Web Crawler Practice

Web Architecture

Dr. Chun-Hsiang Chan

Department of Geography,
National Taiwan Normal University

Outline

- Introduction
- Equipment
- Interconnection Model
- Lab Practice

Introduction

- Before we introduce the static web crawler, you have to know these fundamental knowledge in facilitating the understanding how the network operates.
- We will cover from the physical components to the signal components.

Introduction to Networking

- **Server:** a (high-performance) computer or software provides service and access permissions for multiple users via internet, such as print server, mail server, web server, fax server, and file server, etc.
- **Workstation:** a (personal) computer provides general users to access internet resources.
- **Host:** any device with IP address and an ability to connect to internet, such as printer, laptop, or PDA, etc.

Introduction – Web framework

- According to the resource sharing method, we may simply divides web framework into two categories.
- **Peer to peer web hosting:** all network members could provide and receive resources; therefore, the positions of all members are equal – Peer to peer web. In the same network system, other members can access all resouces with permission.
- **Pros:** simple architecture without server and low cost
- **Cons:** difficult to access resources to multiple users, difficult to manage resources due to distributed resources.

Introduction – Web framework

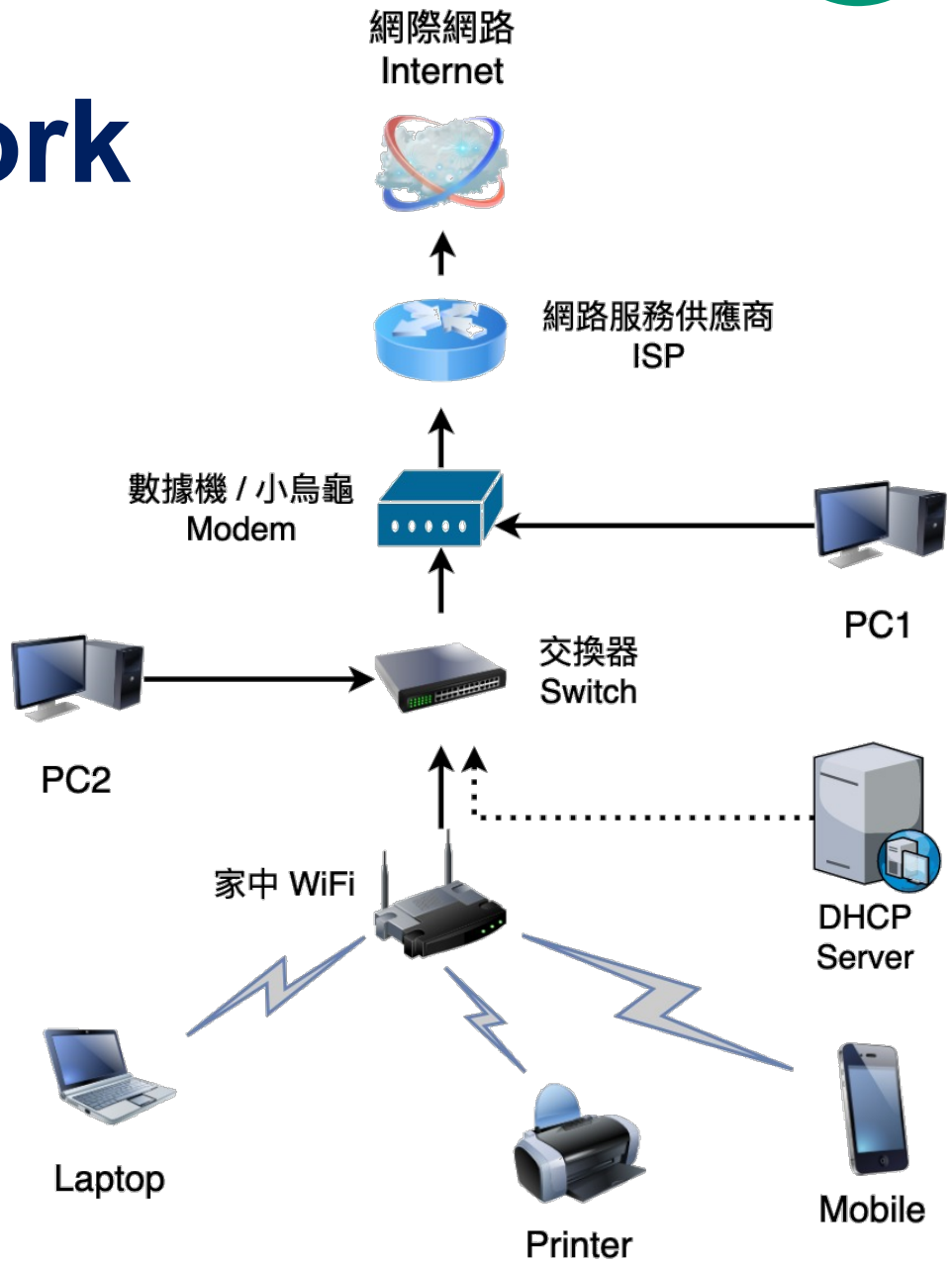
- **Client-server model:** consists of servers providing service or share resource and clients using service or access shared resource. Servers manage shared resource and arrange all requests from clients.
- **Pros:** easy to manage, keep stable service, and allow multiple users' connections
- **Cons:** higher management cost with professional staffs

Introduction – Web Scale

- Web system could be divided into two types by their scales
- **Local Area Network (LAN):** smallest scale; usually connects each other; short-range communication; high-speed communication; low cost.
- **Wide Area Network (WAN):** includes all LANs; relatively low speed and higher cost.

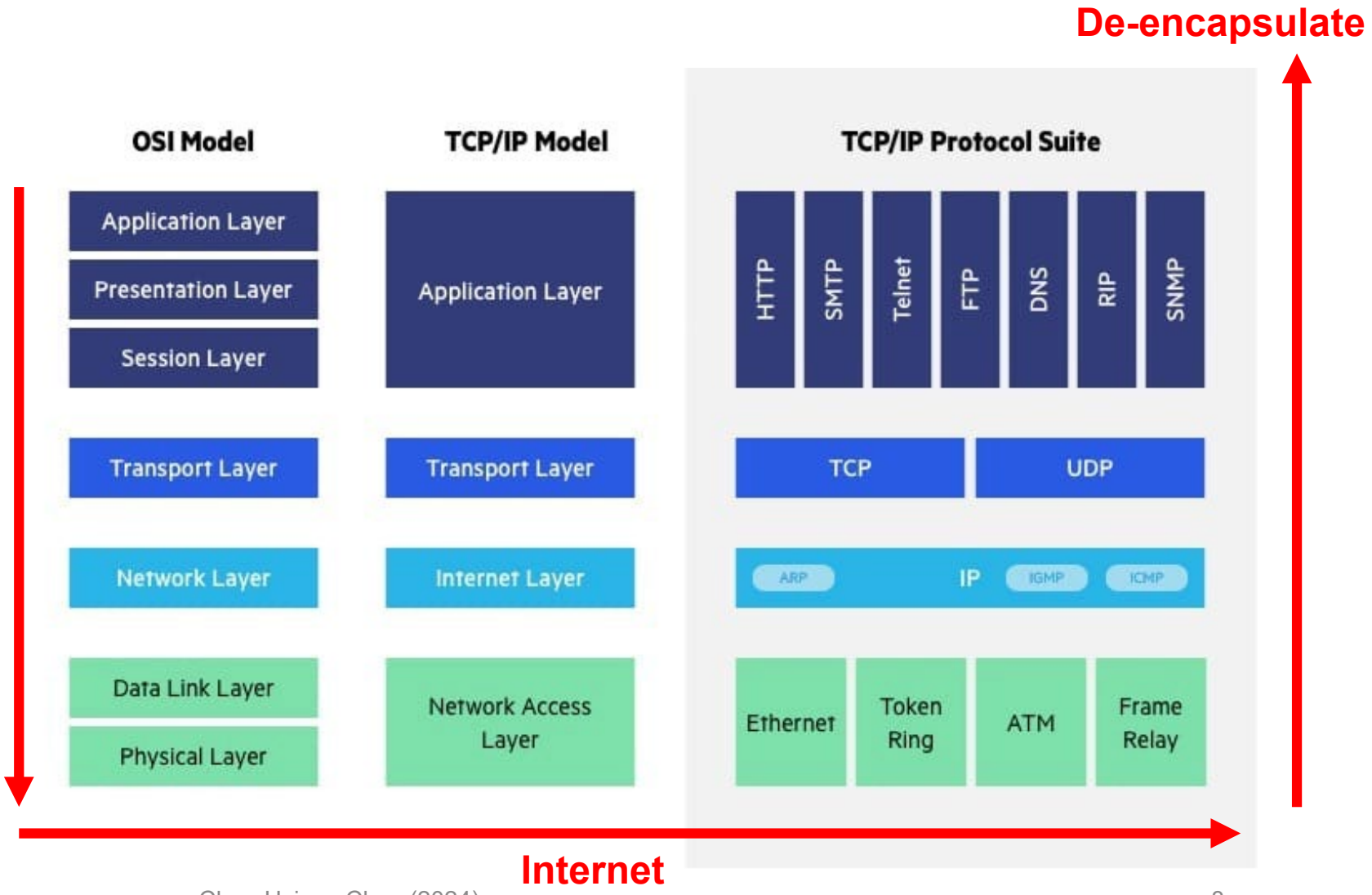
Equipment – Framework

- Internet service providers (ISP)
- Modem
- Switch
- Router
- Access point router (AP router)
- Network interface controller



Interconnection Model

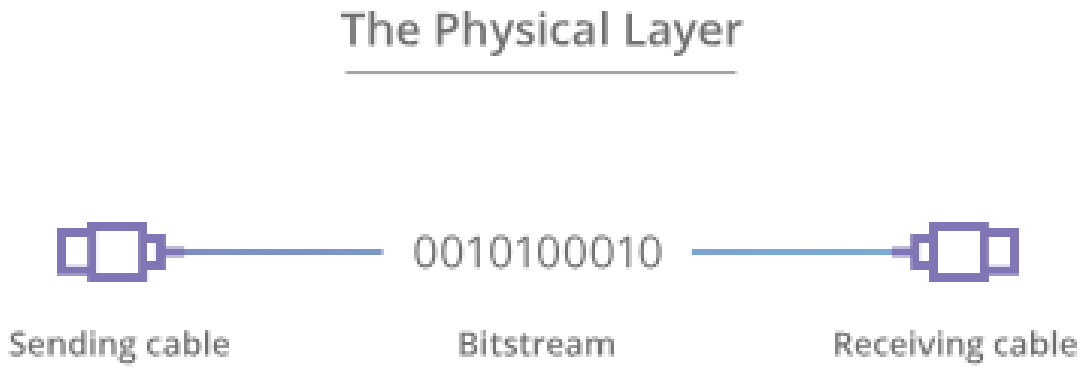
- Open System Interconnection Model (OSI)
- Transfer Control Protocol/Internet Protocol (TCP/IP)



1. Physical Layer

Transmits raw but stream over the physical medium

- The physical layer is responsible for the physical cable or wireless connection between network nodes. It defines the connector, the electrical cable or wireless technology connecting the devices, and is responsible for transmission of the raw data, which is simply a series of 0s and 1s, while taking care of bit rate control.



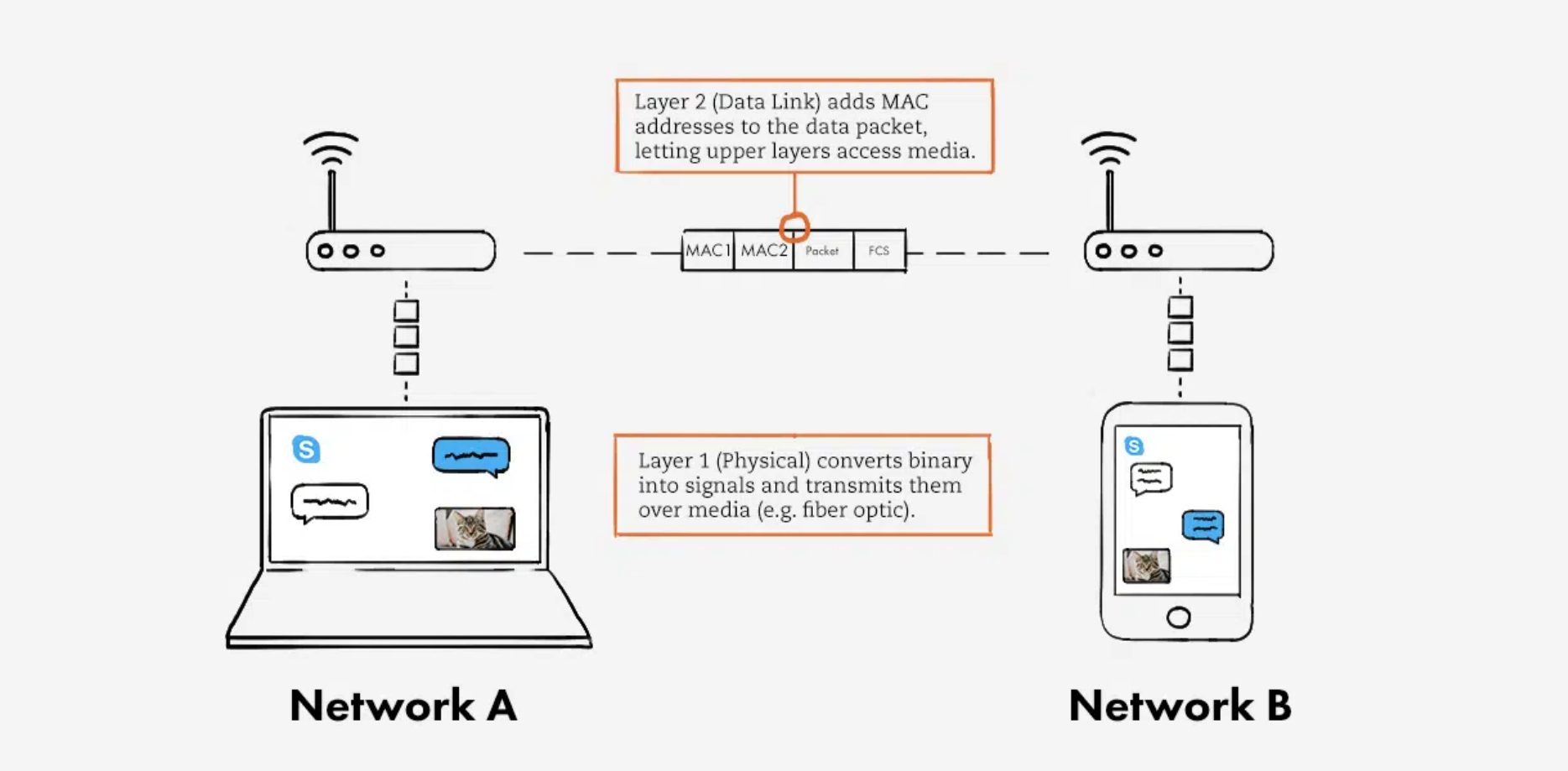
2. Data Link Layer

Defines the format of data on the network

- The data link layer establishes and terminates a connection between two physically-connected nodes on a network. It breaks up packets into frames and sends them from source to destination. This layer is composed of two parts — Logical Link Control (LLC), which identifies network protocols, performs error checking and synchronizes frames, and Media Access Control (MAC) which uses MAC addresses to connect devices and define permissions to transmit and receive data.

2. Data Link Layer

Ethernet (IEEE 802.3)
Wi-Fi (IEEE 802.11)



MAC Address (Media Access Control Address)

- Physical address: In the internet world, network interface controller is the ID which is set at the factory.
- Normally, we will not change the MAC with 48 bits (6 bytes), expressed in hexadecimal.

MAC Encapsulation

- As we know, the maximum communication size of the standard Ethernet is 1500 bytes – that is maximum transmission unit (MTU).
- However, the header information occupies 8 to 10 bytes; therefore, theoretical MTU is 1480 to 1492 bytes.

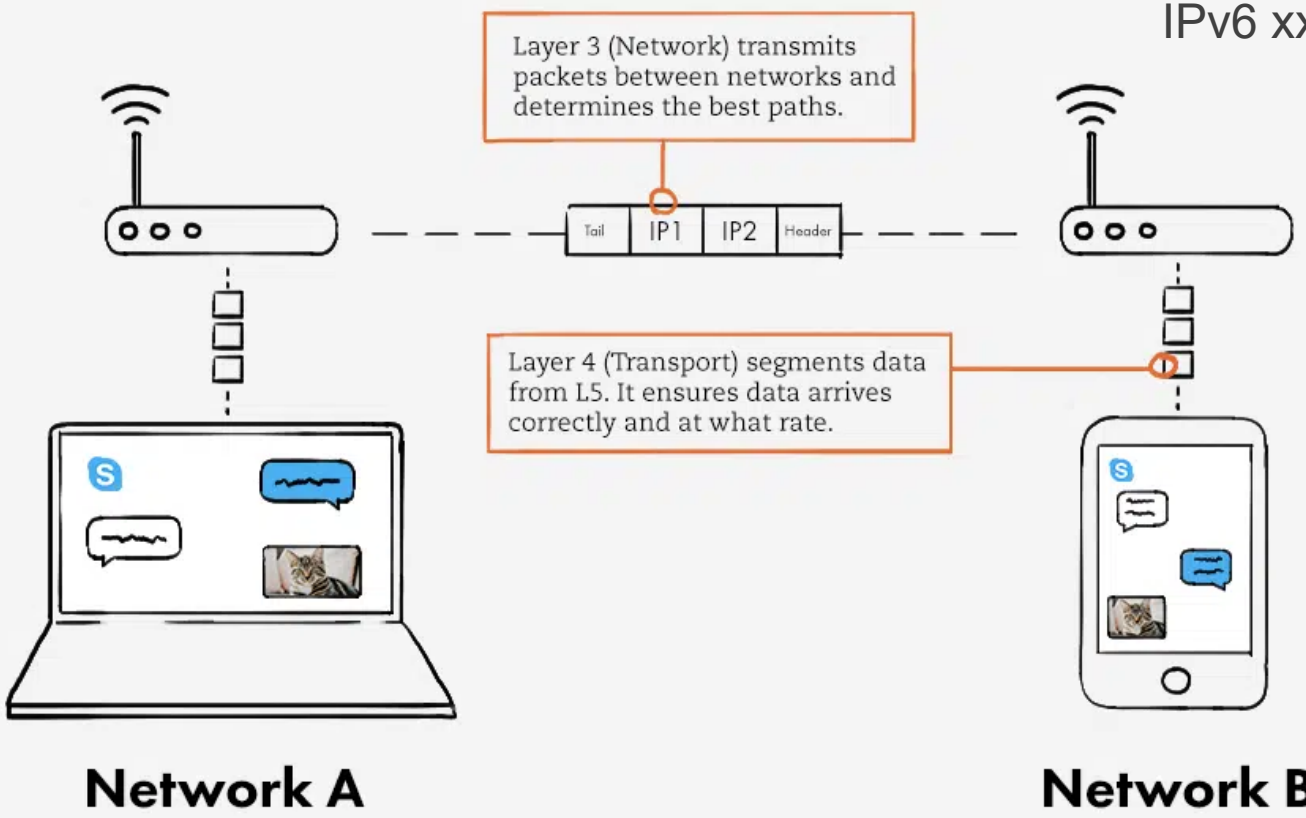
3. Network Layer

Decides which physical path the data will take

- The network layer has two main functions. One is breaking up segments into network packets and reassembling the packets on the receiving end. The other is routing packets by discovering the best path across a physical network. The network layer uses network addresses (typically Internet Protocol addresses) to route packets to a destination node.

3. Network Layer

IPv4 xxx.xxx.xxx.xxx (8 bit)
IPv6 xxx:xxx:xxx:xxx (32 bit)



IP Packet Encapsulation

- The size of IP packet could be 65535 bytes ($0 \sim 2^{16} - 1$).
- Usually, the data size is larger than MAC, and then our operation system will separate the IP into several segments and send them out.
- Once you receive all data packets together, the destination computer will re-construct them back.

IP Packet Encapsulation

4 bits	**4 bits**	**8 bits**	**3 bits**	**13 bits**
Version	Internet Header Length	Type of service	Total length	
Identification			Flags	Fragmentation
Time to live	Protocol		Header Checksum	
Source address				
Destination address				
Options			Padding	
Data				

4. Transport Layer

Transmits data using transmission protocols including TCP and UDP

- The transport layer takes data transferred in the session layer and breaks it into “segments” on the transmitting end. It is responsible for reassembling the segments on the receiving end, turning it back into data that can be used by the session layer. The transport layer carries out flow control, sending data at a rate that matches the connection speed of the receiving device, and error control, checking if data was received incorrectly and if not, requesting it again.

TCP

Three-Way Hand Shake or a TCP 3-way handshake is a process which is used in a TCP/IP network to make a connection between the server and client.

Message	Description
Syn	Used to initiate and establish a connection. It also helps you to synchronize sequence numbers between devices.
ACK	Helps to confirm to the other side that it has received the SYN.
Syn-ACK	SYN message from local device and ACK of the earlier packet.
FIN	Used to terminate a connection.

UDP

User Datagram Protocol

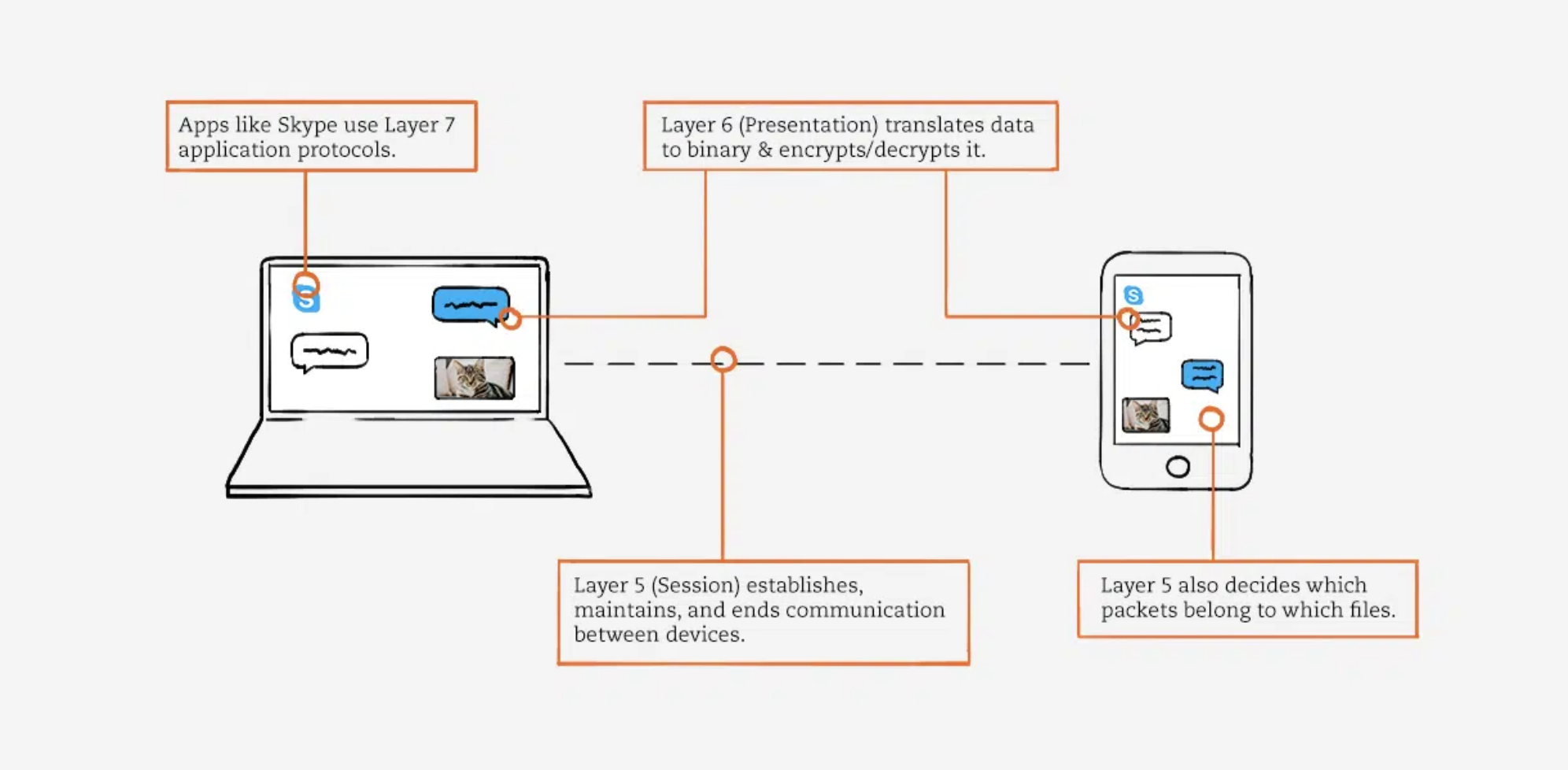
- UDP is a communication protocol used across the Internet for especially time-sensitive transmissions such as video playback or DNS lookups.
- It speeds up communications by not formally establishing a connection before data is transferred.

5. Session Layer

Maintains connections and is responsible for controlling ports and sessions

- The session layer creates communication channels, called sessions, between devices. It is responsible for opening sessions, ensuring they remain open and functional while data is being transferred, and closing them when communication ends. The session layer can also set checkpoints during a data transfer—if the session is interrupted, devices can resume data transfer from the last checkpoint.

5. Session Layer



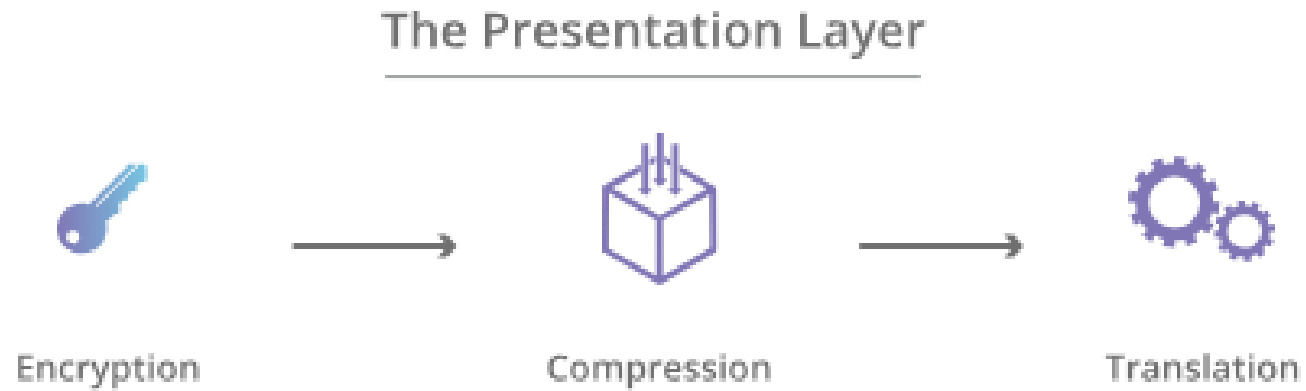
6. Presentation Layer

Ensures that data is in a usable format and is where data encryption occurs.

- The presentation layer prepares data for the application layer. It defines how two devices should encode, encrypt, and compress data so it is received correctly on the other end. The presentation layer takes any data transmitted by the application layer and prepares it for transmission over the session layer.

6. Presentation Layer

- Encoding (decoding), encryption (decryption), and compress (decompression) data.

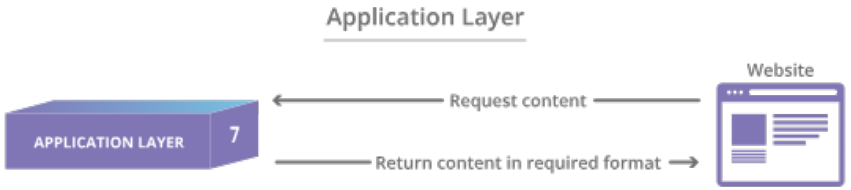


7. Application Layer

Human-computer interaction layer, where applications can access the network services

- The application layer is used by end-user software such as web browsers and email clients. It provides protocols that allow software to send and receive information and present meaningful data to users. A few examples of application layer protocols are the Hypertext Transfer Protocol (HTTP), File Transfer Protocol (FTP), Post Office Protocol (POP), Simple Mail Transfer Protocol (SMTP), and Domain Name System (DNS).

7. Application Layer

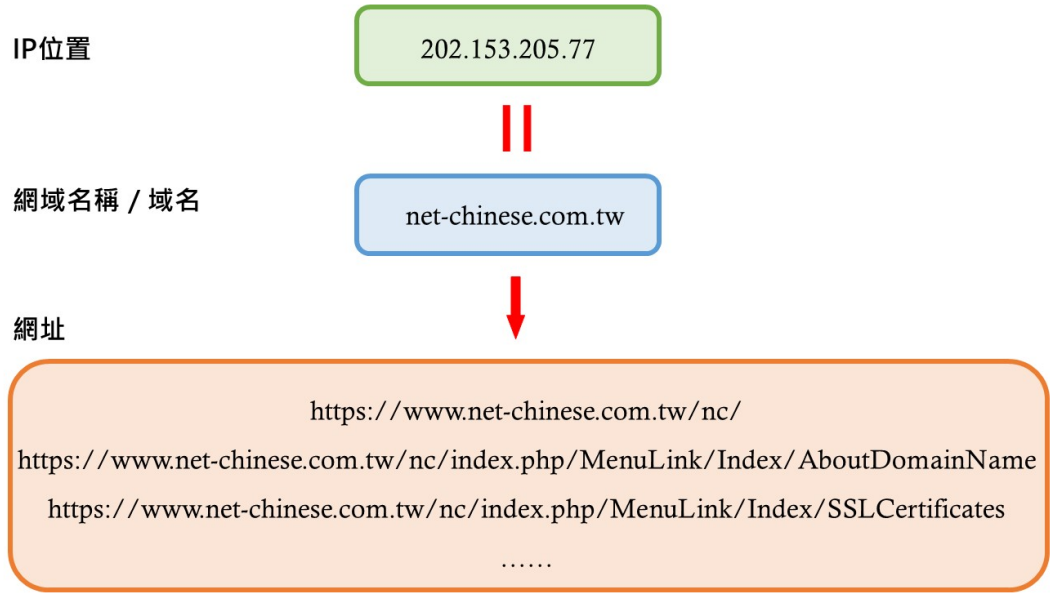


- HTTP (SPDY, HTTP/2), FTP, and DHCP, etc.
- Service: webpage I/O, data exchange, and email, etc.

Port	Service Content
20	FTP-data – data transfer protocols
21	FTP – commands for data transfer protocol
22	SSH – relative safe remote connection server
25	SMTP – simple simple mail transfer protocol
53	DNS – analyze DNS server
80	WWW – World Wide Web
110	POP3 – Post Office Protocol for receiving emails
443	HTTPS - Secured Hypertext Transfer Protocol

Domain Name Service

- It is used to identify the web location of the computer during data transmission. A domain name can be said to be a proxy for an IP address, with the purpose of making it easier to remember the latter.



HTTP Status

Code	Status	Meanings
200	**OK**	**Success**
204	No content	Success, but no content returns
3xx	Redirect	
301	Moved permanently	Resource was moved permanently to the other location; therefore, the browser will directly link to new location in the next request
302	Found (moved temporarily)	Resource was temporally moved to the other location
304	Not modified	The same as previous; get data from cache
4xx	Client error	
400	Bad request	Request syntax error or too large resource
401	Unauthorized	Login or token maybe is required
403	Forbidden	
404	Not found	
5xx	Server error	
500	Internal server error	Server error
502	Bad gateway	Some service on server incorrectly works
503	Service unavailable	Service in maintenance or cannot process request
504	Gateway timeout	No response from the server's service

Source: https://zh.wikipedia.org/zh-tw/HTTP%E7%8A%B6%E6%80%81%E7%A0%81

Lab Practice

- Address Resolution Protocol (ARP): is a network layer address for the web transfer protocols of searching the data link layer address.
- The major usage is to search the other computer in the local internet – simply, it is used to search the corresponding MAC address via IP.
- To understand the network topology, we introduce a new command – “trace route”, which could provide all connection nodes from your location to www.google.com (8.8.8.8).
- Open your terminal and type in the following code:
  1) For windows: tracert
  2) For MacOS: traceroute

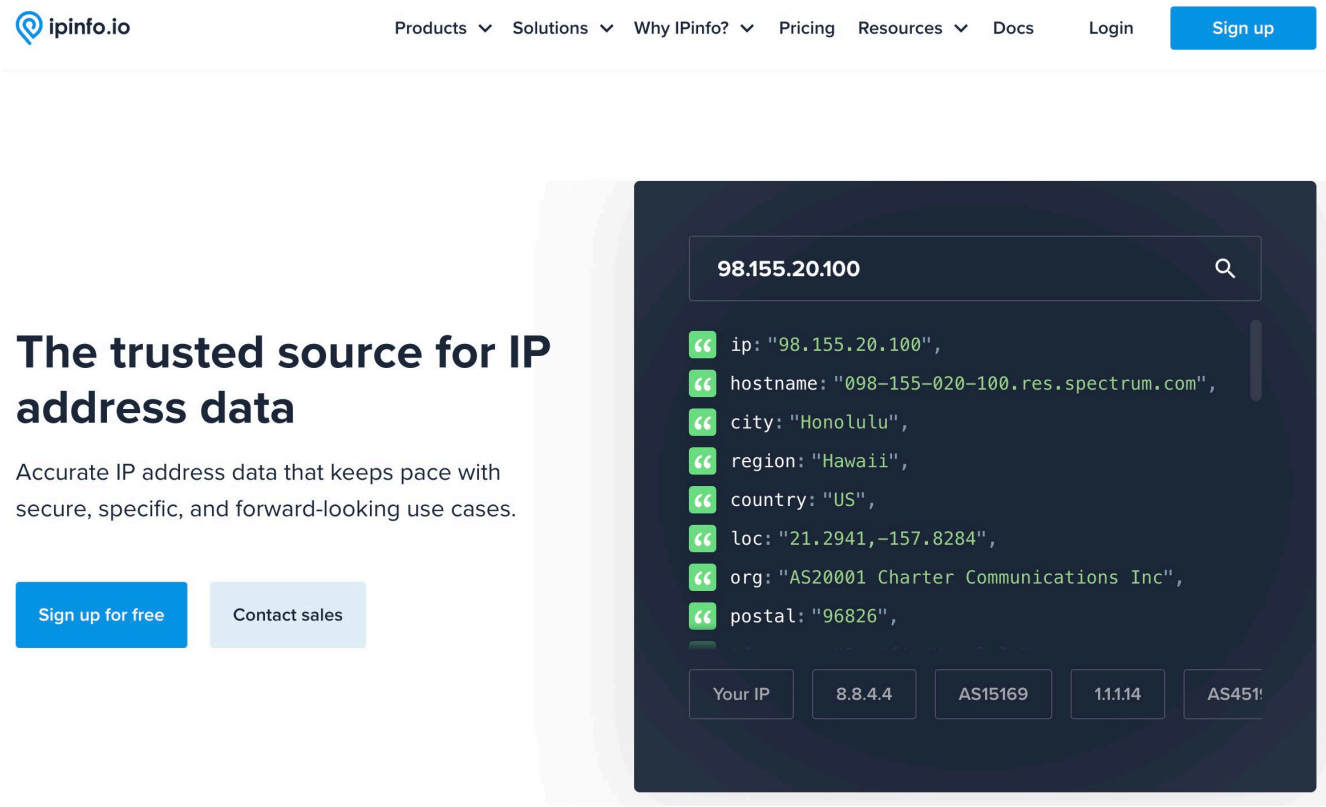
Lab Practice

```
Last login: Sat Apr 13 01:30:11 on ttys000
(base) toodou@TooDous-MBP:~$ traceroute
Version 1.4a12+Darwin
Usage: traceroute [-adDeFInrSvx] [-A as_server] [-f first_ttl] [-g gateway] [-i iface]
        [-M first_ttl] [-m max_ttl] [-p port] [-P proto] [-q nqueries] [-s src_addr]
        [-t tos] [-w waittime] [-z pausemsecs] host [packetlen]
(base) toodou@TooDous-MBP:~$ traceroute 8.8.8.8
traceroute to 8.8.8.8 (8.8.8.8), 64 hops max, 52 byte packets
 1  osync (192.168.1.1)  11.376 ms  5.364 ms  5.631 ms
 2  syn-142-254-191-005.inf.spectrum.com (             )  15.588 ms  15.364 ms  16.303 ms
 3  lag-63.hnllhiku01h.netops.charter.com (             )  29.530 ms  25.946 ms  33.147 ms
 4  lag-37.milnhixd01r.netops.charter.com (             )  17.210 ms  16.377 ms  17.506 ms
 5  lag-31.rcr01lsancarc.netops.charter.com (             )  63.487 ms  61.697 ms  62.385 ms
 6  tge9-1.crlscaij03h.socal.rr.com (             )  63.834 ms  65.853 ms  63.559 ms
 7  * 108.170.247.193 (108.170.247.193)  62.209 ms
    108.170.247.225 (108.170.247.225)  66.104 ms
 8  142.251.60.129 (142.251.60.129)  68.362 ms
    dns.google (8.8.8.8)  64.320 ms
    142.251.60.109 (142.251.60.109)  61.280 ms
```

Lab Practice

- Search ipinfo (https://ipinfo.io/) to obtain the geo-location of IP address.
- The results could be depicted as follows,

IP	coor	location
98.155.20.100	21.2941, -157.8284,	Honolulu, Hawaii, US
…		
…		
…		



References

- 網路概論 https://hackmd.io/@CILS110/SJDLLukJj
- 《 網路架構 》TCP/IP 四層架構 & OSI七層架構
- https://linux.vbird.org/linux_server/centos6/0110network_basic.php#whatisnetwork_tcpip

The End

Thank you for your attention!

Email: chchan@ntnu.edu.tw
Web: toodou.github.io