Web Caching

What is Web Caching?

Cache (memory) is memory that is stored very close to the CPU, say on the same chip as the CPU, to allow fast access. Similarly, a disk cache is memory that is used to store frequently accessed disk pages for fast access. Web caching is the storage of Web objects near the user to allow fast access, thus improving the user experience of the Web surfer. Examples of some Web objects are Web pages (the HTML itself), images in Web pages, etc.

Web objects can be cached locally on the user's computer or on a server on the Web. There are several types of caches for Web objects:

o   Browser cache: Browsers' cache Web objects on the user's machine. A browser first looks for objects in its cache before requesting them from the website. Caching frequently used Web objects speeds up Web surfing. For example, I often use google.com and yahoo.com. If their logos and navigation bars are stored in my browser's cache, then the browser will pick them up from the cache and will not have to get them from the respective websites. Getting the objects from the cache is much faster than getting them from the websites.

The Netscape browser uses both a memory cache and a disk cache, whose sizes on my computer are set to 1 MB and 7.5 MB, respectively. Microsoft's browser, the Internet Explorer, is set to use a disk cache of 63 MB on my computer (there is no mention of a memory cache). A memory cache is faster than a disk cache, and the Netscape browser uses the two to form a small cache hierarchy.

o   Proxy cache: A proxy cache is installed near the Web users, say within an enterprise. Users in the enterprise are told to configure their browsers to use the proxy. Requests for objects from a website are intercepted and handled by the proxy cache. If they are not in the cache, the proxy gets them from another cache or from the website itself.

o   Transparent proxy cache: Using a "normal" proxy cache requires configuring the browser appropriately. A "transparent" proxy cache, on the other hand, intercepts browser Web requests without the browser being aware of the interception. Transparent proxies are placed at "gateways" so that all Web requests automatically go through the proxy. An example of a gateway is the server through which all enterprise Web traffic is funneled out to the Internet and back in.

o   Reverse (inverse) proxy cache: To reduce the load on a website, a proxy cache, called the "reverse" proxy, is placed in front of the website server(s).

The reverse proxy intercepts browser's requests to the websites. If the reverse proxy does not have the requested Web object, it gets the object from another cache or from the website itself.

Web objects can have an expiry time associated with them after which the object is considered to be "stale". A stale object is not used. If the object in the cache is stale, then it is equivalent to the object not being in the cache. An expiry date can be specified in the `http` header of a Web object. The expiry date is specified using `EXPIRES` and `CACHE-CONTROL http` headers.

Proxy caches come in two varieties: software that is installed on servers or separate boxes called appliances. Cache appliances contain only caching software and run on specialized operating systems fine-tuned for caching.

What are the Advantages of Web Caching?

Web caching has the following advantages:

o Faster delivery of Web objects to the end user.
o Reduces bandwidth needs and cost. It benefits the user, the service provider and the website owner.
o Reduces load on the website servers.

The Mechanics of Web Caching

Suppose that a user's browser needs an image for a Web page. The browser is caching, all its requests are funneled through a transparent proxy cache, and the website has a reverse proxy cache sitting in front of it:

o The browser checks to see if the image is cached locally. If yes, and the image is not stale, the browser uses the image from its cache. Otherwise, the browser sends the request for the image to the website. Since there is a transparent proxy cache, the request will be intercepted by the proxy cache.
o The transparent proxy cache checks to see if it has the image. If yes, and the image is not stale, the proxy cache sends the image to the browser, which in addition to using caches it. Otherwise, the proxy cache sends the request for the image to the website where it is intercepted by the reverse proxy cache. When the transparent proxy cache gets the image, it sends it to the browser and also caches it.
o The reverse proxy cache checks to see if it has the image. If yes, and the object is not stale, the reverse proxy cache sends the image to the requesting transparent proxy cache. Otherwise, the reverse proxy cache gets the image from the website, sends it to the requesting proxy cache, and caches the image.

Note that in each case, if the cache size is exceeded, the cache will have to throw out one or more cached objects so as to cache a new object. Typically the objects discarded are the ones that are used infrequently or ones that have not been used for a long time.

What are Some Issues in Caching?

o Cache policies: What objects are cached, when are objects removed from the cache, etc.

- o Cache hit rate: What percentage of the objects are found in the cache?
- o Cache size: How much cache is required for optimal performance?
- o Building a cache aware site: Do not use copies of objects in Web pages, use secure objects minimally since they are typically not cached, etc.
- o Scalability.

Internet Cache Protocols

Web caches use protocols, called the Internet cache protocols, to exchange information about the Web objects cached by them. Caches use this information to decide from where to retrieve a Web object. It can be more expedient to get an object from a neighboring cache rather than from the website. Two such protocols are:

- o ICP (Internet Cache Protocol).
- o HTCP (Hyper Text Caching Protocol) is newer than ICP and is better at predicting hits.

Where Can I Find More Information?

- o Web Caching Resources (http://www.web-cache.com).
- o Web Caching Tutorial & Resources (http://www.web-caching.com).
- o More Web Caching Resources (http://www.caching.com).