CS 5523 Operating Systems: Application-Level Communications

Instructor: Dr. Tongping Liu

Thank Dr. Dakai Zhu, Dr. Palden Lama for providing their slides.

What are the Problems?

- Process A want to send process B a message
- Application-level agreement
- Network: actual message transmission

Outline

- Fundamentals
  - Client/Server communication protocols
    - Request vs. Request-reply vs. Request-reply-acknowledge
    - Invocation semantics
      - Exact once vs. at least once vs. at most once
  - Communication types
    - Transient vs. persistent
    - Synchronous vs. asynchronous
- Models for application communications
  - RPC: remote procedure call
  - Message-oriented communication
  - Stream-Oriented communication
  - Multicast communication

C/S Comm.: Request Protocol (R)

- If service
  - does not have output parameters and
  - does not have a return type
  - client may not want to wait for server to finish.
C/S Comm.: Request-Reply protocol (RR)

- If client expects results from a server
  - Client requests service execution from server through a request message, and
  - Delivery of service results in a reply message

- Most client-server interactions are built on RR protocol

Request-Reply-Acknowledge Protocol (RRA)

- In addition to RR protocol, client sends acknowledgement after it received reply
- Acknowledgement sent asynchronously

What may go wrong?

- Request lost
- Reply lost
- Server down

Typical Failure Handling in RR

- Client side: request retry with time out
  - Client keeps local copy of request
  - If timeout and no reply received, it resends request
  - After N attempts, assumes that server has failed
- Server side: receives a request multiple times
  - Re-execute & resend results; or
  - Store requests & results: filter duplicated requests and re-transmit results

How long should servers keep the requests & results?
Should Servers Re-Do Operations?

- **Idempotent operations**: which can be performed repeatedly with the same effect.
  - Suppose $x$ is input message $\Rightarrow f(f(x)) = f(x)$
  - No state needs to maintain on the server

Are the following operations idempotent?

- HTTP GET: Yes
- UNIX file operations: read, write etc.: No

Server Invocation Semantics in RR

<table>
<thead>
<tr>
<th>Fault tolerance measures</th>
<th>Invocation semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retransmit request message</td>
<td>Duplicate filtering</td>
</tr>
<tr>
<td>No</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
| Yes | Yes | Retransmit reply | At-most-once

Server Invocation Semantics (cont.)

- **Maybe**: if no reply, the client does not know if method was executed or not
- **At least once**: will guarantee that invocation be carried out at least once, but possibly more
- **At most once**: Will guarantee that invocation be carried out at most once, but possibly none at all
  - Detect duplicated requests with sequence numbers
  - No guarantees: When a server crashes, the client gets no help and no promises about what happened

Local invocation: **exactly once** - ideal case

http://lass.cs.umass.edu/~shenoy/courses/spring13/lectures/notes/677_lect09.pdf

Types of Communications

- **Asynchronous communication**
  - Sender continues immediately after it has submitted the request (unblocked, need a local buffer at the sender)

- **Synchronous communication**
  - Sender blocks until the sender receives an OK to continue; where the OK may come?
### Types of communications (cont.)

**Persistent vs. Transient**

- **Persistent**: A message is stored at a communication server as long as it takes to deliver it (e.g., e-mail).
- **Transient**: A message is stored as long as sender and receiver are working at the same time (TCP, UDP, IP routing).

**Persistent:**
- A message is stored at a communication server as long as it takes to deliver it (e.g., e-mail).

**Transient:**
- A message is stored as long as sender and receiver are working at the same time (TCP, UDP, IP routing).

### Type of Communications (cont.)

<table>
<thead>
<tr>
<th>Asynchronous</th>
<th>Synchronous at Submission, delivery, after service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistent</td>
<td>Message-oriented middleware (MOM)</td>
</tr>
<tr>
<td>Transient</td>
<td>Client/Server, RPC, TCP</td>
</tr>
</tbody>
</table>

- **Persistent**
  - Processes send each other messages (queued)
  - Sender does not need to wait for immediate reply
  - Middleware often ensures fault tolerance

- **Transient**
  - Client and server have to be active at time of communication
  - Client issues request and blocks until it receives reply
  - Server essentially waits only for incoming requests, and subsequently processes them.

**Drawbacks of synchronous communication**
- Client cannot do any other work while waiting for reply
- Failures have to be handled immediately; the client is waiting
- The model may simply not be appropriate (mail, news)

### Persistence and Synchronization 1

1. **Persistent asynchronous communication** (e.g., email)
2. **Persistent synchronous communication**

### Persistence and Synchronization 2

1. **A sends message and continues**
2. **B receives message**
3. **B sends request and waits until received**
4. **C sends message and continues**
5. **B receives message**
6. **B runs, but doing something else**
7. **Process request**

**c)** Transient asynchronous communication (e.g., UDP)
**d)** Receipt-based transient synchronous communication (e.g., TCP)
Persistence and Synchronization

Persistence and Synchronization 3

Outline

- Fundamentals
  - Client/Server communication protocols
    - Request vs. Request-reply vs. Request-reply-acknowledge
  - Invocation semantics
    - Exact once vs. at least once vs. at most once
  - Communication types
    - Transient vs. persistent
    - Synchronous vs. asynchronous
- Models for application communications
  - RPC: remote procedure call
  - Message-oriented communication
  - Stream-Oriented communication
  - Multicast communication

High-Level Communication Models

- RPC: remote procedure call
- Message-oriented communication
- Stream-Oriented communication
- Multicast communication

Remote Procedure Calls

- Network communication:
  - Send and receive are not transparent
  - Application developers are good at “procedure call” model
- In 1984, Birrell and Nelson allows programs to call procedures located on remote machines (RPC)
  - RPC is integrated into programming languages
  - Like centralized computing, by allowing remote services to be called as procedures
- Issues: How to pass parameters, semantics in face of errors, failures, etc...

RPC is a high-level model for client-server communication
Review of Local Procedure Call

- Consider a call in C: \( \text{result} = \text{add}(\text{para}_x, \text{para}_y) \)
- Caller pushes parameters onto the stack: in one process
- Stack before (a) and while (b) the called procedure is active is shown below

Parameters Passing in Local Calls

- **Call-by-value**: the parameter value copied to the stack. Modifications do not affect the calling side
- **Call-by-reference**: the address of the parameter is pushed onto the stack (for example, pointers in C; \text{obj ref} in Java). Any modification affects the variable at the calling side.
- **Call-by-copy/restore**: Copy the variable first to the stack (as in call-by-value), and then copy back after the call, overwriting the caller’s original value.
  - In many cases, the same behavior as “call-by-reference”

Remote Procedure Call (RPC)

Client would like to use its own functions as well as the ones in the server. How?

- \text{myAdd()};
- \text{mySub()};
- \text{myMax()};
- \text{myMin()};
- \text{magicAdd()};
- \text{magicSub()};
- \text{magicMax()}
- \text{magicMin()}

Stubs at Client and Server Sides

- Make remote procedure call look like local call
- So that the client can call \text{myAdd()} and \text{magicAdd()} in the same way
  - \text{result1} = \text{myAdd(} \text{para}_x, \text{para}_y \text{);};
  - \text{result2} = \text{magicAdd(} \text{para}_x, \text{para}_y \text{);};

The server stub for \text{magicAdd()} unpacks the parameters from the message and calls the local procedure in a usual way. When complete, the server stub will pack the result and return it to the client.
RPC Steps

A remote procedure call occurs in the following steps:

1. The client procedure calls the client stub.
2. The client stub builds a message and calls the local operating system.
3. The client’s OS sends the message to the remote OS.
4. The remote OS sends the message to the server stub.
5. The server stub unpacks the parameters and calls the server.
6. The server does the work and returns the result to the stub.
7. The server stub packs the result in a message and calls its local OS.
8. The server’s OS sends the message to the client’s OS.
9. The client’s OS gives the message to the client stub.
10. The stub unpacks the result and returns it to the client.

Sounds simple but there are several issues...

RPC Mechanism

Big Endian and Little Endian

- **Big Endian:**
  - Store the most significant byte in the smallest address

- **Little Endian:**
  - Store the least significant byte in the smallest address.

0x 90AB12CD
Problem 1: different data representations

A process on an Intel machine sends a message of an integer and a four-character string ("5, JILL") to another process on a Sun SPARC machine.

(a) original message on Intel (x86, Little Endian)
(b) receipt message on SPARC (Big Endian) "5000, JILL"
(c) simple reverse: message after converted "5, LLIJ"
The number is reversed by the string is not.

Approaches for Exchanging Information

How can clients make servers on different machines understand them?

- With listener’s language
  - English, Chinese, Indian...
- A common external data representation
  - Language defined representation
  - e.g., CORBA CDR
- External self-descriptive data representation
  - e.g., XML, Web Services

Problem 2: Un/Marshaling

- More than just wrapping parameters into a message
- Client and server machines may have different data representations (think of byte ordering)
- Client and server have to agree on a standard representation (e.g., external data representation (XDR))
  - How are basic data values represented (integers, floats, characters)
  - How are complex data values represented (arrays, unions)
- Client and server need to properly interpret messages, and transform them into machine-dependent representations.

Other Problems in Parameter Passing

- Passing reference parameters
  - Pointers, and in general, reference parameters are passed with considerable difficulty
- Solutions
  - Forbid reference parameters
  - Copy the entire data structure (e.g., an entire array may be sent if the size is known). In the case of the server input parameter, it does not need to be copied back.
  - How to handle complex data structures (e.g. general graphs)?
Parameter Passing Semantics in RPC

- RPC parameter passing:
  - RPC assumes copy in/copy out semantics: while procedure is executed, nothing can be assumed about parameter values.
  - RPC assumes all data that is to be operated on is passed by parameters. Excludes passing references to data.

- Conclusion: full access transparency cannot be realized.

- Observation: If we introduce a remote reference, access transparency can be enhanced:
  - Remote reference offers unified access to remote data.
  - Remote references can be passed as parameter in RPCs.

RPC in Practice

- Essence: developers concentrate on client- & server-specific codes, while RPC (generators / libraries) do the rest (actual communication).
- Client stub and server stub communicate over the network to perform RPC.

Generate Stubs

- Hiding a remote procedure call requires that the caller and callee agree on interface:
  - the format of the messages
  - the representation of simple data structures (no complicated data)
  - the transport layer protocol (Connection-oriented or not)

- Using Interface Definition Language (IDL) to specify an interface: services provided by servers.
  - Compile IDL to generate client & server stubs.

DEC RPC Interface

- Type/data definitions (like C).
- Component is described as a PROGRAM.
  - Procedures have a result type, a parameter list and a number.
  - Procedure can be called remotely.

- Used by client or server directly:
  - Locating servers: static vs. dynamic binding
  - Choosing a transport protocol.
  - Authentication and security.
  - Invoking RPCs dynamically.

- Used by stubs for:
  - Generating unique message IDs.
  - Sending messages.
  - Maintaining message history.
**RPC Example**

```c
/* person.x */
const NL=64;
enum sex_type {
FEMALE = 1,MALE = 2};
struct Person {
    string first_name<NL>;  // = 1;
    string last_name<NL>;  // = 1234567;
    sex_type sex;
    string city<NL>;
};
```

**Sun RPC Example**

```c
void * print_0(Person *argp, struct svc_req *rqstp) {
    static char * result;
    printf("%s %s\n", argp->first_name, argp->last_name, argp->city);
    return((void *) &result);
}
```

**Sun RPC Server Implementation**

```c
/* server.c */
void * print_0(Person *argp, struct svc_req * rqstp) {
    static char * result;
    printf("%s %s\n", argp->first_name,
    argp->last_name,
    argp->city);
    return((void *) &result);
}
```

**Sun RPC Client**

```c
/* client.c */
print_person(char * host, Person * pers) {
    CLIENT *clnt;
    clnt = clnt_create(host, PERSONPROG,
    "udp");
    if (clnt == (CLIENT *) NULL) {
        exit(1);
    }
    if (print_0(pers, clnt))==NULL)
        clnt_perror(clnt, "call failed");
    clnt_destroy(clnt);
    printf("%s %s\n", argp->first_name,
    argp->last_name,
    argp->city);
    return((void *) &result);
}
```

**How to locate an RPC server that can execute a given procedure in a network?**
Binding a Client to a Server

- Registration of a server makes it possible for a client to locate the server and bind to it.
- Locate the server’s machine → locate server.

Steps in RPC

- Many RPCs occur between client and server on the same machine.
- Need to optimize RPCs for this special case ⇒ use a lightweight RPC mechanism (LRPC).
- Server exports interface to remote procedures.
- Client on the same machine imports interface.
- OS kernel creates data structures including an argument stack shared between Server and Client.
Lightweight RPCs (cont.)

- RPC execution
  - Push arguments onto stack
  - Trap to kernel
  - Kernel changes mem map of client to server address space
  - Server thread executes procedure (OS upcall)
  - Thread traps to kernel upon completion
  - Kernel changes the address space back and returns control to client

Other RPC Models

- Asynchronous RPC
  - Request-reply behavior often not needed
  - Server can reply as soon as request is received and execute procedure later

- Deferred-synchronous RPC
  - Use two asynchronous RPCs
  - Client needs a reply but can’t wait for it; server sends reply via another asynchronous RPC

- One-way RPC
  - Client does not even wait for an ACK from the server
  - Limitation: reliability not guaranteed (Client does not know if procedure was executed by the server.)

Traditional RPC vs. Asynchronous RPC

- Try to get rid of the strict request-reply behavior, and let the client continue without waiting for an answer from the server.

Deferred Synchronous RPC

- Client can also do a (non)blocking poll at the server to see whether results are available.
- Through two asynchronous RPCs
One-Way RPC

- Client does not know if the request is accepted or not (ch 8 fault tolerance)

RPC Limitations

- Parameters passed by values only and pointer values are possibly not allowed.

- Speed: remote procedure calling (and return) time (i.e., overheads) can be significantly (1 - 3 orders of magnitude) slower than that for local procedure.

  ➢ This may affect real-time design and the programmer should be aware of its impact.

RPC Limitations

- Failure: RPC is more vulnerable to failure (since it involves communication system, another machine and another process).

  ➢ The programmer should be aware of the call semantics, i.e. programs that make use of RPC must have the capability of handling errors that cannot occur in local procedure calls.

Design Issues

- Exception handling
  ➢ Necessary because of possibility of network and nodes failures;
  ➢ RPC uses return value to indicate errors;

- Transparency
  ➢ Syntactic → achievable, exactly the same syntax as a local procedure call;
  ➢ Semantic → impossible because of RPC limitation: failure (similar but not exactly the same);
Outline

- **Fundamentals**
  - Client/Server communication protocols
    - Request vs. Request-reply vs. Request-reply-acknowledge
  - Invocation semantics
    - Exact once vs. at least once vs. at most once
  - Communication types
    - Transient vs. persistent
    - Synchronous vs. asynchronous
- **Models for application communications**
  - RPC: remote procedure call (inherent synchronous)
  - Message-oriented communication
  - Stream-Oriented communication
  - Multicast communication

Message-Oriented Communication

- **Transient Messaging**
- **Message-Queuing System**
- **Message Brokers**
- **Example: IBM Websphere MQ**

Transient Messaging (transport-level sockets)

- Berkeley sockets

<table>
<thead>
<tr>
<th>Primitive</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socket</td>
<td>Create a new communication endpoint</td>
</tr>
<tr>
<td>Bind</td>
<td>Attach a local address to a socket</td>
</tr>
<tr>
<td>Listen</td>
<td>Announce willingness to accept connections</td>
</tr>
<tr>
<td>Accept</td>
<td>Block caller until a connection request arrives</td>
</tr>
<tr>
<td>Connect</td>
<td>Actively attempt to establish a connection</td>
</tr>
<tr>
<td>Send</td>
<td>Send some data over the connection</td>
</tr>
<tr>
<td>Receive</td>
<td>Receive some data over the connection</td>
</tr>
<tr>
<td>Close</td>
<td>Release the connection</td>
</tr>
</tbody>
</table>

Message-Oriented Middleware

- **Asynchronous persistent** communication through support of middleware-level *queues*.
  - Queues correspond to buffers at communication servers
  - Communicated processes may NOT be active at the same time

<table>
<thead>
<tr>
<th>Primitive</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Put</td>
<td>Append a message to a specified queue</td>
</tr>
<tr>
<td>Get</td>
<td>Block until the specified queue is nonempty, and remove the first message</td>
</tr>
<tr>
<td>Poll</td>
<td>Check a specified queue for messages, and remove the first. Never block</td>
</tr>
<tr>
<td>Notify</td>
<td>Install a handler to be called when a message is put into the specified queue</td>
</tr>
</tbody>
</table>
Message-Oriented Middleware (cont.)

- Four different loosely-coupled communications

General Architecture of Message Queue

Message queuing systems assume a common messaging protocol: all applications agree on message format (i.e., structure and data representation)

Message Broker

- Message queuing systems assume a common messaging protocol: all applications agree on message format (i.e., structure and data representation)

- **Message broker:** Centralized component that takes care of application heterogeneity in an MQ system:
  - Transforms incoming messages to target format
  - Very often acts as an application gateway
  - May provide subject-based routing capabilities
  - Enterprise Application Integration
Continuous Media

**Observation:** All discussed communication facilities are based on a *discrete* or *time-independent* exchanging of information

**Continuous media:** Characterized by the fact that values are *time dependent*:
- Audio
- Video
- Animations
- Sensor data (temperature, pressure, etc.)

Transmission of Continuous Media

- Different timing guarantees: 3 types of transmission
- **Asynchronous:** no restrictions with respect to when data is to be delivered
- **Synchronous:** define a maximum end-to-end delay for individual data packets
- **Isochronous:** define a maximum and minimum end-to-end delay (*jitter* is bounded)
Stream

- **Definition:** A (continuous) data stream is a connection-oriented communication facility that supports isochronous data transmission.

- **Common stream characteristics**
  - Streams are unidirectional
  - A single source, and one or more sinks
  - Often, either the sink and/or source is a wrapper around hardware (e.g., CD device, TV monitor, dedicated storage)

- **Two types of streams:**
  - Simple: single flow of data, e.g., audio or video
  - Complex: multiple data flows, e.g., stereo audio or combination audio/video

Streams and QoS

- Streams are all about timely delivery of data.

- **Quality of Service (QoS) specifications**
  - The required bit rate at which data should be transported.
  - The maximum delay until a session has been set up (i.e., when an application can start sending data).
  - The maximum end-to-end delay (i.e., how long it will take until a data unit makes it to a recipient).
  - The maximum delay variance, or jitter.
  - The maximum round-trip delay.

**Enforce QoS**

- Network-level tools: prioritize packets (differentiated services)
- Use buffers to reduce jitter

**Enforce QoS: Example**

- Packet departs source
- Packet arrives at buffer
- Packet removed from buffer
- Time in buffer
- Gap in playback

Diagram showing multimedia server, client, stream synchronization, QoS control, and buffer timing with packets and delays.
Enforce QoS: Example

Interleaving: multiple samples in a single packet: the effects of packet loss spread the samples

---

Application-level Multicast

Multicasting: multiple receivers

- Organize nodes of a distributed system into an overlay network and use that network to disseminate data
- Network routers are not a part of groups

Epidemic-based data dissemination

---

Epidemic Algorithms

Basic idea: rapidly propagate info among a network using only local info (Infect all nodes AFAP)

- Update operations are initially performed at a few nodes
- A node passes its updated state to a limited neighbors
- Update propagation is lazy, i.e., not immediate
- Eventually, each update should reach every node

Propagation models

- Anti-entropy (entropy: lack of order or predictability): chooses another node at random, and exchanges state differences, leading to identical states at both
- Gossiping: A replica which has just been updated (i.e., has been contaminated), tells a number of other replicas about its update (contaminating them as well).

---

Anti-Entropy Propagation Model

Node P selects node Q from the system at random

Schemes to propagates updates

- Push: P only sends its updates to Q
- Pull: P only retrieves updates from Q
- Push-Pull: P and Q exchange mutual updates (after which they hold the same information).

Susceptible: not-infected

Observation: pure push-based is not effective:

- Updates can be propagated by infected nodes
- When many nodes are infected, the chance to pickup a susceptible node is very small.
Anti-Entropy Propagation Model

- **Pull-based**: works better when many nodes are infected. Spreading is triggered by susceptible nodes.
- **Push-pull**: best approach of anti-entropy.
  - $O(\log(N))$ rounds to disseminate updates to all $N$ nodes (round = when every node as taken the initiative to start an exchange).
  - Propagating is fast and scalable

Gossiping or Rumor Spreading

- A server $S$ having an update to report, contacts other servers. If a server is contacted to which the update has already propagated, $S$ stops contacting other servers with probability $1/k$.
- **Gossiping in reality**:
  - Bob has hot news, he may tell Alice
  - Alice will spread the gossip to her friends as well.
  - When Alice tell Charlie about this news, if Charlie said that he already heard this news, thus Alice will be disappointed and may stop to spread this news

Gossiping vs. Epidemic

- **Advantage of Gossiping**: Scalability.
- **Epidemic algorithm**: Advantage: good for spreading updates
  - But spreading the removal is hard since the deletion of a data item destroys all information on that item: when a node receive old copies of the data item, they will think of it as updates on something that it didn’t have before.

If $k = 4$, then $\ln(s) = -4.97$, then $s$ is less than $0.007$. Less than 0.7% of the nodes remain susceptible.
Deleting Values

**Fundamental problem:** cannot remove an old value from a server and expect the removal to propagate.

**Death certificate:** Removal has to be registered as a special update.

**When to remove a death certificate?**
- Run a global algorithm to detect whether the removal is known everywhere, and then collect the death certificates (looks like garbage collection).
- Assume death certificates propagate in finite time, and associate a maximum lifetime for a certificate (can be done at risk of not reaching all servers).

Summary

**Fundamentals**
- Client/Server communication protocols
- Invocation semantics
- Communication types
- RPC: remote procedure call
- Message-oriented communication
- Stream-Oriented communication
- Multicast communication