Lecture 17 - TCP/IP (last updated: 2023-11-16)

Internet protocol layers

The Internet is built on a layer of protocols:

Layer	Purpose	Examples
L5: Application	interpret transported data	HTTP, SMTP, SSH
L4: Transport	manage flow of network packets	TCP, UDP
L3: Network	route packets over networks of links	IP
L2: Link	send packets over physical connections	Ethernet, Wi-Fi
L1: Physical	provide physical connections	IEEE 802.3

Layer 4, the transport layer, provides a virtual pipeline called a TCP connection, through which a stream of bytes can be sent and received reliably between any two computers on the Internet. Internet applications treat L4 (and all the layers under it) as a black box, and interacts with L4 by using a set of C functions defined in the Sockets API (application programming interface).

There is quite a lot of boilerplate (repetitive) code that you need to write when using the Sockets API. Instead of explaining that code in detail, we will outline the steps taken to establish a TCP connection, and focus on context and background useful for navigating the Sockets API. Make sure to study the accompanying code examples for more details about how the API should be used.

TCP is the representative protocol in L4, and IP is the only protocol in L3, so the networking technology for the Internet is sometimes referred to as the "TCP/IP" networking. Also, when we write "IP" we are referring to IPv4, the Internet Protocol version 4. IPv6 addresses are structured slightly differently and are beyond the scope of this course.

TCP/IP networking

TCP/IP networking provides a reliable, two-way connection between computers (usually called "hosts" or "peers" in the context of networking). Here, "reliable" means that, as long as two hosts are connected, bytes sent between them are never dropped, duplicated, corrupted, or reordered (as might happen with other networking protocols); "two-way" means that both hosts can send or receive data.

Hosts are identified by an IP address, a 4-byte (32-bit) integer usually written in "dotted-quad" notation, where four decimal numbers, each representing a byte, are separated by periods, e.g., 34.145.159.110 and 127.0.0.1.

Hosts are also identified via hostnames, which can be a fully qualified domain name like "clac.cs.columbia.edu", or just "clac" within its local network. Hostnames are translated to IP addresses using Domain Name System (DNS).

Different programs running on the same hosts are distinguished by port numbers, a 2-byte (16-bit) integer. Well-known applications typically use fixed port numbers, e.g., port 80 for HTTP web servers, port 22 for SSH servers. Port numbers below 1024 are typically reserved for well-known applications.

TCP/IP follows a client-server model, where one host acts a server and the other as a client. At a high level, a TCP connection is established as follows:

- a server program "listens" on a certain port number
- a client program "connects" to the server using the server’s IP address and the port number the server program is listening on
- from that point onward, a bidirectional, two-way connection is established

netcat: TCP/IP swiss army knife

nc ("netcat") is a handy command-line tool for TCP/IP networking that exposes TCP connections to stdin/stdout. nc is convenient for quickly making TCP connections and is perfect for experimenting with networking.

You can use nc as either the server or the client in a TCP connection. For clarity, I will write "server$" as the prompt for the server machine and "client$" for the client machine (though they may be the same machine).

To use nc as a server, use the -l flag (short for "listen"):

```
    server$ nc -l <port-number>
```

To use nc as a client, simply specify the host and port number:

```
    client$ nc <hostname-or-IP-address> <port-number>
```

If you connect two instances of nc, what you type on one end should be visible on the other.

Some other useful flags:

-N: close the network connection upon EOF on stdin.

-C: translate \n to \r\n; useful for interacting with protocols that expect lines to end with \r\n, like HTTP.

File descriptors

Recall that the I/O functions in the Standard C library represents open files with ’FILE*’, and that stdin, stdout, and stderr are global FILE* objects representing the standard input, the standard output, and the standard error, respectively.

In UNIX, the native file handles are integers called file descriptors, and the Standard C library file handles are wrappers on top of UNIX file descriptors. The Standard C library functions fopen(), fread(), and fwrite() eventually call the native UNIX counterparts, open(), read(), and write() system calls. The UNIX systems calls take file descriptors instead of FILE* handles.

The file descriptors for the standard input, the standard output, and the standard error are 0, 1, and 2, respectively, and the subsequent calls to the open() system call will return file descriptors starting from 3.

Here is a program that reads bytes from the standard input and writes them to the standard output after capitalizing them (if it falls in lowercase alphabet):

```
int main() {
    char c;

    FILE *fp_in  = stdin;
    FILE *fp_out = stdout;

    while (fread(&c, 1, 1, fp_in) == 1) {
        c = toupper(c);
        fwrite(&c, 1, 1, fp_out);
    }
}
```

Here is the same program written using the native UNIX system calls:

```
int main() {
    char c;

    int fd_in  = 0;
    int fd_out = 1;

    while (read(fd_in, &c, 1) == 1) {
        c = toupper(c);
        write(fd_out, &c, 1);
    }
}
```

Sockets API

On POSIX systems, networking capabilities are made available to programs in the form of internet sockets. In particular, stream sockets represent an endpoint for reliable, bidirectional connections such as TCP connections. Stream sockets are associated with file descriptors that you can use with I/O system calls such as read(), write(), and close(). (There are also other kinds of sockets used for other kinds of connections, beyond the scope of this course.)

You can create a socket and obtain a file descriptor for it using the socket() system call (arguments omitted for brevity; see sample code for details):

```
int fd = socket(...);
```

fd is a file descriptor referring to the newly created socket, though this socket is not yet associated with any TCP connection. How we connect that socket depends on whether we are using it as a TCP client or server.

A TCP client forms a socket connection with a server by connect()ing the socket file descriptor to the server address:

```
int fd = socket(...);

connect(fd, ... /* server address */);

// Communicate with the server by read()ing from and write()ing to fd.

close(fd);
```

A TCP server is a little more complicated. We first need to bind() the socket to a server address and port, and then tell it to listen() for incoming connections; finally, we can accept() incoming connections:

```
    int serv_fd = socket(...);

    bind(serv_fd, ... /* server address */);

    listen(serv_fd, ... /* max pending connections */);

    for (;;) {
        int clnt_fd = accept(serv_fd, ...);

        // Communicate with the client by read()ing from and write()ing to
        // clnt_fd, NOT serv_fd.

        close(clnt_fd);
    }
```

Several things to note here:

- accept() will block until a client connects to the server.

- accept() returns the file descriptor for a NEW socket (clnt_fd) for communicating with the connected client.

- We can terminate the TCP connection with the client by close()ing clnt_fd.

- The server accept()s new connections in an infinite loop because that's how servers typically operate; we don't need to loop if we only care about connecting with a single client.

Socket address structures and network byte order

The connect() and bind() system calls require you to specify the server's internet address and port using a socket address structure, defined as follows:

```
    struct sockaddr_in {
        sa_family_t    sin_family;          // address family: AF_INET
        uint16_t       sin_port;            // port in network byte order
        struct in_addr sin_addr;            // internet address
    };

    struct in_addr {
        uint32_t       s_addr;              // address in network byte order
    };
```

Here, "network byte order" refers to the endianness of the 4-byte address and 2-byte port number. Network byte order is always big-endian, and contrasts "host byte order" which differs from machine to machine. The byte order of the address and port number must be explicitly specified so that it is consistent across networks where hosts do not all have the same endianness. The following functions convert 2- and 4-byte integers between network and host byte order:

```
    uint16_t htons(uint16_t host);          // "host-to-network, short"
    uint16_t ntohs(uint16_t net);           // "network-to-host, short"
    uint32_t htonl(uint32_t host);          // "host-to-network, long"
    uint32_t ntohl(uint32_t net);           // "network-to-host, long"
```

On big-endian hosts, these functions do nothing; on little-endian hosts, these functions convert the byte order accordingly.

We do not normally need to populate socket address structures ourselves, since nowadays helper functions will do so for us (e.g., getaddrinfo()), but it is still illustrative to see how one might do so (as one would "in the old days"):

```
    // Use 3157 as an example port number:
    uint16_t ip_port = 3157;

    // Construct 34.145.159.110 as a 4-byte integer:
    uint32_t ip_addr = 34 << 24 | 145 << 16 | 159 << 8 | 110;

    struct sockaddr_in addr;                // Define socket address struct

    addr.sin_family = AF_INET;              // Set internet address family
    addr.sin_port = htons(ip_port);         // Set port number
    addr.sin_addr.s_addr = htonl(ip_addr);  // Set IP address
```

Socket address polymorphism in C (optional)

Astute readers may have noticed from the man pages that the type signature of connect() does not take internet socket addresses of type struct sockaddr_in:

```
    int connect(int sockfd, const struct sockaddr *addr, socklen_t addrlen);
```

Instead, the second parameter accepts pointers to struct sockaddr! You will find a similar type mismatch with the bind() system call.

The sockaddr structure exists because sockets can be used to form connections other than IPv4 connections (e.g., IPv6), which use different kinds of socket address structures. Rather than make a different connect() and bind() system call for each kind of socket connection, POSIX exposes a single "polymorphic" interface that supports all socket address types, whose pointers are to be cast to the generic sockaddr structure:

```
    struct sockaddr {
        sa_family_t sa_family;      // address family
        char        sa_data[14];    // interpretation depends on sa_family
    };
```

The "address family" (i.e., sa_family) of all specific socket address structures (e.g., struct sockaddr_in, whose address family is AF_INET) is recorded in their first field, and is used by connect() and bind() to distinguish socket address types from one another (e.g., if they see addr->sa_family == AF_INET, they will cast the pointer type to struct sockaddr_in *).

Thus, when we manually build an internet socket address structure, we cast its pointer to struct sockaddr * before passing it to connect() or bind(), e.g.,:

```
    // Same ip_port and ip_addr as before

    struct sockaddr_in addr;                // Define socket address struct
    memset(&addr, 0, sizeof(addr));         // Zero-initialize addr’s bytes

    addr.sin_family = AF_INET;              // Set internet address family
    addr.sin_port = htons(ip_port);         // Set port number
    addr.sin_addr.s_addr = htonl(ip_addr);  // Set IP address

    // Obtain stream socket for TCP/IP connections, and connect() it to addr:
    int fd = socket(AF_INET, SOCK_STREAM, IPPROTO_TCP);
    connect(fd, (struct sockaddr *) &addr, sizeof(addr));
```

Sockets I/O with file descriptors and FILE pointers

Once you’ve established a TCP/IP connection on a socket, you can communicate with your peer using the socket file descriptor with the write() and read() system calls, e.g.:

```
    int write(int fd, const void *buf, size_t len);
    int read(int fd, void *buf, size_t len);
```

Since sockets encapsulate more complicated behaviors than regular files, POSIX also provides the send() and recv() system calls, which work like write() and read(), except they take an additional flags argument:

```
    int send(int sockfd, const void *buf, size_t len, int flags);
    int recv(int sockfd, void *buf, size_t len, int flags);
```

All of these system calls can be used interchangeably with TCP stream sockets; write() and read() are equivalent to send() and recv(), with the flag set to 0. We can also pass non-zero flags to send() and recv() to customize their behavior with respect to the underlying socket connection. For example, the MSG_WAITALL flag tells recv() to block until the entire buffer is filled (or until the connection is interrupted or disconnected); the MSG_DONTWAIT tells recv() not to block, and only read from buffered TCP packets the kernel has already received.

However, file descriptors and direct system calls can be more cumbersome to use than their FILE pointer counterparts. For instance, there isn’t any system call equivalent to fgets() or fprintf(). To overcome this limitation, we can "wrap" a file descriptor using fdopen(), which will give us a FILE pointer we can use with those functions:

```
    FILE *sock_fp = fdopen(sock_fd, "wb");

    fprintf(sock_fp, "Sending a formatted number: %4d\n", 42);

    // ...

    fclose(sock_fp);
```

When we fclose() the socket FILE pointer, fclose() will close() the underlying socket file descriptor for us, so there’s no need to do so separately.

Note that FILE pointers are block-buffered by default, so you may need to call fflush() to ensure any buffered output is actually sent through the socket connection to your peer (or just turn off buffering altogether using setbuf()).

Duplicating file descriptors to use fdopen() for reading & writing (optional)

One wrinkle with using FILE pointers to encapsulate socket file descriptors is that FILE pointers aren’t really designed for duplex file streams. In particular, mixing read and write operations with the same FILE pointer can be problematic in the presence of output buffering, so it’s best to create separate FILE pointers, one for reading, and one for writing:

```
    FILE *sock_fpw = fdopen(sock_fd, "wb");       // write-only FILE pointer
    FILE *sock_fpr = fdopen(dup(sock_fd), "rb");  // read-only FILE pointer

    // ...

    fclose(sock_fpr);
    fclose(sock_fpw);
```

Note that we first duplicate the socket file descriptor by calling dup() on sock_fd; this creates a new file descriptor referring to the same stream socket, which we give to fdopen(). Since we now have two file descriptors, we need to make sure we close() both of them by calling fclose() on their wrapping FILE pointers.