Cisco Cache Engine Frequently Asked Questions

Contents

See Also:

- **More Internet Products Tips**
- **Cisco Cache Engine Product Information**

Background

Q. What is web caching?

A. Web caching is the technique of storing frequently accessed information in a location close to the requester on storage that is physically or logically closer to the user--closer and faster than a lookup across the Internet or intranet. When a user accesses a web page, the caching system transparently saves the page’s graphics and HTML text on a local network storage device. That content is now "cached." When another user later accesses the same web page, the cached content is pulled from the local cache. This process improves download time for the user, and reduces bandwidth use on the network.

Q. What is the difference between the cache that’s built into my browser and a shared network caching system?

A. Internet browser applications allow an individual user to cache web content as files on the user’s local hard disk. A user can configure how much disk space should be devoted to caching web content. This method serves one user well, but does not benefit other users on the same network who are accessing the same web sites, so the overall benefits of reduced wide-area network (WAN) and server access traffic are negated.

In contrast, shared network caching stores the content on a network-level device. Here’s how this works. When user A visits a web site, the content is cached on the local network. When users B or C point to the same web page, the content can be provided from the local cache, improving response time and freeing up bandwidth that would have been used to access the web server and download the content. All users benefit by the fact that content is cached in the local network, and demand for WAN bandwidth decreases.

Q. How do I know that the content in the cache is "up-to-date"?

A. Any caching system must ensure that the user sees the same content from the cache as they would from the Internet or intranet. Both browser-based and network-level caches must address this issue.

The Cache Engine is compliant with the Hypertext Transfer Protocol (HTTP) 1.1 (RFC 2068). Cisco's compliance with this protocol enables cached data to be fresh. In addition, content providers using HTTP 1.1 have direct control over how long content is to be cached. The content author can set caching attributes for each object on a web page, as follows:

- Non-cachable
- Explicit time out
- OK to cache (this is the default setting)

HTTP 1.0 offers authors a "Pragma: no cache" option for web objects. In addition, users can explicitly refresh content at any time by using the browser's Reload button. All other data remains cached until it times out in the caching system.

Q. How effective is caching of real-time data, for example, pages containing stock quotes?

A. Many of the objects that make up a web page are static and can therefore be cached (for example, the banners and buttons). Other objects on the page are dynamic, and should not ever be cached (for example, stock prices). Thus, with the Cache Engine, only a few dynamic objects need to be retrieved from the host site while the static objects are fulfilled locally. This method provides an effective way to increase performance while maintaining real-time data.

Q. Why would anyone want to use web caching? What are the benefits?

A. By reducing the amount of traffic on WAN links and on overburdened web servers, caching provides significant benefits to Internet service providers (ISPs), enterprise networks, and end users. There are two key benefits:

- **Cost savings due to WAN bandwidth reduction--**Internet Service Providers can place Cache Engines at strategic points on their networks to improve response times and lower the bandwidth demand on their backbones. ISPs can station Cache Engines at strategic WAN access points to serve web requests from a local disk rather than by accessing the (distant or overrun) web servers. In enterprise networks, the dramatic reduction in bandwidth usage due to web caching would allow a lower-bandwidth (lower cost) WAN link to serve the same user base. Alternatively, the organization could add users or add more services to use the available bandwidth on the existing WAN link.

- **Improved productivity for end users--**The response of a local web cache is often three times faster than the download time for the same content over the WAN. End users see dramatic improvements in response times, and the implementation is completely transparent to them.

In addition, the Cache Engine provides network administrators with a simple, secure method to enforce a site-wide access policy through URL filtering.

Q. Who uses web cache solutions today?

A. Both large and small web caching solutions are in use today.

Large web caches are used primarily by service providers and large enterprises or institutions with thousands of users--for example, universities and Internet gateway sites around the world. Because so much Internet content is currently located in North America, institutions in Europe use web caching to extract maximum benefit from costly transatlantic links. Small and medium businesses and branch offices benefit from small-scale local web caching. Even sites with fewer than 20 users can use web caching to improve network response times and reduce WAN link congestion and costs.

Q. What are the problems with existing web cache solutions?

A. There are two fundamental problems with today's web cache solutions: *lack of transparency* and *lack of scalability*.

The first, and perhaps biggest, problem is the *lack of transparent solutions*. For current systems, in order to direct web requests to the local cache rather than to the intended server, someone must configure the client's web browser to point all requests to the cache. Although this isn't difficult for an experienced user, it involves touching the client, an activity that most network managers wish to avoid. It is just too difficult and costly to manage, support, and enforce client uniformity for any large-scale installation.

The second problem is *scalability.* Today's systems are limited in terms of cache size, growth, performance, and fault tolerance. Current web cache systems evolved from firewalling proxy servers, and network administrators face many of the same scaling constraints as they did with early firewalls. As a result, current web cache systems do not offer a growth path to many thousands of simultaneous connections, terabytes of page storage, and redundancy--they have been designed as single, centralized caches. Some have recently added simple multiple cache support as an afterthought.

Q. What are the features I should look for in a web cache?

A. The key features are:

1. **Transparency:** Transparent to user--no changes to browser required. Browser does not have to be in "proxy server mode."

2. **Scalability:** You should be able to add cache devices and see a linear improvement in both storage capacity and connection-level performance. The solution should also easily support hierarchical architectures.

3. **Ease of deployment and management:** Plug and play. Hierarchical organization should be simple to implement.

4. **Fault-tolerance:** If one of your web caches fails, other caches should pick up the load.

Caching should happen "off to the side of the main backbone," otherwise it creates a dangerous bottleneck ("bump in the cable") along the critical path, and loss of web connection if the cache fails.

Cisco Cache Engine

Q. What is the Cisco Cache Engine?

A. Cisco's Cache Engine is the only truly transparent, highly scalable, and easily deployed network solution that reduces WAN usage costs and accelerates web traffic.

The Cache Engine solution comprises two primary components:

- The Web Cache Control Protocol--a standard feature of the Cisco Internetwork Operating System (Cisco IOS).
- One or more Cisco Cache Engines that store the data in the local network to provide shared network caching. The Cache Engine is a specially designed, dedicated network appliance that performs content caching and retrieval using highly optimized algorithms. The engine communicates with the Cisco IOS router using the Web Cache Control Protocol, and uses special purpose thread-oriented real time software and a customized networking stack to achieve extremely high performance for individual Cache Engines. To achieve even higher performance, multiple engines can be added to a network, creating a cache farm.

Q. What is the Web Cache Control Protocol?

A. The Web Cache Control Protocol is a standard part of Cisco IOS that enables the communications between Cache Engines and the router. The Web Cache Control Protocol does not interfere with normal router operations. Using the Web Cache Control Protocol, the router directs all web requests to the Cache Engines (rather than to the intended host sites). The Web Cache Control Protocol also enables the Cache Engine to balance the traffic load across a cache farm and to ensure fault-tolerant and fail-safe operation.

The router also determines Cache Engine availability, and redirects requests to new engines as they are added to a cache farm. Each Cache Engine alerts the router of its presence when it comes on line upon installation or after a failure.

Q. What changes do users need to make in order to benefit from the Cache Engine?

A. None. No client configuration changes are required; the Cache Engine is a completely transparent solution. Client browsers operate exactly as they did before the system was deployed. The user notices only that web page access times improve (threefold in most cases).

To achieve this transparency, the Cisco IOS router uses the Web Cache Control Protocol to route only web requests--traffic on TCP port 80--to the Cache Engine. For all other traffic, the router functions entirely in its traditional role. The network routing intelligence is provided by the router, not by the client browser.

Q. How scalable is a Cache Engine installation?

A. The Cache Engine was designed from the ground up as a highly scalable system. The Cache Engine is a high-performance network appliance that uses special purpose caching and retrieval software to fulfill content requests. The real-time, embedded operating system running on the Cache Engine is designed to securely handle the many thousands of simultaneous, short lived transactions. This performance is required to enable the system to scale to a very high level. The system is not based on a general purpose computer--it does not have the administrative burden and the high context per process inherent in a UNIX system. UNIX systems cannot scale to large numbers of simultaneous processes in a graceful fashion. This is especially true of a networking system with performance requirements like those of the Cache Engine.

The Cache Engine single engine solution (the Cache Engine CE2050) provides 24 GB of cache storage. When placing a Cache Engine at a large Internet gateway router or at a central enterprise site, customers might wish to make use of a larger caching solution. The loosely coupled architecture of the system allows customers to add Cache Engines to a cache farm, scaling performance and storage linearly as Cache Engines are added. Installing a single Cache Engine CE2050 provides 24 GB of storage and 900 concurrent sessions; adding a second Cache Engine CE2050 provides 48 GB of storage and 1,800 concurrent sessions. The system supports cache farms of up to 32 Cache Engines. Thus, a single cache farm can provide 768 GB of cache storage and 28,800 simultaneous connections. Cache Engines in a farm need not be on the same subnet--the only requirement is that each subnet is directly attached to the router.

Because the Cache Engine is installed adjacent to a Cisco router, it does not interfere or become a bottleneck for non-web traffic. The router redirects web requests to the Cache Engine, but routes other types of traffic normally. In contrast, putting a web cache in the path of all traffic is not a scalable solution.

Q. What happens to my users' network access if a Cache Engine fails?

A. Because the Cache Engine is implemented with caching in parallel with the data--not in line with the data--network operation is not affected if a Cache Engine fails. Here's how the system handles partial failure (fault tolerant operation) and total failure (fail safe operation):

- **Fault-tolerant operation--**If any Cache Engine in a Cache Engine cache farm fails, the farm "heals" itself automatically. Under direction from the Web Cache Control Protocol, the remaining Cache Engines apportion the failed engine's load evenly among them. The cache farm continues operation using one less Cache Engine, but otherwise operation is unaffected.

  When the Cache Engine comes back online, it sends a Web Cache Control Protocol message to alert the router of its presence, and the Web Cache Control Protocol reassigns the engine its portion of the IP address space. The engine rebuilds its directory structure, and, as it receives requests, communicates with each other engine in the farm to send it content that was cached while it was down. This use of engine-to-engine data transfer speeds the recovery process by a factor of 100 over network access.

- **Fail-safe operation--**If an entire cache farm fails, the router falls back to normal operation, sending users' web requests to the actual destination web site in the traditional fashion. This loss of the entire cache farm can appear to users as an increase in download time for web content, but has no other effect on their work. This designed-in fail-safe response is made possible because the cache farm is not in line with the users' other network traffic.

When the cache farm comes back on line, the Cache Engines alert the router of their availability using the Web Cache Control Protocol, and the router resumes diverting web requests to the cache farm.

If the router fails, existing Cisco IOS fault tolerance and fail-safe mechanisms are applied. For example, a hot-standby router could take over operations, directing web requests to the cache farm.

Q. Is there an additional charge for the Web Cache Control Protocol in the Cisco IOS?

A. No. The Web Cache Control Protocol is included in Cisco IOS.

Q. How do I configure and manage the Cache Engine?

A. The system is designed to be extremely easy to configure and manage. To use the Web Cache Control Protocol, IP must be configured on the interface connected to the Internet and the interface connected to the Cisco Cache Engine. The interface connected to the Cisco Cache Engine must be an Ethernet or Fast Ethernet interface. The installation process involves connecting the Cache Engine to an Ethernet port on the router and, to initialize configuration, the Cache Engine prompts for information: assign an IP address, a subnet mask, and a default router to the Cache Engine. (See your Cisco IOS documentation for details about how to configure and enable the Web Cache Control Protocol on your router.)

After you have assigned those values, the Web Cache Control Protocol between the Cache Engine and the router automatically starts up, and you set additional configuration options using a standard web browser interface. Adding additional engines involves the same, simple process with no special configurations.

Q. Can I take advantage of this technology even if my network does not have Cisco routers?

A. At least one router must be Web Cache Control Protocol-aware; therefore it must be a Cisco IOS device. The other routers in the network need not be Cisco IOS devices.

Q. Do I need to upgrade all of my routers?

A. No. There must be at least one Web Cache Control Protocol-aware router in the data path. Consider the case of an enterprise network with a single router connection to the Internet. If that gateway router is Web Cache Control Protocol-aware, it, in conjunction with a cache farm, can serve the entire user community behind it. The other routers in the enterprise network do not need to be Web Cache Control Protocol-aware.

Q. I already have some web caching working in my network. How can I take advantage of the Cache Engine without throwing away the investment that I've already made?

A. The Cache Engine can operate alongside existing caching implementations. As long as you have a Web Cache Control Protocol-enabled router and at least one Cache Engine, the surrounding environment is not important. In addition, Cache Engines can operate in the standard "Proxy Mode." Please note that this eliminates many of the scalability and performance benefits unique to the Cache Engine.

Q. How can the Cisco Cache Engine be integrated into existing proxy cache environments?

A. The solution allows proxy made for inbound and outbound traffic. Network administrators can configure the cache farm to forward all requests to a traditional proxy server. The cache farm also accepts proxy requests. Thus, the cache farm can be placed at the tail, middle, or head of a proxy cache chain. In addition, the solution will support the Internet Cache Protocol (ICP) in the future.

Solutions and Applications

Q. Where do I deploy the Cache Engine in my service provider infrastructure?

A. Install a large cache farm at your main point of access to the Internet--all of your Points of Presence (POPs) will benefit. Client requests are fulfilled at the cache farm, avoiding traffic on your main Internet access links.

To further improve service to clients, you should deploy a Cache Engine or cache farm at each of your POPs. When a client serviced by that POP accesses the Internet, the request is routed to the POP cache farm. If the POP cache farm fulfills the request from local storage, the link to the central access point experiences less traffic.

If the POP cache farm is unable to fulfill the request, it makes a normal web request. This request, in turn, is routed to the cache farm at the main access point. If the request is filled by that cache farm, the traffic on the Internet main access line is avoided, and the client still experiences improved performance.

Q. Where do I deploy the Cache Engine in my enterprise network?

A. The Cache Engine is designed to operate in a hierarchical environment, providing the benefits of caching at every level in your network.

Install a Cache Engine near the router that is directly connected to the Internet at the top of your network hierarchy to reduce Internet access costs. You can also place a smaller system at remote offices that are connected to the main office via WAN links to reduce congestion at the central servers and to reduce bandwidth demand on those inter-site links.

This hierarchical arrangement is made possible because the system is transparent--clients do not have to point at any particular cache device in order to benefit from web caching.

Posted: Mon Apr 13 16:07:08 PDT 1998