Lab Assignment: Web Proxy

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 used for many purposes. Sometimes proxies are used in firewalls, such that the proxy is the only way for a browser inside the firewall to contact an end server outside. The proxy may do translation on the page, for instance, to make it viewable on a Web-enabled cell 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, by storing a copy of, say, an image when a request for it is first made, and then serving that image in response to future requests rather than going to the end server.

In this lab, you will write a concurrent Web proxy that logs requests. In the first part of the lab, you will write a simple sequential proxy that repeatedly waits for a request, forwards the request to the end server, and returns the result back to the browser, keeping a log of such requests in a disk file. This part will help you understand basics about network programming and the HTTP protocol.

In the second part of the lab, you will upgrade your proxy so that it uses threads to deal with multiple clients concurrently. This part will give you some experience with concurrency and synchronization, which are crucial computer systems concepts.

Hand Out Instructions

Start by creating a new directory called, say, `proxylab` and then moving to it.

```
$ mkdir proxylab
$ cd proxylab
```

Then copy `proxylab-handout.tar` to this directory and unpack it:

```
$ cp /home/lperkovic/public/proxylab-handout.tar .
$ tar xvf proxylab-handout.tar
```

This will cause a number of files to be unpacked in the directory:

- `proxy.c`: This is the only file you will be modifying and handing in. It contains the bulk of the logic for your proxy.
- `csapp.c`: This is the file of the same name that is described in the CS:APP textbook. It contains error handling wrappers and helper functions such as the RIO (Robust I/O) package, `open_clientfd`, and `open_listenfd`, all of which we covered in class.
- `csapp.h`: This file contains a few manifest constants, type definitions, and prototypes for the functions in `csapp.c`.
- `Makefile`: Compiles and links `proxy.c` and `csapp.c` into the executable `proxy`.

Your `proxy.c` file may call any function in the `csapp.c` file. However, since you are only handing in a single `proxy.c` file, please don't modify the `csapp.c` file. If you want different versions of functions in in `csapp.c` (see the Hints section), write new functions in the `proxy.c` file.

Part I: Implementing a Sequential Web Proxy

In this part you will implement a sequential logging proxy. Your proxy should open a socket and listen for a connection request. When it receives a connection request, it should accept the connection, read the HTTP request, and parse it to determine the name of the end server. It should then open a connection to the end server, send it the request, receive the reply, and forward the reply to the browser if the request is not blocked.

Since your proxy is a middleman between client and end server, it will have elements of both. It will act as a server to the web browser, and as a client to the end server. Thus you will get experience with both client and server programming.

Logging

Your proxy should keep track of all requests in a log file named `proxy.log`. Each log file entry should be of the form:

```
Date: browserIP URL size
```

where `browserIP` is the IP address of the browser, `URL` is the URL asked for, `size` is the size in bytes of the object that was returned. For instance:

```
Sun 21 Feb 2016 23:15:52 CST: 99.120.120.51 http://www.bbc.co.uk/wwscripts/flag\
 417
```

Note that `size` is essentially the number of bytes received from the end server, from the time the connection is opened to the time it is closed. Only requests that are met by a response from an end server should be logged. We have provided the function `format_log_entry` in `csapp.c` to create a log entry in the required format.

Port Numbers

You proxy should listen for its connection requests on the port number passed in on the command line:

```
$ ./proxy 23456
```

In order to avoid conflicts with other students, I recommend that you use a 5 digit port number consisting of 2 followed by the last four digits of your student ID.

Part II: Dealing with multiple requests concurrently

Real proxies do not process requests sequentially. They deal with multiple requests concurrently. Once you have a working sequential logging proxy, you should alter it to handle multiple requests concurrently. The simplest approach is to create a new thread to deal with each new connection request that arrives.

With this approach, it is possible for multiple peer threads to access the log file concurrently. Thus, you will need to use a semaphore to synchronize access to the file such that only one peer thread can modify it at a time. If you do not synchronize the threads, the log file might be corrupted. For instance, one line in the file might begin in the middle of another.

Evaluation

- Basic proxy functionality (75 points). Your sequential proxy should correctly accept connections, forward the requests to the end server, and pass the response back to the browser, making a log entry for each request. Your program should be able to proxy browser requests to the following Web sites and (usually) correctly log the requests:
  - `http://reed.cs.depaul.edu/lperkovic/csc407/`
  - `http://www.cdm.depaul.edu`
  - `http://www.depaul.edu`
  - `http://www.google.com`
  - `http://www.bbc.com`
- Handling concurrent requests (25 points).

  Your proxy should be able to handle multiple concurrent connections. We will determine this using the following test: (1) Open a connection to your proxy using `telnet`, and then leave it open without typing in any data. (2) Use a Web browser (pointed at your proxy) to request content from some end server.

  Furthermore, your proxy should be thread-safe, protecting all updates of the log file and protecting calls to any thread unsafe functions such as `gethostbyaddr`. We will determine this by inspection during the demo.

Hints

- The best way to get going on your proxy is to start with the basic echo server and then gradually add functionality that turns the server into a proxy.
- Initially, you should debug your proxy using telnet as the client. To do this, login to perko406.cdm.depaul.edu and start the proxy as follows:

  ```
  $ ./proxy <your port number>
  ```

  Then, login to perko406.cdm.depaul.edu using another SSH session and initiate a telnet connection to your proxy web server:

  ```
  $ telnet 127.0.0.1 <your port number>
  Trying 127.0.0.1...
  Connected to localhost.
  Escape character is '^]'.
  ```

  and then type a HTTP request as:

  ```
  GET http://reed.cs.depaul.edu/lperkovic/csc407/ HTTP/1.0
  Host: reed.cs.depaul.edu
  ```

  Your proxy server should output:

  ```
  $ ./proxy <your port number>
  Thread 0: Received request from localhost (127.0.0.1):
  GET http://reed.cs.depaul.edu/lperkovic/csc407/ HTTP/1.0
  Host: perko406.cdm.depaul.edu

  *** End of Request ***

  Thread 0: Forwarding request to end server:
  GET /lperkovic/csc407/ HTTP/1.0
  Host: perko406.cdm.depaul.edu

  *** End of Request ***

  Thread 0: Forwarded 8192 bytes from end server to client
  Thread 0: Forwarded 5081 bytes from end server to client
  ```

  which shows the proxy receiving the HTTP get request from the client (at IP address 127.0.0.1), forwarding it to the end server (in this case perko406.cdm.depaul.edu), and forwarding the response from the end server back to the client. The response is output on the client:

```
HTTP/1.1 200 OK
Server: Apache-Coyote/1.1
Set-Cookie: JSESSIONID=DF8488CE680D3F29ADC72E7CC33F3659; Path=/lperkovic/; Http
Accept-Ranges: bytes
ETag: W/"14493-1454424723000"
Last-Modified: Tue, 02 Feb 2016 14:52:03 GMT
Content-Type: text/html
Content-Length: 14493
Date: Mon, 22 Feb 2016 04:21:52 GMT
Connection: close

<!DOCTYPE html PUBLIC "-//w3c//dtd html 4.0 transitional//en">
<html>
  <head>
    <meta http-equiv="Content-Type" content="text/html;
      charset=windows-1252">
    <meta http-equiv="Expires" content="Mon, 10 Jan 2000 12:00:00 GMT">
    <meta name="GENERATOR" content="Mozilla/4.75 [en] (X11; U; Linux
      2.2.14-5.0 i686) [Netscape]">
    <title>Systems II</title>
  </head>
  <body style="background-color: rgb(255, 255, 255);">
    <center>
      <h1>Systems II<br>
      </h1>
    </center>
    <center>
      <h1> Sections 801 and 810,  Winter 2016</h1>
    </center>
...
...
...
    <br>
    <br>
  </body>
</html>
Connection closed by foreign host.
```

The log file `proxy.log` will contain just one entry:

```
Sun 21 Feb 2016 22:28:34 CST: 127.0.0.1 http://reed.cs.depaul.edu/lperkovic/cs
07/ 14827
```

- Later, test your proxy with a real browser. You will need to change the setting for your browser to the the host and port of your proxy server. To do this with Firefox, go to Options, then Advanced, then Network, then Settings, then Manual Proxy Configuration, and then write down the proxy server hostname (perko407.cdm.depaul.edu) and proxy server port (e.g., ¡your port number¿). In Internet

Explorer, choose Tools, then Internet Options, then Connections, then LAN Settings. Check 'Use proxy server', and set your HTTP proxy. Safari, Chrome, and other browsers are set up in similar ways.

- Since we want you to focus on network programming issues for this lab, we have provided you with two additional helper routines: `parse_uri`, which extracts the hostname, path, and port components from a URI, and `format_log_entry`, which constructs an entry for the log file in the proper format.
- Be careful about memory leaks. When the processing for an HTTP request fails for any reason, the thread must close all open socket descriptors and free all memory resources before terminating.
- You will find it very useful to assign each thread a small unique integer ID (such as the current request number) and then pass this ID as one of the arguments to the thread routine. If you display this ID in each of your debugging output statements, then you can accurately track the activity of each thread.
- To avoid a potentially fatal memory leak, your threads should run as detached, not joinable.
- Since the log file is being written to by multiple threads, you must protect it with mutual exclusion semaphores whenever you write to it.
- Be very careful about calling thread-unsafe functions such as `inet_ntoa`, `gethostbyname`, and `gethostbyaddr` inside a thread. In particular, the `open_clientfd` function in `csapp.c` is thread-unsafe because it calls `gethostbyaddr`, a Class-3 thread unsafe function. You will need to write a thread-safe version of `open_clientfd`, called `open_clientfd_ts`, that uses the lock-and-copy technique when it calls `gethostbyaddr`.
- Use the RIO (Robust I/O) package 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.
- The `Rio_readn`, `Rio_readlineb`, and `Rio_writen` error checking wrappers in `csapp.c` are not appropriate for a realistic proxy because they terminate the process when they encounter an error. Instead, you should write new wrappers called `Rio_readn_w`, `Rio_readlineb_w`, and `Rio_writen_w` that simply return after printing a warning message when I/O fails. When either of the read wrappers detects an error, it should return 0, as though it encountered EOF on the socket.
- Reads and writes can fail for a variety of reasons. The most common read failure is an `errno = ECONNRESET` error caused by reading from a connection that has already been closed by the peer on the other end, typically an overloaded end server. The most common write failure is an `errno = EPIPE` error caused by writing to a connection that has been closed by its peer on the other end. This can occur for example, when a user hits their browser's Stop button during a long transfer.
- Writing to connection that has been closed by the peer first time elicits an error with errno set to EPIPE. Writing to such a connection a second time elicits a SIGPIPE signal whose default action is to terminate the process. To keep your proxy from crashing you can use the SIG_IGN argument to the signal function to explicitly ignore these SIGPIPE signals

1 Hand In Instructions

When you have completed the lab, you will submit it as follows:

```
$ make handin
```

Before you do, make sure you:

- Remove any extraneous print statements;
- Make sure that you have included your identifying information in `proxy.c`.