Spring 2021
Lab Assignment #6: Writing a Caching Web Proxy
Assigned: June 2
Deadline: June 16, 23:59:00

1 Introduction

A Web proxy is a program that acts as a middleman between a Web browser and an *end server*. Instead of contacting the end server directly to get a Web page, the browser contacts the proxy, which forwards the request on to the end server. When the end server replies to the proxy, the proxy sends the reply on to the browser.

Proxies are useful for many purposes. Sometimes proxies are used in firewalls, so that browsers behind a firewall can only contact a server beyond the firewall via the proxy. Proxies can also act as anonymizers: by stripping requests of all identifying information, a proxy can make the browser anonymous to Web servers. Proxies can even be used to cache web objects by storing local copies of objects from servers then responding to future requests by reading them out of its cache rather than by communicating again with remote servers.

In this lab, you will write a simple HTTP proxy that caches web objects. For the first part of the lab, you will set up the proxy to accept incoming connections, read and parse requests, forward requests to web servers, read the servers' responses, and forward those responses to the corresponding clients. This first part will involve learning about basic HTTP operation and how to use sockets to write programs that communicate over network connections. In the second part, you will upgrade your proxy to deal with multiple concurrent connections. This will introduce you to dealing with concurrency, a crucial systems concept. In the third and last part, you will add caching to your proxy using a simple main memory cache of recently accessed web content.

2 Handout instructions

Start by downloading `proxylab.tar` from eTL to your working directory of sp machine. Then give the command:

```
linux> tar xvf proxylab.tar
```

This will generate a handout directory called `proxylab`. The README file describes the various files.

3 Part I: Implementing a sequential web proxy

The first step is implementing a basic sequential proxy that handles HTTP/1.0 GET requests. Other requests type, such as POST, are strictly optional.

When started, your proxy should listen for incoming connections on a port whose number will be specified on the command line. Once a connection is established, your proxy should read the entirety of the request from the client and parse the request. It should determine whether the client has sent a valid HTTP request; if so, it can then establish its own connection to the appropriate web server then request the object the client specified. Finally, your proxy should read the server's response and forward it to the client.

3.1 HTTP/1.0 GET requests

When an end user enters a URL such as `http://www.cmu.edu/hub/index.html` into the address bar of a web browser, the browser will send an HTTP request to the proxy that begins with a line that might resemble the following:

```
GET http://www.cmu.edu/hub/index.html HTTP/1.1
```

In that case, the proxy should parse the request into at least the following fields: the hostname, `www.cmu.edu`; and the path or query and everything following it, `/hub/index.html`. That way, the proxy can determine that it should open a connection to `www.cmu.edu` and send an HTTP request of its own starting with a line of the following form:

```
GET /hub/index.html HTTP/1.0
```

Note that all lines in an HTTP request end with a carriage return, `'\r'`, followed by a newline, `'\n'`. Also important is that every HTTP request is terminated by an empty line: `"\r\n"`.

You should notice in the above example that the web browser's request line ends with `HTTP/1.1`, while the proxy's request line ends with `HTTP/1.0`. Modern web browsers will generate HTTP/1.1 requests, but your proxy should handle them and forward them as HTTP/1.0 requests.

It is important to consider that HTTP requests, even just the subset of HTTP/1.0 GET requests, can be incredibly complicated. The textbook describes certain details of HTTP transactions, but you should refer

to RFC 1945 for the complete HTTP/1.0 specification. Ideally your HTTP request parser will be fully robust according to the relevant sections of RFC 1945, except for one detail: while the specification allows for multiline request fields, your proxy is not required to properly handle them. Of course, your proxy should never prematurely abort due to a malformed request.

3.2 Request headers

The important request headers for this lab are the `Host`, `User-Agent`, `Connection`, and `Proxy-Connection` headers:

- Always send a `Host` header. While this behavior is technically not sanctioned by the HTTP/1.0 specification, it is necessary to coax sensible responses out of certain Web servers, especially those that use virtual hosting.

  The `Host` header describes the hostname of the end server. For example, to access `http://www.cmu.edu/hub/index.html`, your proxy would send the following header:

  ```
  Host: www.cmu.edu
  ```

  It is possible that web browsers will attach their own `Host` headers to their HTTP requests. If that is the case, your proxy should use the same `Host` header as the browser.

- You *may* choose to always send the following `User-Agent` header:

  ```
  User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:10.0.3)
      Gecko/20120305 Firefox/10.0.3
  ```

  The header is provided on two separate lines because it does not fit as a single line in the writeup, but your proxy should send the header as a single line.

  The `User-Agent` header identifies the client (in terms of parameters such as the operating system and browser), and web servers often use the identifying information to manipulate the content they serve. Sending this particular User-Agent: string may improve, in content and diversity, the material that you get back during simple telnet-style testing.

- Always send the following `Connection` header:

  ```
  Connection: close
  ```

- Always send the following `Proxy-Connection` header:

  ```
  Proxy-Connection: close
  ```

The `Connection` and `Proxy-Connection` headers are used to specify whether a connection will be kept alive after the first request/response exchange is completed. It is perfectly acceptable (and suggested) to have your proxy open a new connection for each request. Specifying `close` as the value of these headers alerts web servers that your proxy intends to close connections after the first request/response exchange.

For your convenience, the values of the described `User-Agent` header is provided to you as a string constant in `proxy.c`.

Finally, if a browser sends any additional request headers as part of an HTTP request, your proxy should forward them unchanged.

3.3 Port numbers

There are two significant classes of port numbers for this lab: HTTP request ports and your proxy's listening port.

The HTTP request port is an optional field in the URL of an HTTP request. That is, the URL may be of the form, `http://www.cmu.edu:8080/hub/index.html`, in which case your proxy should connect to the host `www.cmu.edu` on port 8080 instead of the default HTTP port, which is port 80. Your proxy must properly function whether or not the port number is included in the URL.

The listening port is the port on which your proxy should listen for incoming connections. Your proxy should accept a command line argument specifying the listening port number for your proxy. For example, with the following command, your proxy should listen for connections on port 15213:

```
linux> ./proxy 15213
```

You may select any non-privileged listening port (greater than 4500 and less than 65000) as long as it is not used by other processes. Since each proxy must use a unique listening port and many people will simultaneously be working on each machine, the script `port-for-user.pl` is provided to help you pick your own personal port number. Use it to generate port number based on your user ID:

```
linux> ./port-for-user.pl stu110
stu110: 24358
```

The port, p, returned by `port-for-user.pl` is always an even number. So if you need an additional port number, say for the `Tiny server`, you can safely use ports p and $p + 1$.

Please don't pick your own random port. If you do, you run the risk of interfering with another user.

4 Part II: Dealing with multiple concurrent requests

Once you have a working sequential proxy, you should alter it to simultaneously handle multiple requests. The simplest way to implement a concurrent server is to spawn a new thread to handle each new connection

request. Other designs are also possible, such as the prethreaded server described in Section 12.5.5 of your textbook.

- Note that your threads should run in detached mode to avoid memory leaks.
- The `open_clientfd` and `open_listenfd` functions described in the CS:APP3e textbook are based on the modern and protocol-independent `getaddrinfo` function, and thus are thread safe.

5 Part III: Caching 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 by which web servers can give instructions as to how the objects they serve should be cached and clients can specify how caches should be used on their behalf. However, your proxy will adopt a simplified approach.

When your proxy receives a web object from a server, it should cache it in memory as it transmits the object to the client. If another client requests the same object from the same server, your proxy need not reconnect to the server; it can simply resend the cached object.

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

5.1 Maximum cache size

The entirety of your proxy's cache should have the following maximum size:

```
MAX_CACHE_SIZE = 1 MiB
```

When calculating the size of its cache, your proxy must only count bytes used to store the actual web objects; any extraneous bytes, including metadata, should be ignored.

5.2 Maximum object size

Your proxy should only cache web objects that do not exceed the following maximum size:

```
MAX_OBJECT_SIZE = 100 KiB
```

For your convenience, both size limits are provided as macros in `proxy.c`.

The easiest way to implement a correct cache is to allocate a buffer for each active connection and accumulate data as it is received from the server. If the size of the buffer ever exceeds the maximum object size, the

buffer can be discarded. If the entirety of the web server's response is read before the maximum object size is exceeded, then the object can be cached. Using this scheme, the maximum amount of data your proxy will ever use for web objects is the following, where `T` is the maximum number of active connections:

```
MAX_CACHE_SIZE + T * MAX_OBJECT_SIZE
```

5.3 Eviction policy

Your proxy's cache should employ an eviction policy that approximates a least-recently-used (LRU) eviction policy. It doesn't have to be strictly LRU, but it should be something reasonably close. Note that both reading an object and writing it count as using the object.

5.4 Synchronization

Accesses to the cache must be thread-safe, and ensuring that cache access is free of race conditions will likely be the more interesting aspect of this part of the lab. As a matter of fact, there is a special requirement that multiple threads must be able to simultaneously read from the cache. Of course, only one thread should be permitted to write to the cache at a time, but that restriction must not exist for readers.

As such, protecting accesses to the cache with one large exclusive lock is not an acceptable solution. You may want to explore options such as partitioning the cache, using Pthreads readers-writers locks, or using semaphores to implement your own readers-writers solution. In either case, the fact that you don't have to implement a strictly LRU eviction policy will give you some flexibility in supporting multiple readers.

6 Evaluation

This assignment will be graded out of a total of 80 points:

- *Basic Correctness*: 40 points for basic proxy operation (autograded)
- *Concurrency*: 15 points for handling concurrent requests (autograded)
- *Cache*: 15 points for a working cache (autograded)
- *Report*: 10 points

6.1 Autograding

Your handout materials include an autograder, called `driver.sh`, that your instructor will use to assign scores for *BasicCorrectness*, *Concurrency*, and *Cache*. From the `proxylab-handout` directory:

```
linux> ./driver.sh
```

You must run the driver on a Linux machine.

6.2 Robustness

As always, you must deliver a program that is robust to errors and even malformed or malicious input. Servers are typically long-running processes, and web proxies are no exception. Think carefully about how long-running processes should react to different types of errors. For many kinds of errors, it is certainly inappropriate for your proxy to immediately exit.

Robustness implies other requirements as well, including invulnerability to error cases like segmentation faults and a lack of memory leaks and file descriptor leaks.

6.3 Test manually using curl

For checking your proxy working correctly with real pages, you can test your proxy manually. First, run your proxy with port allocated by `./port-for-user.pl`. Then using curl, you can retrieve the contents of some real pages. Only http pages are working correctly (Not for pages provided using https protocol). The lab lecture material give you some details about manual testing with curl.

7 Testing and debugging

Besides the simple autograder, you will not have any sample inputs or a test program to test your implementation. You will have to come up with your own tests and perhaps even your own testing harness to help you debug your code and decide when you have a correct implementation. This is a valuable skill in the real world, where exact operating conditions are rarely known and reference solutions are often unavailable.

Fortunately there are many tools you can use to debug and test your proxy. Be sure to exercise all code paths and test a representative set of inputs, including base cases, typical cases, and edge cases.

7.1 Tiny web server

Your handout directory the source code for the CS:APP Tiny web server. While not as powerful as `thttpd`, the CS:APP Tiny web server will be easy for you to modify as you see fit. It's also a reasonable starting point for your proxy code. And it's the server that the driver code uses to fetch pages.

7.2 `telnet`

As described in your textbook (11.5.3), you can use `telnet` to open a connection to your proxy and send it HTTP requests.

7.3 `curl`

You can use `curl` to generate HTTP requests to any server, including your own proxy. It is an extremely useful debugging tool. For example, if your proxy and Tiny are both running on the local machine, Tiny is

listening on port 15213, and proxy is listening on port 15214, then you can request a page from Tiny via your proxy using the following `curl` command:

```
linux> curl -v --proxy http://localhost:15214 http://localhost:15213/home.html
* About to connect() to proxy localhost port 15214 (#0)
*   Trying 127.0.0.1... connected
* Connected to localhost (127.0.0.1) port 15214 (#0)
> GET http://localhost:15213/home.html HTTP/1.1
> User-Agent: curl/7.19.7 (x86_64-redhat-linux-gnu)...
> Host: localhost:15213
> Accept: */*
> Proxy-Connection: Keep-Alive
>
* HTTP 1.0, assume close after body
< HTTP/1.0 200 OK
< Server: Tiny Web Server
< Content-length: 120
< Content-type: text/html
<
<html>
<head><title>test</title></head>
<body>
<img align="middle" src="godzilla.gif">
Dave O'Hallaron
</body>
</html>
* Closing connection #0
```

7.4 `netcat`

`netcat`, also known as `nc`, is a versatile network utility. You can use `netcat` just like `telnet`, to open connections to servers. Hence, imagining that your proxy were running on `sp1` using port 12345 you can do something like the following to manually test your proxy:

```
sh> nc sp1.snucse.org 12345
GET http://www.cmu.edu/hub/index.html HTTP/1.0

HTTP/1.1 200 OK
...
```

In addition to being able to connect to Web servers, `netcat` can also operate as a server itself. With the following command, you can run `netcat` as a server listening on port 12345:

```
sh> nc -l 12345
```

Once you have set up a `netcat` server, you can generate a request to a phony object on it through your proxy, and you will be able to inspect the exact request that your proxy sent to `netcat`.

7.5 Web browsers

Eventually you should test your proxy using the *most recent version* of Mozilla Firefox. Visiting `About Firefox` will automatically update your browser to the most recent version.

To configure Firefox to work with a proxy, visit

```
Preferences>Network>Settings
```

It will be very exciting to see your proxy working through a real Web browser. Although the functionality of your proxy will be limited, you will notice that you are able to browse the vast majority of websites through your proxy.

An important caveat is that you must be very careful when testing caching using a Web browser. All modern Web browsers have caches of their own, which you should disable before attempting to test your proxy's cache.

8 Handin instructions

The provided Makefile includes functionality to build your final handin file.

Before build your final handin file, you should modify **STUNO** defined in `Makefile`

Issue the following command from your working directory:

```
linux> make handin
```

The output is the file `../student-number-proxylab-handin.tar`, which you can then handin.

To submit your work, make a zip file with your tarball of your implementation and a report. Upload a zip file of your submission by eTL.

- Chapters 10-12 of the textbook contains useful information on system-level I/O, network programming, HTTP protocols, and concurrent programming.
- RFC 1945 (`http://www.ietf.org/rfc/rfc1945.txt`) is the complete specification for the HTTP/1.0 protocol.

9 Hints

- As discussed in Section 10.11 of your textbook, using standard I/O functions for socket input and output is a problem. Instead, we recommend that you use the Robust I/O (RIO) package, which is provided in the `csapp.c` file in the handout directory.
- The error-handling functions provide in `csapp.c` are not appropriate for your proxy because once a server begins accepting connections, it is not supposed to terminate. You'll need to modify them or write your own.

- You are free to modify the files in the handout directory any way you like. For example, for the sake of good modularity, you might implement your cache functions as a library in files called `cache.c` and `cache.h`. Of course, adding new files will require you to update the provided Makefile.
- As discussed in the Aside on page 964 of the CS:APP3e text, your proxy must ignore SIGPIPE signals and should deal gracefully with `write` operations that return EPIPE errors.
- Sometimes, calling `read` to receive bytes from a socket that has been prematurely closed will cause `read` to return `-1` with `errno` set to `ECONNRESET`. Your proxy should not terminate due to this error either.
- Remember that not all content on the web is ASCII text. Much of the content on the web is binary data, such as images and video. Ensure that you account for binary data when selecting and using functions for network I/O.
- Forward all requests as HTTP/1.0 even if the original request was HTTP/1.1.

Have Fun!