ICS(II), Spring 2021
Lab 3: Web Proxy
Assigned: 2021 May 18, Due: 2021 June 13, 11:59PM

Xie Dongfang(20212010037@fudan.edu.cn) is the lead person for this assignment. Due to some unknown legacy problem, the third part is deleted by some former TA. But considering the importance of cache, we added the third part at last. To make a compensation, we give you more time to finish it. Good luck!

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.

In the third and last part, you will add caching to your proxy using a simple main memory cache of recently accessed web content.

Logistics

You will work on this lab alone. Any clarifications and revisions to the assignment will be posted on the course Web page.

Hand Out Instructions

Check out the lab from the svn server. Username is your student ID and password has been sent to you before by email. Command: svn checkout svn://10.176.63.161/labInfo - -username=[Your student ID]. You can find all th following files in labInfo/proxylab/proxylab/.

- `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 (CS:APP 11.4), `open_clientfd` (CS:APP 12.4.4), and `open_listenfd` (CS:APP 12.4.7).
- `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.

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 a file 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 27 Oct 2002 02:51:02 EST: 128.2.111.38 http://www.cs.cmu.edu/ 34314
```

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:

```
unix> ./proxy 15213
```

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 (including other students' proxies). See `/etc/services` for a list of the port numbers reserved by other system services.

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 (CSAPP 13.3.8).

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.

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.

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.

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
```

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.

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.

Evaluation

This assignment will be graded out of a total of 70 points:

- Basic proxy functionality (30 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 correctly log the requests:
  - `http://www.baidu.com`
  - `http://www.sohu.com`
  - `http://www.qq.com`

- Handling concurrent requests (20 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`.

- Providing a working cache (10 points).

- Style (10 points). Up to 10 points will be awarded for code that is readable and well commented. Your code should begin with a comment block that describes in a general way how your proxy works. Furthermore, each function should have a comment block describing what that function does. Furthermore, your threads should run detached, and your code should not have any memory leaks.

Hints

- The best way to get going on your proxy is to start with the basic echo server (CS:APP 12.4.9) and then gradually add functionality that turns the server into a proxy.

- Initially, you should debug your proxy using telnet as the client (CS:APP 12.5.3).

- Later, test your proxy with a real browser. Explore the browser settings until you find "proxies", then enter the host and port where you're running yours. With Netscape, choose Edit, then Preferences, then Advanced, then Proxies, then Manual Proxy Configuration. In Internet Explorer, choose Tools, then Options, then Connections, then LAN Settings. Check 'Use proxy server,' and click Advanced. Just set your HTTP proxy, because that's all your code is going to be able to handle.

- 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 (CS:APP 13.3.6).

- Since the log file is being written to by multiple threads, you must protect it with mutual exclusion semaphores whenever you write to it (CS:APP 13.5.2 and 13.5.3).

- 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 (CSAPP 13.7.1). You will need to write a thread-safe version of `open_clientfd`, called `open_clientfd_ts`, that uses the lock-and-copy technique (CS:APP 13.7.1) when it calls `gethostbyaddr`.

- Use the RIO (Robust I/O) package (CS:APP 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.
- 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 (CS:APP 8.5.3) to explicitly ignore these SIGPIPE signals