PKU-ICS, Fall 2014
Proxy Lab: Writing a Caching Web Proxy

1 Introduction

A proxy server is a computer program that acts as an intermediary between clients making requests to access resources and the servers that satisfy those requests by serving content. A web proxy is a special type of proxy server whose clients are typically web browsers and whose servers are the same servers that browsers use. When a web browser uses a proxy, it contacts the proxy instead of communicating directly with the web server; the proxy forwards its client's request to the web server, reads the server's response, then forwards the response to the client.

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. A proxy may also perform translations on pages, for example, to make them viewable on web-enabled phones. Importantly, proxies are used 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

As usual, start by downloading the lab handout (`proxylab-handout.tar`) from Autolab and extracting it into the directory in which you plan to work—issue the following command:

```
sh> tar xvf proxylab-handout.tar
```

If possible, extract the files directly onto a linux machine; some operating systems and file transfer programs clobber Unix file system permission bits.

2.1 Robust I/O package

The handout contains the files `csapp.c` and `csapp.h`, which comprise the CS:APP package your textbook discusses. In addition to various error-handling wrapper functions and helper functions, the CS:APP package includes the robust I/O (RIO) package. You should use the RIO package instead of vanilla POSIX I/O functions, such as `read`, `write`, `fread`, and `fwrite`.

Do not feel obligated to use the RIO functions or anything at all from the CS:APP package. In fact, feel free to create your own version of a robust I/O package if you find the provided code deficient in any way. Moreover, keep in mind that the error-handling functions in CS:APP may not be appropriate for use in your proxy. Before blindly using wrapper functions or writing any of your own, carefully consider the proper action a server should take on each particular error (more about error handling can be found in section 7.2).

2.2 Modularity

The skeleton file `proxy.c` is provided in the handout. `proxy.c` contains a `main` function that does practically nothing, and you should fill in that file with the guts of your implementation. Modularity, though, should be an important consideration, and it is permissible and encouraged for you to separate the individual modules of your implementation into different files. For example, your cache should be largely (or completely) decoupled from the rest of your proxy, so one popular idea is to move the implementation of the cache into separate files like `cache.c` and `cache.h`.

2.3 Makefile

Since you are free to add your own source files for this lab, you are responsible for updating the makefile. The entire project should compile without warnings (you may want to use the `-Werror` flag), and you will want to determine the appropriate set of compilation flags (including optimization, linking, and debugging flags) for your final submission.

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 POSIX sockets

You will use the POSIX sockets API to handle network I/O in your web proxy. The `socket` function creates a socket to be used for network communication and returns a file descriptor. Once you have created a socket, you can manipulate it in various ways by passing its associated file descriptors to other functions in the socket library. These include `bind`, `listen`, `accept`, and `connect`. Some of the functions are to be used by servers and others by clients; the `man` page for `listen` has details.

Since the file descriptors that `socket` returns are no different from those that `open` do, you can use `read` and `write` as usual to send and receive data. Keep in mind that there are wrapper functions in the CS:APP package.

Also remember that a socket is to be used for two-way communication. You can and should read from and write to the same socket to interface with one client or one server.

3.2 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 a 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.

3.3 Request headers

Request headers are very important elements of an HTTP request. Headers are essentially key-value pairs provided line-by-line following the first request line of an HTTP request. Of particular importance for this lab are the `Host`, `User-Agent`, `Accept`, `Accept-Encoding`, `Connection`, and `Proxy-Connection` headers. Your proxy must perform the following operations with regard to the listed HTTP request 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 web server your proxy is trying to access. 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.

- 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. Your proxy must always send the provided `User-Agent` header to ensure consistent, well-behaved replies from web servers.

- Always send the following `Accept` header:

  ```
  Accept: text/html,application/xhtml+xml,
      application/xml;q=0.9,*/*;q=0.8
  ```

  Again, 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 `Accept` header identifies the types of content that the client is willing to accept, and web servers use the information when preparing replies. If your proxy does not send an `Accept` header, it is likely that many servers will only respond with text-only content, which is both boring and improper. The provided `Accept` header ensures that your proxy will receive content in a variety of interesting formats that servers support.

- Always send the following `Accept-Encoding` header:

  ```
  Accept-Encoding: gzip, deflate
  ```

  The `Accept-Encoding` header is related to the `Accept` header. This header alerts web servers that the client is willing to receive content that is compressed using the mechanisms listed in the header.

- 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.

With the exception of the `Host` header, your proxy should ignore the values of the request headers described above; instead, your proxy should always send the headers this document specifies.

For your convenience, the values of the described `User-Agent`, `Accept`, and `Accept-Encoding` headers are provided to you as string constants in `proxy.c`.

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

3.4 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 12345:

```
sh> ./proxy 12345
```

You will have to supply a port number every time you wish to test your proxy by running it. You may select any non-privileged port (greater than 1,000 and less than 64,000) as long as it is not used by other processes; using a port in the upper thousands (like 3070 or 8104) is suggested. 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 a reasonable port number. Use it to generate a port number based on your User ID:

```
sh> ./port_for_user.pl gxt@pku.edu.cn
gxt@pku.edu.cn: 44988
```

Consistently using the port number generated for you instead of using a random one each time you run your proxy will help you avoid trampling other students' ports.

4 Part II: Dealing with multiple concurrent requests

Production web proxies usually do not process requests sequentially; they process multiple requests in parallel. This is particularly important when handling a single request can involve a lengthy delay (as it might when contacting a remote web server). While your proxy waits for a response from the remote web server, it can work on a pending request from another client. Thus, once you have a working sequential proxy, you should alter it to simultaneously handle multiple requests.

4.1 POSIX Threads

You will use the POSIX Threads (Pthreads) library to spawn threads that will execute in parallel to serve multiple simultaneous requests. A simple way to implement concurrent request service is to spawn a new thread to process each new incoming request. In this architecture, the main server thread simply accepts connections and spawns off independent worker threads that deal with each request to completion and terminate when they are done. Other designs are also viable: you might alternatively decide to have your proxy create a pool of worker threads from the start. You may use any architecture you wish as long as your proxy exhibits true concurrency, but spawning a new worker thread for each request is the simplest and historically most common method.

The basic usage of Pthreads involves the implementation of a function that will serve as the start routine for new threads. Once a start routine exists, you can use the `pthread_create` function to create and start a new thread. New threads are by default joinable, which means that another thread must clean up spare resources after the thread exits, similar to how an exited process must be reaped by a call to `wait`. Luckily, it is possible to detach threads, meaning spare resources are automatically reaped upon thread exit. To properly detach threads, the first line of the start routine should be as follows:

```
pthread_detach(pthread_self());
```

4.2 Race conditions

While multithreading will almost certainly improve the performance of your web proxy, concurrency comes at a price: the threat of race conditions. Race conditions most often arise when there is a shared resource between multiple threads. You must find ways to avoid race conditions in your concurrent proxy. That will likely involve both minimizing shared resources and synchronizing access to shared resources. Synchronization involves the use of objects called locks, which come in many varieties. The Pthreads library contains all of the locking primitives you might need for synchronization in your proxy, including mutexes and semaphores. The `pthreads man` page has details.

As an example, one problem you will encounter is that of domain name resolution: to connect to web servers, your proxy will have to translate text hostnames to numeric IP addresses. The textbook suggests that you might use the function `gethostbyname` for host name resolution. Unfortunately `gethostbyname` happens to be thread-unsafe, meaning you will either have to find an alternative or create your own thread-safe variant. Since `gethostbyname` is obsolete anyway, it is probably best to use a newer alternative like `getaddrinfo`, but coming up with a thread-safe variant of `gethostbyname` is somewhat of an interesting exercise: to do it correctly, you will have to carefully consider why it is thread-unsafe in the first place.

5 Part III: Caching web objects

For the final part of the lab, you will add a cache to your proxy that will keep 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 will 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 must employ a least-recently-used eviction policy. 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.

A readers-writer lock may be instrumental for solving the concurrent cache access problem, but you should not attempt to implement your own. Instead, carefully search the libraries that are available for your use.

6 Evaluation

This assignment will be graded out of a total of 100 points, which will be awarded based on the following criteria:

- 30 points for basic proxy operation
- 30 points for handling concurrent requests
- 30 points for caching
- 10 points for style

6.1 Thread safety

A very important grading criterion is thread safety. Obviously, thread safety will play a significant role in your score for the second part of the lab, but be aware that it will also determine a large fraction of your score for the third part.

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 Style

Style points will be awarded based on the usual criteria. Proper error handling is as important as ever, and modularity is of particular importance for this lab, as there will be a significant amount of code. You should also strive for portability.

7 Testing and debugging

For this lab, 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 `curl`

You can use `curl` to generate HTTP requests to any server, including your own proxy. Because `curl` allows you to specify arbitrary request headers, you should be able to perfectly emulate any browser on any operating system. Moreover, with `curl`, you should be able to generate HTTP requests as if a browser were using your proxy to access web servers, giving you a neat way to determine what responses your proxy should be sending to its clients.

7.2 netcat

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

```
sh> nc catshark.ics.cs.cmu.edu 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`.

7.3 thttpd

`thttpd` is a tiny HTTP server that is available for download on the web. You can run `thttpd` on any port given you have sufficient permissions, and it will serve files as requested. You can use `thttpd` to test your proxy's ability to handle arbitrary content that you can make available as files to `thttpd`.

7.4 Tiny web server

The CS:APP code base includes the source code for its own tiny web server. While not as powerful as `thttpd`, the CS:APP web server will be easy for you to modify as you see fit.

7.5 Web browsers

Eventually you should test your proxy using actual web browsers, like Mozilla Firefox. It is possible to configure any modern browser to use an HTTP proxy, but since the provided user agent string is for Firefox, you should use Firefox for best results. In particular, the important HTTP request headers are drawn from the version of Firefox that is present on the Linux cluster machines and the shark machines. As a result, it may be best to perform at least your final testing on a Linux cluster machine.

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 submission for you. Issue the following command from your working directory:

```
sh> make submit
```

The output is the file `proxylab.tar.gz`. Please handin it in your autolab.

9 Resources

- 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.

10 Hints

- Remember that network byte order is big endian.
- Your proxy will have to do something about the generation of the `SIGPIPE` signal. The kernel will sometimes deliver a `SIGPIPE` to a process that has a handle to a socket which has been broken. Although the default action for a process that receives `SIGPIPE` is to terminate, your proxy should not terminate due to that signal.
- 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.
- Sometimes, calling `write` to send bytes on a socket that has been prematurely closed will cause `write` to return `-1` with `errno` set to `EPIPE`. Your proxy should not terminate due to this error.
- 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.