SQUID CACHING PROXY

HEZRON MWANGI
Systems Administrator
hmwangi@kenet.or.ke

19th August 2013

kenet
Kenya Education Network

Introduction

- Squid is a caching proxy for the Web supporting HTTP, HTTPS, FTP, and more.
- Supports transparent proxying.
- Supports proxy hierarchies (ICP protocol).
- Squid is not an origin server!

Other proxies

- Apache with mod_proxy
- Internet Information Services
- nginx
- Privoxy
- WinGate
- Netscape Proxy
- Microsoft Proxy Server
- NetAppliance’s NetCache
- CacheFlow
- Cisco Cache Engine

What is a proxy?

- Is a server or an application that acts as an intermediary for requests from clients seeking resources from other servers.
- Internal users communicate with the proxy, which in turn talks to the Internet.
- Gates private address space (RFC 1918) into publicly routable address space.
- Allows one to implement policies:
  - Restrict who can access the Internet.
  - Restrict what sites users can access.
  - Provides detailed logs of user activity.

What is a caching proxy?

- Stores a local copy of objects fetched.
  - Subsequent accesses by other users in the organization are served from the local cache, rather than the origin server.
  - Reduces network bandwidth.
  - Users experience faster web access.

How proxies work (user request)

- User requests a page: http://training.kenet.or.ke/
- Browser forwards request to proxy.
- Proxy optionally verifies user’s identity and checks policy for right to access training.kenet.or.ke.
- Assuming right is granted, fetches page and returns it to user.

How proxies work (configuration)

- User configures web browser to use proxy instead of connecting directly to origin servers.
  - Manual configuration for older PC based browsers, and many UNIX browsers (e.g., Lynx).
  - Proxy auto-configuration file for Netscape 2.x+ or Internet Explorer 4.x+.
    - Far more flexible caching policy.
    - Simplifies user configuration, help desk support, etc.

How proxies work (configuration)

- User configures web browser to use proxy instead of connecting directly to origin servers.
    - Manual configuration for older PC based browsers, and many UNIX browsers (e.g., Lynx).
    - Proxy auto-configuration file for Netscape 2.x+ or Internet Explorer 4.x+.
        - Far more flexible caching policy.
        - Simplifies user configuration, help desk support, etc.

Squid’s page fetch algorithm

- Check cache for existing copy of object (lookup based on MD5 hash of URL).
- If it exists in cache.
  - Check object’s expire time; if expired, fall back to origin server.
  - Check object’s refresh rule; if expired, perform an If-Modified-Since against origin server.
  - If object still considered fresh, return cached object to requester.

kenet
Kenya Education Network

Squid’s page fetch algorithm cont'd

- If object is not in cache, expired, or otherwise invalidated.
  - Fetch object from origin server.
  - If 500 error from origin server, and expired object available, returns expired object.
  - Test object for cacheability; if cacheable, store local copy.

Cacheable objects

- HTTP
  - Must have a Last-Modified: tag.
  - If origin server required HTTP authentication for request, must have Cache-Control: public tag.
  - Ideally also has an Expires or Cache-Control: max-age tag.
  - Content provider decides what header tags to include.
  - Web servers can auto-generate some tags, such as Last-Modified and Content-Length, under certain conditions.
- FTP
  - Squid sets Expires time to fetch timestamp + 2 days.

Non-cacheable objects

- HTTPS, WAIS
- HTTP
  - No Last-Modified: tag.
  - Authenticated objects.
  - Cache-Control: private, no-cache, and no-store tags.
  - URLs with cgi-bin or ? in them.
  - POST method (form submission).

Implications for content providers

- Caching is a good thing for you!
- Make cgi and other dynamic content generators return Last-Modified and Expires/Cache-Control tags whenever possible.
  - If at all possible, also include a Content-Length tag to enable use of persistent connections.
- Consider using Cache-Control: public, must-revalidate for authenticated web sites.

Implications for content providers cont'd

- If you need a page hit counter, make one small object on the page non-cacheable.
- FTP sites, due to lack of Last-Modified timestamps, are inherently non-cacheable. Put (large) downloads on your web site instead of on, or in addition to, an FTP site.

Implications for content providers cont'd

- Microsoft’s IIS with ASP generates non-cacheable pages by default.
- Other scripting suites (e.g., Cold Fusion) also require special work to make cacheable.
- Squid doesn’t implement support for Vary: tag yet; considers object non-cacheable.
- Squid currently treats Cache-Control: must-revalidate as Cache-Control: private.

Transparent proxying

- Router forwards all traffic to port 80 to proxy machine using a route policy.
- Advantages.
  - Requires no explicit proxy configuration in the user’s browser.

Transparent proxying cont'd

- Disadvantages
  - Route policies put excessive CPU load on routers on many (Cisco) platforms.
  - Kernel hacks to support it on the proxy machine are still unstable.
  - Often leads to mysterious page retrieval failures.
  - Only proxies HTTP traffic on port 80; not FTP or HTTP on other ports.
  - No redundancy in case of failure of the proxy.

kenet
Kenya Education Network

Transparent proxying cont'd

- Recommendation: Don’t use it!
- Create a proxy auto-configuration file and instruct users to point at it.
- If you want to force users to use your proxy, either
  - Block all traffic to port 80.
  - Use a route policy to redirect port 80 traffic to an origin web server and return a page explaining how to configure the various web browsers to access the proxy.

squid.conf runtime settings

- Default squid.conf file is heavily commented! Read it!
- Must set:
  - cache_dir (one per disk).
  - cache_peer (one per peer) if participating in a hierarchy.
  - cache_mem (8-16M preferred, even for large caches).
  - acl rules (default rules mostly work, but must reflect your address space).

squid.conf runtime settings cont'd

- Recommendations
  - ipcache_size, fqdncache_size to 4096.
  - log_fqdn off (use Apache’s logresolve offline).
  - Increase dns_children, redirect_children, authenticate_children based on usage statistics (see cachemgr.cgi front-end).
  - Tweak refresh_pattern rules

squid.conf runtime settings cont'd

- Recommendations (cont'd).
- quick_abort_min 128 KB, quick_abort_max 4096 KB, quick_abort_pct 75.
- Tailor based on your bandwidth to the Internet.
- By default, squid will complete retrieval of any object requested, regardless of size; can burn considerable amounts of bandwidth!

Creating a proxy auto-configuration file

```
function FindProxyForURL(url, host)
{
        if (isPlainHostName(host) ||
                dnsDomainIs(host, ".cawtech.com"))
                return "DIRECT";

        if ((url.substring(0, 5) == "http:") ||
            (url.substring(0, 6) == "https:") ||
            (url.substring(0, 4) == "ftp:") ||
            (url.substring(0, 7) == "gopher:"))
                return "PROXY proxy.cawtech.com:3128; DIRECT";

        return "DIRECT";
}
```

Managing Squid

- Use Calamaris logfile analysis script, available at http://calamaris.cord.de/.
- Use modified MRTG/Cacti with Squid’s SNMP support to monitor.

Q&A.

?

kenet

Kenya Education Network

THANK YOU!