M1522.000800 System Programming, Fall 2019
Proxy Lab: HTTP and Proxy Servers with Cache
Assigned: Tue, Nov. 19, Due: Tue, Dec.3, 16:59

Introduction

An HTTP server (or web server) is a program that receives HTTP requests from a web browser, parses this information, then sends back the according HTTP response. This could be a simple html page, an image or something more complex. When we type a web address in the address bar of our browser we are sending an HTTP GET request to an HTTP server asking for the page with the given address. Examples of common HTTP servers in use today are apache or nginx.

Proxies are used for many purposes in the real world. Sometimes, proxies are used in firewalls, here the proxy is the only way for a web browser inside the firewall to contact a web server outside the firewall. A proxy may make modifications to external pages, for instance, by formatting them so they are viewable on a mobile phone. Proxies are also used as *anonymizers*. By stripping a request of all identifying information, a proxy can make the browser anonymous to the end server. Proxies can even be used to cache web objects, here they store a copy of a requested file when it is first requested, in future requests it can server the cached web object instead of going to the end server.

In this lab, you will be implementing two servers: a simple HTTP server and a simple proxy server.

- In the first part of the lab, you will write a simple sequential HTTP server that serves static content, e.g. html pages and images. This server waits repeatedly for HTTP GET requests, parses the requests, and, if the requested web object exists, sends it back to the client (browser) as part of an HTTP response. This part will give you some experience with network programming and the HTTP protocol.
- In the second part of the lab, you will implement a simple proxy server that will accept HTTP requests from the client (browser), forward these to the HTTP server, and then return the response of the HTTP server back to the client. Here the proxy is acting as a middleman between the two entities, anonymizing the client from the HTTP server.
- In the third part of the lab, you will add caching to your proxy that stores web objects in memory, and keeps a log of cache status of requested URI in a file in disk. The web objects being stored on a local proxy server are delivered much faster and uses less bandwidth that would require to download it from the destination server. If proxy caches are full, you need to implement the cache replacement policy to get new contents.

Hand Out Instructions

You can clone the proxylab repository from GIT (http://git.csap.snu.ac.kr/sysprog/proxy-lab.git). The four files you will be modifying are http.c (the HTTP server), proxy.c (the proxy server), cache.c and cache.h (the proxy cache). When you have completed the lab, you will commit all files containing your solution as well as a hard-copy of the report.

The proxylab directory contains the following code:

- `http.c`: This is the skeleton code of the HTTP server you will be modifying and handing in.
- `proxy.c`: This is the skeleton code of the proxy server you will be modifying and handing in.
- `cache.c`: This is the skeleton code of the proxy cache you will be modifying and handing in.
- `cache.h`: This file contains a few manifest constants, type definitions, and prototypes for the functions in `cache.c`. For the sake of good modularity, you can use `proxy.c` as a library.
- `csapp.c`: This is the file of the same name as the one described in the textbook. It contains error handling wrappers and helper functions such as the RIO (Robust I/O) package (textbook[1] 10.5), `open_clientfd` (textbook 12.4.4), and `open_listenfd` (textbook 12.4.7).
- `csapp.h`: This file contains a few manifest constants, type definitions, and prototypes for the functions in `csapp.c`.
- `Makefile`: Combines and links `proxy.c` and `csapp.c` into the executable `proxy`. Combines and links `http.c` and `csapp.c` into the executable `http`.

The `http.c` and `proxy.c` contain comments to guide you in your solution as well as some variable declarations that have been commented out. You can use these variables in your solution, you should not need any more to implement the requested functionality, but feel free to add any additional variables you feel are necessary. Your `http.c` and `proxy.c` files may call any function in the `csapp.c` file. However, since only the `http.c` and `proxy.c` files will be evaluated please **do NOT** modify the `csapp.c` file. If you want different versions of the functions found in `csapp.c`, write a new function in the `proxy.c` or `http.c` file.

Part I: Implementing an HTTP Server

In this part you will implement a sequential HTTP server. Your HTTP server should open a socket and listen for a connection request from a web browser. You can start your HTTP server by running:

```
unix> ./http 1234
```

[1] All textbook references are based on the second edition of the textbook

You may use any port number p, where $1024 \leq p \leq 65536$, and where p is not currently being used by any other system or user services. See `/etc/services` for a list of the port numbers reserved by other services on your system. You can send an HTTP request to your HTTP server by typing the following address into the browser address bar:

```
http://127.0.0.1:1234
```

Instead of 127.0.0.1 you may type localhost, this is a loopback address pointing to your own computer.

Before implementing the full HTTP server it is useful to change some settings in your browser, for simplicity please just use firefox for testing your solutions. Modern browsers often do automatic error correction on mis-typed address, or addresses that don't respond. This can be inconvenient when building your own web server. To disable this autocorrection type `about:config` into the address bar search for: `browser.fixup.alternate.enabled` and `network.captive-portal-service.enabled` and turn off both of these flags. This will stop the browser from autocorrecting your requests.

Your first task is to accept requests from the browser. Write a loop in the `main function` to accept connections and pass this to the `doit` function to handle them. Once this has been implemented, when sending a request to 127.0.0.1:1234 you should see the HTTP request being printed (probably multiple times) in your terminal. As your webserver doesn't respond to requests yet, the browser keeps sending requests as it believes there is a connection error. The next step is to parse the request sent by the browser. HTTP requests arrive in the form:

```
method URI version
Host: <host>:<port>
```

You are implementing the GET method, here you must find the file specified by the URI and send it back to the browser. You are also expected to handle 3 types of error:

```
501 - if the HTTP method is not GET
404 - if the requested page does not exist
403 - if the user requests a directory rather than a file
```

Firstly, you should implement these errors by inspecting the requested file on the local computer then passing the correct arguments to the `clienterror` function. Additionally a `parse_uri` function is provided to turn the requested URI to a path on your local machine e.g. ./pages/index.html. You can use this to find the requested file. Please do not modify the `clienterror` or `parse_uri` functions.

Once you can handle the aforementioned errors, if the file exists you should send the a requested file back to the web browser. This should be done by implementing the `serve_static` function.

Firstly you should build the response header, then send the file data (content). The format of the response should be:

```
version status-code status-message
Server: <serverName>
Content-Length: <contentLength>
Content-type: <contentType>
<content>
```

Since HTTP defines at last two ways the client can provide information about the length of the data it is sending (a "Content-Length" header or "Transfer-Encoding"), detecting end of file is probably easier. This http server uses "Content-Length". You can use the Content-Lenght in the hear to know client how large the file you are being sending.

Note that the two lines between `<contentType>` and `<content>` is represented as the string '\r\n'. The format in which the HTTP protocol should be implemented is specified in a Request for Comments (RFC) document, you can check: `https://tools.ietf.org/html/rfc2616` for a specification of the HTTP 1.1 protocol. You can use the `get_filetype` function to determine the type of the requested file, currently only html files are supported, please extend this function to support images: jpgs, pngs, and gifs. Once the response has been sent the client connection should be closed. You should now see the webpage or image displayed in your browser.

A directory called pages has been included in the handout, this contains a number of pages with which you can test your web browser. However please feel free to test with your own html and image files. The defined of behaviour some of these files is:

```
http://127.0.0.1:1234/pages/index.html -> send index.html to the browser
http://127.0.0.1:1234/pages/image.jpg -> send image.jpg to the browser
http://127.0.0.1:1234/pages/abcd.html -> 404 error
http://127.0.0.1:1234/pages/ -> 403 error
```

You can test your webserver with some of the files in the pages folder along with some of your own files. You have now built an HTTP server!

Part II: Implementing a Proxy Server

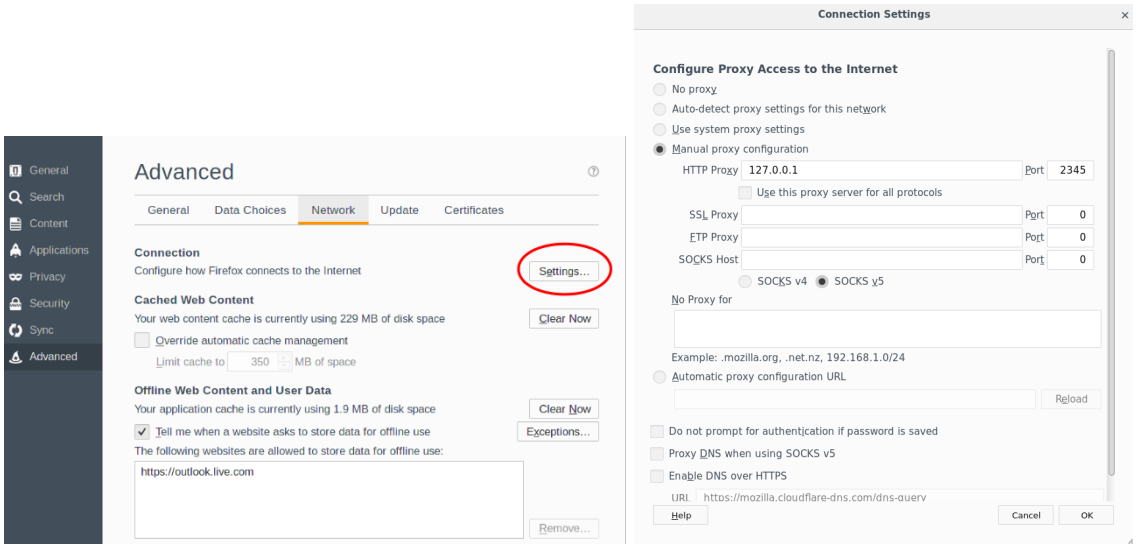
Now that you have built a web server that you can communicate with using your browser, you will implement a proxy server to communicate between the browser and http servers. Since your proxy is a middleman between the browser and the HTTP server, it will act as a server to the browser and a client to the HTTP server. Thus you will get experience with both client and server programming.

You can run the proxy file as shown below, you will have to choose a different port to your HTTP server as you will be running both simultaneously.

```
unix> ./proxy 5678
```

In order to get the browser to connect to your proxy server rather than other web servers directly you will have to change some of your browser settings. In Firefox, go to `Preferences -> Advanced -> Network` and click on the settings button next to connection. Change your settings as follows:

Click the 'Manual proxy configuration' checkbox. Set the HTTP proxy as 127.0.0.1 (your local machine) and set the port to the port number that you will run your proxy on. Don't click the checkbox 'use this proxy server for all protocols'. The proxy server in this lab session doesn't https protocol. Finally, remove '127.0.0.1' and 'localhost' from the 'No proxy for:' box. Now, firefox will send all HTTP requests to your



proxy server instead of connecting directly to the outside world, to test this you can try to visit any web page with firefox, you should no longer be able to connect to them.

You should now implement your proxy server to accept connections from your browser, forward these to your HTTP server, wait for a response, and then send this response back to the browser. To test local http server you made, you should start both your HTTP and your proxy server on different ports. You should then make a request to the HTTP server in your browser as in Section 1, your proxy will intercept this request. Accepting connections will work in the same way as with the HTTP server.

You should send a request to your HTTP server as before (http://127.0.0.1:1234/pages/index.html). The browser will send an HTTP request in the same format as before. Your first job is to parse this to determine the end destination of the request, you should then open a connection to that destination and then forward the request to there.

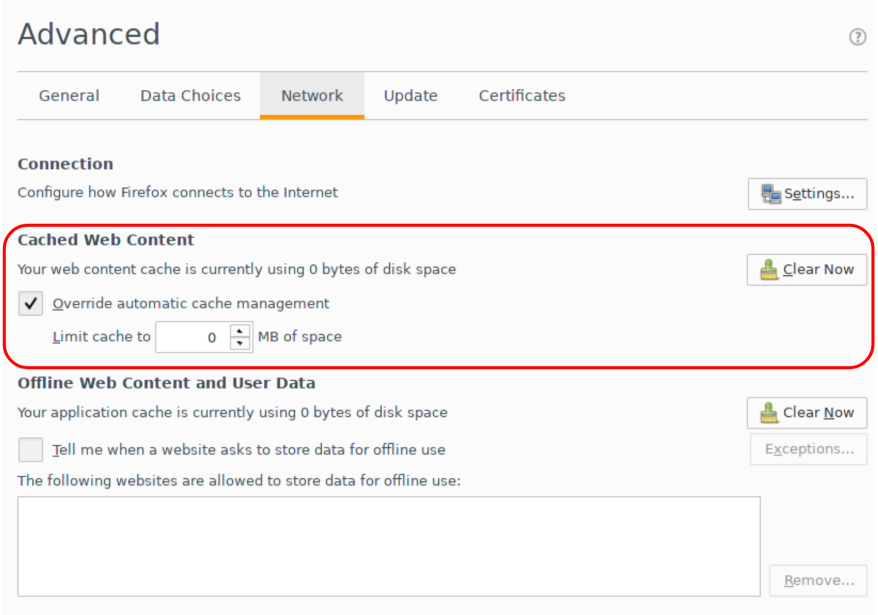
After forwarding the request from the client to the requested host, we should read the HTTP response sent back by the server. This is the example of response header sent back from web HTTP server in `http://info.cern.ch`:

```
HTTP/1.1 200 OK
Date: Thu, 11 Nov 2019 01:28:50 GMT
Server: Apache
Last-Modified: Wed, 05 Feb 2014 16:00:31 GMT
....
Content-Length: 646
Connection: close
Content-Type: text/html
...
```

After implementing the proxy you should be able to use your HTTP server in the same way as in part 1, i.e. by making requests with your browser configured to point at the proxy. This has the effect of anonymizing

the client from the end server, i.e. the HTTP server is not aware of your browser. You are currently running these locally on your machine but with different ports, but it should be possible to run your HTTP server on one machine, your proxy on another and your browser on yet another this would still work.

Part III: Caching and logging web objects



For the final part of the lab, you will add a cache to your proxy that stores recently used web objects in memory. HTTP actually defines a fairly complex model enabling web servers to specify how the objects they provide should be cached and clients can share their cache policy. However, your proxy will adopt a simplified approach.

When receiving a web object from a server, your proxy should cache it in memory as it is being transmitted to the client. If the client requests the same object from the same server, again, your proxy can simply resend the cached object without having to reconnect to the server.

Obviously, if your proxy were to cache every object that is ever requested, it would require an unlimited amount of memory. Moreover, since some web objects are larger than others, a single giant object could consume the entire cache, preventing other objects from being cached. To avoid those problems, your proxy should have both a maximum cache size and a maximum cache object size.

Your proxys cache should have the following maximum size:

`MAX_CACHE_SIZE = 1MB`

When calculating the size of its cache, your proxy must only count bytes used to store the actual web objects. Don not include additional bytes like metadata. Your proxy should only cache web objects that do not exceed the following maximum size:

`MAX_OBJECT_SIZE = 200kB`

For your convenience, both size limits are provided as macros in `cache.h`.You should use FIFO order for your cache replacement policy. You can use any data structure to implement your proxy server's cache such as linked-list, hash table.

In order to disable the automatic web cache provided by Firefox, you will have to change some of your browser settings. In Firefox, go to `Preferences -> Advanced -> Network`. Click the 'Override automatic cache management' checkbox. Set the 'Limit cache to '0'MB of the space.

Your proxy should keep track of the cache's activity in log files named `proxy.log`. Log entries should follow the following form:

`cacheStatus date requestedURL contentLength`

where `cacheStatus` provides information about whether the URL is currently cached, `requestedURL` indicates the URL requested by the client and `contentLength` is the size in bytes of the contents in the requested URL which is get from the response header. This is a simple example of how the log files from your proxy server should look like:

```
[uncached] Thu 14 Nov 2019 14:35:43 KST: http://localhost:1234/index.html 35
[cached] Thu 14 Nov 2019 14:35:45 KST: http://localhost:1234/index.html 35
```

The proxy server records the log when it is newly cached or cached contents are sent to the client. Note that size is essentially the number of bytes received from the end server, from the time the connection opened to the time it was closed.

Evaluation

- (15 Points) HTTP server. Your HTTP server should correctly accept connections from the browser, if the path is valid it should serve static html or image content to the browser. Based on the request it should also do some basic error handling (404, 403, 501 errors).
- (20 Points) Proxy server. Your proxy server should correctly accept connections from the browser and forward this to the correct server. It should then read the response from the end server and send this back to the browser.
- (30 Points) Proxy cache. Your proxy server should correctly insert, delete cache blocks according to cache replacement policy. Not only FIFO order but also other policies such as LRU are allowed to implement. Just make sure write comment in your code. Also, you have to take care the current size of the proxy cache.
- (10 Points) Cache log. Your proxy server should correctly log the status of cache when the client request to the proxy server. The format described above should be matched, otherwise your score can be penalized.
- (5 points) Code style. You can get up to ten points for well written and commented code. Your code should begin with a short block describing how your proxy works. Also, each function should have

a comment block describing what that functions does. Make sure you free any memory you allocate dynamically.

- (20 points) Report. In your report you should describe your implementation. Also, you should add a section regarding difficulties and suggestions for this lab, similar to the previous lab (shell lab).

Hand-in Instructions

To submit your solution simply commit, and push the changes to your remote repository.

```
devel@gentoo  /proxylab $ git add <list of modified files>
devel@gentoo  /proxylab $ git commit -m <commit message>
devel@gentoo  /proxylab $ git push origin master
```

We wons deadline. If for some reason you cannot push to the git server and there are 5 minutes before the deadline then, and only then, create a tar file of your submission and send it via email to the TAs.

Hints

- Use the RIO (Robust I/O) package (textbook 11.4) for all I/O on sockets. Do not use standard I/O on sockets. You will quickly run into problems if you do. However, standard I/O calls such as fopen and fwrite are fine for I/O on the log file.
- You will have to perform a lot of string manipulation in this lab, the sprintf and sscanf functions will be very helpful to you. When using sscanf you can use the string `"%[^:]"` to read in a string up until the `:` character.
- You can use the `stat` system call to find out information about a file based on it's file path.
- This is the one of the example web site that can test Content-Length in the http server. Don't too much focus on corner case.
  `http://www.columbia.edu/~fdc/sample.html -> Content-Length`
- You can check request/response header, network activity and detailed information in Firefox. Press hamburger bar on the right corner then click `Developer->Network` or type `Ctrl+Shift+Q`.