TCP/IP and Socket Programming

CS3214 Instructors

Virginia Tech

April 15, 2024
Goal: obtain working knowledge of TCP/IP (+UDP), including IPv4/IPv6, to become productive with writing simple network applications

Transport layer protocols: TCP and UDP
  - Use of ports
  - Demultiplexing in TCP/UDP

IPv4 addressing & routing
  - including subnets & CIDR

Protocol independence (IPv6)

BSD socket API
  - including utility functions for DNS name resolution
Transport and Network Layer

- Transport Layer Protocols: UDP and TCP
  - TCP: reliable data transmission
  - UDP: unreliable (best effort) data transmission
  - Port numbers are used to address applications
- Network Layer Protocols: IPv4 and IPv6
  - IP addresses are used to address hosts (*)
- Both protocols are designed to work with IP, hence the terms TCP/IP and UDP/IP

(*) technically, network interfaces - will explain difference shortly
User Datagram Protocol - UDP

- Specified in RFC 768 (1980)
- **simple**: specification is 2 pages
- **datagram oriented**: up to 64K messages
- **connectionless**: no connection setup required
- **unreliable**: best effort, makes no attempt to compensate for packet loss
- supports multicast

---

<table>
<thead>
<tr>
<th>Offsets</th>
<th>Octet</th>
<th>Bit</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Source port**
- **Destination port**
- **Length**
- **Checksum**

---

**Figure 2**: Source: WikiPedia
Transmission Control Protocol - TCP

Specification


- **point-to-point**: one sender, one receiver
- **reliable, in-order byte stream**: no “message boundaries”
- **pipelined**: transmission proceeds even while partially unack’ed data
- **send & receive buffers**: to hold this data
- **full duplex data**: bi-directional data flow in same connection
- **connection-oriented**: handshaking (exchange of control msgs) before data exchange
- **flow controlled**: sender will not overwhelm receiver
- **congestion controlled**: protects the network
### TCP Segment Header

<table>
<thead>
<tr>
<th>Offsets</th>
<th>Octet</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bit</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td>Source port</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td></td>
<td>Sequence number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td></td>
<td>Acknowledgment number (if ACK set)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>96</td>
<td>Data offset</td>
<td></td>
<td>Reserved</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>000</td>
<td>NS</td>
<td>C</td>
<td>W</td>
</tr>
<tr>
<td>16</td>
<td>128</td>
<td>Checksum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>160</td>
<td>Options (if data offset &gt; 5. Padded at the end with &quot;0&quot; bytes if necessary.)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3:** TCP Segment Header. Source: WikiPedia
Question:

How does process $A$ on host $H_1$ communicate with process $B$ on host $H_2$?

- Each stream is characterized by a quadruple $(A_s, P_s, A_d, P_d)$ where
  - $A_s, A_d$ are source and destination addresses - either a 32-bit IPv4 address or a 128-bit IPv6 address, e.g. 172.217.9.196 or 2607:f8b0:4004:807::2004
  - $P_s, P_d$ are 16-bit port numbers - there is one namespace per address + protocol combination, e.g. 80/tcp, 80/tcp6, 53/udp, 53/udp6. See /etc/services for commonly used port numbers

- Local vs remote/peer addresses are pairs $(A_s, P_s)$ or $(A_d, P_d)$ respectively, depending on perspective

- Demultiplexing (determining where to deliver incoming packets) requires full quadruple for TCP, but only $(A_d, P_d)$ for UDP
Figure 4: What’s wrong with this picture? Source: http://i.imgur.com/zXR0qAN.png
IPv4 Addressing

- IP addresses do not denote hosts, they denote interfaces (a host may have more than 1)
- Connected interfaces form a subnet whose addresses must share a common prefix
- Subnets are routing destinations
- No routing within subnet - can reach destination directly
- CIDR allows for up to 31 prefix bits:
  - 223.1.1.0/24 includes 223.1.1.0 – 223.1.1.255 (netmask 255.255.255.0)
  - 223.1.7.0/30 includes 223.1.7.0 – 223.1.7.3 (netmask 255.255.255.252)

Figure 5: Subnetting in IPv4
You are hired as a network administrator by a small company. You are given a small block of 256 addresses at 191.23.25.0/24. Hypothetically. As of 2020, all available IPv4 address space is assigned, and this belongs to Telefônica Brasil. You have to connect 2 LANs with 60/120 machines at 2 separate sites via PPP to an edge router at your ISP. Assign IP addresses to each subnet!
IPv4 Address Space Subdivision: Solution

Internet

R1

PPP Link 1
191.23.25.196/30

191.23.25.197

PPP Link 2
191.23.25.193

191.23.25.194

R2

Subnet address: 191.23.25.128/26
Default gateway: 191.23.25.129

191.23.25.129

R3

Subnet address: 191.23.25.0/25
Default gateway: 191.23.25.1

Ethernet

LAN 1
60 Machines

191.23.25.198

191.23.25.192/30

191.23.25.196/30

PPP Link 2

191.23.25.193

PPP Link 1

191.23.25.194

191.23.25.192/30

191.23.25.196/30

Ethernet

LAN 2
120 Machines

191.23.25.1

191.23.25.197
first introduced in BSD 4.1 Unix (1981), now de facto standard on all platforms
as a general interprocess communication (IPC) facility:
  • a host-local, application-created, OS-controlled interface (a “door”) into which application
    process can both send and receive messages to/from another application process
when used for network communication:
  • a door between application process and end-to-end transport protocol (UDP/TCP)
in Unix, sockets are file descriptors, so read(2), write(2), close(2) and
  others work
Bindings exist in many higher-level languages: e.g. java.net.Socket, Python
socket
Figure 6: Socket API calls used in typical UDP communication scenario
socket(2)

Usage:

```c
int socket(int domain, int type, int protocol);
```

- **domain**: PF_INET, PF_UNIX, PF_INET6, ...
- **type**: SOCK_DGRAM (for UDP), SOCK_STREAM (for TCP), ...
- **protocol**: 0 for Unspecified (or IPPROTO_UDP or IPPROTO_TCP)
- **returns integer file descriptor**
- **entirely between process and OS – no network actions involved whatsoever**
- **man pages**: ip(7), udp(7), tcp(7), socket(2), socket (7), unix(7) type “man 2 socket”, “man 7 socket”
bind(2)

Usage:

```c
int bind(int sockfd, struct sockaddr *my_addr, socklen_t addrlen);
```

- sockfd: return by socket()
- my_addr: “socket address” - this is the local address (destination for receive, source for send)
- addrlen length of address (address is variable-sized data structure)
- “binds” (reserves, associates with) socket to (local) address specified in the protocol’s namespace
- no information is transmitted over network
- one socket can be bound to one protocol/port, exceptions are
  1. multicast
  2. dual-bind same socket can bind to IPv4 and IPv6
struct sockaddr {
    /* GENERIC TYPE, should be "abstract" */
    sa_family_t sa_family;  /* address family */
    char sa_data[14];      /* address data */
};

/* This is the concrete "subtype" for IPv4 */
struct sockaddr_in {
    sa_family_t sin_family;  /* address family: AF_INET */
    u_int16_t sin_port;      /* port in network byte order */
    struct in_addr sin_addr; /* internet address */
};

struct sockaddr_storage {
    /* large enough to store addresses */
    sa_family_t sa_family;  /* address family */
    char sa_data[?];        /* address data */
};
IPv4 vs IPv6 addresses

/* Internet IPv4 address. */
struct in_addr {
    u_int32_t s_addr; /* address in network byte order */
};

/* IPv6 address */
struct in6_addr {
    union {
        uint8_t u6_addr8[16];
        uint16_t u6_addr16[8];
        uint32_t u6_addr32[4];
    } in6_u;
};

Good News

RFC 3493 functions for address manipulation mostly hide internal representations from the casual and professional socket programmer.
sendto(2), recvfrom(2), send(2), recv(2), connect(2)

```c
ssize_t sendto(int s, const void *buf, size_t len, int flags,
               const struct sockaddr *to, socklen_t tolen);

ssize_t recvfrom(int s, void *buf, size_t len, int flags,
                  struct sockaddr *from, socklen_t *fromlen);
```

- `s`, `buf`, `len` as in read/write
- `flags`: MSG_OOB, MSG_PEEK – mostly 0
- `to/from` are remote/peer addresses: where did the datagram come from, where should it be sent to
- NB: `fromlen` is value-result!
- Side note: can use `connect(2)` to set default address, then `send(2)/recv(2)`. 
TCP Socket API Call Sequence

- Left: client ("connecting socket"), Right: server ("listening socket")
- Server may accept multiple clients via multiple calls to accept, either sequentially or concurrently
- Independent directions: `read(2)/write(2)` may be used in any order.
- `read(2)/write(2)` or `recv(2)/send(2)` may be used
- Not shown: `shutdown(2)` for shutting down one direction
connect(2)

Usage:

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

- sockfd: returned by `socket()`
- peeraddr: peer address
- initiates handshake with server, sending SYN packet
- successful completion indicates successful handshake
listen(2), accept(2)

Usage:

```c
int listen(int s, int backlog);
int accept(int s, struct sockaddr *addr, int *addrlen);
```

- `addr`: accepted peer’s (aka client) address
- `listen()` must precede call to `accept()`
  - No network action, but informs OS to start queuing connection requests
- `accept()` blocks until client is pending, then returns new socket representing connection to this client; the passed in socket is ready to accept more clients on subsequent calls
The IPv6 Challenge

- IPv4 provides only 4 billion addresses, leading to address space exhaustion
- IPv6 was designed as a successor in the 1990’s
- ... but IPv6 is a separate network
  - A host may be connected via IPv4
  - ... or via IPv4 and IPv6
  - ... or only via IPv6
- Your network application must work in either case
  - Do not embed addresses or make assumptions about their size/format in your socket code
  - Let system tell you which address(es) you should use (as a client)/you should support (as a server)
Servers provide both IPv4 and IPv6, clients prefer IPv6 to IPv4 when both are available, eventually IPv4 connections will die out ... will it happen?

IPv6 adoption among users accessing Google services, Feb 24 2020
int getaddrinfo(const char *node, const char *service, 
const struct addrinfo *hints, struct addrinfo **res);

- Use `getaddrinfo()` to obtain information about suitable address families and addresses
  - For servers to bind to (IPv4, or IPv6, or both): if AI_PASSIVE is set and node == NULL
  - For clients to connect to (based on DNS name or specified address notation); based on RFC 3484 (now RFC 6724) ordering
- Use `getnameinfo()` to transform addresses in printable form
- Mostly correct tutorial at http://www.akkadia.org/drepper/userapi-ipv6.html, except for pesky issue of how to support both families as a server
  - can use so-called dual-bind feature (with care, Linux-only)
  - portable solution is to use 2 separate sockets.
References

RFC 7414, February 2015.