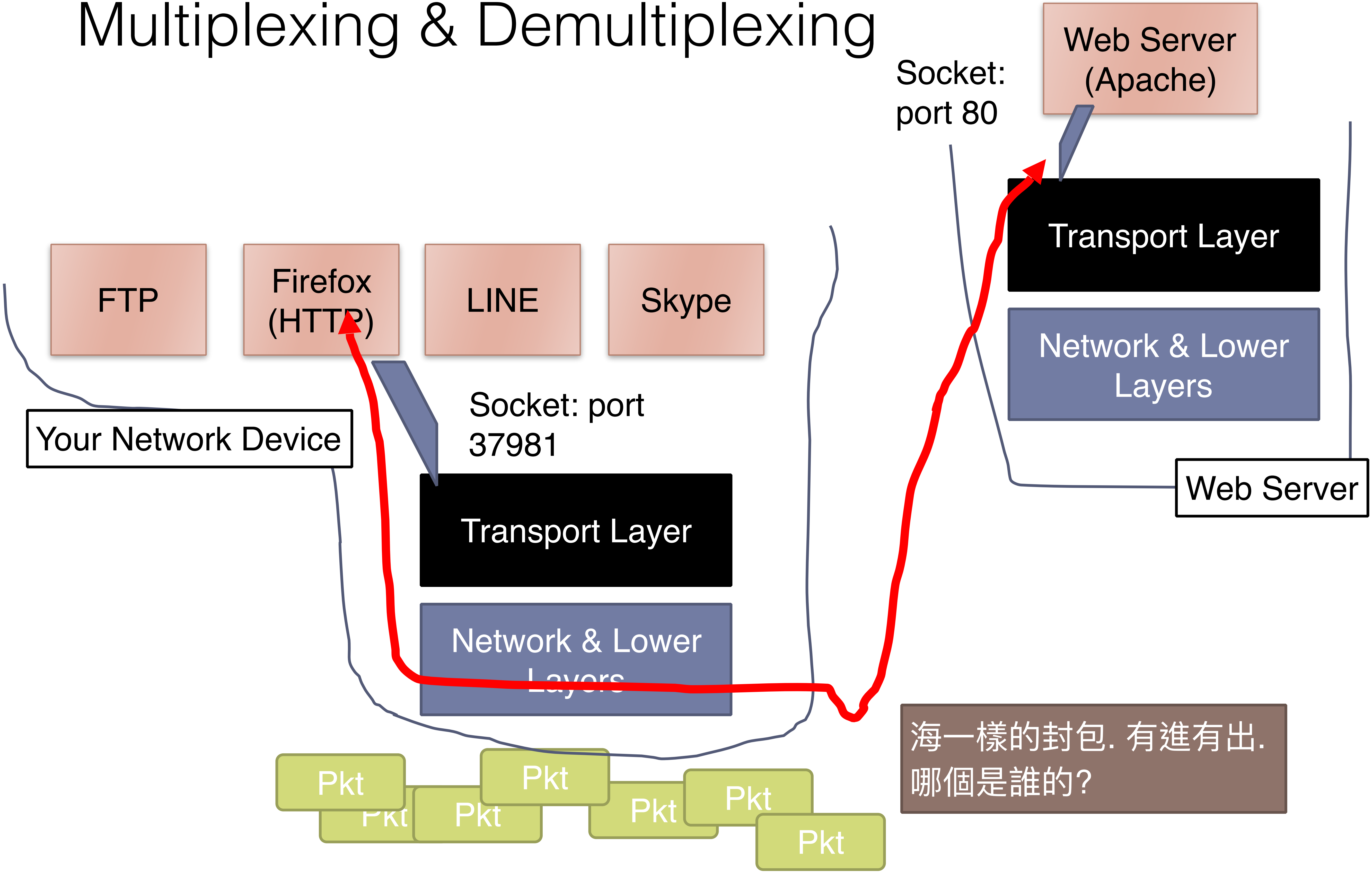
# IP Services + Transport Layer

Michael Tsai
2019/05/13

# Transport Layer

# Multiplexing & Demultiplexing



海一樣的封包. 有進有出.
哪個是誰的?

# Well-Known Port Number

| Service/ Application | Port Number | Service/ Application | Port Number |
|---|---|---|---|
| FTP | 20 (data) / 21 (command) | HTTP | 80 / 8080 |
| SSH | 22 | HTTPS | 443 |
| Telnet | 23 | POP3 (mail client -> server) | 110 |
| SMTP (mail server->server) | 25 | IMAP (mail client -> server) | 143 |
| DNS | 53 | NFS | 2049 |
| DHCP | 67/68 | PPTP (VPN) | 1723 |

# UDP: Connection-less Transport

- Simplest design

- Header has only 4 fields:
  - Source and destination port numbers
  - Length
  - Checksum

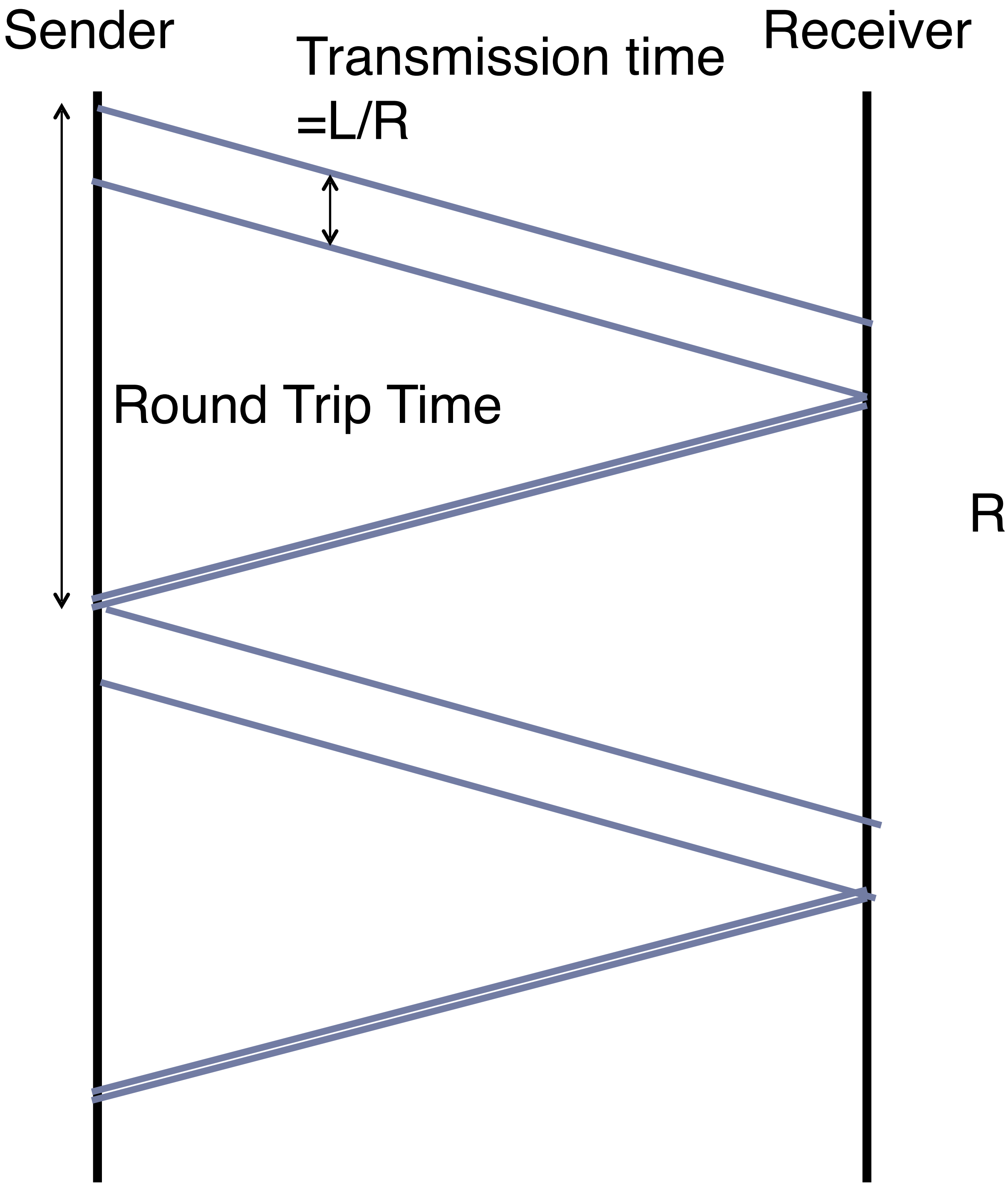
- Why do we need checksum for end-to-end connections?
  - Possible: no bottom layer protocol does it
    - IP only has header checksum (no data checksum)
    - Layer 2 CAN have NO error detection scheme (though most do)
  - Could be caused by relaying hosts (router / switch) (e.g., memory corruption)

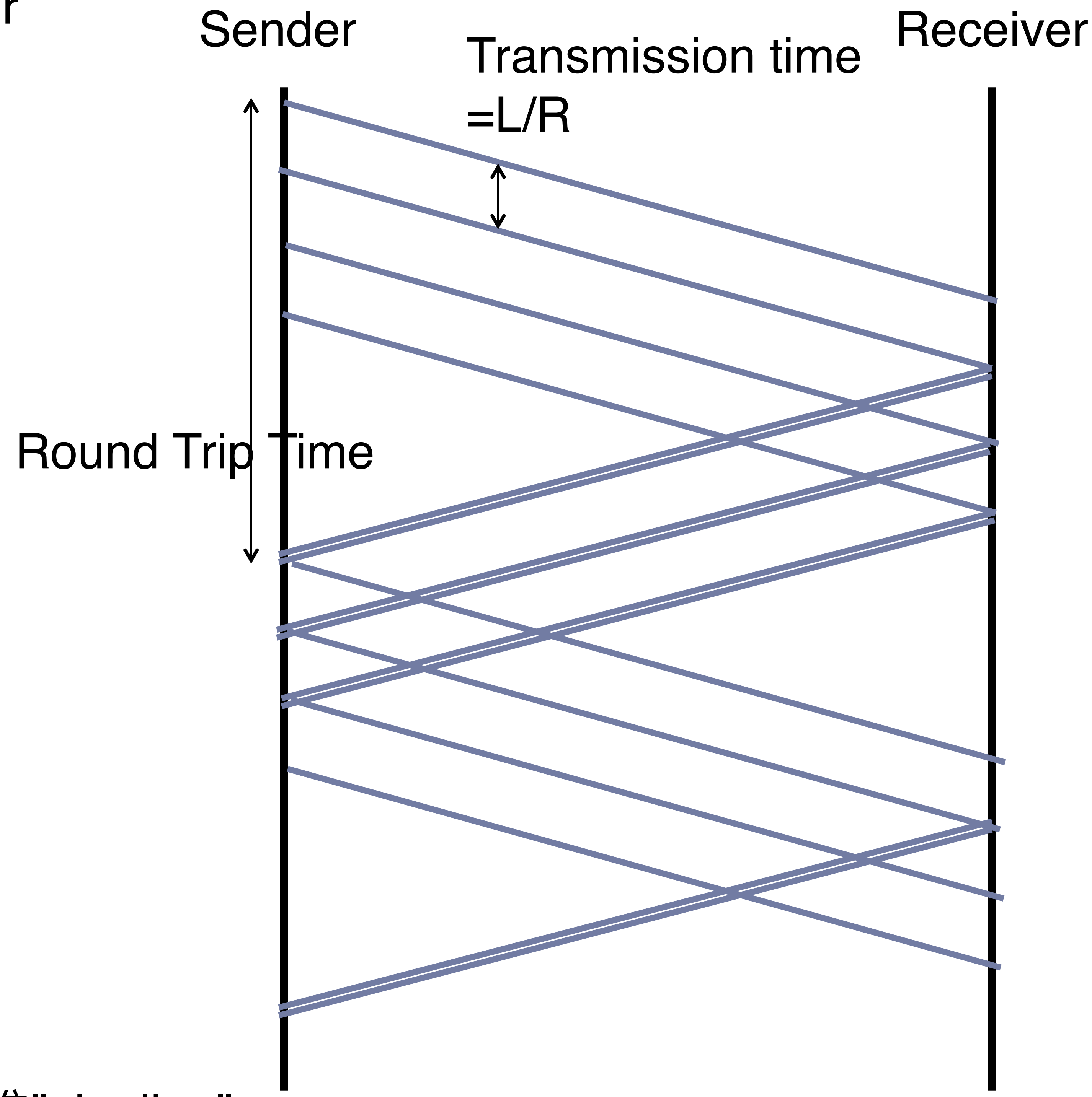
# TCP: Reliable Transport

- Why?
  - IP is not reliable
    - Lost packets, erroneous packets
    - No idea about latency - varying over time

- Basic ingredients:
  - Error detection: verify checksum
  - ACK or NACK:
    notify the other side whether a packet is correctly received
  - Re-transmission: if not, do it again
  - Sequence number: check for missing or redundant packets
  - Time-out: estimate the latency. If not received within a time period, then do it again.
    - Need an accurate estimation of round-trip time (RTT). What if it is too short or too long?

# TCP: Pipelined Transmissions





- 最多可以丟N個”未被ack”的封包進”pipeline”.
- 收到ack代表前面的都收到了

# TCP Designs

- Round-trip Time Estimation
  - EstimatedRTT=0.875 * EstimatedRTT + 0.125 * SampleRTT
  - DevRTT=0.75*DevRTT+0.25* |SampleRTT – EstimatedRTT|
  - TimeoutInterval=EstimatedRTT+4*DevRTT

- Flow Control
  - Make sure the buffer at the receiver is not filled up **(the top layer is too slow)**
  - Receiver will provide the information of the remaining size of the buffer → **receive window**
  - Transmitter makes sure that the transmitted data will not exceed the size of the receive window

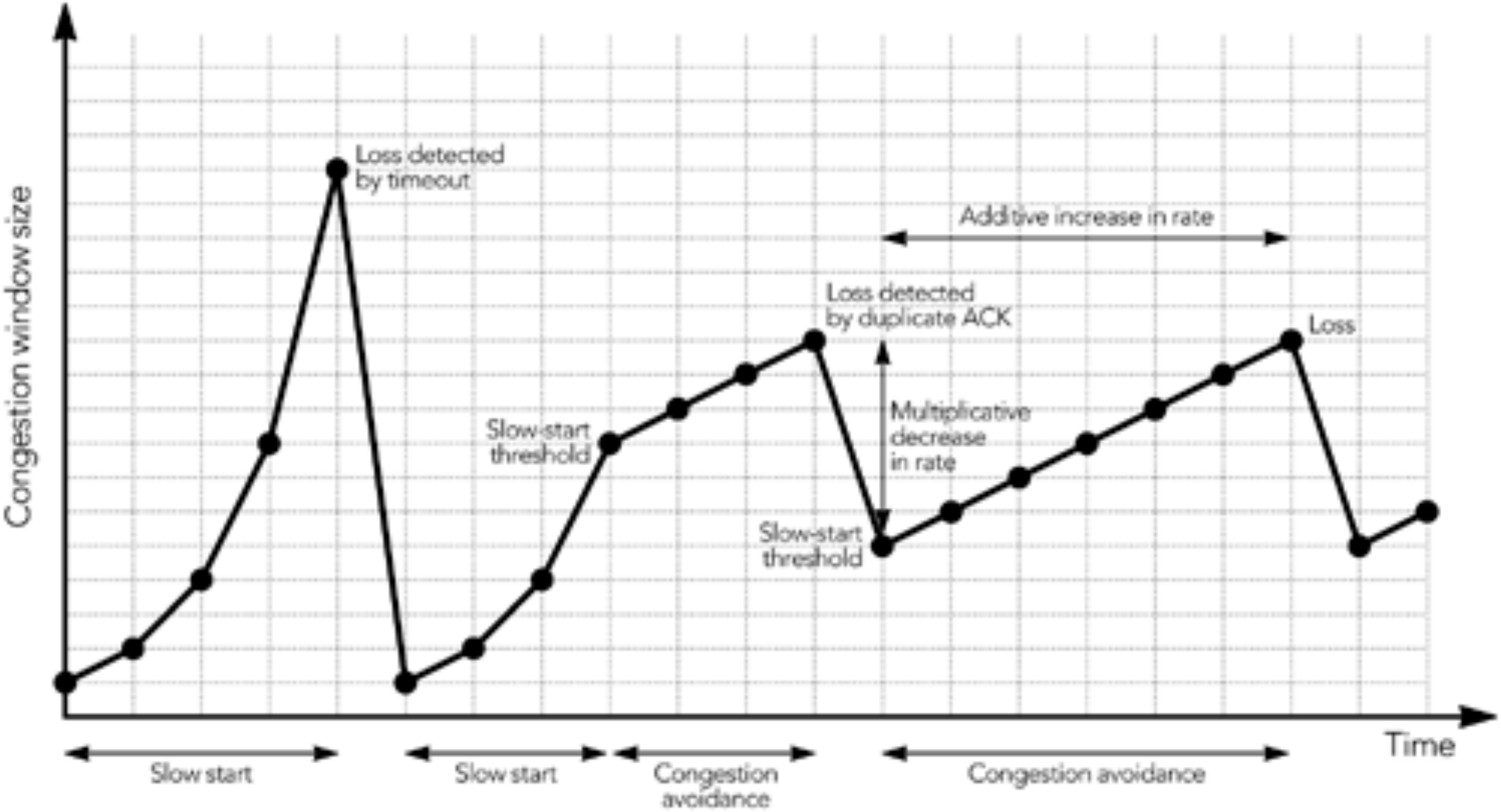
# TCP Designs

- Connection Management
  - **Establish**: SYN (Client initial sequence number) →SYN-ACK →ACK
  - **End**: FIN→ACK

- Congestion Control
  - 因為”一次送多個未ack的封包”而造成的網路更塞車
  - 封包掉 → 塞車了 → 減少傳送速度
  - 收到ACK → 通順的網路 → 增加傳送速度
  - Slow start: 從1開始, 緩慢地增加(2x)一次可以傳送未ack封包個數
  - Congestion avoidance: 超過一個特定的threshold (ssthresh)後，一次加一個未ack封包數

# TCP Congestion Control

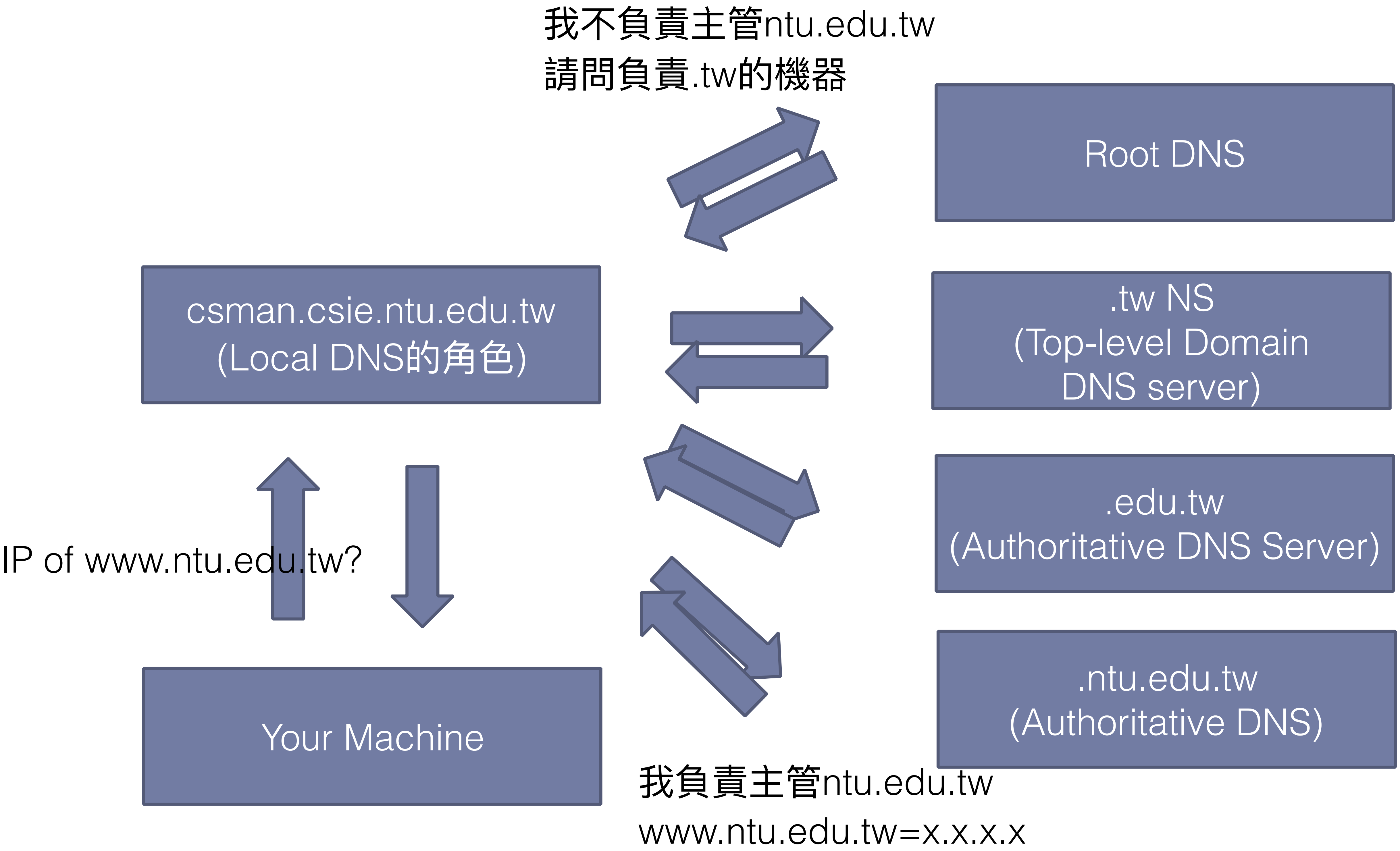
Read: https://flylib.com/books/en/4.245.1.75/1/



# DNS (Domain Name Service)

- 一言以蔽之: 將名稱轉為IP的服務
- 常見的轉換種類:
  - Domain name -> IP (type A):
    ntucsv.csie.ntu.edu.tw -> 140.112.30.28
  - @domainname的mail server (type MX):
    csie.ntu.edu.tw -> ms.csie.ntu.edu.tw
  - Domain name -> domain name (type CNAME):
    www.csie.ntu.edu.tw -> ntucsv.csie.ntu.edu.tw
  - IP -> domain name (type PTR)
    140.112.30.21 -> csman.csie.ntu.edu.tw
- 可以多重宣告: 增加可靠度或分散性.
  - 例如www.google.com的A指到了6個IP!

# 分散式的架構: 分層負責 (recursive query)



我不負責主管ntu.edu.tw
請問負責.tw的機器

Root DNS

csman.csie.ntu.edu.tw
(Local DNS的角色)

.tw NS
(Top-level Domain
DNS server)

.edu.tw
(Authoritative DNS Server)

IP of www.ntu.edu.tw?

.ntu.edu.tw
(Authoritative DNS)

Your Machine

我負責主管ntu.edu.tw
www.ntu.edu.tw=x.x.x.x

# DNS的細節

- 如果local DNS本身主管被查詢的domain，則可以直接回覆。
  - 例如140.112.30.21如果被查詢www.csie.ntu.edu.tw
- Local DNS可以暫存之前查詢過的結果。
  - 主要用來減輕主管DNS server及網路的負擔。
  - 每筆在主管DNS server上的紀錄都有對應的TTL值，規範可以被占存多久。

# /etc/resolv.conf

```
nameserver 140.112.30.21
nameserver 140.112.254.4
nameserver 140.112.2.2
search csie.ntu.edu.tw
```

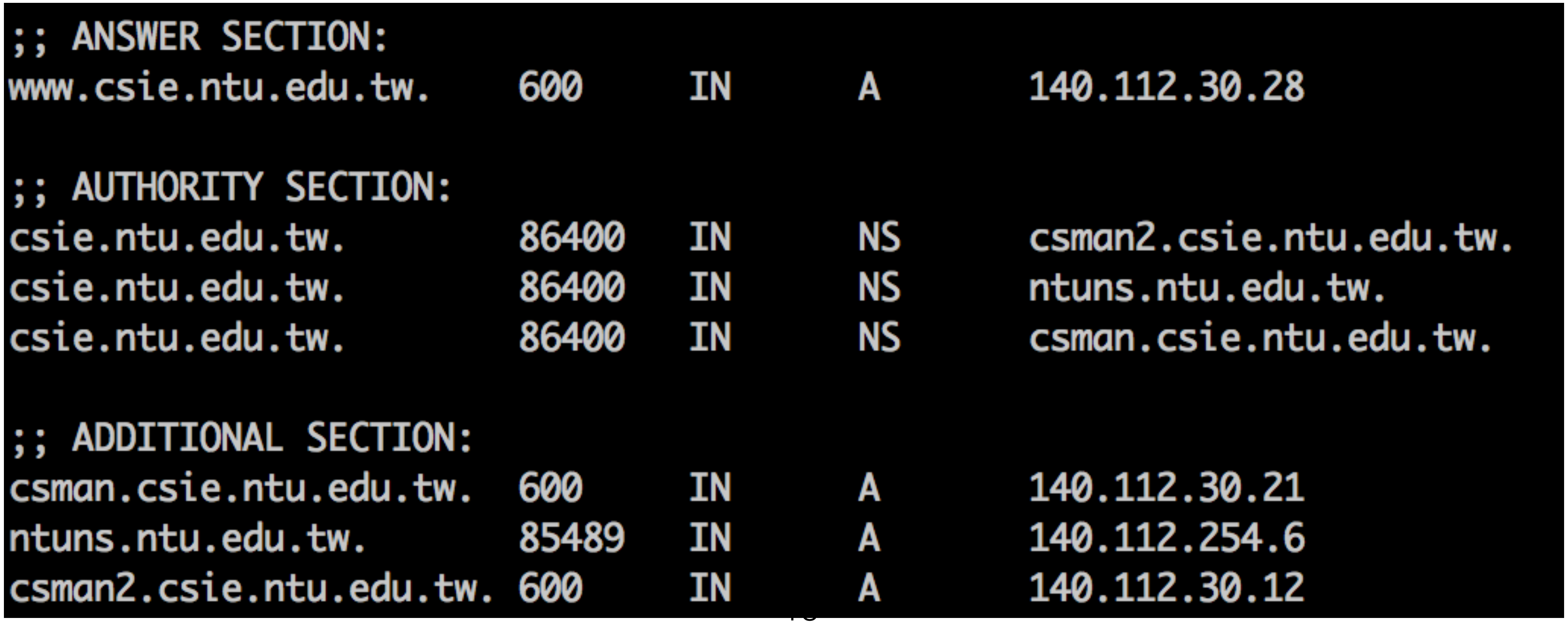
- search
  —> resolve incomplete names (linux1 —> linux1.csie.ntu.edu.tw)
- nameserver —> specify the address of the DNS server

# DNS延伸閱讀

- Top 10 DNS attacks: http://www.networkworld.com/article/2886283/security0/top-10-dns-attacks-likely-to-infiltrate-your-network.html#slide1
- Distributed Reflection DoS attack
- Cache poisoning / DNS hijacking (sol: DNSSEC)
- TCP SYN floods

# 常用DNS指令

- Examples:
  - dig @8.8.8.8 -t MX csie.ntu.edu.tw
  - dig @140.112.30.21 www.csie.ntu.edu.tw



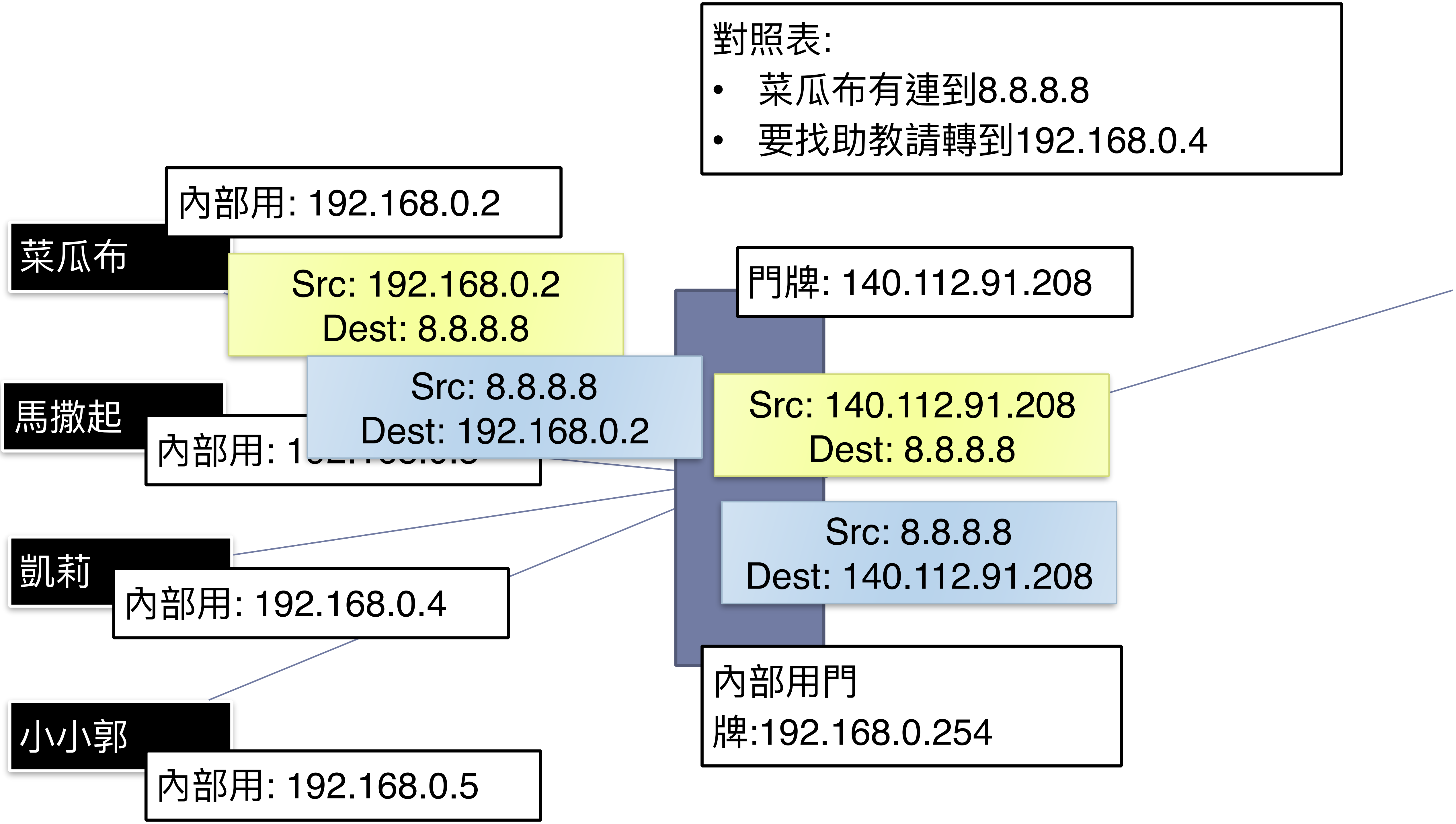
```
;; ANSWER SECTION:
www.csie.ntu.edu.tw.    600     IN      A       140.112.30.28

;; AUTHORITY SECTION:
csie.ntu.edu.tw.        86400   IN      NS      csman2.csie.ntu.edu.tw.
csie.ntu.edu.tw.        86400   IN      NS      ntuns.ntu.edu.tw.
csie.ntu.edu.tw.        86400   IN      NS      csman.csie.ntu.edu.tw.

;; ADDITIONAL SECTION:
csman.csie.ntu.edu.tw.  600     IN      A       140.112.30.21
ntuns.ntu.edu.tw.       85489   IN      A       140.112.254.6
csman2.csie.ntu.edu.tw. 600     IN      A       140.112.30.12
```

# NAT

# NAT (Network Address Translation) Revisited



對照表:
- 菜瓜布有連到8.8.8.8
- 要找助教請轉到192.168.0.4

菜瓜布

內部用: 192.168.0.2

Src: 192.168.0.2
Dest: 8.8.8.8

Src: 8.8.8.8
Dest: 192.168.0.2

馬撒起

內部用: 192.168.0.3

凱莉

內部用: 192.168.0.4

小小郭

內部用: 192.168.0.5

門牌: 140.112.91.208

Src: 140.112.91.208
Dest: 8.8.8.8

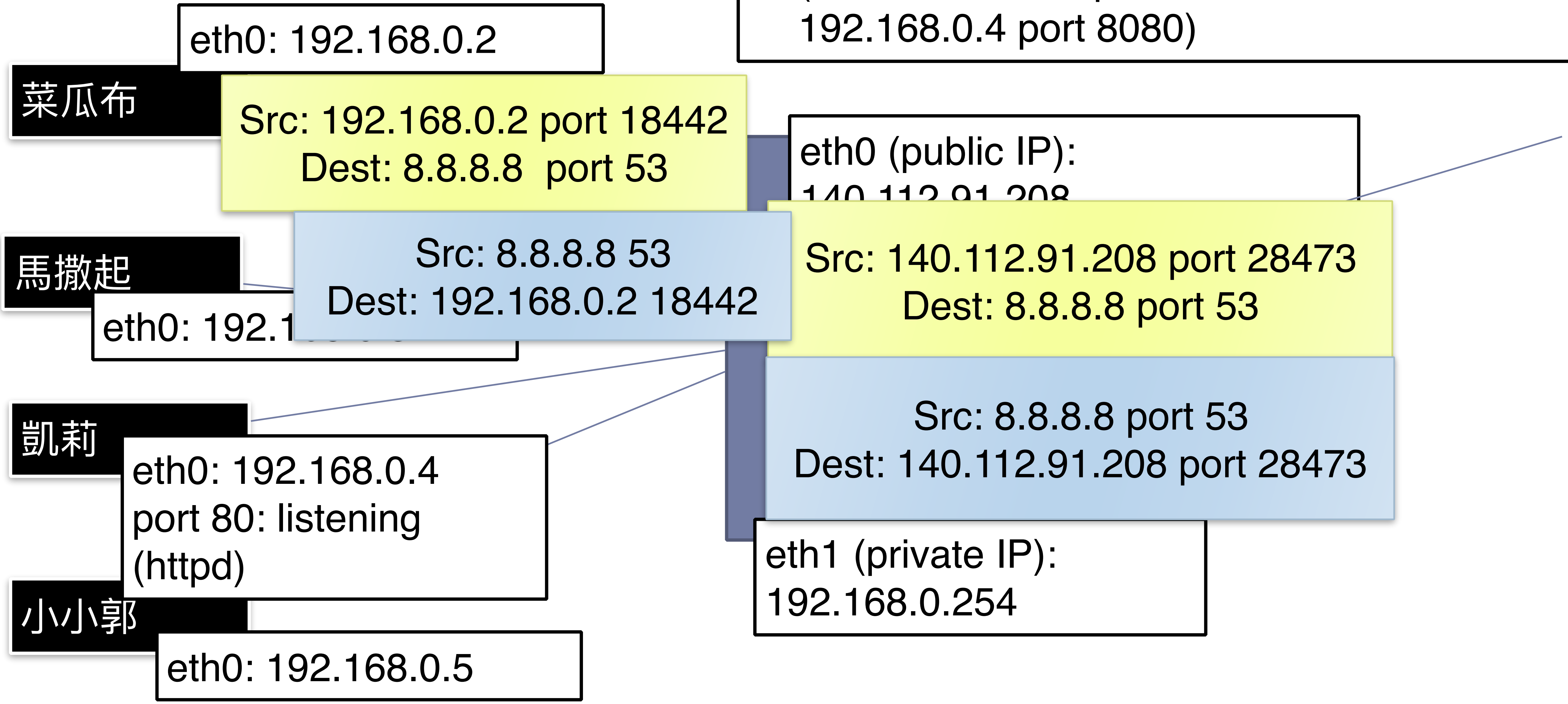
Src: 8.8.8.8
Dest: 140.112.91.208

內部用門牌:192.168.0.254

# NAT (Network Address Translation) Revisited

對照表:

- 菜瓜布有連到8.8.8.8 (192.168.0.2 port 18442 —> 140.112.91.208 port 28473 —> 8.8.8.8 port 53)
- 要找助教請轉到192.168.0.4 (140.112.91.208 port 80 redirects to 192.168.0.4 port 8080)



eth0: 192.168.0.2

菜瓜布

Src: 192.168.0.2 port 18442
Dest: 8.8.8.8 port 53

馬撒起

eth0: 192.1

Src: 8.8.8.8 53
Dest: 192.168.0.2 18442

凱莉

eth0: 192.168.0.4
port 80: listening
(httpd)

小小郭

eth0: 192.168.0.5

eth0 (public IP):
140.112.91.208

Src: 140.112.91.208 port 28473
Dest: 8.8.8.8 port 53

Src: 8.8.8.8 port 53
Dest: 140.112.91.208 port 28473

eth1 (private IP):
192.168.0.254

# Check all the connections

- netstat:
  print network connections, routing tables, and interface statistics

- Example:
  netstat -na --inet
  (list connections on all interfaces, using numerical addresses, listing only IP connections)
  netstat -s (print the statistics)
  netstat -i (list interface statistics)

# DHCP

# DHCP
# (Dynamic Host Configuration Protocol)

- 每個地方有自己的subnet及IP設定
- 到一個新的地方，一開始怎麼取得此一subnet的IP呢?
- 通常同一個subnet中會設置一台DHCP server
- 此server將負責”接待”新來的機器，分發未使用的IP給它們
- 想像全系如果都需要手動設定IP, 會發生什麼事情?
  - 網管需要分配IP給所有電腦 (全系有多少電腦???)
  - IP衝突 (同樣的IP被不同的電腦使用)

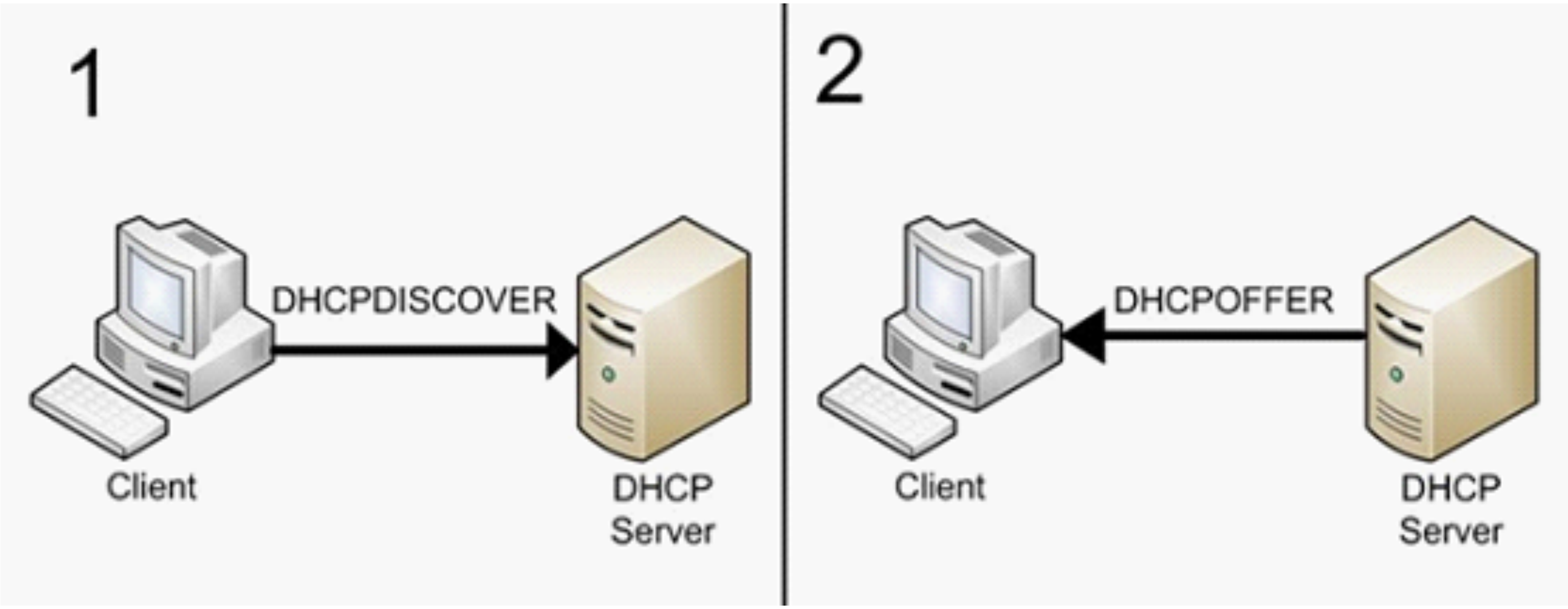
# DHCP 4部曲

DHCP Offer:我這邊有一組IP看看你要不要用.

DHCP Discover: 請問有人可以發IP給我嗎?

Src: 0.0.0.0, 68
Dest: 255.255.255.255, 67
DHCPDISCOVER
Yiaddr: 0.0.0.0
Transaction ID: 654
Request:
Subnet Mask, Router, Domain Name Server

Src: 192.168.55.254, 67
Dest: 255.255.255.255, 68
DHCPOFFER
Yiaddr: 192.168.48.15
DHCP server ID: 192.168.55.254
Transaction ID: 654
Lifetime: 4 hrs
Netmask: 255.255.248.0
Router: 192.168.55.254
DNS: 140.112.30.21, 140.112.254.4



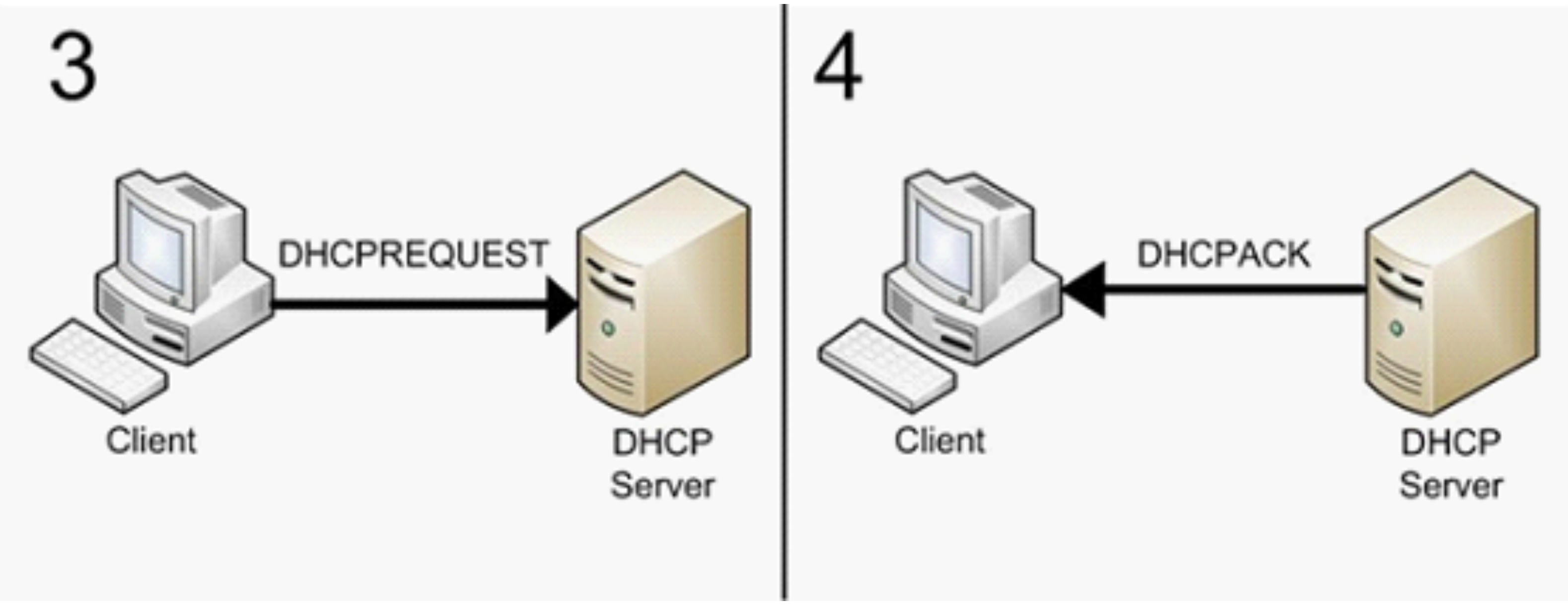
# DHCP 4部曲

DHCP Ack: 沒問題. 請用.

Src: 192.168.55.254, 67
Dest: 255.255.255.255, 68
DHCPACK
Yiaddr: 192.168.48.15
DHCP server ID: 192.168.55.254
Transaction ID: 655
Lifetime: 4 hrs
Netmask: 255.255.248.0
Router: 192.168.55.254
DNS: 140.112.30.21, 140.112.254.4

DHCP Request:那我要把這組IP拿走囉!

Src: 0.0.0.0, 68
Dest: 255.255.255.255, 67
DHCPREQUEST
Yiaddr: 192.168.48.15
Transaction ID: 655
DHCP server ID: 192.168.55.254
Lifetime: 4 hrs



# Some additional facts

- Originally designed as BOOTP - for diskless workstations.
  - Today, You can still use PXEBOOT on various network interface card for diskless operation.
- A server keeps track of “lease” (of IPs), which would expire after a predefined time period
  - A client usually “renews” the lease when the time is half over
  - The lease information must survive reboot for network stability
  - A typical lease entry:

```
lease 192.168.20.4 {
   starts 6 2009/06/27 00:40:00;
   ends 6 2009/06/27 12:40:00;
   hardware ethernet 00:00:00:00:00:00;
   uid 00:00:00:00:00:00;
   client-hostname "examle-workstation1";
}
```

# DHCP 的細節

- 一個subnet上可能有多個DHCP server. 因此發出DHCPREQUEST之後，可能收到多個DHCPOFFER。
- Client可以要求使用之前使用過的IP，但DHCP server可以拒絕(可能根本已經不在同一個網段，或是已經被別的client使用中)
- Authoritative & non-authoritative: 有主管權的DHCP server可以發出”拒絕”client使用某IP的要求，而沒有主管權的DHCP server則會忽略該要求(沒有回應)
- 想想看: DHCP server的安全漏洞. 如果有人接在系上網路上且開啟DHCP server，會發生什麼事情?

# A typical DHCP server configuration file

- The following dhcpd.conf is used by ISC DHCPD

```
default-lease-time 600;
max-lease-time 7200;
option subnet-mask 255.255.255.0;
option broadcast-address 192.168.1.255;
option routers 192.168.1.254;
option domain-name-servers 192.168.1.1, 192.168.1.2;
option domain-search "example.com";
subnet 192.168.1.0 netmask 255.255.255.0 {
   range 192.168.1.10 192.168.1.100;
}

subnet 140.112.31.0 netmask 255.255.255.0 {
}

host apex {
   option host-name "apex.example.com";
   hardware ethernet 00:A0:78:8E:9E:AA;
   fixed-address 192.168.1.4;
}
```