the retrieval of non-cached pages. This is due to the fact that the fresh content must first be stored in the disk cache then delivered to the requester. Vicomsoft solves this problem through a technique known as concurrent caching and delivery⁴. Other vendors may also be using similar techniques, so potential users may wish to take this into account in their evaluation criteria.

Streaming media by their nature cannot be cached. Some caching servers will not allow streaming media to pass through. Users behind such servers may find they do not have access to any streaming audio or video. Other servers will allow streaming media to pass through, but only to a single client. As soon as one user opens a connection to a streaming server, other



users would be denied access to any streaming media. There are some products however that allow multiple users to access streaming media through a caching server. Vicomssoft RapidCache is among them.

Many proxy caching servers may not allow access to any type of media other than web content. Again, users behind such proxy servers may find they do not have access to email, FTP or other Internet resources. Be certain to verify the support of any and all services you require before making a decision.

15. What is the bottom line? What does Vicomssoft recommend?

The advantages of using a caching server on a LAN to optimize network efficiency are significant:

By using a caching server, many businesses or schools may find they do not need to upgrade their Internet connection.

Improved content delivery can add up to hours per month of time saved by knowledge workers.

The Internet is far more pleasant to use when response times are prompt.

By using caching servers, enterprises and schools are being good Internet citizens, and avoiding wasteful bandwidth usage.

Internet congestion is going to get worse, not better. The most optimistic forecasts concerning improvements in server hardware, router hardware and bandwidth will not permit us to believe that this will solve the global Internet traffic jam. Intelligent resource allocation is necessary. Web caching is an essential part of any such strategy.

Most vendors of web caching server software offer free trial versions. We strongly recommend you try these for yourself before making a decision.

Notes to the text

1. Moore's Law, postulated in 1965 by Gordon Moore, co-founder of Intel Corp.
2. Study performed using Vicomssoft InterGate on an ethernet 10Mb/s LAN comprising 32 computers, all of which were connected to the Internet through a 256Kb/s leased line.
3. Source: ZD Labs.