Sockets and Parallel Computing

CS439: Principles of Computer Systems
November 14, 2018
Bringing It Together

• Networks are collections of boxes (computers and specialized computers) and cables that move data
  – Networks themselves are disparate and have their own protocols
    • These protocols move data within a network
  – Internets smooth out those differences and allow us to communicate between the different networks
    • Internet protocols move data from source machine to destination machine
  – The Internet is an example of an internet

• When data arrives at a machine, the OS is responsible for delivering it to the correct process
  – UDP and TCP/IP help with this
  – TCP provides an illusion of reliability---the OS is responsible for implementing this reliability
  – With reliability comes danger: if we aren’t careful, we could overload an already congested network and cause it to collapse
  – TCP has algorithmic pieces to help with this: additive increase, multiplicative decrease; slow start; reaction to timeout events; round trip time variance (initial timer setting); and retransmit backoff
Transferring Data Over an Internet

LAN1

Host A

client

protocol software

LAN1 adapter

(1) data

internet packet

LAN1 frame

(2) data PH FH1

Host B

server

protocol software

LAN2 adapter

(8) data

(7) data PH FH2

Router

LAN1 adapter

LAN2 adapter

(6) data PH FH2

(5) data PH FH2

(4) data PH FH1

(3) data PH FH1

PH: Internet packet header
FH: LAN frame header
Network Costs

• Overhead: CPU time to put the data on or pull it off the wire
• Latency: time for one byte to go from one place to another
  – Latency from New York to San Francisco:
    • $3000 \text{ miles} \times \frac{1 \text{ sec}}{186,282 \text{ miles}} = 15 \text{ ms}$
• Throughput: Maximum bytes per second (bandwidth)

• *Note*: *Bandwidth is not the whole story!*
• Key to good performance:
  – LAN: minimize overhead
  – WAN: keep the pipeline full!
Transferring Data Over an internet: Text Description

• Example has 2 LANs connected by a router
• Terminology:
  – PH: Internet packet header
  – FH: LAN frame header
• Steps to transfer:
  1. Data is sent from Host A’s client application to protocol software
  2. The PH and FH1 are appended to the data and sent along to the LAN1 adapter
  3. LAN1 adapter sends Data + PH + FH1 along to the router’s LAN1 adapter
  4. Data + PH + FH1 sent to router’s protocol software, which strips off FH1 and creates FH2
  5. Data + PH + FH2 sent to router’s LAN2 adapter
  6. Data + PH + FH2 sent to LAN2 adapter in LAN2
  7. Data + PH + FH2 sent along to protocol software where the PH and FH2 are stripped off
  8. Data goes along to Host B’s server application
Today’s Additions

• Accessing the Network
  – Sockets
    • Client-Side Programming
    • Server-Side Programming

• Parallel Computing
  – Description of Parallel Computing
  – Parallel Programming Models
    • Shared Memory
    • Message Passing
Hardware and Software Organization of an Internet Application

Hardware and firmware

Global IP Internet

Internet client host

Internet server host

TCP/IP

Client

User code

TCP/IP

Server

Network adapter

Hardware and firmware

Network adapter
Internet Connections

• Clients and servers communicate by sending streams of bytes over connections
• A *socket* is an endpoint of a connection
  – Underlying basis for all Internet applications
  – Created in the early 80s as a part of the original Berkeley distribution of UNIX that contained an early version of the Internet protocols
  – Socket address is an IP address:port pair
• A *port* is a 16-bit integer that identifies a process:
  – *Ephemeral port*: Assigned automatically by client kernel when client makes a connection request.
  – *Well-known port*: Associated with some service provided by a server (e.g., port 80 is associated with Web servers)
Anatomy of an Internet Connection

Client socket address
128.2.194.242:51213

Server socket address
208.216.181.15:80

Connection socket pair
(128.2.194.242:51213, 208.216.181.15:80)

Client host address
128.2.194.242

Server host address
208.216.181.15

51213 is an ephemeral port allocated by the kernel

80 is a well-known port associated with Web servers
Well-known Ports and Service Names

• Popular services have permanently assigned *well-known ports and corresponding well-known service names*:
  – echo server: 7/echo
  – ssh servers: 22/ssh
  – email server: 25/smtp
  – Web servers: 80/http

• Mappings between well-known ports and service names is contained in the file `/etc/services` on each Linux machine.
Sockets: Implementation Details

• What is a socket?
  – To the kernel, a socket is an endpoint of communication
  – To an application, a socket is a file descriptor that lets the application read/write from/to the network
    • Remember: All Unix I/O devices, including networks, are modeled as files

• Clients and servers communicate with each other by reading from and writing to socket descriptors

• The main distinction between regular file I/O and socket I/O is how the application “opens” the socket descriptors
Interfacing with DNS

Functions for retrieving host entries from DNS:

**getaddrinfo()**: query key is a DNS domain name, returns:

```c
struct addrinfo {
    int           ai_flags;
    int           ai_family;      /* desired address family */
    int           ai_socktype;    /* desired Layer 4 protocol */
    int           ai_protocol;
    socklen_t     ai_addrlen;
    struct sockaddr *ai_addr;
    char           *ai_canonname;
    struct addrinfo *ai_next;
};
```

**getnameinfo()**: query key is an IP address, returns a sockaddr struct (more soon)
Using getaddrinfo()

```c
int clientfd; /**< socket descriptor */

/* setup information*/
struct addrinfo hints;
memset(&hints, 0, sizeof(struct addrinfo));
hints.ai_family = AF_UNSPEC; /**< Allow IPv4 or IPv6 */
hints.ai_socktype = SOCK_DGRAM; /**< Datagram socket */

/* fill in the server's IP address and port */
struct addrinfo *result = NULL; /**< DNS host entry */
int s = 0;
if ((s = getaddrinfo(hostname, port, &hints, &result)) != 0)
    return -1;
```

Note: This is untested code.
Client / Server Session

Client

- getaddrinfo
- socket
- connect
- send()
- recv()
- close

Server

- getaddrinfo
- socket
- bind
- listen
- accept
- recv()
- send()
- recv()
- close

Await connection request from next client

Sockets Interface

open_clientfd

open_listenedfd
Client: `socket(2)`

`socket(2)` creates a socket descriptor on the client

- Allocates and initializes some internal data structures
- `AF_UNSPEC`: indicates that the socket is associated with Internet protocols but either IPv4 or IPv6 is fine
- `SOCK_STREAM`: selects a reliable byte stream connection
  - Bi-directional pipes
  - Gives you TCP
- `SOCK_DGRAM` results in UDP

```c
int clientfd; /* socket descriptor */

if ((clientfd = socket(AF_UNSPEC, SOCK_STREAM, 0) <0)
    return -1; /* check errno for cause of error */
...
```
Socket Implementation in the OS

• Each socket fd has associated socket structure with:
  – Send and receive buffers
  – Queues of incoming connections (on listen socket)
  – A protocol control block (PCB)
  – A protocol handle

• PCB contains protocol-specific information, such as:
  – Pointer to IP Transmission Control Block with source/destination IP address and port
  – Information about received packets and position in stream
  – Information about unacknowledged sent packets
  – Information about timeouts
  – Information about connection state (setup/teardown)
**Sockets Interface**

**Client**
- `getaddrinfo`
- `socket`
- `connect`
- `send()`
- `recv()`
- `close`

**Server**
- `getaddrinfo`
- `socket`
- `bind`
- `listen`
- `accept`
- `recv()`
- `send()`
- `close`

- `open_clientfd`
- `open_listenfd`
- `Await connection request from next client`
Finally the client creates a connection with the server

– Client process blocks until the connection is created

– After resuming, the client is ready to begin exchanging messages with the server via Unix I/O calls (typically `send/recev`) on descriptor `clientfd`

```c
int clientfd;  /* socket descriptor */
struct addrinfo *result;  /* DNS host entry initialized in * call to `getaddrbyinfo()` */
typedef struct sockaddr SA;  /* generic sockaddr */

...  

/* Establish a connection with the server */
if (connect(clientfd, result->ai_addr, result->ai_addrlen) < 0)
    return -1;
```
**Server: bind(2)**

`bind()` associates the socket descriptor with the socket address (created similarly to that of the client)

```c
int listenfd;
struct addrinfo *result;

/* listening socket */
/* DNS host entry initialized in * call to getaddrbyinfo() */

... /* listenfd will be an endpoint for all requests to port 
on any IP address for this host */
if (bind(listenfd, rp->ai_addr, rp->ai_addrlen) < 0)
    return -1;
```
Sockets Interface

Client
- getaddrinfo
- socket
- connect
- send()
- recv()
- close

Server
- getaddrinfo
- socket
- bind
- listen
- accept
- recv()
- send()
- recv()
- close

open_clientfd
open_listeningfd

Connection request
Await connection request from next client
Server: `listen(2)`

- **`listen()` indicates that this socket will accept connection (`connect`) requests from clients**
  - Kernel assumes an active socket on the client end by default
- **`LISTENQ` is a constant indicating how many pending requests are allowed**

```c
int listenfd; /* listening socket */

... /* Make it a listening socket ready to accept connection requests */
if (listen(listenfd, LISTENQ) < 0)
    return -1;
```

- We’re finally ready to enter the main server loop that accepts and processes client connection requests.
Client

- getaddrinfo
- socket
- connect
- send()
- recv()
- close

Server

- getaddrinfo
- socket
- bind
- listen
- accept
- recv()
- send()
- close

open_clientfd
open_listenfd

Await connection request from next client
Server: `accept(2)`

- `accept()` blocks waiting for a connection request

```c
int listenfd; /* listening descriptor */
int connfd; /* connected descriptor */
struct sockaddr_in clientaddr;
int clientlen;

clientlen = sizeof(clientaddr);
connfd = accept(listenfd, (SA *)&clientaddr, &clientlen);
```

- `accept()` returns a *connected descriptor* (`connfd`) with the same properties as the *listening descriptor* (`listenfd`)
  - Returns when the connection between client and server is created and ready for I/O transfers
  - All I/O with the client will be done via the connected socket

- `accept()` also fills in client’s IP address
Connected vs. Listening Descriptors

• Listening descriptor
  – End point for client connection requests
  – Created once and exists for lifetime of the server

• Connected descriptor
  – End point of the connection between client and server
  – A new descriptor is created each time the server accepts a connection request from a client
  – Exists only as long as it takes to service client

• Why the distinction?
  – Allows for concurrent servers that can communicate over many client connections simultaneously
    • E.g., Each time we receive a new request, we create a new thread to handle the request
accept Illustrated

1. Server blocks in `accept`, waiting for connection request on listening descriptor `listenfd`

2. Client makes connection request by calling and blocking in `connect`

3. Server returns `connfd` from `accept`. Client returns from `connect`. Connection is now established between `clientfd` and `connfd`
iClicker Question

The listen() call is listening on a different port than the port the server will eventually use to send data to the client.

A. True
B. False
1. Start server

Server

getaddrinfo
socket
bind
listen
accept
recv()
send()
recv()
close

2. Start client

Client

getaddrinfo
socket
connect
send()
recv()
close

open_clientfd
Connection request
open_listenfd

3. Exchange data

Await connection request from next client

4. Disconnect client

5. Drop client
Parallel Computing
Parallel Computing: Description

• Tightly-coupled systems
• Processors share:
  – Clock
  – Memory (a single physical address space)
  – A single OS
• Examples: Multicore systems, Symmetric Multi Processor (SMP) systems (and you can make SMP multicore processors...)
Single CPU

Processor

Cache memories

Register file

ALU

Bus

Main memory

Bus interface
Symmetric Multi Processor

Processor 0
- Registers
- L1 d-cache
- L1 i-cache
- L2 unified cache
- L3 unified cache

ALU

Processor 3
- Registers
- L1 d-cache
- L1 i-cache
- L2 unified cache
- L3 unified cache

ALU

Main memory

Bus
Multicore Processor

Processor package

Core 0
- Registers
- L1 d-cache
- L1 i-cache
- L2 unified cache
- L3 unified cache (shared by all cores)

Core 3
- Registers
- L1 d-cache
- L1 i-cache
- L2 unified cache

ALU

Main memory
Parallel Programming
Why a different type of programming?

• Sequential programs get no benefit from multiple processors
• Many applications can be (re)designed/coded/compiled to generate cooperating, parallel instruction streams
• No automated compiler/language exists to automate this “parallelization” process.
• Key property is how much communication per unit of computation.
  • The less communication per unit computation the better the scaling properties of the algorithm.
Parallel Programming

• Parallel programming involves:
  – Decomposing an algorithm into parts
  – Distributing the parts as tasks which are worked on by multiple processes simultaneously
  – Coordinating work and communication of those processes
    • Synchronization

• Parallel programming considerations:
  – Type of parallel architecture being used
  – Type of process communication used
Parallel Programming Models

Communication and synchronization based on either:

- Shared memory
  - Interprocess communication is implicit
  - Synchronization is explicit
  - Assume processes/threads can read & write a set of shared memory locations
  - Difficult to provide across machine boundaries

- Message passing
  - Interprocess communication is explicit
  - Synchronization is implicit
  - Extensible to communication in distributed systems

send(message)  receive(message)
Shared Memory Programming Model

- Programs/threads communicate/cooperate via loads/stores to memory locations they share.
- Communication is therefore at memory access speed (very fast) and is implicit.
- Cooperating pieces must all execute on the same system (computer).
- OS services and/or libraries used for creating tasks (processes/threads) and coordination (semaphores/barriers/locks.)
Example Shared Memory Code

fork N processes
each process has a number, p, and computes
    istart[p], iend[p], jstart[p], jend[p]
for (s = 0; s < STEPS; s++) {
    k = s & 1; m = k ^ 1;
    forall (i = istart[p]; i <= iend[p]; i++) {
        forall (j = jstart[p]; j <= jend[p]; j++) {
            a[k][i][j] = c1*a[m][i][j] + c2*a[m][i-1][j] +
                c3*a[m][i+1][j] + c4*a[m][i][j-1] +
                c5*a[m][i][j+1]; // implicit comm
        }
    }
}
Message Passing Programming Model

- “Shared” data is communicated using send/receive services (across an external network).
- Shared data must be formatted into message chunks for distribution
  - Unlike shared memory
- Coordination is also via sending/receiving messages.
- Program components can be run on the same or different systems, so can use many of processes.
- Standard libraries exist to encapsulate messages:
  - Parasoft's Express (commercial)
  - PVM (standing for Parallel Virtual Machine, non-commercial)
  - MPI (Message Passing Interface, also non-commercial).
Message Passing Logistics: Synchronization

When do **send/receive** operations terminate?

**Blocking (aka Synchronous):**
Sender waits until its message is received
Receiver waits if no message is available

**Non-blocking (aka Asynchronous):**
Send operation “immediately” returns
Receive operation returns if message is available or not (polling)

**Partially blocking/non-blocking:**
`send()`/`receive()` with *timeout*
Limitations of Message Passing

Easy for OS, hard for programmer

– Programmer must code communication
– Programmer may have to code format conversions, flow control, error control
– No dynamic resource discovery
Summary

• Accessing the network from user code is detailed but often formulaic
• Parallel Computing: tightly-coupled systems
• Parallel Programming Models
  – Shared Memory (OpenMP)
    • Sharing is implicit, synchronization is explicit
  – Message Passing (MPI)
    • Sharing is explicit, synchronization is implicit
Announcements

• Exams are graded, scores are not entered
  – Will be returned in discussion section this week
• Project 3 due Friday
• Project 4 out Friday
  – Group registration is due Monday, 11/19
• Problem Set 9 is due Friday in section