Chapter 3
Processes

Da-Wei Chang

CSIE.NCKU

Source: Abraham Silberschatz, Peter B. Galvin, and Greg Gagne, "Operating System Concepts", 10th Edition, Wiley.

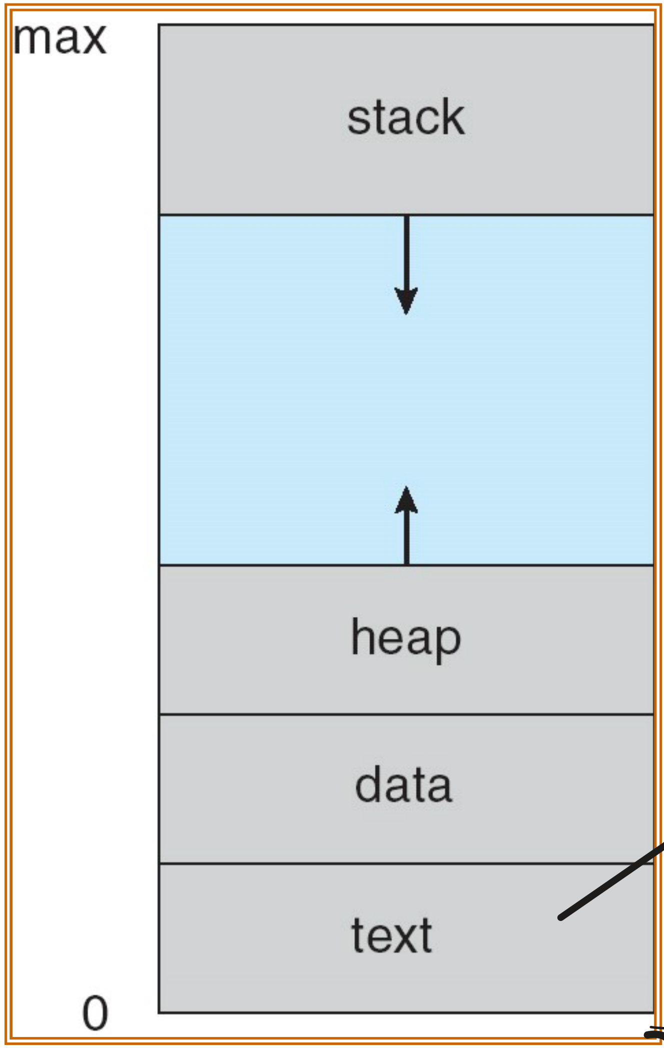
Outline

- Process Concept
- Process Scheduling
- Operations on Processes
- Inter-Process Communication (IPC)
- Examples of IPC Systems
- Communication in Client-Server Systems

Process Concept

- An operating system executes user programs
  - ~~Batch system – jobs~~
  - Time-sharing systems – processes or tasks
- We use the terms *job* and *process* almost interchangeably
- Process – a program in execution; process execution must progress in sequential fashion
- A process includes 一個process至少有一個thread
  - text (i.e., code section)
  - data section
  - heap
  - stack
  - program counter and the content of the processor registers
- Program – ***passive*** ; Process – ***active***

A Process in Memory

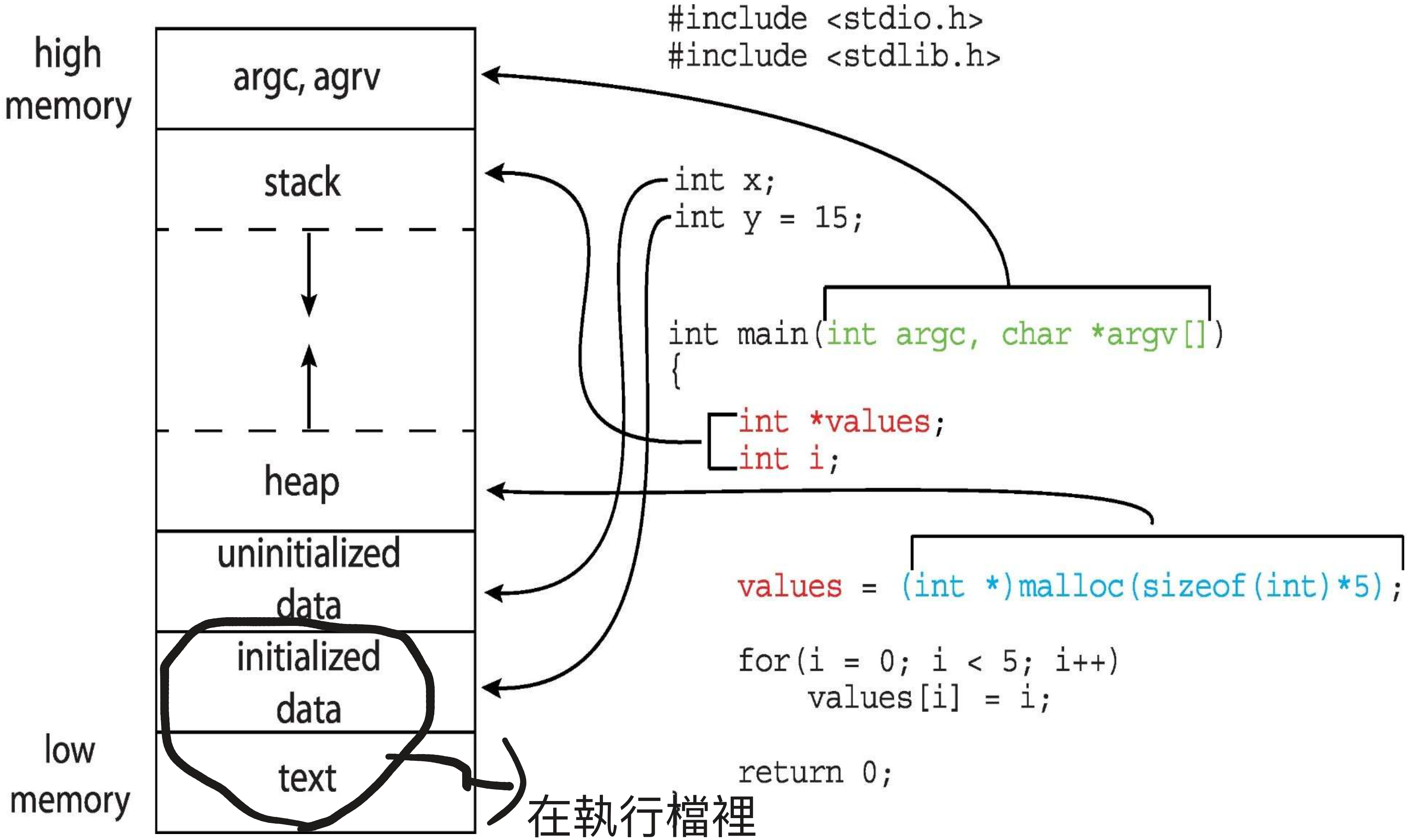


More than one processes can be associated with the same program

Which parts are the same for these processes? 兩個process的Text區域會一樣

Process佔這些區域

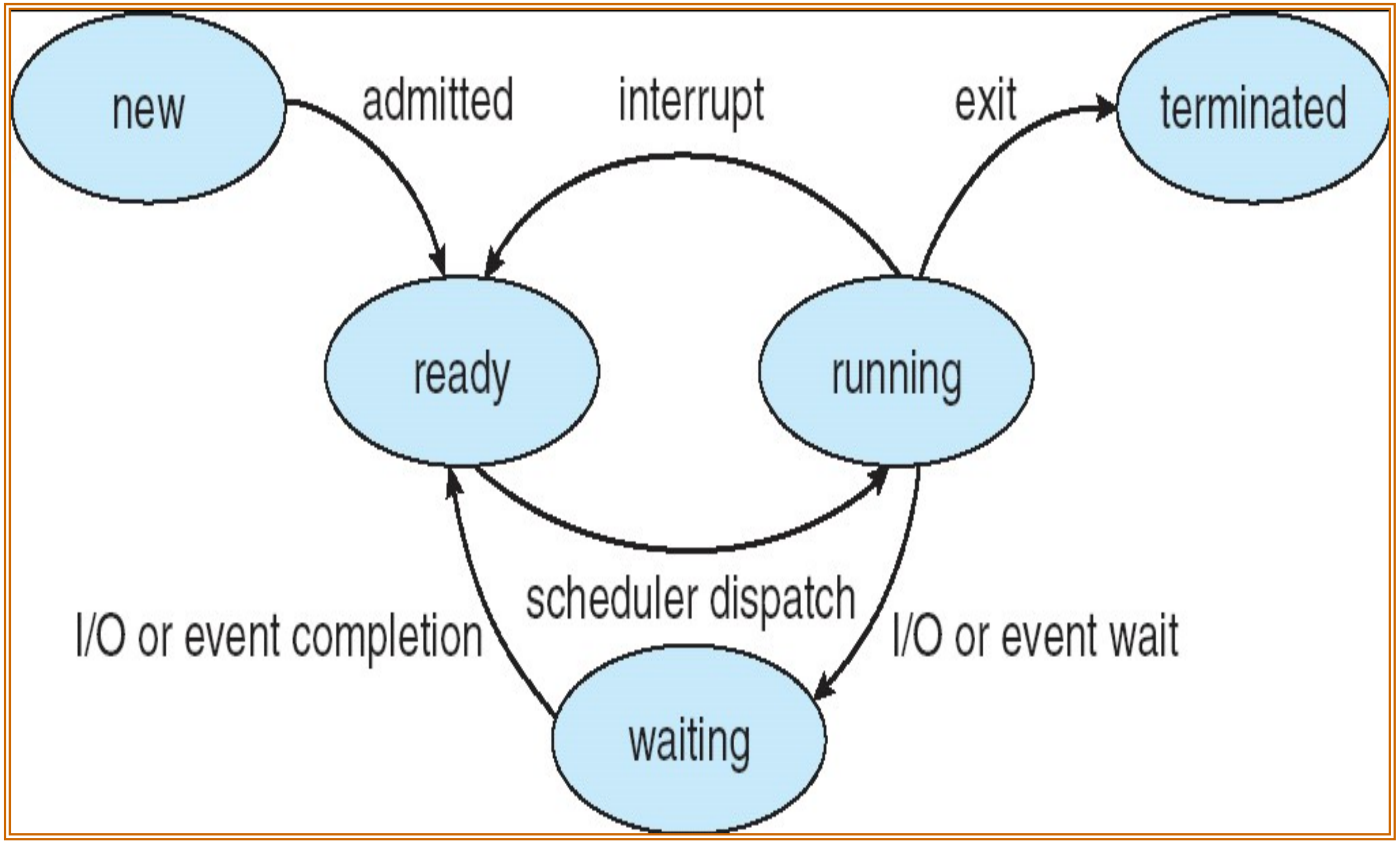
Memory Layout of a C Program



Process State

- As a process executes, it changes its *state*
  - **new**: The process is being created
  - **ready**: The process is waiting to be assigned to a CPU
    - The process is **runnable**
  - **running**: Instructions are being executed **(i.e., owns the CPU)**
  - **waiting**: The process is waiting for some event to occur
  - **terminated**: The process has finished execution

Diagram of Process State



Each process has its own process state diagram!!!

Process Control Block (PCB)

Information associated with each process

PCB 包含

- Process state
- CPU registers
  - Vary in number and type, depending on the processor architecture
    - Accumulators, index registers, stack pointers, general purpose registers…
    - Program counter
      - The address of the ***next*** instruction
  - The reason?
  - Saved when an interrupt occurs
- CPU scheduling information
  - Priority, pointers to scheduling queues, other scheduling parameters…

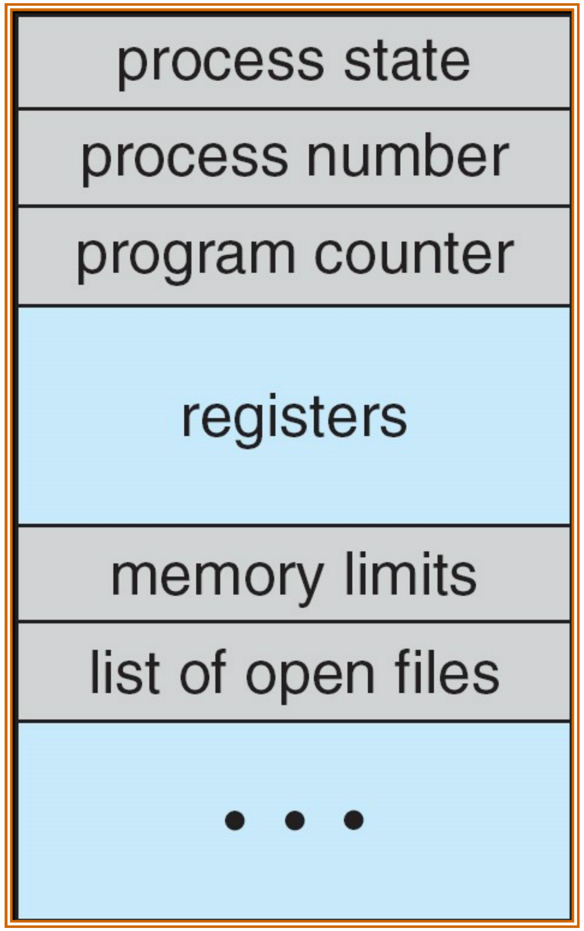
Process Control Block (PCB)

Information associated with each process (cont.)

- Memory-management information
  - Will be introduced later
- Acc會計ounting & identification information
  - CPU time and real time used
  - Time limits
  - Process IDs/numbers
- I/O status information
  - Open files and devices…

Process Control Block (PCB)

Also called **Task Control Block (TCB)**



task_struct in Linux: /include/linux/sched.h

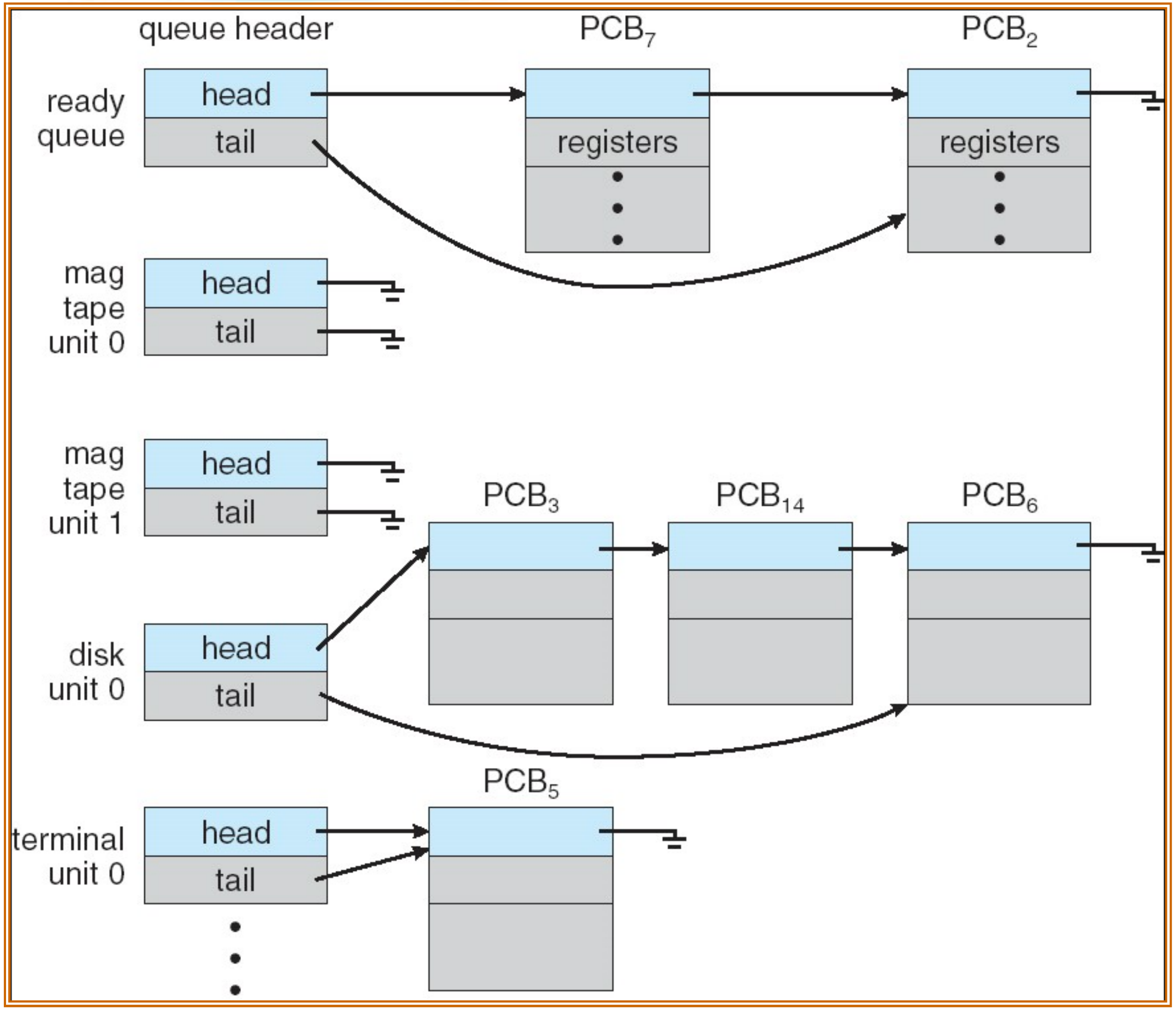
Threads

- So far, a process is a program that performs a **single thread** of execution
  - Process executes in sequential fashion
  - Many modern operating systems support multi-threaded processes
  - Discussed in Chapter 4

Process Queues

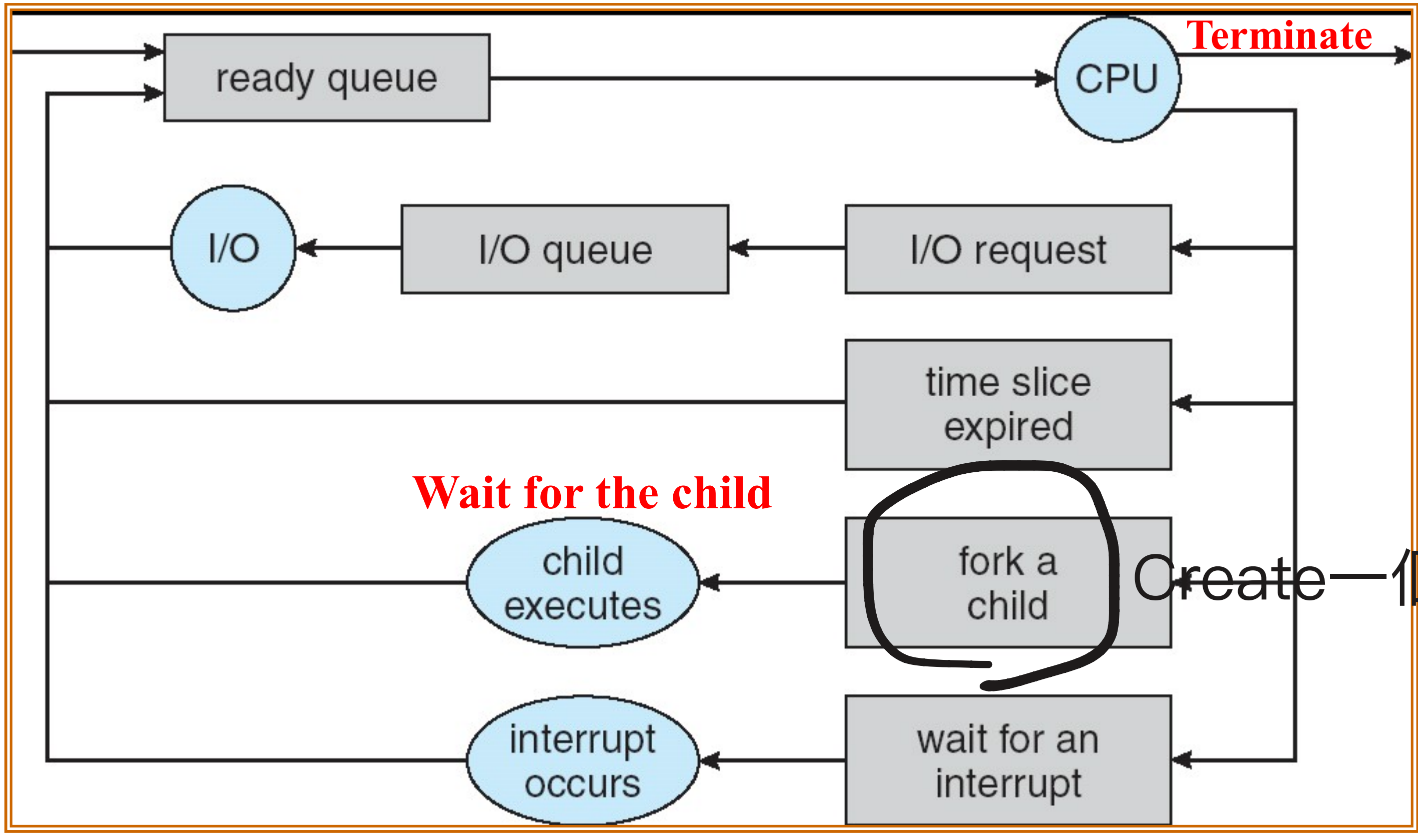
- **Job queue** – set of all processes in the system Maintain所有process
- **Ready queue** (or called run queue) – set of all processes residing in main memory, ready and waiting to execute
- **Device queues** – set of processes waiting for an I/O device
- **Event queues** – set of processes waiting for an event
- **Processes migrate among the queues**

Ready Queue and Various I/O Device Queues



Queuing-Diagram Representation of Process Scheduling

Create



Free PCB and resources

Create一個新的process

Process Scheduling

- Goals of
  - Multiprogramming
    - High CPU utilization
  - Time-sharing
    - Short response time

  **(introduced in Chapter 1)*

- A process scheduler is responsible for selecting a process to run, trying to achieve the above goals

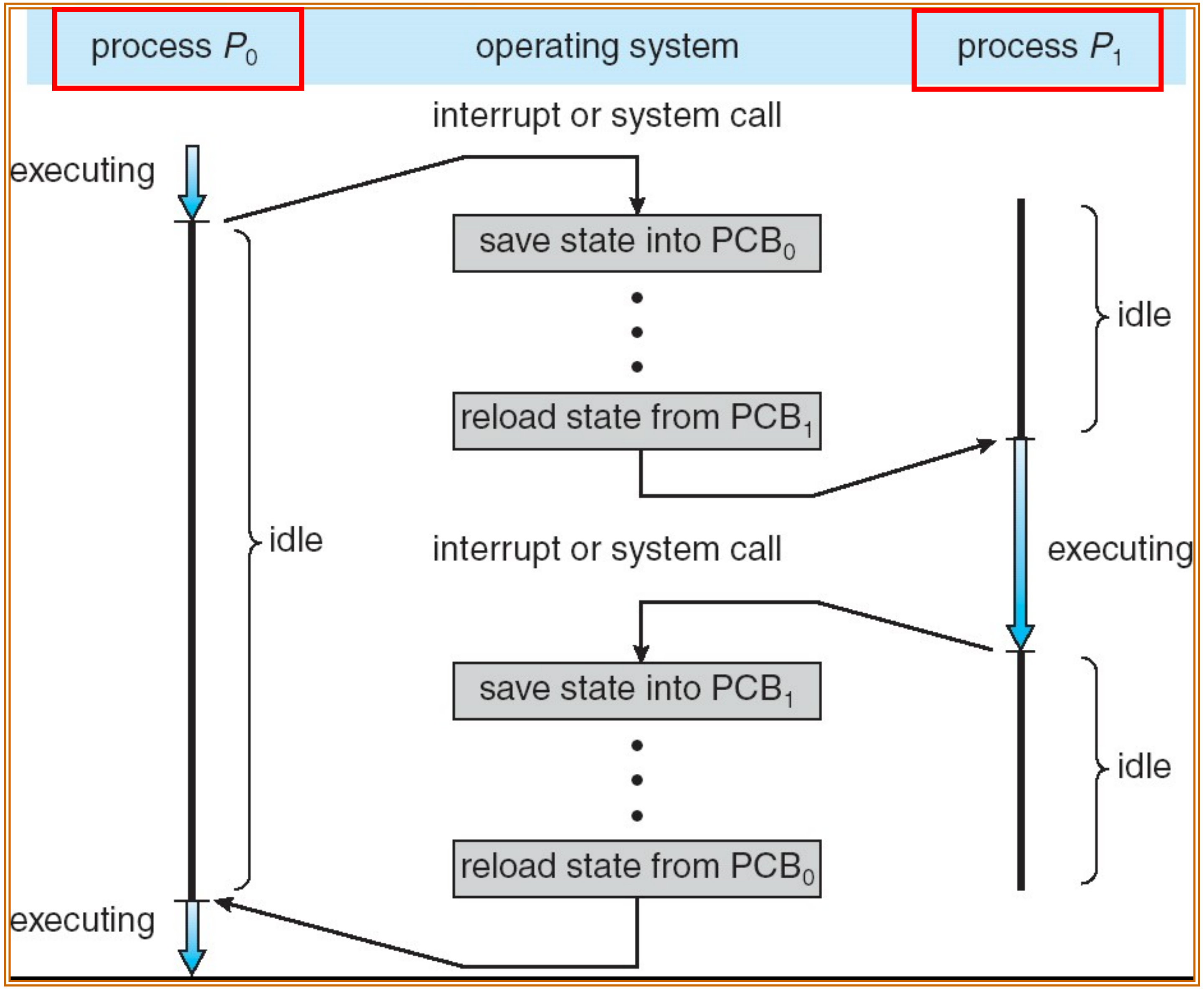
Context Switch

- When CPU switches to another process, the system must save the state of the old process in its PCB and load the saved state for the new process
  - This is called context switch

要切換到下一個process要先把當前process的狀態存起來

- Context-switch time is overhead
  - The system does no *useful* work while switching
  - The time can be reduced with hardware support
    - e.g., SUN UltraSPARC provide multiple set of registers
      - Context switch == change the pointer to the current register set
  - Typically, context switch requires a few microseconds

CPU Switch from Process to Process



Schedulers

- **Long-term scheduler** (or Job scheduler) – determines which processes should be brought into the ready queue from the job queue Jobs queue -> ready queue

- **Short-term scheduler** (or CPU scheduler) – determines which process in the ready queue should be executed next (on the CPU) Ready queue -> running

Schedulers (Cont.)

- Short-term scheduler is invoked frequently (milliseconds) ⇒ (must be fast)
  - E.g., allocate 100 ms for each process, 10ms for performing scheduling
    - 10 / (100+10) = 9% CPU time wasted on the scheduling

- Long-term scheduler is invoked infrequently (seconds, minutes) ⇒ (can be slow)

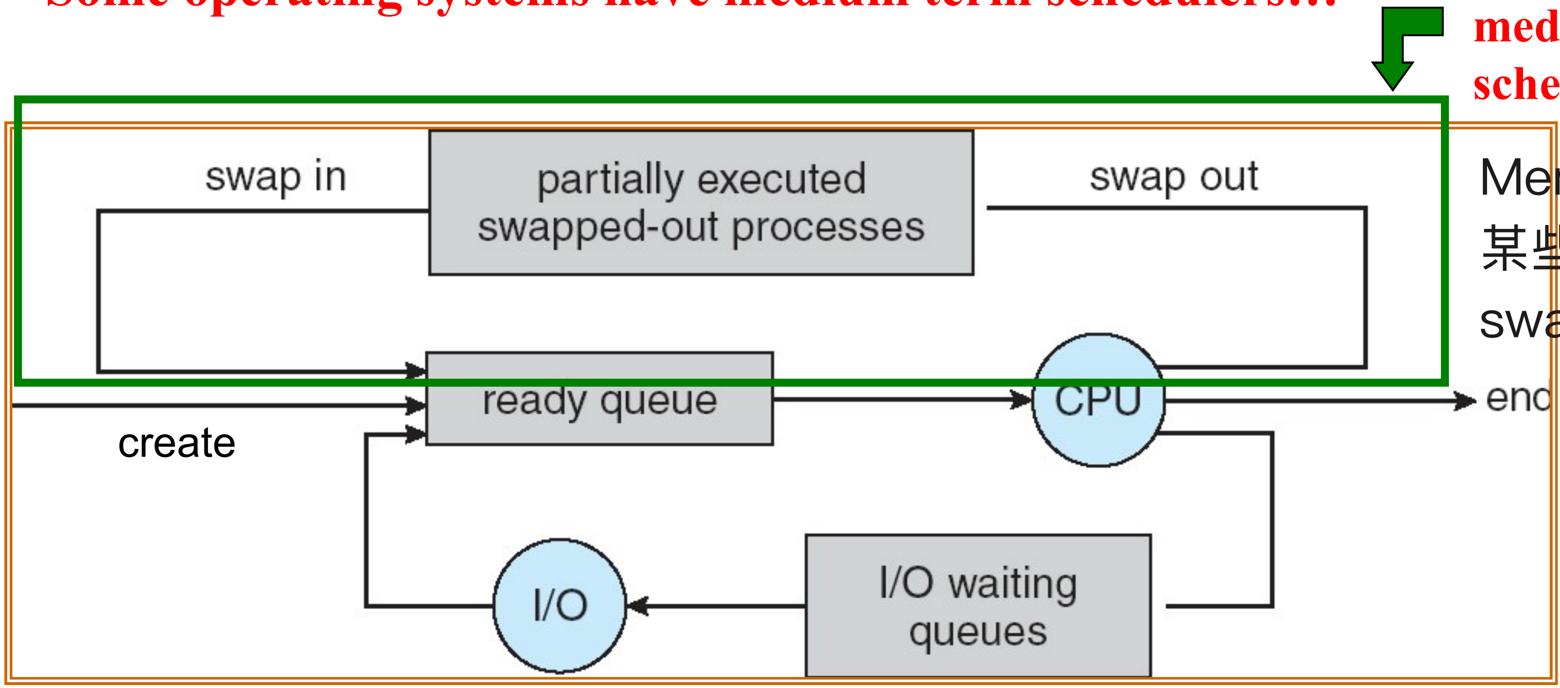
Schedulers (Cont.)

- The long-term scheduler controls the *degree of multiprogramming* Process 的總數
- Processes can be described as either
  - **I/O-bound process** – spends more time doing I/O than computations ➔ many short CPU bursts
  - **CPU-bound process** – spends more time doing computations ➔ few very long CPU bursts
- A long-term scheduler should select a good process mix
  - All IO-bound? All CPU-bound? 都會有
- Windows and UNIX have no long-term schedulers
  - Put all the jobs into memory

Addition of Medium Term Scheduling

Some operating systems have medium term schedulers…

medium term scheduler



Memory不夠用，某些process先swap到硬碟

Needed

for improving the process mix

when a change of memory requirement has overcommitted the available memory

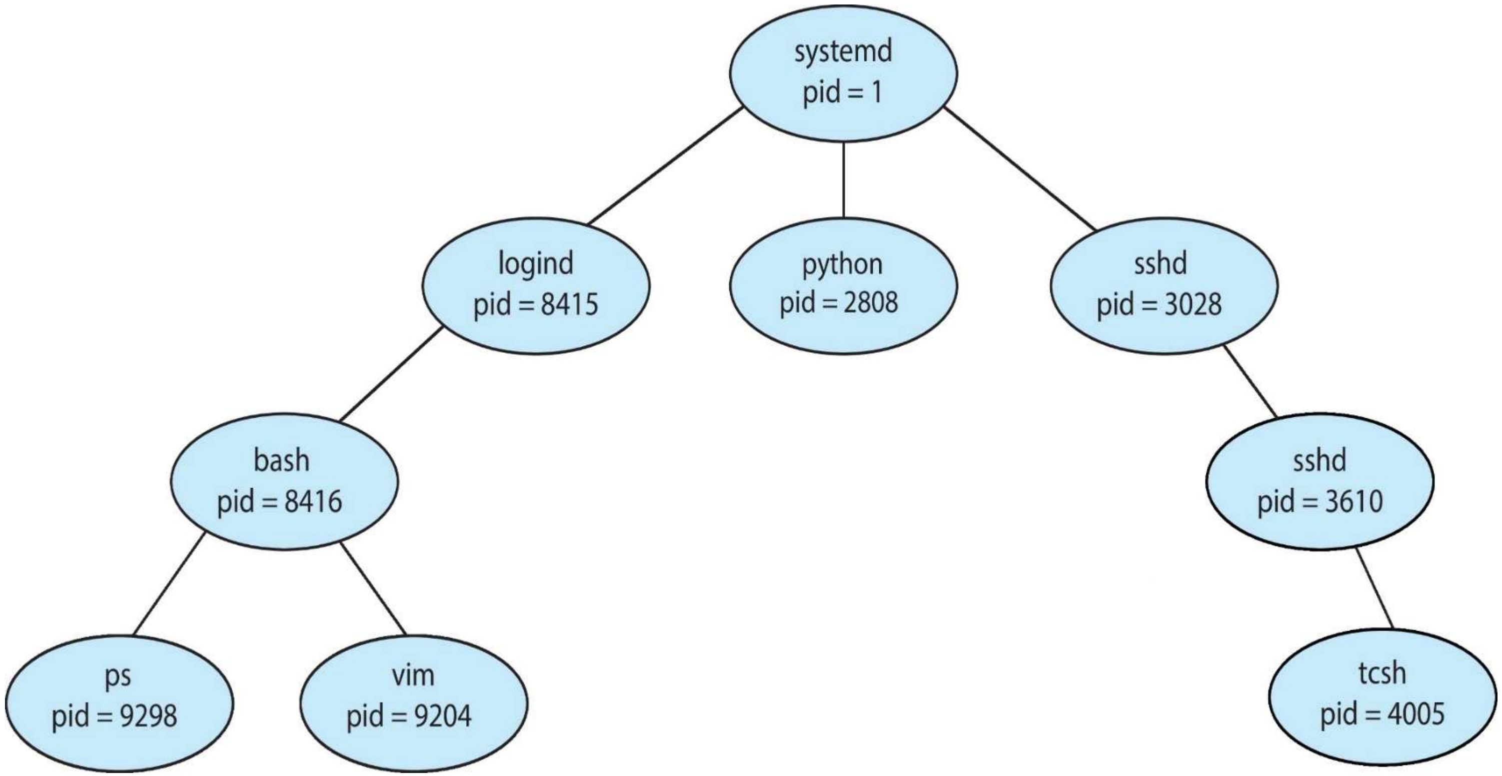
超額分配

Operations on Processes

- We introduce two operations here
  - Creation
  - Termination

Process Creation

- Parent process create children processes, which, in turn create other processes, forming a tree of processes



Generally, a process identified via a **process identifier** (**pid**)

Process Creation (Cont.)

- Parent and Child
  - Resource allocation
    - A child can obtain its resources (CPU time, mem, files, IO dev…) from
      a) OS
      b) Its parent
    - If a child can **only** obtain its resources from its parent
      - Prevent any process from overloading the system by creating too many processes

Process Creation (Cont.)

- Parent and Child
  - Resource sharing
    - Parent and child share all resources
    - Child shares subset of parent's resources
    - Parent and child share no resources
  - Execution
    - Parent and child execute concurrently, or
    - Parent waits until child terminates
  - Address space第4頁的圖
    - Child runs in the same address space as its parent, or
    - Child has its own address space

Process Creation (Cont.)

- UNIX examples
  - **fork** system call creates new process
  - **exec** system call 更改child process的程式
    - **replace** the process' memory space with a new program
    - usually used after a **fork**

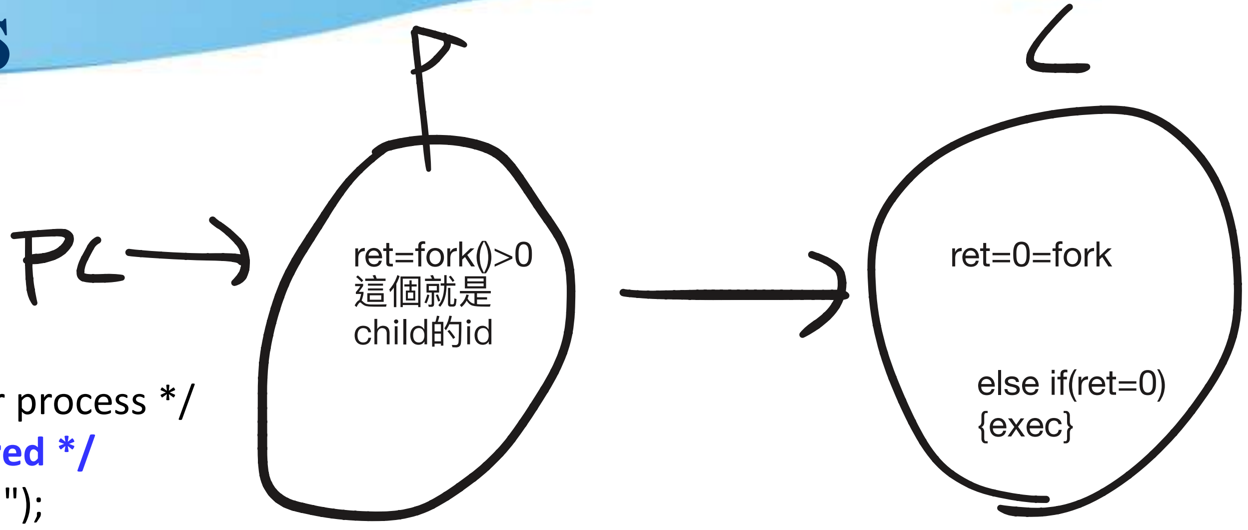
生成一個跟parent一模一樣的process

兩個指令一起使用，就可以生成新的process

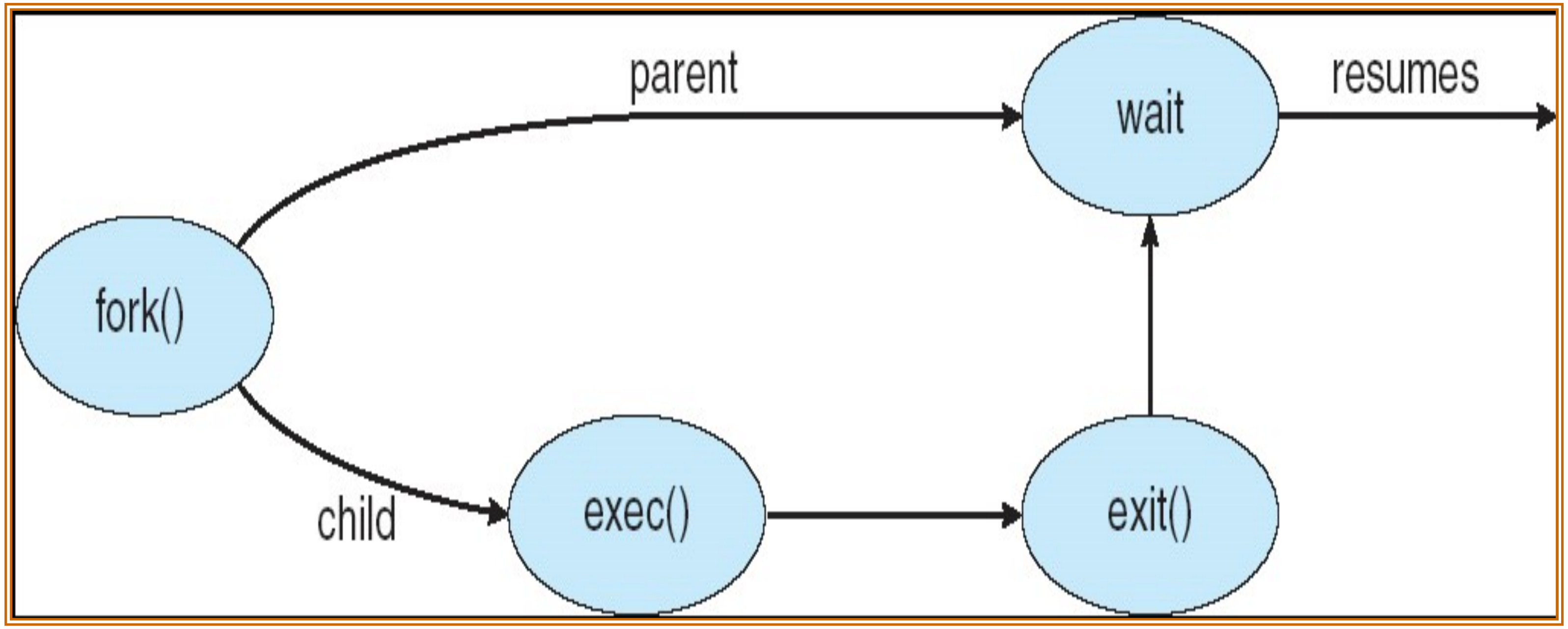
C Program Forking Another Process

```
int main()
{
	pid_t ret, dead;
	int status;

	ret = fork();		/* fork another process */
	if (ret < 0) {		/* error occurred */
		fprintf(stderr, "Fork Failed");
		exit(-1);
	}
	else if (ret == 0) { /* child process */
		execlp("/bin/ls", "ls", NULL); 這行之後寫什麼都沒作用
	}
	else {	/* parent process */	// ret == child's pid
		/* parent will wait for the child to complete */
		dead = wait (&status); // dead == the pid of the child that has died
		printf ("Child Complete");
		exit(0);
	}
}
```



Execution Flow of the Program



Process Termination

1
- Process executes last statement and asks the operating system to delete it (via the **exit()** system call) 正常終止
  - Output data/status to its parent
    - The parent can get the exit status via the **wait()** system call
  - Process' resources are deallocated by operating system

2
- Parent may terminate execution of children processes
  - Reasons Parent終止child
    - Child has exceeded allocated resources
    - Task assigned to child is no longer required

3
- If parent is exiting
  - Some operating system (such as VMS) do not allow child to continue if its parent terminates
    - All children terminated - *cascading termination* 級連、層遞性的
  - In UNIX, cascading termination is not required
    - Child's parent is set to the init/systemd process (pid = 1)

Inter-Process Communication (IPC)

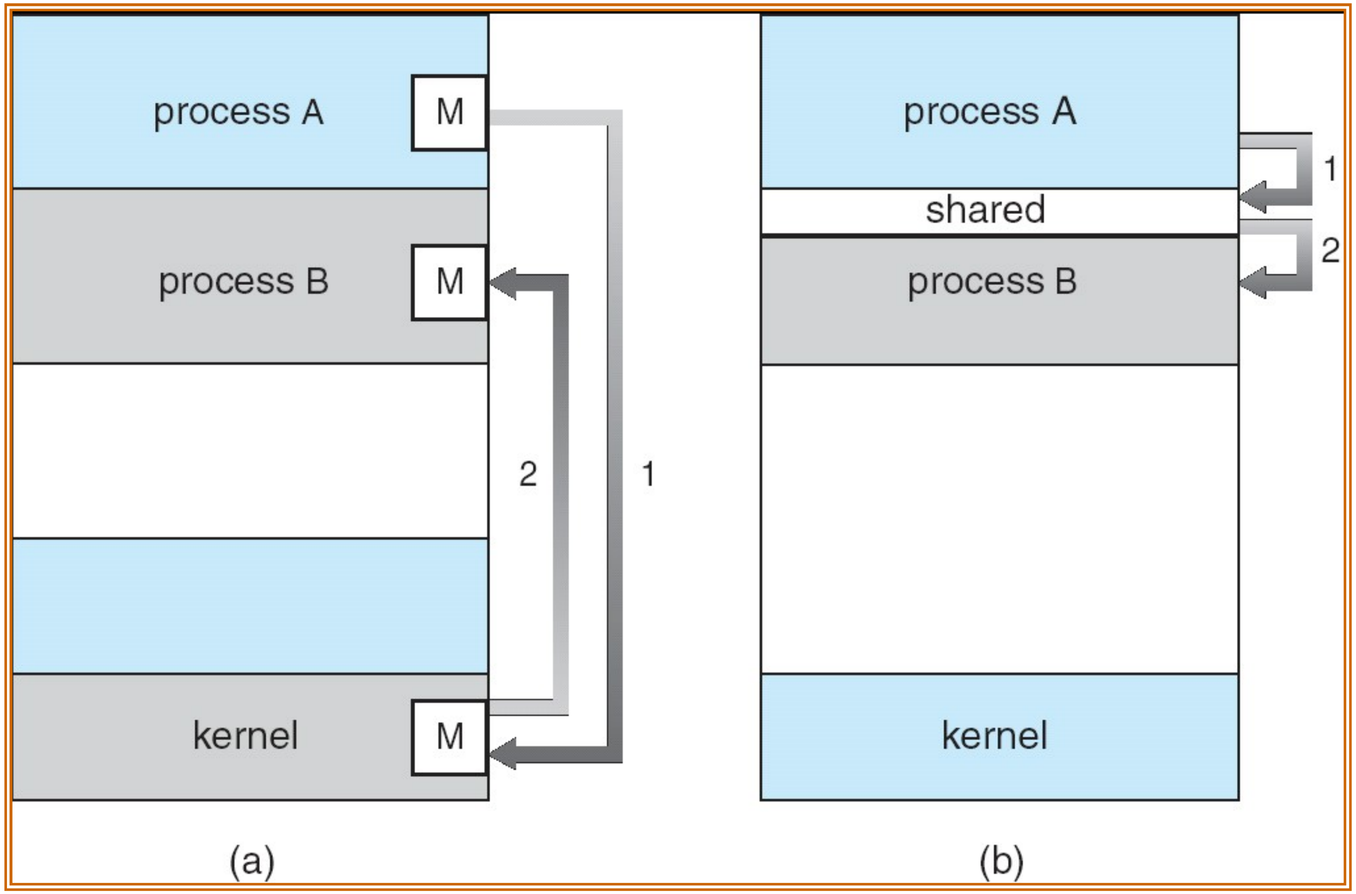
CPU之間的交流

- **Independent** process cannot affect or be affected by the execution of another process
- **Cooperating** process can affect or be affected by the execution of another process
  - Any processes that share data with others are cooperating processes
- Advantages of process cooperation
  - Information sharing
  - Computation speed-up
    - Divide the jobs and run them in parallel on a MP system
  - Modularity

Inter-Process Communication (IPC)

- IPC allows processes to exchange data
- Two fundamental models
  - Shared memory
    - Faster
      - Requires **no** system calls when reading/writing data
  - Message passing
    - Useful for exchanging smaller data
    - Easier to implement
    - Slower; requires system calls for sending/receiving messages

Communications Models



Message passing **Shared memory**

Shared Memory

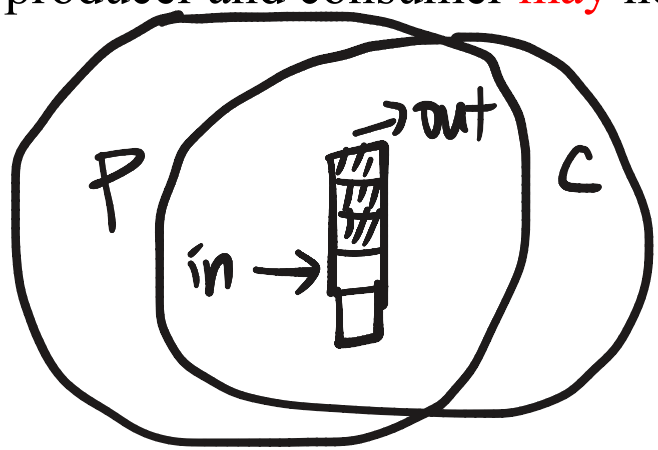
1

- A shared memory region must be created first
- Then, it is attached to the address spaces of the processes
- Memory access can be done freely in the region
  - Originally, OS prevents the address space of a process to be accessed by the other processes.
  - However, the users/processes of the shared memory remove this restriction
  - OS do not care about the data written in the shared memory region
- The region can be detached if it is no longer needed by a process

2

Producer-Consumer Problem

- Paradigm for cooperating processes, *producer* process produces information (i.e. data items) that is consumed by a *consumer* process
  - Use a buffer to contain the data items
  - ***unbounded-buffer*** places no practical limit on the size of the buffer
    - Producer never has to wait; consumer may need to wait
  - ***bounded-buffer*** assumes that there is a fixed buffer size
    - Both producer and consumer may need to wait



交集的就是producer跟consumer的shared memory/buffer

Bounded-Buffer – Shared-Memory Solution

- Shared data

```
#define BUFFER_SIZE 10
typedef struct {
    . . .
} item;

item buffer[BUFFER_SIZE];
int in = 0;
int out = 0;
// in == out : empty,  (in+1) mod BUFFER_SIZE == out: full
```

  – Can only use **BUFFER_SIZE-1** elements

Bounded-Buffer – Producer

```
while (true)
{
   ...Produce an item...
   while ( ((in + 1) % BUFFER SIZE)  == out )
      ;   /* do nothing -- no free buffers */
   buffer[in] = item;
   in = (in + 1) % BUFFER SIZE;
}
```

Bounded Buffer – Consumer

```
while (true)
{
     while (in == out)
          ; // do nothing -- nothing to consume

     // remove an item from the buffer
     item = buffer[out];
     out = (out + 1) % BUFFER SIZE;
     ...Process the item...
}
```

Shared Memory APIs

- System V API
  - shmget()
  - shmat()
  - shmdt()
  - shmctl()
- POSIX (Portable Operating System Interface) Shared Memory API 例如Linux
  - shm_open(), mmap(), munmap(), close(), shm_unlink()…

POSIX Shared Memory Example

- POSIX Shared Memory
  - Process first creates shared memory segment
    shm_fd = shm_open(name, O_CREAT | O_RDWR, 0666); Shared memory open
    - Also used to open an existing segment
  - Set the size of the object

    ftruncate(shm_fd, 4096);
  - Use **mmap()** to memory-map the shared memory object to the address space
  - Reading and writing to shared memory is done by using the pointer returned by mmap()

POSIX Shared Memory Example – Writing a String

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <fcntl.h>
#include <sys/shm.h>
#include <sys/stat.h>

int main()
{
/* the size (in bytes) of shared memory object */
const int SIZE = 4096;
/* name of the shared memory object */
const char *name = "OS";
/* strings written to shared memory */
const char *message_0 = "Hello";
const char *message_1 = "World!";

/* shared memory file descriptor */
int shm_fd;
/* pointer to shared memory obect */
void *ptr;

   /* create the shared memory object */
   shm_fd = shm_open(name, O_CREAT | O_RDWR, 0666);

   /* configure the size of the shared memory object */
   ftruncate(shm_fd, SIZE);

   /* memory map the shared memory object */
   ptr = mmap(0, SIZE, PROT_WRITE, MAP_SHARED, shm_fd, 0);

   /* write to the shared memory object */
   sprintf(ptr,"%s",message_0);
   ptr += strlen(message_0);
   sprintf(ptr,"%s",message_1);
   ptr += strlen(message_1);

   return 0;
}
```

POSIX Shared Memory Example – Reading a String

```
#include <stdio.h>
#include <stdlib.h>
#include <fcntl.h>
#include <sys/shm.h>
#include <sys/stat.h>

int main()
{
/* the size (in bytes) of shared memory object */
const int SIZE = 4096;
/* name of the shared memory object */
const char *name = "OS";
/* shared memory file descriptor */
int shm_fd;
/* pointer to shared memory obect */
void *ptr;

    /* open the shared memory object */
    shm_fd = shm_open(name, O_RDONLY, 0666);

    /* memory map the shared memory object */
    ptr = mmap(0, SIZE, PROT_READ, MAP_SHARED, shm_fd, 0);

    /* read from the shared memory object */
    printf("%s",(char *)ptr);

    /* remove the shared memory object */
    shm_unlink(name);

    return 0;
}
```

Message Passing

- Message system – processes communicate with each other by sending/receiving messages
  - without resorting to shared variables/memory
- Two major operations
  - **send**(*message*)
  - **receive**(*message*)
- Message size can be fixed or variable
- If P and Q wish to communicate, they need to
  - have a *communication link* between them
  - exchange messages via send/receive

Implementation Questions

- How are links established?
- Can a link be associated with more than two processes?
- How many links can there be between every pair of communicating processes?
- What is the capacity of a link?
- Is the size of a message fixed or variable?
- Is a link unidirectional or bi-directional?

Direct Communication

明確指定對方的id

- Processes must name each other explicitly:
  - **send** (*P, message*) – send a message to process P
  - **receive**(*Q, &message*) – receive a message from process Q
- Properties of communication link
  - Links are established automatically
  - Need to know each other's identity
  - A link is associated with exactly one pair of communicating processes
  - Between each pair there exists exactly one link
- The API above is symmetric in addressing 送的時候指定 收的時候也指定
- Asymmetric addressing
  - **send** (*P, message*) – send a message to process P
  - **receive**(*&id , &message*) – receive a message, the id returns the sender
- Drawback
  - Hard-coding process id
    - Changing process id would cause problems…

直接通訊等於說寫死了

Indirect Communication

- Messages are sent to and received from mailboxes (also referred to as ports)
  Indirect的關鍵
  - Each mailbox has a unique id
  - Processes can communicate only if they share a mailbox
- Properties of communication link
  - Link established only if processes share a common mailbox
  - A link may be associated with many processes
  - Each pair of processes may share several communication links
  - Link may be unidirectional or bi-directional

Link可以是單向或雙向

Indirect Communication

- Operations
  - create a new mailbox
  - send and receive messages through mailbox
  - destroy a mailbox
- Primitives are defined as:

  send(*A, message*) **–** send a message to mailbox A

  receive(*A, &message*) **–** receive a message from mailbox A

Mailbox Ownership

- A mailbox is owned by the creator or the OS
  - Mailboxes owned by the Creator
    - Mailbox is part of the address space of the creator
    - The creator is the only receiver
    - Creator terminates ➔ mailbox disappears…
  - Mailboxes owned by the OS
    - May be still existed after the creator has terminated
    - OS should provide a mechanism for users to delete the mailbox explicitly

Indirect Communication

- Mailbox sharing
  - P_1, P_2, and P_3 share mailbox A
  - P_1 sends; P_2 and P_3 receive
  - Who gets the message?

不確定性

- Solutions
  - Allow a link to be associated with at most two processes
  - Allow only one process at a time to execute a receive operation
  - Allow the system to select arbitrarily the receiver. Sender is notified who the receiver was.

Synchronization

- Message passing may be either blocking or non-blocking
- **Blocking** is considered **synchronous** 阻塞式被視為同步
  – **Blocking send** has the sender block until the message is received 必須同步完成，否則要等待
  – **Blocking receive** has the receiver block until a message is available
- **Non-blocking** is considered **asynchronous** 非阻塞式是非同步
  – **Non-blocking send** has the sender send the message and continue → does not wait for the receiver Sender送完就走
  – **Non-blocking receive** has the receiver receive a valid message or null

Message Buffering

- A queue of messages attached to the link 訊息先放到一個序列中，等人來拿
- Implemented in one of three ways
  - Zero capacity – 0 message
    Sender has to wait for receiver
    - No buffering
  - Bounded capacity – finite length of *n* messages
    Sender has to wait if link full
    - Automatic buffering
  - Unbounded capacity – infinite length
    Sender does not have to wait
    - Automatic buffering
- Receivers have to wait if there is no message (in the case of blocking receive)

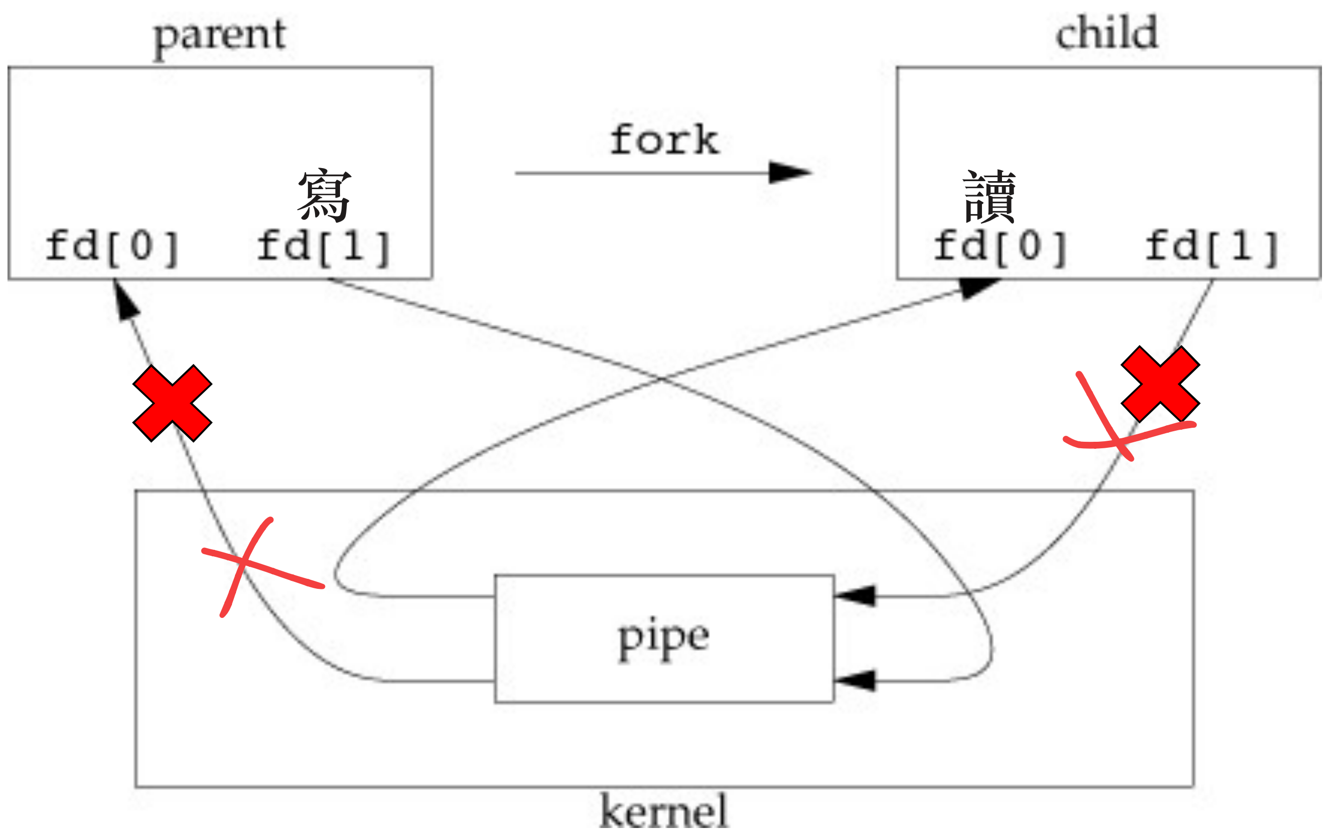
Message Passing APIs

- System V API
  - msgget()
  - msgsnd()
  - msgrcv()
  - msgctl()
- POSIX Message Queue API
  - mq_xxx() functions
    - mq_open(), mq_send(), mq_receive(), mq_close(), mq_unlink()….

Pipes

- A container for byte stream Pipe是用來傳byte stream的容器
- Producer writes to one end (the **write-end** of the pipe) 寫端
- Consumer reads from the other end (the **read-end** of the pipe) 讀端
- Pipes are typically **uni**directional 單向的
  - According to POSIX.1-2001
  - On some systems, pipes are bidirectional
- Require parent-child relationship between communicating processes 使用pipe是，必須是父子關係
- Windows calls these **anonymous pipes**

IPC Using Pipes



Source: http://www.cs.columbia.edu/~jae/4118/L05-ipc.html

IPC Using Pipes – An Example

```
…
int main(int argc, char *argv[])
{
    int pipefd[2];
    pid_t cpid;
    char buf;
    if (argc != 2) {
        fprintf(stderr, "Usage: %s <string>\n", argv[0]);
        exit(EXIT_FAILURE);
    }
    if (pipe(pipefd) == -1) {
        perror("pipe");
        exit(EXIT_FAILURE);
    }
    cpid = fork();
    if (cpid == -1) {
        perror("fork");
        exit(EXIT_FAILURE);
    }
```

IPC Using Pipes – An Example

```
if (cpid == 0) {    /* Child reads from pipe */
    close(pipefd[1]);          /* Close unused write end */
    while (read(pipefd[0], &buf, 1) > 0)
        write(STDOUT_FILENO, &buf, 1);
    write(STDOUT_FILENO, "\n", 1);
    close(pipefd[0]);
    exit(EXIT_SUCCESS);

} else {            /* Parent writes argv[1] to pipe */
    close(pipefd[0]);          /* Close unused read end */
    write(pipefd[1], argv[1], strlen(argv[1]));
    close(pipefd[1]);          /* Reader will see EOF */
    wait(NULL);                /* Wait for child */
    exit(EXIT_SUCCESS);
}

}
```

Named Pipes

- Also called FIFOs
  - On Linux, name pipes are created by **mkfifo**()
- Named pipes are more powerful than ordinary pipes
  - No parent-child relationship is necessary between the communicating processes 跟 匿名pipe 的差別
  - Any process can open the FIFO for reading or writing since it has a name
- Provided on both UNIX and Windows systems

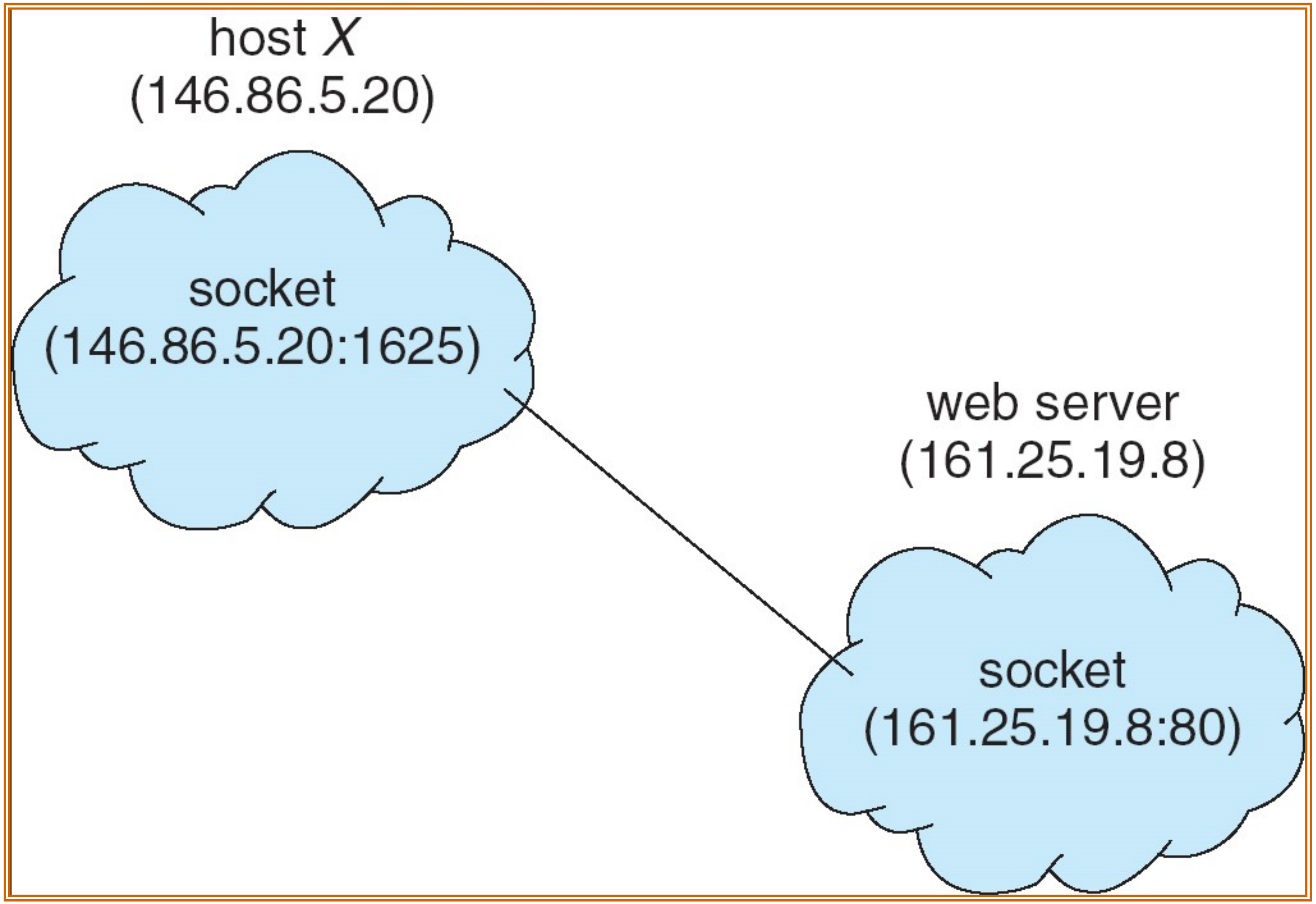
Client-Server Communication

- Sockets
- Remote Procedure Calls (RPC)
- Remote Method Invocation (RMI, for Java)

Sockets

- A socket is defined as an *endpoint for communication*
- Concatenation of IP address and port
- The socket **161.25.19.8:1625** refers to port **1625** on host **161.25.19.8**
  - Some well known ports
    - FTP server: 21
    - HTTP server: 80
- Communication consists between a pair of sockets

Socket Communication



Remote Procedure Calls

- Remote procedure call (RPC) abstracts procedure 抽象化 calls between **processes on networked systems**
- Allow a client to **invoke a procedure** ***on a remote host*** as it would invoke a procedure locally
  Client可以呼叫遠端主機的程式
- Based on message exchange
  - Each message is sent to the RPC daemon listening to a port
    - A message contains function id, and function parameters
  - Results are sent back as separated messages

Remote Procedure Calls

- **Stubs**
  - transforms procedure calls to msgs, and vice versa.
  - Hide the communication details
- The client-side stub locates the server and *marshals* the parameters 找出server，封裝參數
- The server-side stub receives this message, unpacks the marshaled parameters, and performs the procedure on the server 解開封裝的參數，呼叫client要的程式
- The return values are processed in a similar manner

再返還值

Issues about RPC

- Data representation
  - Little Endian? Big Endian?
  - External Data Representation (XDR)
- Error handling
  - RPC can fail, or be executed more than once due to network error

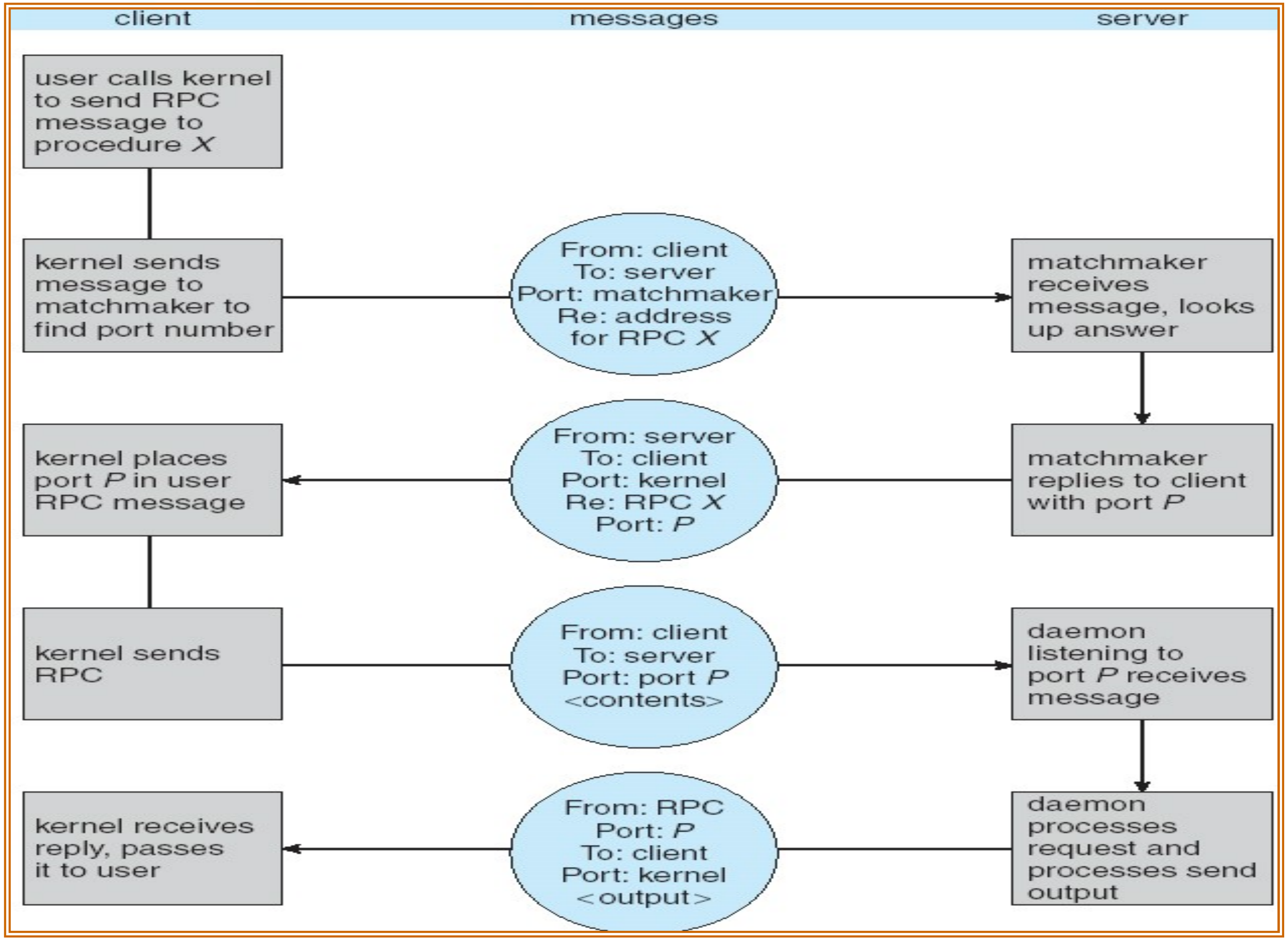
    OS確保只執行一次
  - OS ensures the "exactly once" semantic by (ALO && AMO)
    - At least once (ALO)
      - Server acks the client when the client msg has been received & executed
      - Client msg transmission repeats until ACKed Client一直傳直到收到ack
    - At most once (AMO)
      - Each msg has a timestamp
      - Drop the msg if its timestamp is in the *history* on the server

每個請求都有timestamp，所以被已經接受過的訊息不會重複接受

Issues about RPC

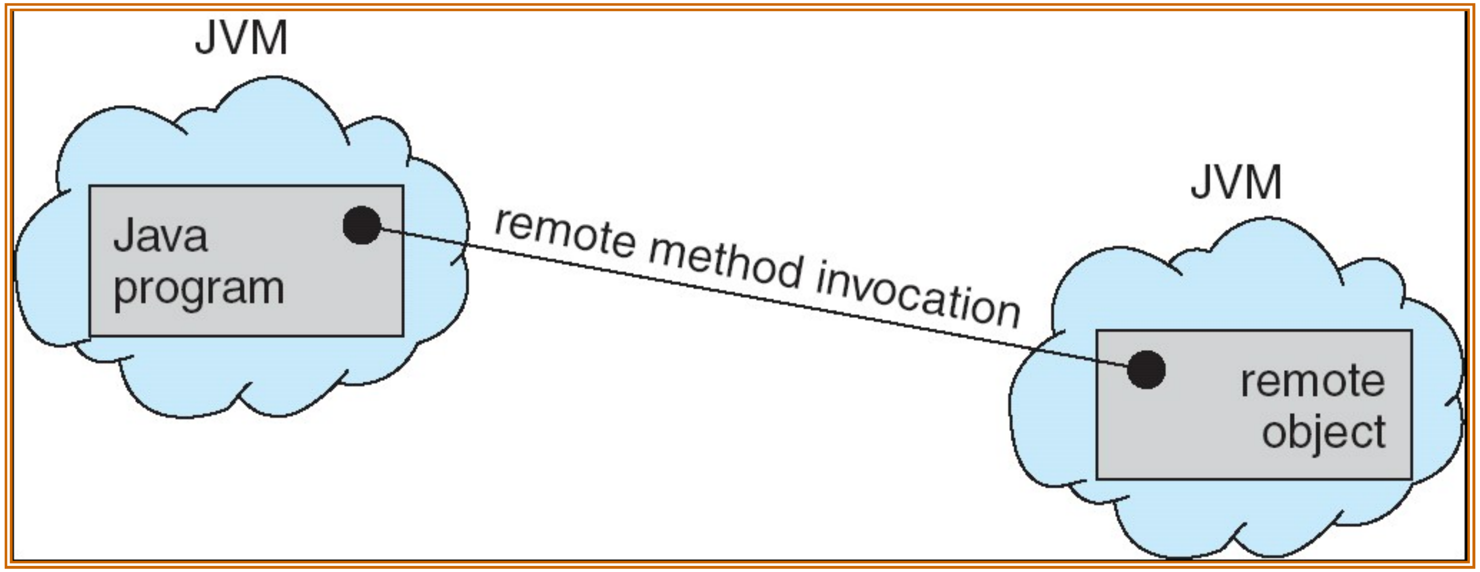
- Binding
  - How does the client knows the port number of the server?
    - Fixed port number 固定port number
      - Inflexible
        - » Server can not change its port easily
    - Dynamic binding Port mapper
      - A rendezvous daemon (a matchmaker) listening on a fixed port to find out the port of the server
      - See the next slide

Execution of RPC



Remote Method Invocation

- Remote Method Invocation (RMI) is a Java mechanism similar to RPCs
- RMI allows a Java program on one machine to invoke a method on a remote object
- Different from RPC, RMI is
  - Object-based
  - Possible to pass objects as parameters



Marshalling Parameters