インターネットアーキテクチャ
Internet Architecture

ソフトウェア・クラウド開発プロジェクト実践III

浅井大史

2016年5月6日

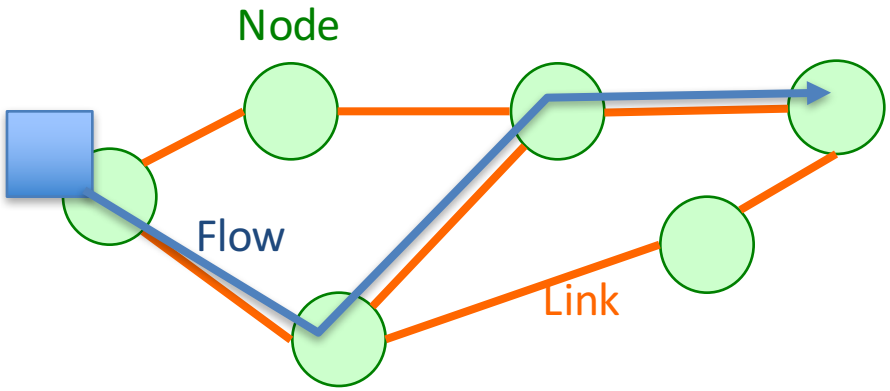
The Internet

- Inter-network
  - Network of networks

Components of “Network”

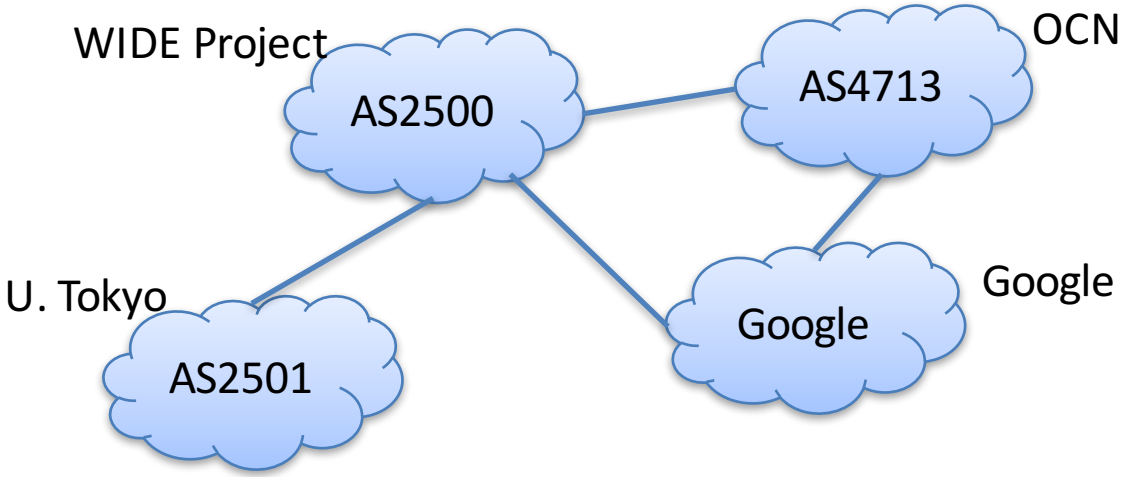
= Graph + Flow

- Vertex (Node)
- Edge (Link)
- Flow

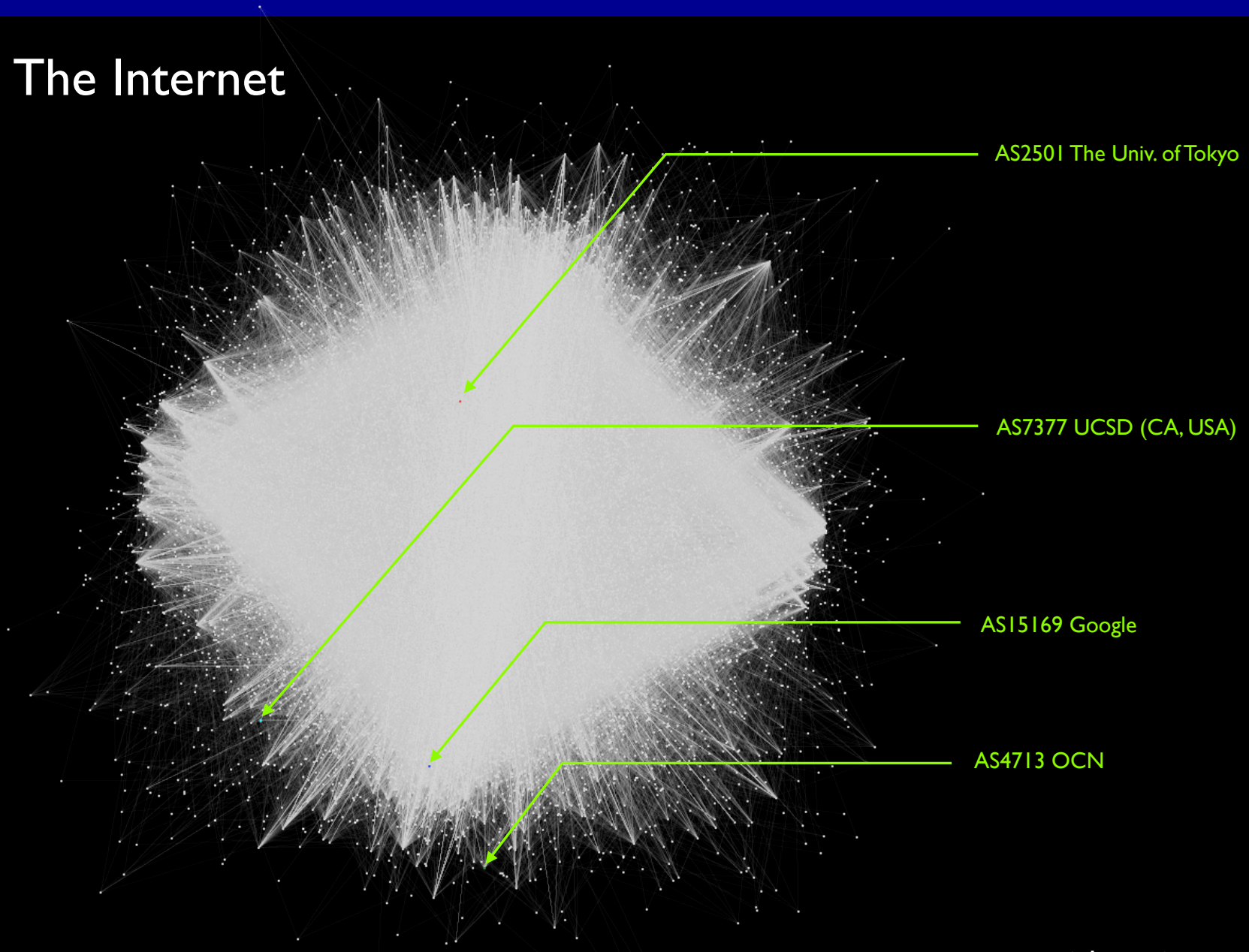


The Internet

- Autonomous System (AS)
  - Administrative domain
    - Internet service providers (ISP)
    - Company
    - University
  - Identified with AS number (4-octet)

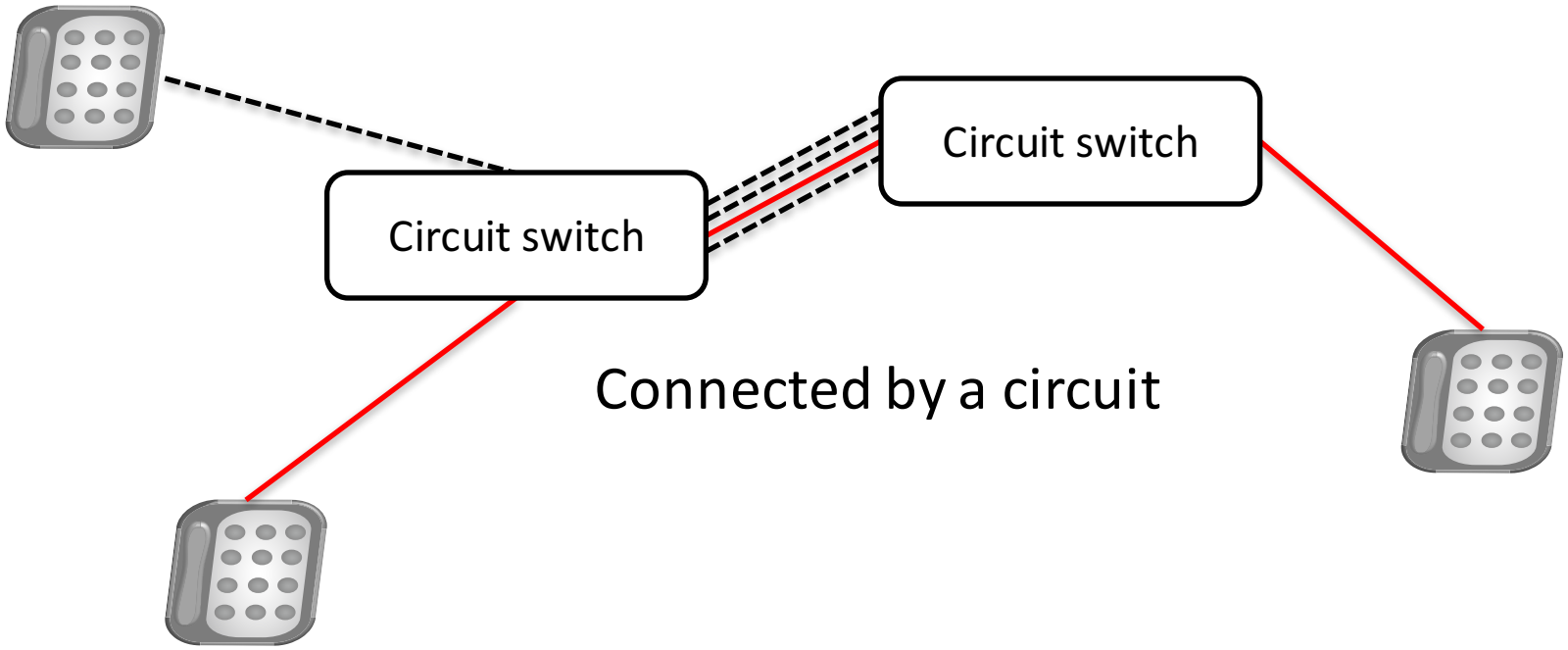


The Internet

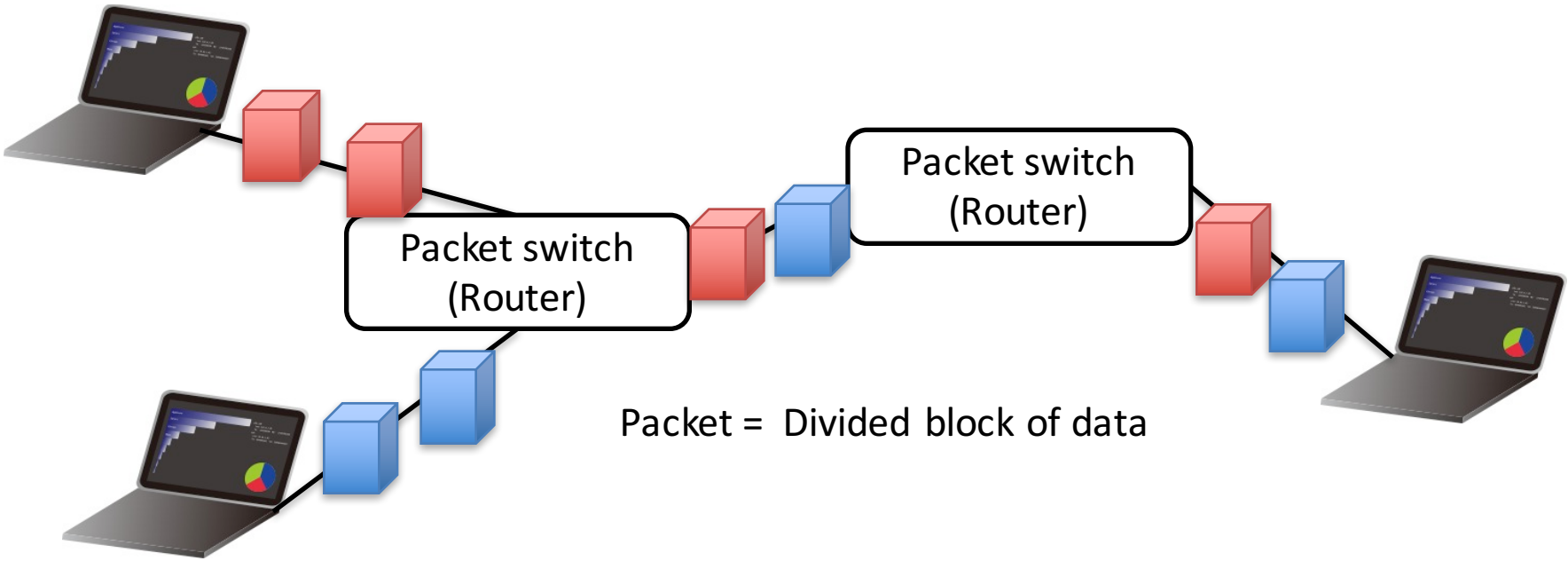


Drawn with spring model

Circuit switching（回線交換方式）



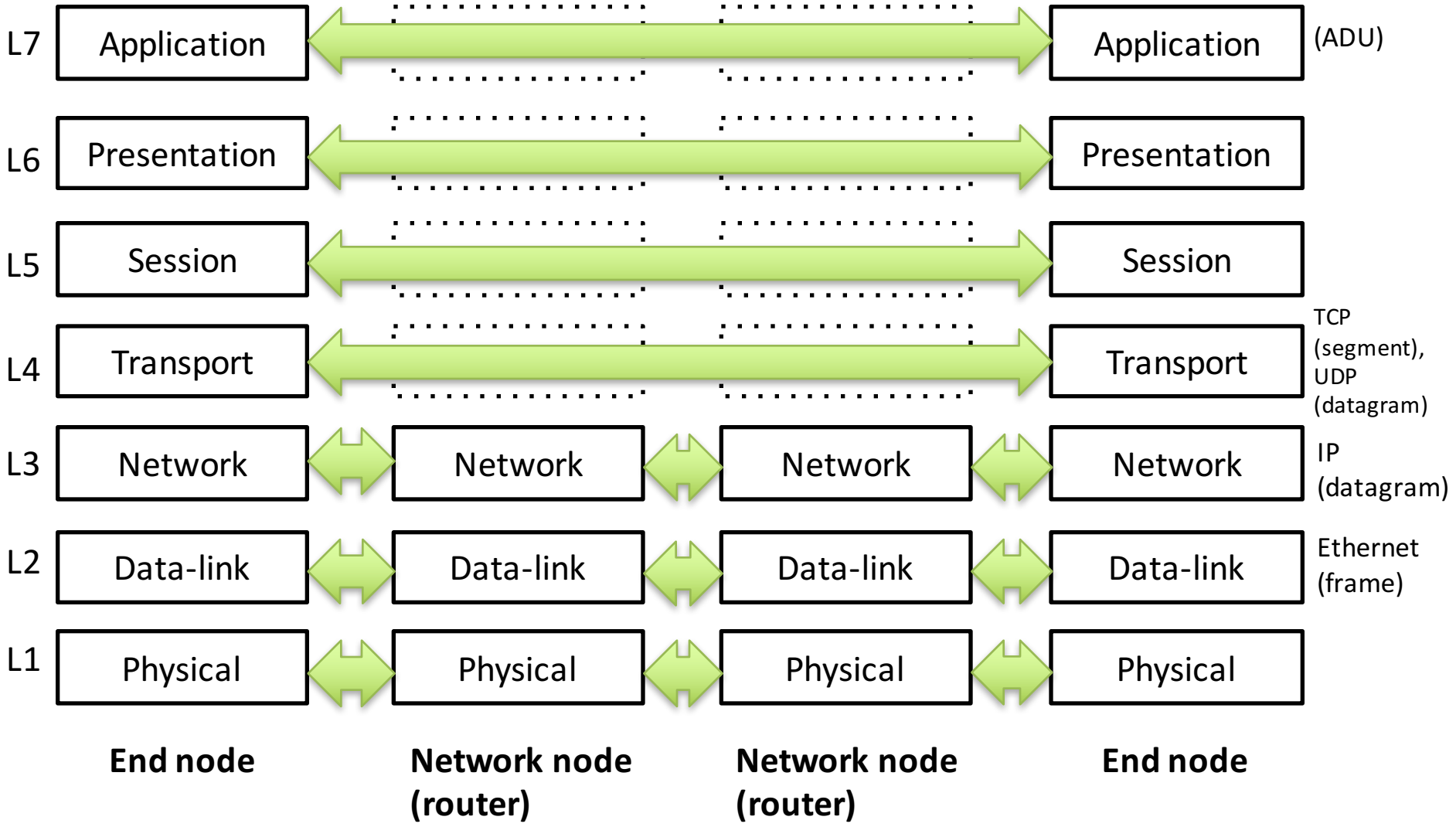
Packet switching（パケット交換方式）



Packet = Divided block of data

More efficient compared to circuit switching

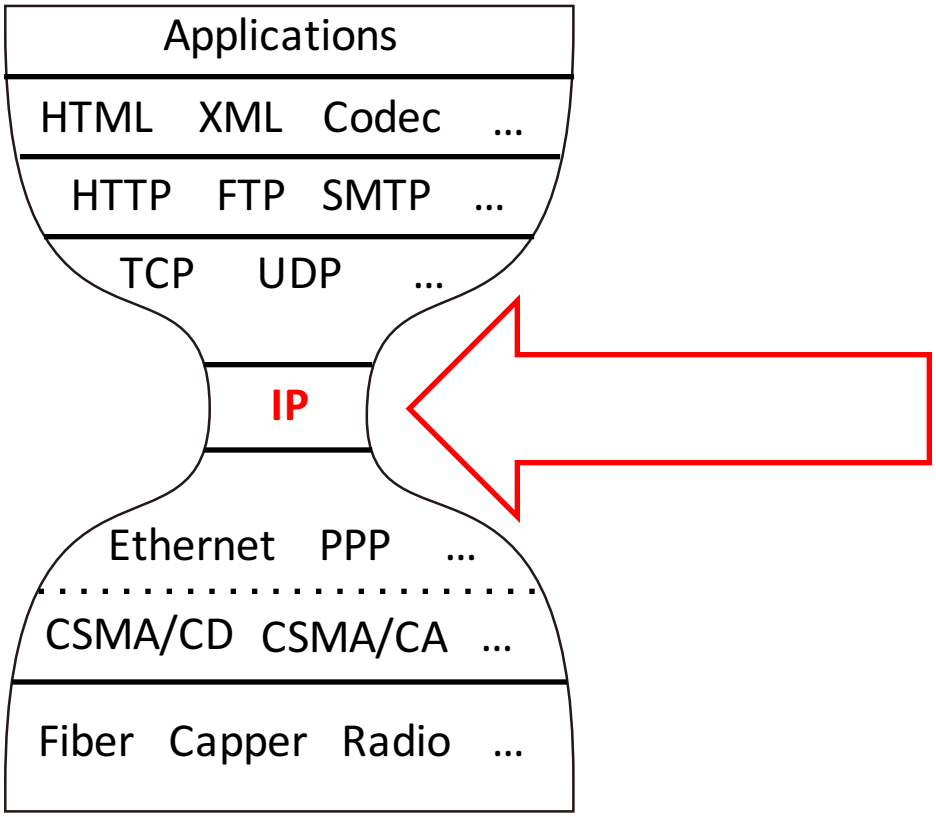
OSI reference model



Protocol

- Rules governing semantics
  - Examples
    - Internet Protocol (L3)
    - Transmission Control Protocol (L4)
    - File Transfer Protocol (L5-L7)
    - HyperText Transfer Protocol (L5-L7)
    - Routing protocols (to exchange routing tables)
      - Routing Information Protocol (using UDP)
      - Open Shortest Path First Protocol (using IP)
      - Border Gateway Protocol (using TCP)

Hourglass Model



Hourglass model

“The” Internet

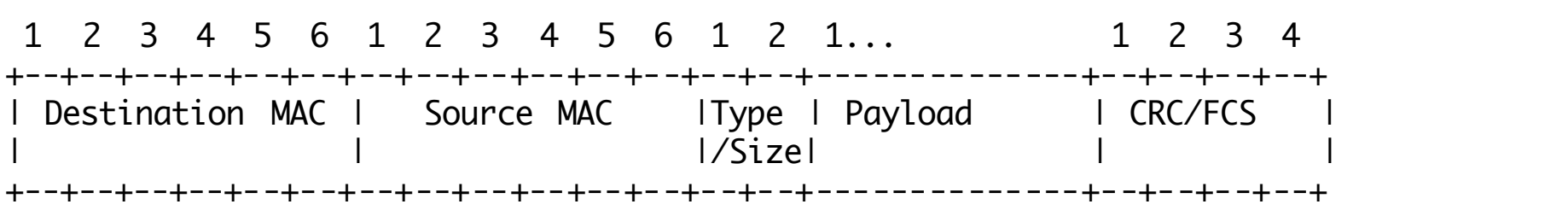
L1: Physical Layer

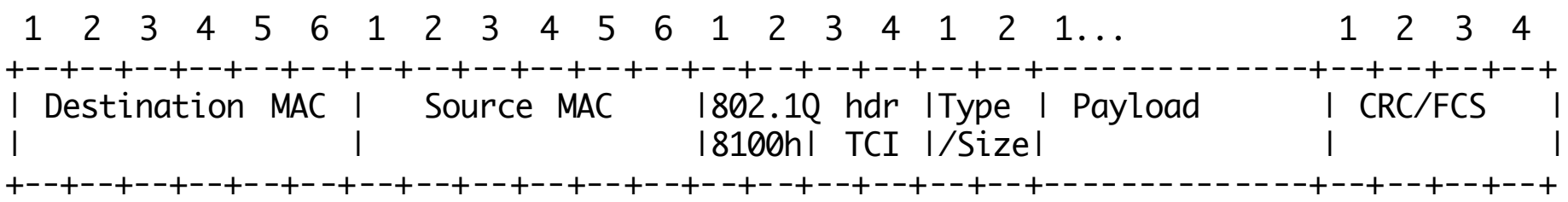
- Media
  - Optical fiber
    - Single mode fiber (SMF)
    - Multi mode fiber (MMF)
    - Dispersion shifted fiber (DSF)
  - Metal
    - UTP (Unshielded Twisted Pair) cable
    - RS-232C
- Connectors
  - XFP (SC), SFP (LC), SFP+ (LC), QSFP+ (LC), QSFP+, CFP, CFP2, QSFP28
  - RJ11, RJ45, ARJ45, TERA
- Functionalities
  - Connection
  - Multiplexing, Modulation
  - Encoding

L2: Data-link layer

- Media Access Control (MAC)
  - Multiple access w/ collision detection
    - CSMA/CD : Ethernet (IEEE 802.3) etc.
  - Multiple access w/ collision avoidance
    - CSMA/CA : WiFi (IEEE 802.11) etc.
  - Token (shows who has the privilege of sending data)
    - Token bus
    - Token ring
- Functionalities
  - Flow control
  - Error correction

VLAN





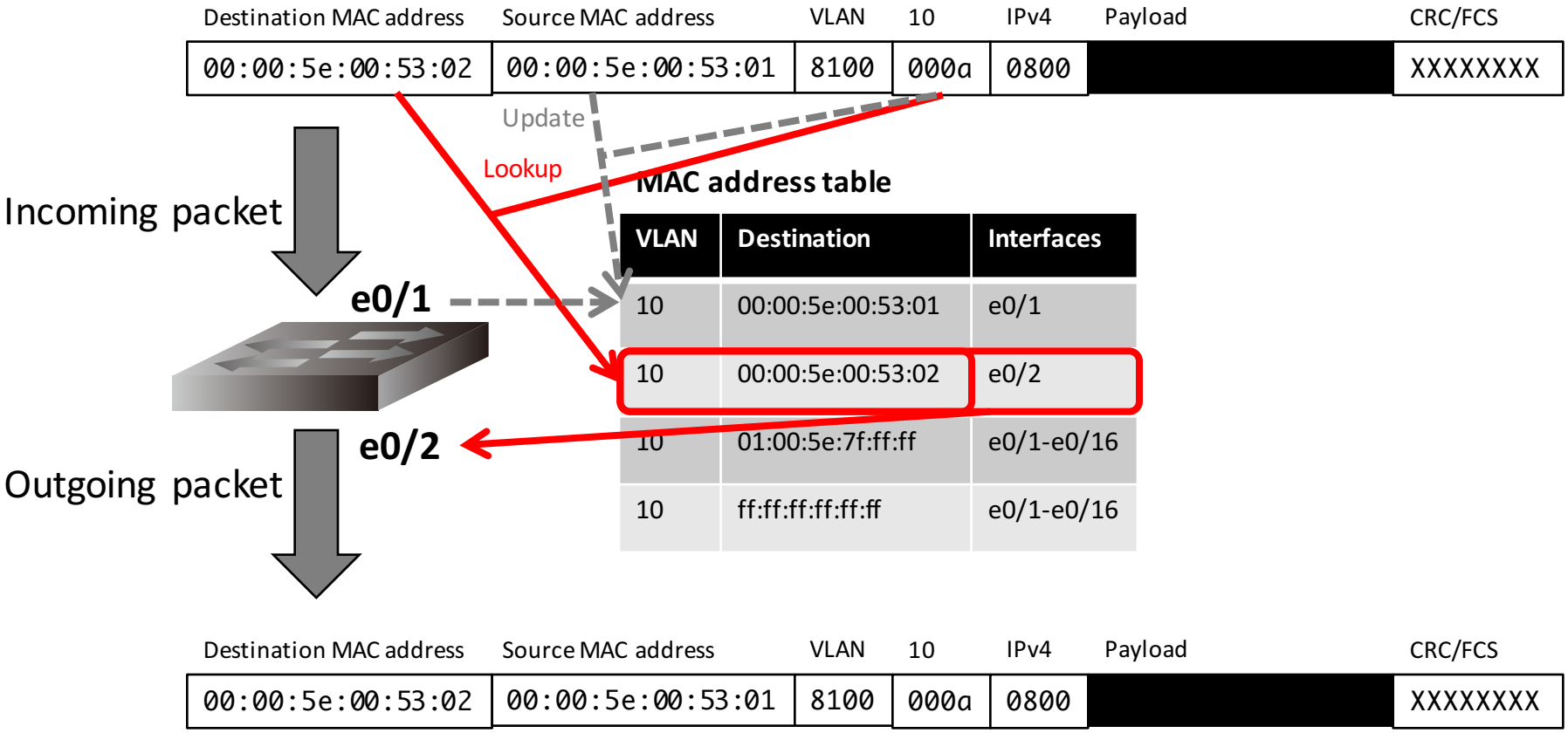
802.1Q tag control information (TCI)

PCP 3bit (*Priority Code Point*)

DEI 1bit (*Drop Eligible Indicator*)

VID 12bit (*VLAN Identifier)*

Ethernet Switching



Ethernet misc.

- Frame size
  - Minimum: 64 byte (including a 4-byte FCS)
    - Link utilization
      - 1 packet: 8 (preamble) + 60 (frame) + 4 (FCS) + 12 (inter-frame gap) = 84 [bytes]
      - Utilization: 60/84 ≒ 71.4%
      - Line rate: (utilization / (60 * 8)) / (speed) [packet per second]
        - » 1GbE: 1.488Mpps
        - » 10GbE: 14.88Mpps
        - » 100GbE: 148.8Mpps
  - Maximum
    - 1518 byte (including FCS)
    - 1522 byte (including FCS) for 802.1Q
    - Jumbo frame (>1518-byte frame) may be supported

L3: Network layer

- Main functionalities
  - Addressing
    - 32bit for IPv4
    - 128bit for IPv6
  - Packet delivery
    - Routing
      - Static routing
      - Dynamic routing
        - » Distance vector (DV): RIPv2 (IPv4) / RIPv3 (IPv6)
        - » Link state (LS): OSPFv2 (IPv4) / OSPFv3 (IPv6)
        - » Path vector (PV): BGP4 (IPv4) / BGP4+ (IPv6)
    - Fragmentation
      - Note: Fragmentation is not allowed at middleboxes in IPv6

Internet Protocol (IP)

- Two versions
  - IPv4 (version 4)
  - IPv6 (version 6)
    - No compatibility between IPv4 and IPv6

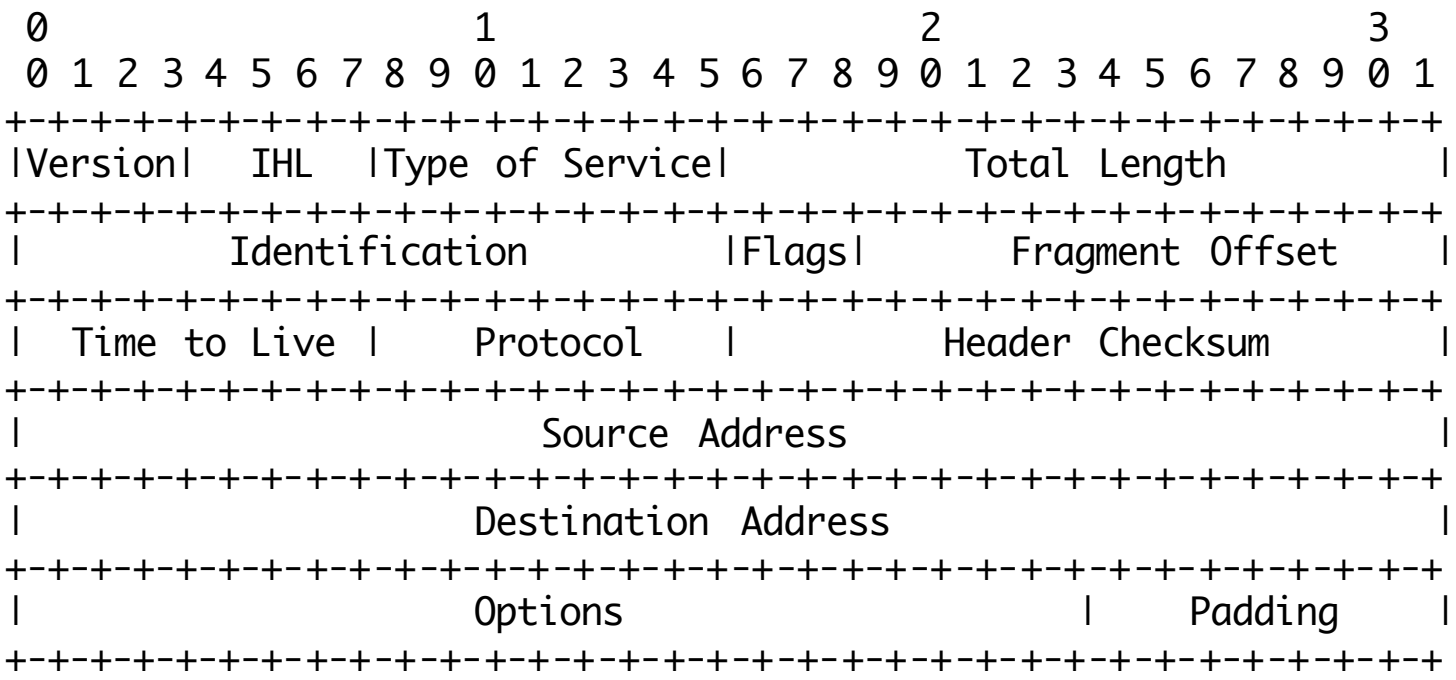
Table. Key differences of IPv4 and IPv6

	IPv4	IPv6
Address space	32bit	128bit
Minimum MTU	576 byte	1280 byte
Fragmentation at middleboxes (intermediate routers)	Allowed	Not allowed
Private address/NAPT	Yes	No
Neighbor discovery	ARP	ICMPv6

IPv4 address

- 32bit address space
  - Textual representation d.d.d.d
    - 4 one-to-three-digit (in decimal) segments separated by dot
    - Examples
      - 192.0.2.1
      - 133.11.169.1
      - 203.178.135.1

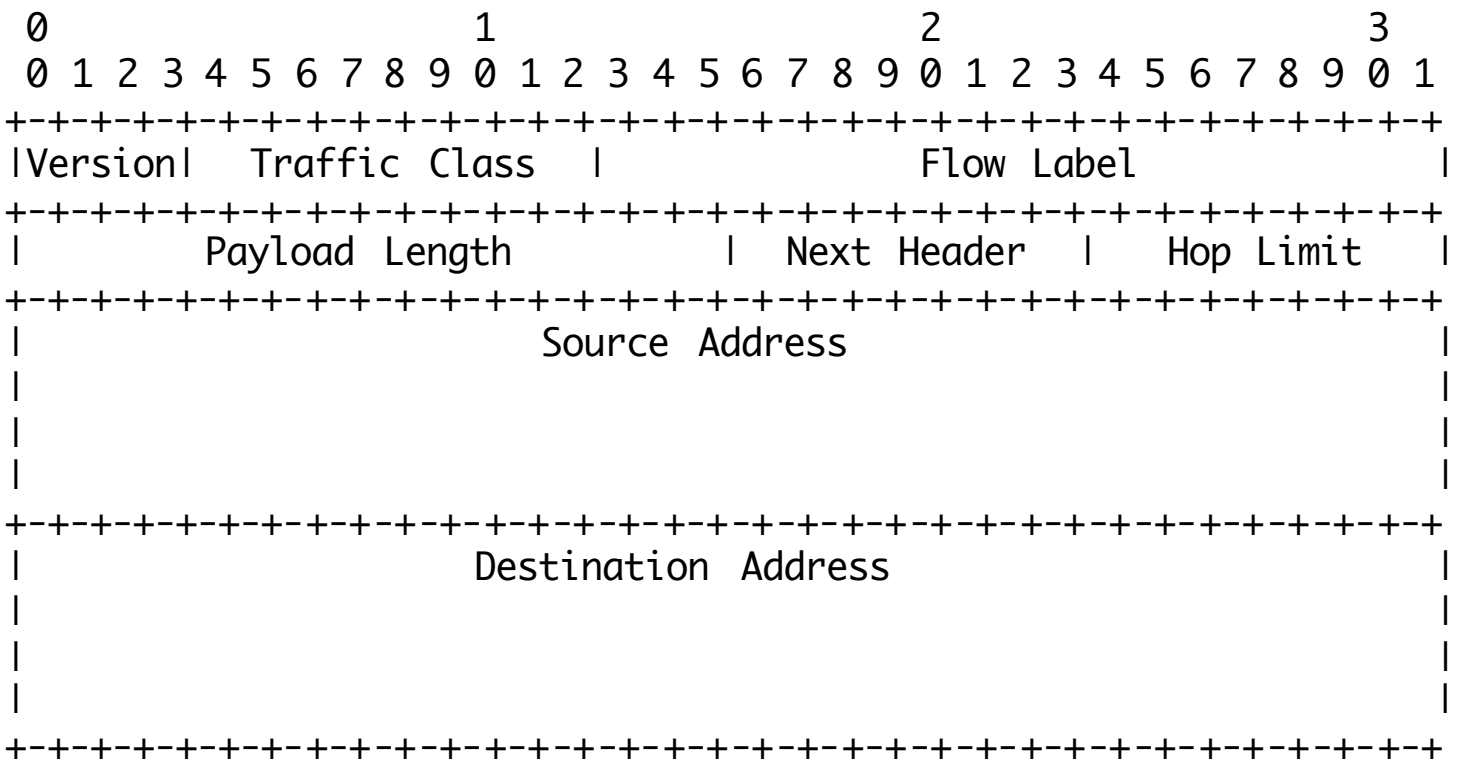
IPv4 datagram header format



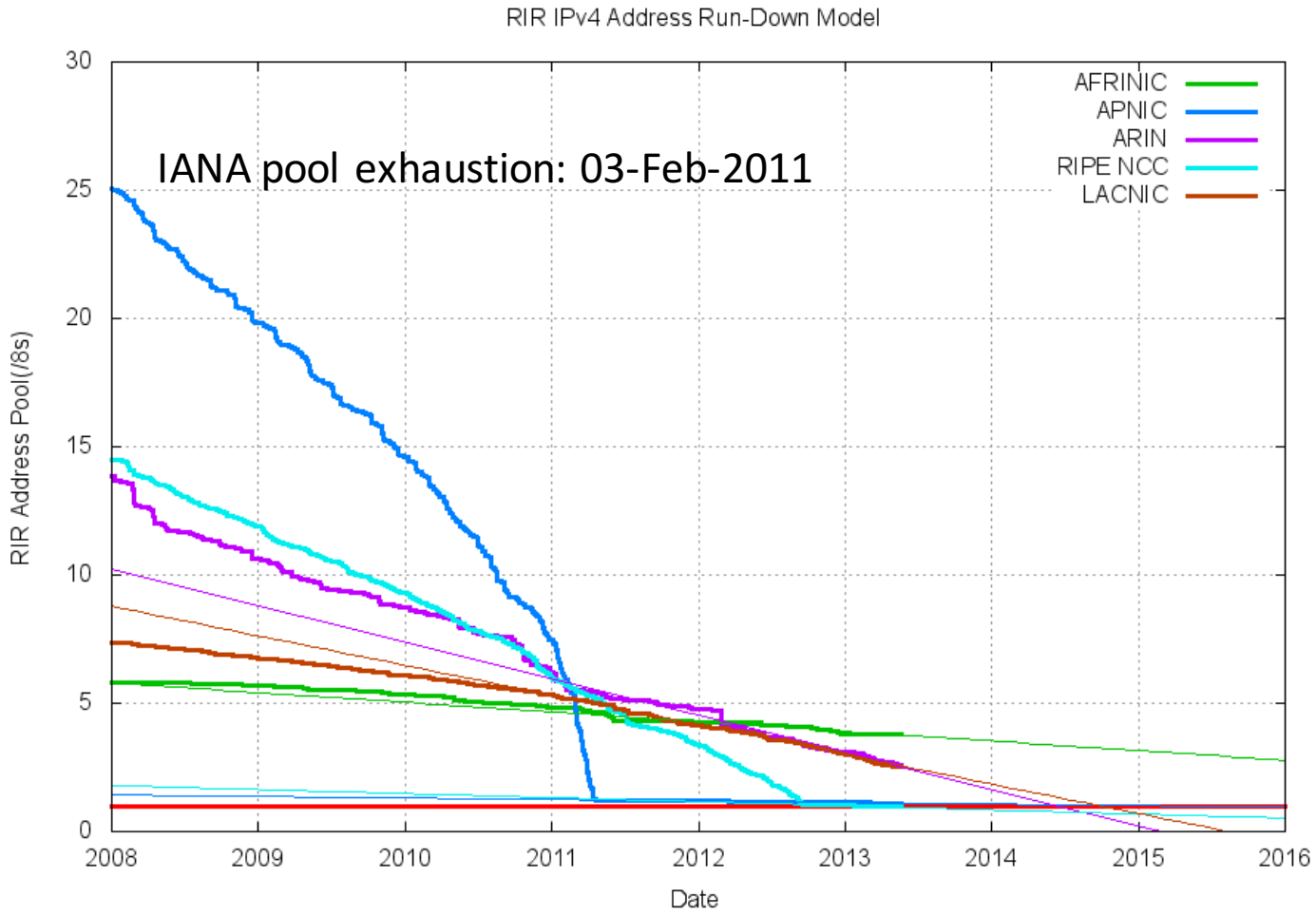
IPv6 address

- 128bit address space
  - Textual representation x:x:x:x:x:x:x:x
    - 8 four-digit (in hexadecimal) segments separated by colon
    - Leading zeros can be omitted
      - » Example
        - 2001:0db8:0003:0004:0005:cafe:dead:beef
          → 2001:db8:3:4:5:cafe:dead:beef
    - One set of continuous zero segments can be compressed by "::"
      - » Example
        - 2001:0db8:0000:0000:beef:0000:0003:0bed
          → 2001:db8::beef:0:3:bed
  - Recommendation
    - See RFC 5952 for details

IPv6 datagram header format



Note: IPv4 address exhaustion



Source: Geoff Huston's report http://www.potaroo.net/tools/ipv4/

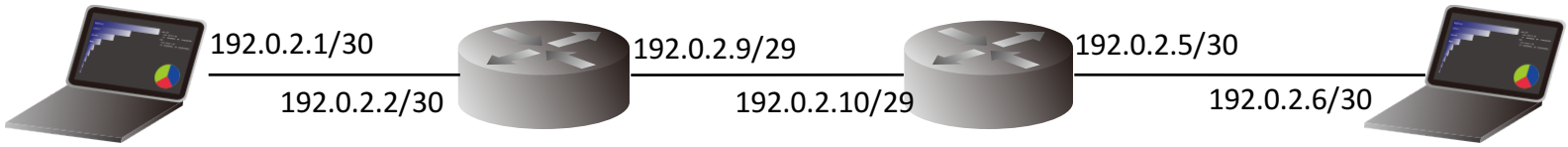
Subnet

- Subdivision of IP network
  - Examples
    - 192.0.2.0/24
      - 11000000 00000000 00000010 xxxxxxxx
    - 2001:db8::/32
      - 0010 0000 0000 0001 0000 1101 1011 1000 xxxx xxxx…
  - Network address (IPv4)
    - Host #=0b0..0
      - 192.0.2.128/25 => 192.0.2.128 (11000000 00000000 00000010 10000000)
  - Broadcast address (IPv4)
    - Host #=0b1..1
      - 192.0.2.0/25 => 192.0.2.127 (11000000 00000000 00000010 01111111)
  - Netmask (IPv4)
    - Network prefix=0b1..1, Host #=0b0..0
      - 192.0.2.0/25 => 255.255.255.128 (11111111 11111111 11111111 10000000)
  - Prefix length (IPv6)
    - 2001:db8::/32 => 32bit prefix
    - 2001:db8::/48 => 48bit prefix

Routing

Destination	Next hop
192.0.2.0/30	Link
0.0.0.0/0	192.0.2.2

Destination	Next hop
192.0.2.4/30	Link
192.0.2.0/30	192.0.2.9
192.0.2.8/29	Link
0.0.0.0/0	192.0.2.11



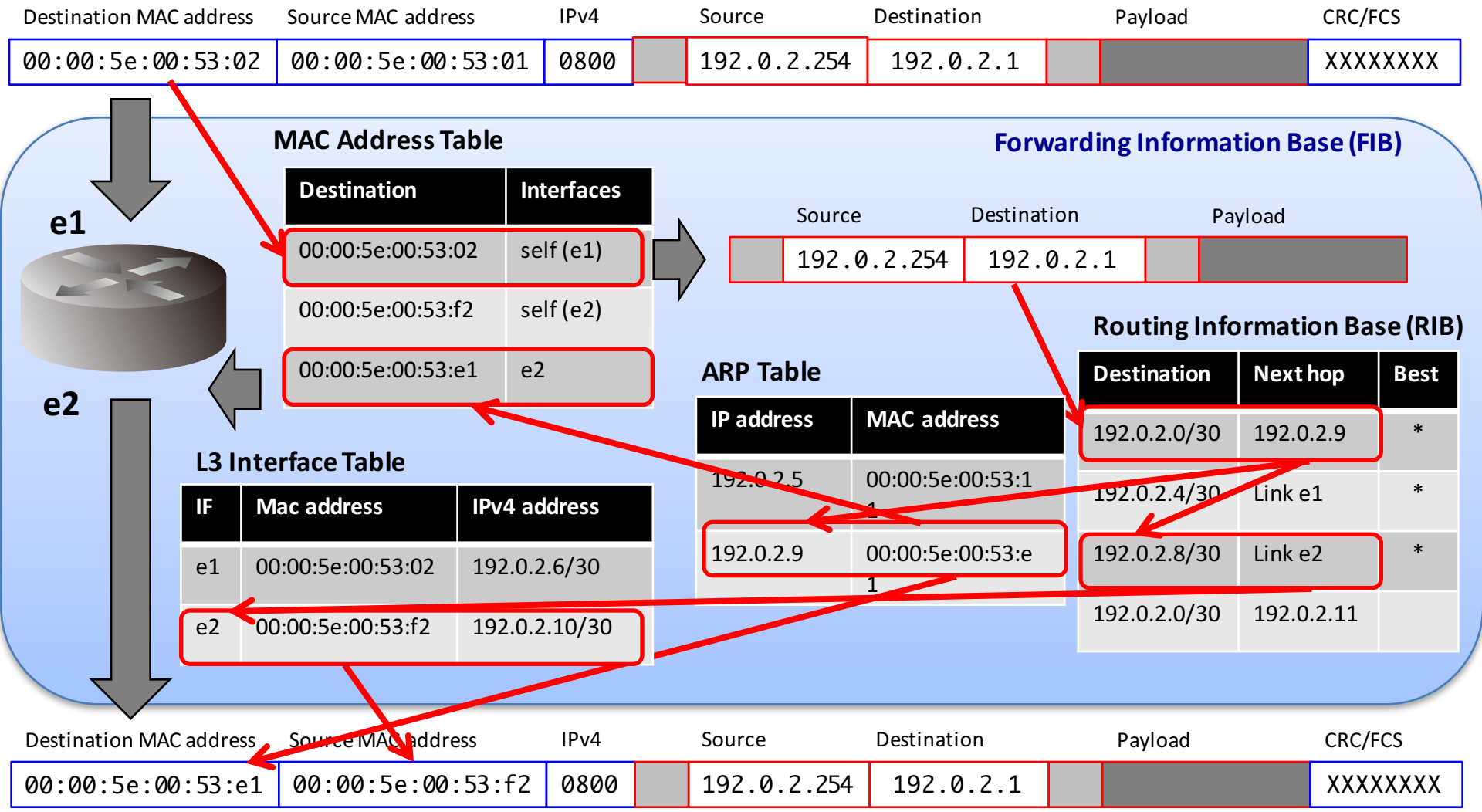
Longest match =>

Destination	Next hop
192.0.2.0/30	Link
192.0.2.4/30	192.0.2.10
192.0.2.8/29	Link
192.0.2.0/25	192.0.2.11
0.0.0.0/0	192.0.2.11

Destination	Next hop
192.0.2.4/30	Link
0.0.0.0/0	192.0.2.5

Dynamic routing protocol
→ Exchange routing tables

Routing


Destination MAC address
Source MAC address
IPv4
Source
Destination
Payload
CRC/FCS
00:00:5e:00:53:02
00:00:5e:00:53:01
0800
192.0.2.254
192.0.2.1
XXXXXXXX
MAC Address Table
Destination
Interfaces
00:00:5e:00:53:02
self (e1)
00:00:5e:00:53:f2
self (e2)
00:00:5e:00:53:e1
e2
Forwarding Information Base (FIB)
Source
Destination
Payload
192.0.2.254
192.0.2.1
e1
e2
Routing Information Base (RIB)
Destination
Next hop
Best
192.0.2.0/30
192.0.2.9
*
192.0.2.4/30
Link e1
*
192.0.2.8/30
Link e2
*
192.0.2.0/30
192.0.2.11
ARP Table
IP address
MAC address
192.0.2.5
00:00:5e:00:53:11
192.0.2.9
00:00:5e:00:53:e1
L3 Interface Table
IF
Mac address
IPv4 address
e1
00:00:5e:00:53:02
192.0.2.6/30
e2
00:00:5e:00:53:f2
192.0.2.10/30
Destination MAC address
Source MAC address
IPv4
Source
Destination
Payload
CRC/FCS
00:00:5e:00:53:e1
00:00:5e:00:53:f2
0800
192.0.2.254
192.0.2.1
XXXXXXXX

(BREAK) Routing Table Lookup

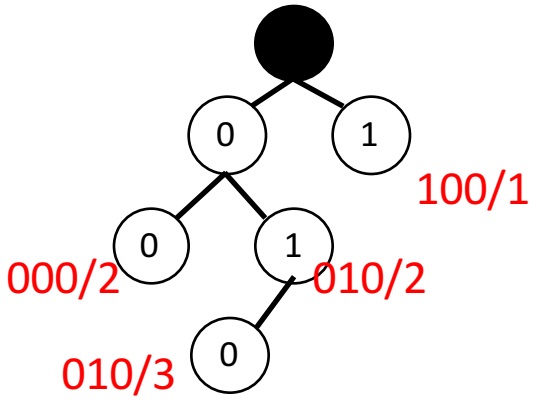
- @ASBR (AS Border Router)
  - # of entries: 512k+
- Routing table lookup
  - Hardware
    - tCAM
      (Ternary Content Addressable Memory)
  - Software
    - Linear search
    - Crit-bit-trie (Radix tree)

512k+

Destination	Next hop
192.0.2.0/30	Link
192.0.2.4/30	192.0.2.10
192.0.2.8/29	Link
192.0.2.0/25	192.0.2.11
0.0.0.0/0	192.0.2.11

Address	Data
1000 10**	0100 1011
**** ****	0100 0101

*: Don't care bit



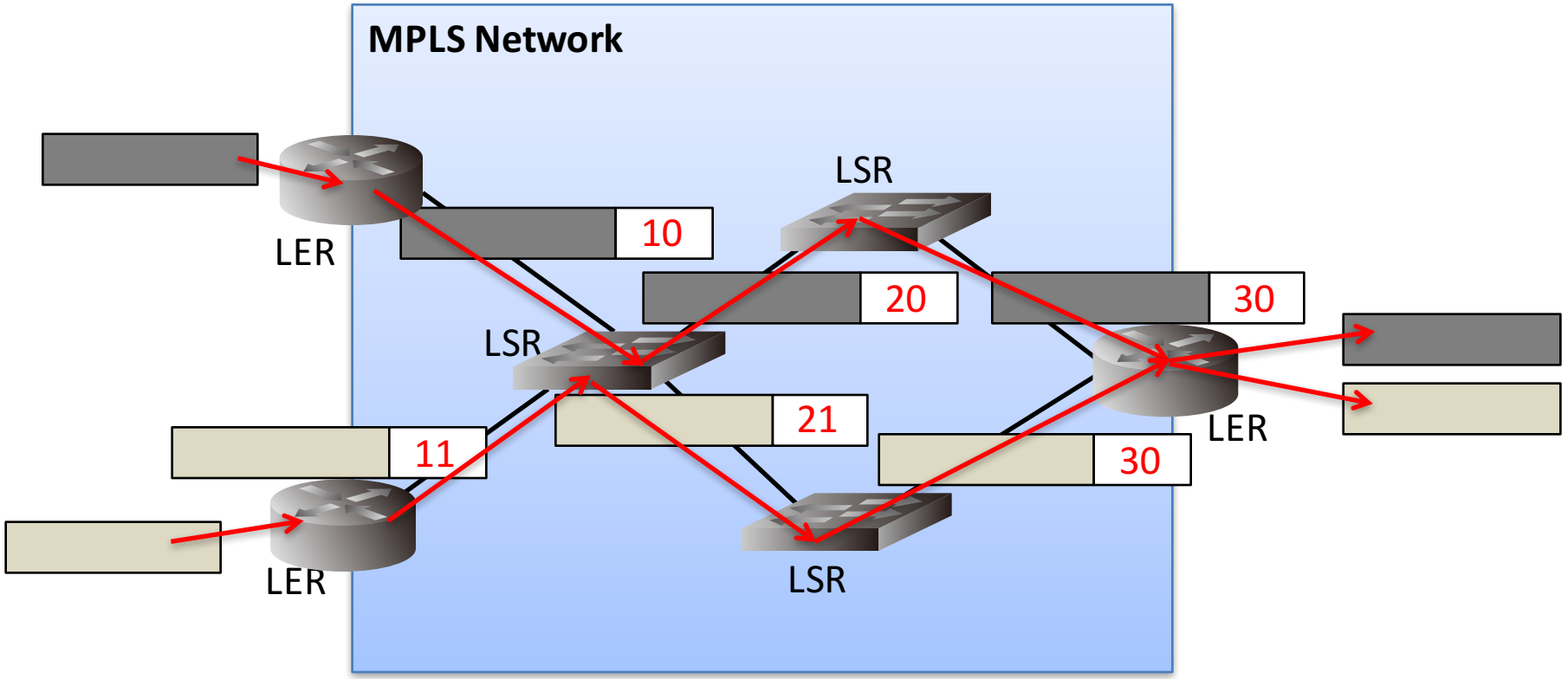
(BREAK) Dynamic Routing

- Dynamic routing protocols
  - Interior Gateway Protocol (IGP)
    - RIP
      - Distant vector algorithm
    - OSPF / IS-IS
      - Link-state protocol
      - Shortest path algorithm (Dijkstra)
  - Exterior Gateway Protocol (EGP)
    - BGP
      - Path vector algorithm
      - Policy-based routing

(BREAK) Label Switching

- Multiprotocol Label Switching (MPLS)
  - Structure
    - MPLS header followed by IP header
      - Label, 20bit
      - EXP (Experimental), 3bit: Used for QoS
      - S (Bottom-of-Stack), 1bit: Last label or not
      - TTL, 8bit
  - Feature
    - Faster than IP routing (maybe traditional feature)
      - MPLS header: 4-byte < IP header: 20-byte
    - Virtual Private Network (VPN)MPLS header
    - Traffic Engineering (TE)
      - Fast reroute
      - Label switching not rely on destination

(BREAK) Label Switching



LER: Label Edge Router
LSR: Label Switch Router

L4: Transport layer

- TCP (Transmission Control Protocol)
  - Connection-full
  - Ordered data transfer
  - Retransmission of lost packets / Discarding duplicate packets
  - Flow control
  - Congestion control
    - Algorithms: Reno, New Reno, Vegas, CUBIC, Compound TCP etc.
- UDP (User Datagram Protocol)
  - IP + port (src, dst)
  - Connection-less
- QUIC (Quick UDP Internet Connections)
  - Connection-full protocol with multiplexed UDP

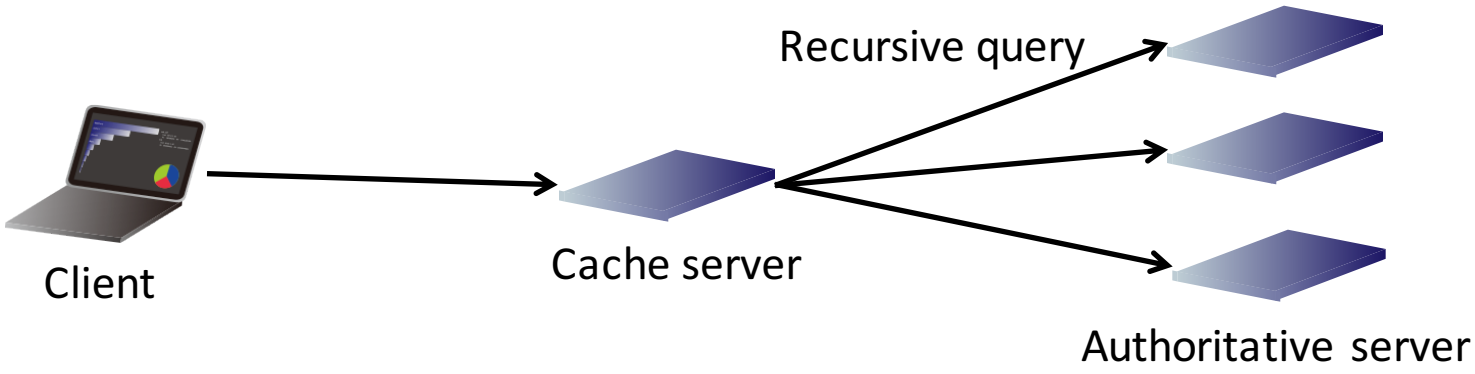
L7: Application Layer

- Web
  - HTTP (Hypertext Transfer Protocol)
    - HTTP/1.1
    - HTTP/2
- Mail
  - SMTP (Simple Mail Transfer Protocol)
  - POP (Post Office Protocol)
  - IMAP (Internet Message Access Protocol)
- etc...

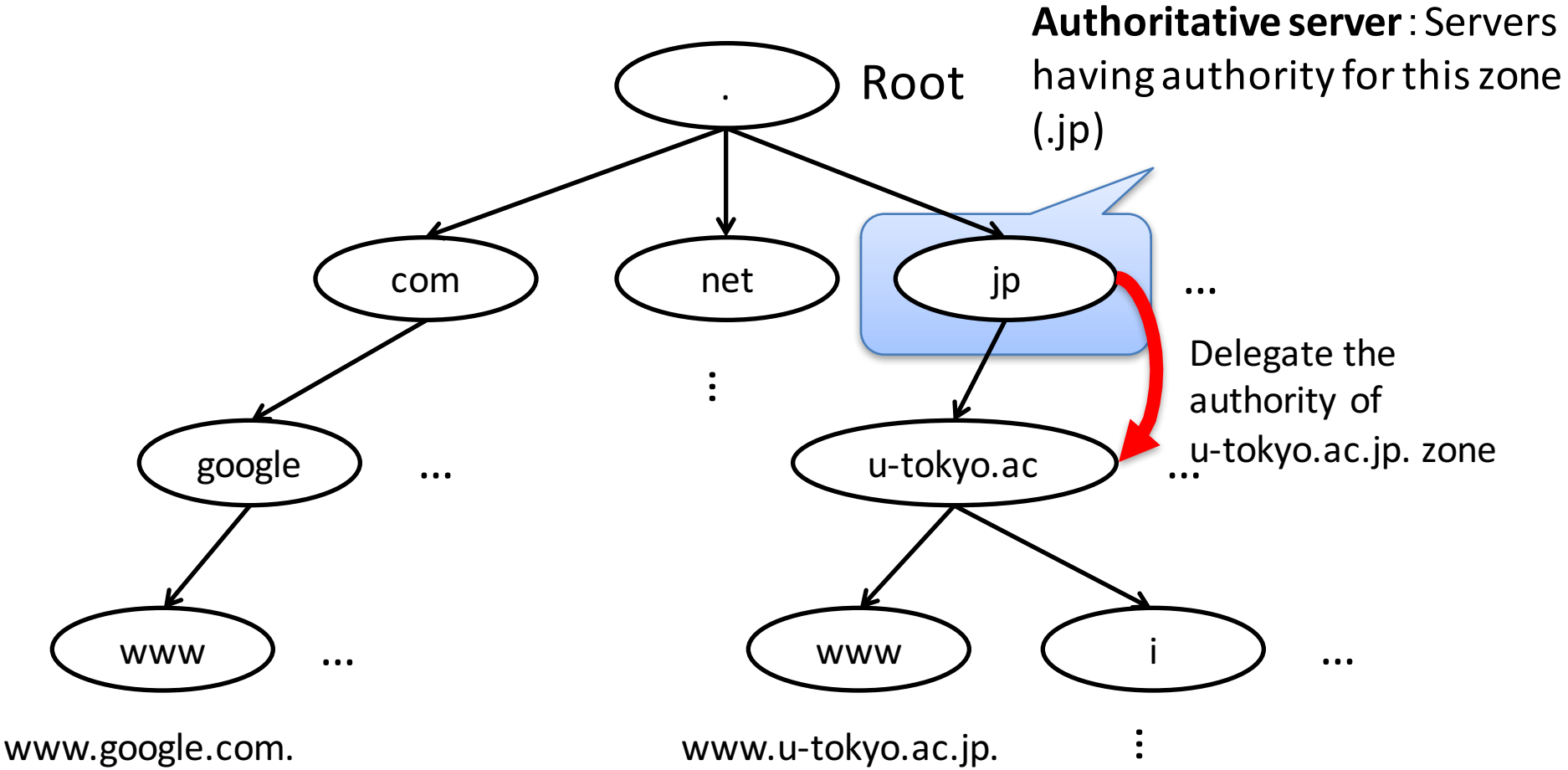
OTHER KEY PROTOCOLS & TECHNOLOGIES

DNS: Domain Name System

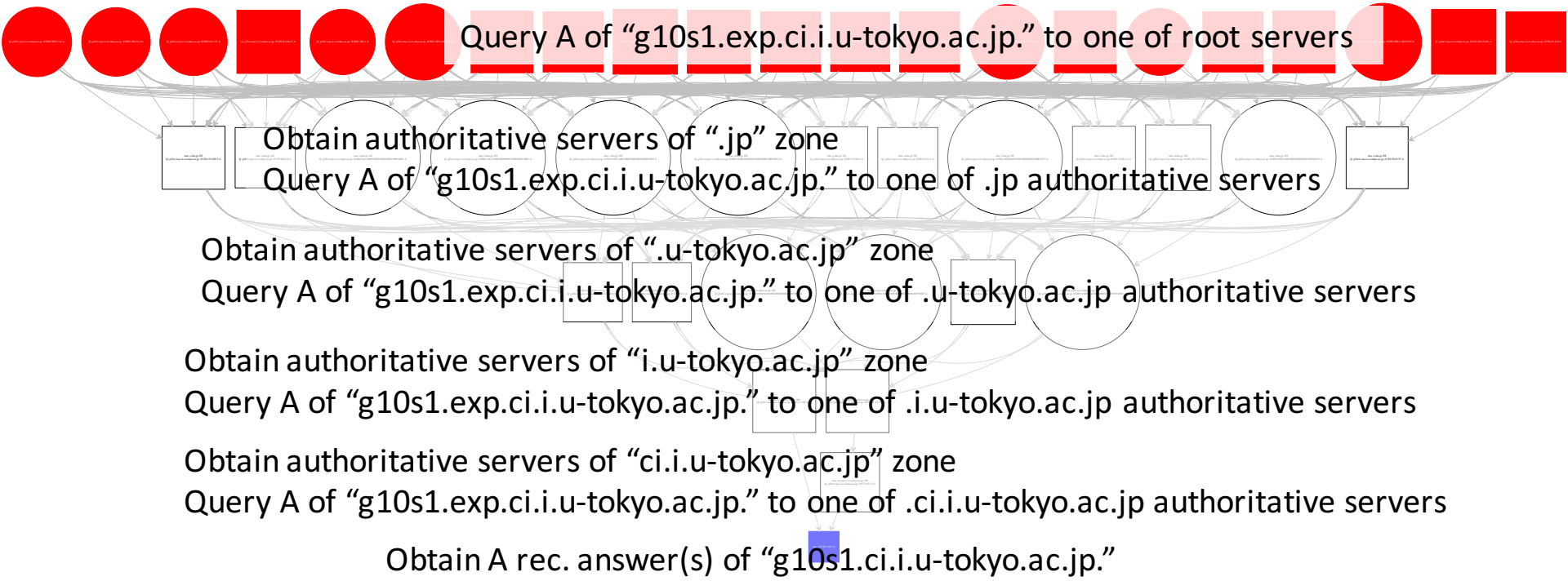
- Tree structure database
  - To resolve resources (e.g., IPv4 address, IPv6 address, mail server domain name) from domain name
  - Root servers
    - 13 IP addresses [a-m].root-servers.net
- Two types of servers
  - Cache server (a.k.a. recursive server)
  - Authoritative server (a.k.a. Content server)



DNS delegation



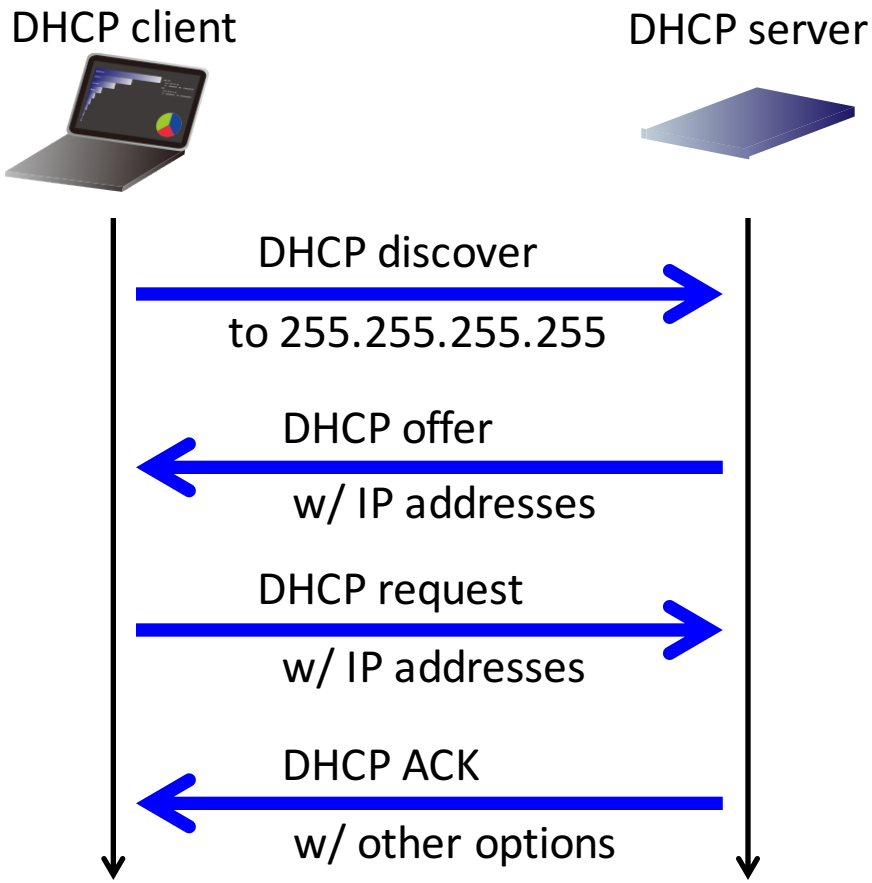
An example of DNS lookup



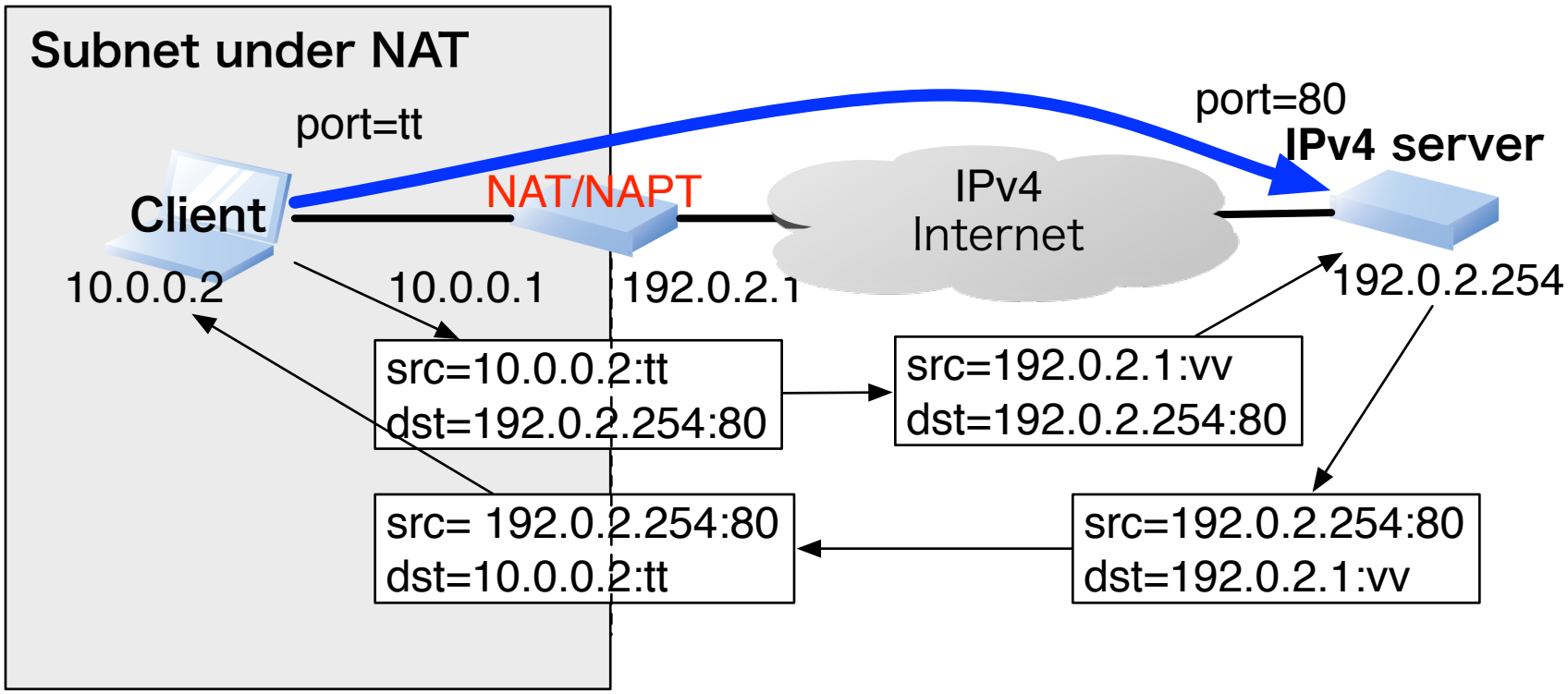
DHCP: Dynamic Host Configuration Protocol

- Auto-configuration
  - Assigning an IP address (w/ netmask)
  - Configuring default gateway
  - Configuring DNS cache servers
  - etc.
- Note
  - DHCP
    - for IPv4
    - Stateful
  - DHCPv6
    - for IPv6
    - Stateless vs. Stateful

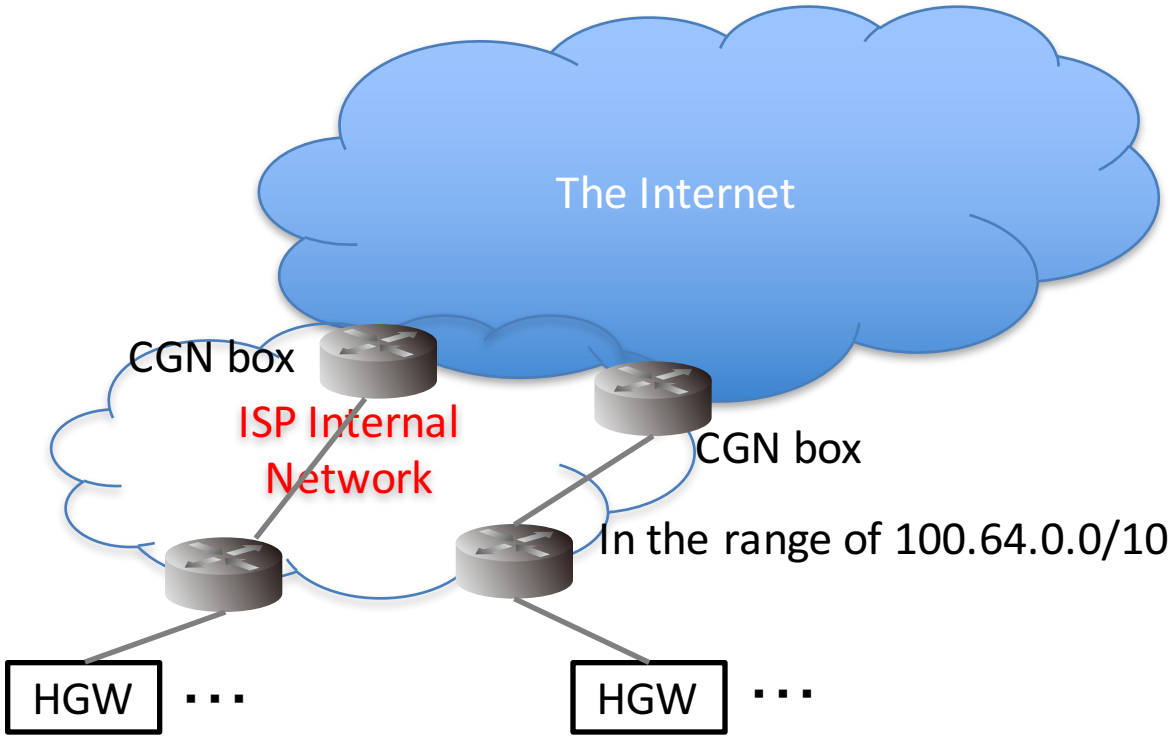
DHCP (cont.)


DHCP client
DHCP server
DHCP discover
to 255.255.255.255
DHCP offer
w/ IP addresses
DHCP request
w/ IP addresses
DHCP ACK
w/ other options

NAT/NAPT


Subnet under NAT
port=tt
port=80
IPv4 server
Client
NAT/NAPT
IPv4
Internet
10.0.0.2
10.0.0.1
192.0.2.1
192.0.2.254
src=10.0.0.2:tt
dst=192.0.2.254:80
src=192.0.2.1:vv
dst=192.0.2.254:80
src= 192.0.2.254:80
dst=10.0.0.2:tt
src=192.0.2.254:80
dst=192.0.2.1:vv

Carrier Grade NAT (CGN)



Private address range

- 10.0.0.0/8
- 172.16.0.0/12
- 192.168.0.0/16

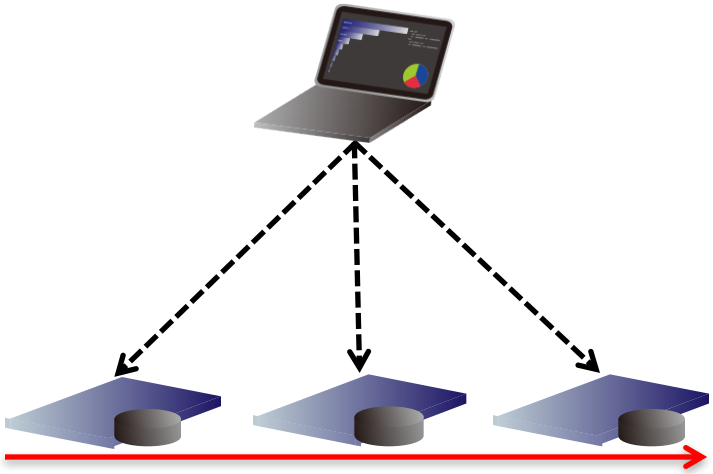
TECHNOLOGIES RELATED TO CLOUD COMPUTING

Network for Cloud Computing

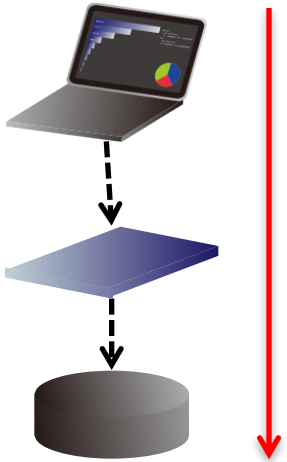
- Load balance / Scaling
  - Horizontal vs. Vertical
- Network infrastructure
  - Traffic engineering / Virtualization
    - VXLAN
    - OpenFlow

Fundamentals of load balance

- Horizontal scaling
- Vertical scaling
  - Web server + DB server + contents server



Horizontal scaling

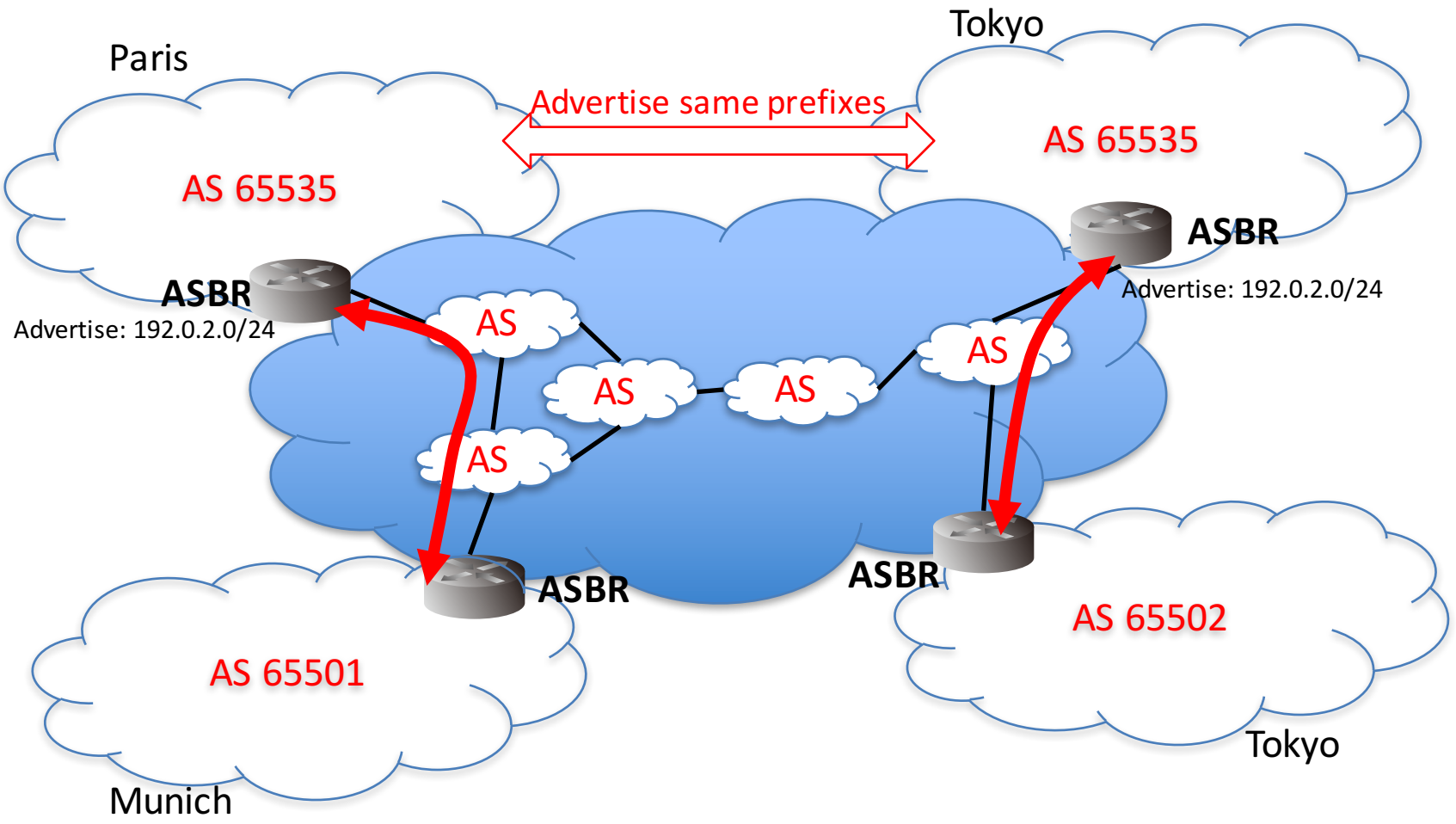


Vertical scaling

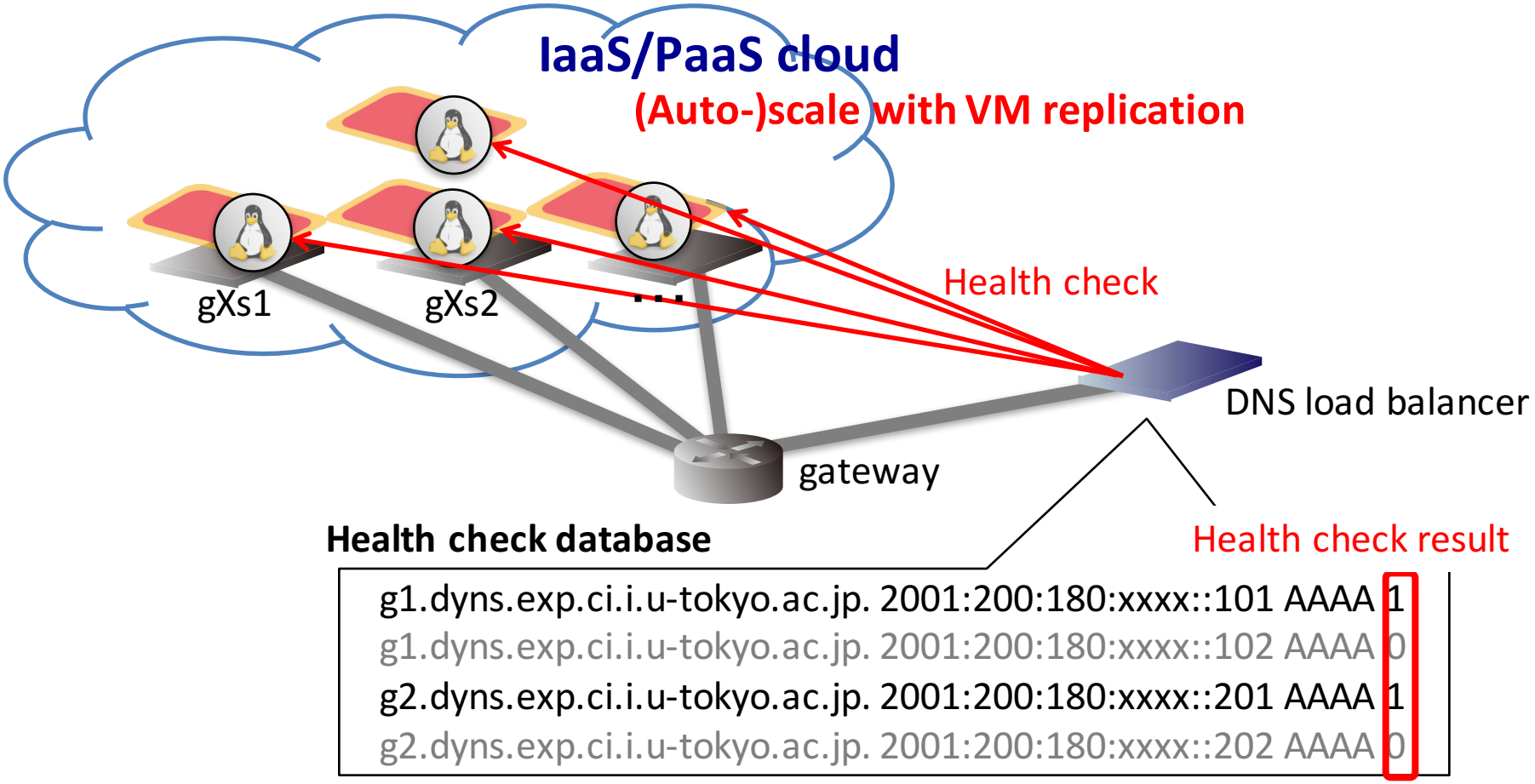
Load balancing technologies

- Horizontal load balancing
  - Local area
    - load balancer
    - OSPF anycast
  - Wide area
    - BGP anycast
    - DNS round robin
    - Localization with DNS

BGP Anycast


Paris
Tokyo
Advertise same prefixes
AS 65535
AS 65535
ASBR
Advertise: 192.0.2.0/24
ASBR
Advertise: 192.0.2.0/24
AS
AS
AS
AS
AS
ASBR
ASBR
AS 65501
AS 65502
Munich
Tokyo

DNS-based Load Balancing



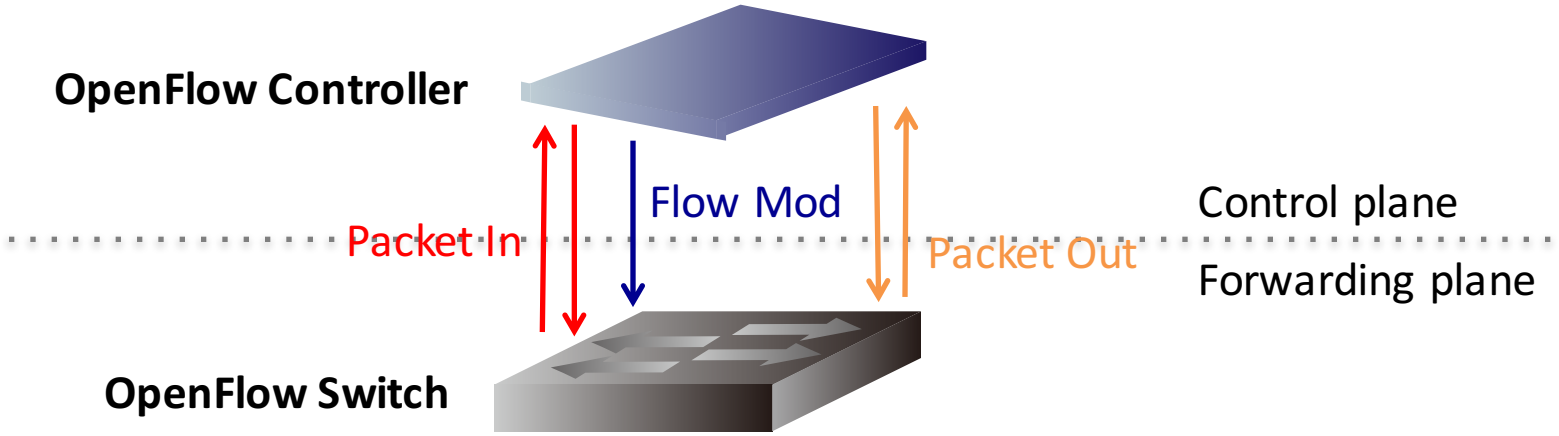
For scaling out and down: Need to add *enable flag* or a function to add/delete records
(N.B., Health check and TTL take some minutes to disable the record.)

Software Defined Networking (SDN)

- Legacy IP routing facilities
  - Traffic engineering by operators' hands
    - Destination-based routing cannot establish "path"
    - Proactive routing cannot change path in accordance with current traffic volume
- Software Defined Networking
  - Traffic engineering by software
    - Flexible path control
    - Reactive path control (On-demand control)

OpenFlow

- OpenFlow is one of SDN technologies
  - Separation of
    - control plane
    - forwarding plane



Flow Table Entry

Rule	Action	Stats

Rule: 9-tuple

In Port	Dst MAC	Src MAC	Eth Type	VLAN ID	Src IP	Dst IP	Src TCP port	Dst TCP port

Actions
- Forward
- Encapsulate and forward
- Drop

NFV/SFC

- NFV: Network Functionality Virtualization

  (Standardization activity in ETSI)

  – Implement and deploy network functionalities on virtual machines
    - Dedicated hardware to software
- SFC: Service Function Chaining

  (Standardization activity in IETF)

  – Provide a protocol to chain network service functions