### Communication Networks Spring 2018



Laurent Vanbever nsg.ee.ethz.ch

ETH Zürich (D-ITET) March 26 2018

Materials inspired from Scott Shenker & Jennifer Rexford



#### Last week on Communication Networks

#### How do local computers communicate?





### **Communication Networks** Part 2: The Link Layer



- #2 How do we identify link adapters?
- #3 How do we share a network medium?
- #4 What is Ethernet?
- #5 How do we interconnect segments at the link layer?

### Communication Networks Part 2: The Link Layer



What is a link?

#2 How do we identify link adapters?

How do we share a network medium?

What is Ethernet?

How do we interconnect segments at the link layer?

MAC addresses...

### identify the sender & receiver adapters used within a link

#### are uniquely assigned

hard-coded into the adapter when built

#### use a flat space of 48 bits

allocated hierarchically

#### Why don't we simply use IP addresses?

Links can support any protocol (not just IP) different addresses on different kind of links

Adapters may move to different locations

cannot assign static IP address, it has to change

Adapters must be identified during bootstrap

need to talk to an adapter to give it an IP address

You need to solve two problems when you bootstrap an adapter

Who am I? MAC-to-IP binding How do I acquire an IP address?

Who are you? IP-to-MAC binding Given an IP address reachable on a link, How do I find out what MAC to use? Who am I? MAC-to-IP binding How do I acquire an IP address? Dynamic Host Configuration Protocol

Who are you? IP-to-MAC binding Given an IP address reachable on a link, How do I find out what MAC to use?

**Address Resolution Protocol** 

### Communication Networks Part 2: The Link Layer



What is a link?

How do we identify link adapters?

#3 How do we share a network medium?

What is Ethernet?

How do we interconnect segments at the link layer?

## In practice, Carrier-Sense Multiple Access (CSMA) is used to govern shared medium access

carrier-sense

*listen* before speaking, don't interrupt

collision detection

*stop* if someone else starts talking ensure everyone is aware of the collision

randomness

don't talk again right away

### **Communication Networks** Part 2: The Link Layer



What is a link?

How do we identify link adapters?

How do we share a network medium?

#4 What is Ethernet?

How do we interconnect segments at the link layer?

Ethernet...

#### was invented as a broadcast technology

each packet was received by all attached hosts

is now *the* dominant wired LAN technology by far the most widely used

has managed to keep up with the speed race from 10 Mbps to 400 Gbps

### **Communication Networks** Part 2: The Link Layer



What is a link?

How do we identify link adapters?

How do we share a network medium?

What is Ethernet?

#5 How do we interconnect segments at the link layer?

## Switches connect two or more LANs together at the Link layer, acting as L2 gateways

Switches are "store-and-forward" devices, they

- extract the destination MAC from the frame
- look up the MAC in a table (using exact match)
- forward the frame on the appropriate interface

Switches are similar to IP routers, except that they operate one layer below While flooding enables automatic discovery of hosts, it also creates problems when the networks has loops



Each frame leads to the creation of *at least two new frames!* exponential increase, with no TTL to remove looping frames...

#### This week on Communication Networks

Link Layer

Network Layer

The End

The Beginning

Link Layer

Network Layer

The End

The Local Area Networks we have considered so far define single broadcast domains

If one user broadcast a frame, every other user receives it

#### As the network scales, network operators like to segment their LANs

Why? Improves security

smaller attack surface (visibility & injection)

Improves performance

limit the overhead of broadcast traffic (e.g. ARP)

Improves logistics

separates traffic by role (e.g. staff, students, visitors)

# Organizational changes are too frequent to segment networks purely physically—rewiring is a major pain

What about doing this in software though?

#### Enters "Virtual Local Area Networks" (VLANs)

Definition

A VLAN logically identifies a set of ports attached to one (or more) Ethernet switches, forming one broadcast domain

## A VLAN identifies a set of ports attached to one or more Ethernet switches



Switches need configuration tables telling them which VLANs are accessible via which interfaces

Switches need configuration tables telling them which VLANs are accessible via which interfaces



#### Consider that A sends a broadcast frame

say, an ARP request



## That frame should be received by all staff members: i.e. C and F, and *only* them







# To identify VLAN, switches add new header when forwarding traffic to another switch

#### Without VLAN

preamble	dest	src	/pe		CRC
	address	address	t,		



802.1q Header (4 bytes)

(4 bits missing)

### With VLANs, Ethernet links are divided in two sets: access and trunks (inter switches) links



## Access links belong to one VLAN they do not carry 802.1q headers



# Trunk links carry traffic for more than one VLAN and as such carry 801.1q tagged frames



#### Each switch runs one MAC learning algorithm for each VLAN

When a switch receives a frame with an unknown or a broadcast destination,

it forwards it over all the ports that belong to the same VLAN

When a switch learns a source address on a port

it associates it to the VLAN of this port and only uses it when forwarding frames on this VLAN


Switches can also compute per-VLAN spanning-tree allowing a distinct SPT for each VLAN

allow the operators to use more of their links









Any communication between the red hosts on switch 5 and 6 need to go via switch 1...







Now any communication between the red hosts on switch 5 and 6 go via the direct link



Link Layer

Network Layer

The Beginning

### Moving on to IP and the network layer



### Internet Protocol and Forwarding



1 IP addresses

use, structure, allocation

- 2 IP forwarding longest prefix match rule
- 3 IP header IPv4 and IPv6, wire format

source: Boardwatch Magazine

### Internet Protocol and Forwarding



#### IP addresses

1

use, structure, allocation

IP forwarding longest prefix match rule

IP header IPv4 and IPv6, wire format IPv4 addresses are unique 32-bits number associated to a network interface (on a host, a router, ...)

IP addresses are usually written using dotted-quad notation



## Routers forwards IP packets based on their destination IP address

If IP addresses were assigned arbitrarily, routers would require forwarding entries for all of them



WAN Wide Area Network

# 17.1 billion

estimated\* # of Internet connected devices in 2016

\* Cisco Visual Networking Index 2016-2021

# 27.1 billion

estimated\* # of Internet connected devices in 2021

\* Cisco Visual Networking Index 2016-2021

Two universal tricks you can apply to any computer sciences problem

When you need...more flexibility,you add...a layer of indirection

When you need...more scalability,you add...a hierarchical structure

When you need...more scalability,you add...a hierarchical structure

# IP addresses are hierarchically allocated, similarly to the postal service

Address	
Zip	8092
Street	Gloriastrasse
Building	35 (ETZ)
Location in building	G 90
Name	Laurent Vanbever

### Nobody in the Swiss mail system knows where every single house or building is

principleRouting tables are separatedat each level of the hierarchy

each one with a manageable scale

### Forwarding in the Swiss mail

in 4 steps

- 1 Deliver the letter to the post office responsible for the zip code
- 2 Assign letter to the mail person covering the street
- 3 Drop letter into the mailbox attached to the building
- 4 Hand in the letter to the appropriate person

IP addressing is hierarchical, composed of a prefix (network address) and a suffix (host address)



## Each prefix has a given length, usually written using a "slash notation"

### IP prefix 82.130.102.0 /24

prefix length (in bits)

Here, a /24 means that we have 8 bits left to address hosts address, enough for 256 hosts

#### 82.130.102.0 /24

prefix part	host part	IP address
01010010.10000010.01100110.	0000000	82.130.102.0
01010010.10000010.01100110.	0000001	82.130.102.1
01010010.10000010.01100110.	00000010	82.130.102.2
01010010.10000010.01100110.	11111110	82.130.102.254
01010010.10000010.01100110.	11111111	82.130.102.255

# In practice, the first and last IP address of a prefix are not usable

 prefix part
 IP address

 01010010.10000010.01100110.
 00000000
 82.130.102.0

01010010.10000010.01100110. 11111111 82.130.102.255

# The address with the host part being all 0s identifies the network itself

 prefix part
 host part
 IP address

 01010010.10000010.01100110.
 00000000
 82.130.102.0

# The address with the host part being all 1s identifies the broadcast address

C.	
prefix	part

host part

IP address

01010010.10000010.01100110. 11111111 82.1

82.130.102.255

# A /24 has therefore only 254 addresses that can be allocated to hosts

Prefixes are also sometimes specified using an address and a mask

#### Address 82.130.102.0

01010010.10000010.01100110. 0000000

Mask 255.255.255.0

# ANDing the address and the mask gives you the prefix

#### Address 82.130.102.0

01010010.10000010.01100110. 0000000

111111111111111111111111111100000000

Mask 255.255.255.0

#### Given this IP prefix

### 82.130.0.0/17

#### Compute

# of addressable hosts

the prefix mask

network address

1st host address

last host address

broadcast address

Routers forward packet to their destination according to the network part, *not* the host part

# Doing so enables to scale the forwarding tables



WAN Wide Area Network

forwarding table

Hierarchical addressing enables to add new hosts without changing or adding forwarding rules



forwarding table
# Originally, there were only 5 fixed allocation sizes, (or classes)—known as classful networking

	leading bits	prefix length	# hosts	start address	end address
class A	0	8	2 <sup>24</sup>	0.0.0.0	127.255.255.255
class B	10	16	216	128.0.0.0	191.255.255.255
class C	110	24	28	192.0.0.0	223.255.255.255
class D multicast	1110			224.0.0.0	239.255.255.255
class E reserved	1111			240.0.0.0	255.255.255.255

# Classful networking was quite wasteful leading to IP address exhaustion

problem Class C was too small, so everybody requested class B which where: *i*) too big and *ii*) too few (wasted space)

solution Classless Inter-Domain Routing (CIDR) introduced in 1993

# CIDR enabled flexible division between network and hosts addresses

### CIDR must specify both the address and the mask classful was communicating this in the first address bits

#### Masks are carried by the routing algorithms

it is *not* implicitly carried in the address

Say that an organization needs 500 addresses...

with	it gets a	leading to a waste of
classful	class B (/16)	99%
CIDR	/23 (=2 class C's)	2%

With CIDR, the max. waste is bounded to 50% (why?)

Today, addresses are allocated in contiguous chunks



# As of now, the Internet has around 710,000 IPv4 prefixes



source http://www.cidr-report.org/

### The allocation process of IP address is also hierarchical

The root is held by Internet Corporation for Assigned Names and Numbers, aka ICANN



ICANN allocates large prefixes blocks to Regional Internet Registries (RIRs)



RIRs allocates parts of these prefixes blocks to Internet Service Providers (ISPs) and large institutions



ISPs and large institutions may, in turn, allocate even smaller prefixes to their own customers





 ICANN gives RIPE
 82.0.0/8

 Prefix
 01010010



RIPE gives ETHZ Prefix

82.130.64.0/18 010100101000001001

7	<b>=1</b> :	
h		

ETHZ gives ITET/TIK82.130.102.0/23Prefix01010010100000100110011



ITET gives me Address 82.130.102.254 0101001010000010011001101111110



3.88.0/21
)

- 2 129.132.0.0/16 7 192.33.96.0/21
- 3 148.187.192.0/19 8 192.33.104.0/22
- 4 195.176.96.0/19 9 192.33.108.0/23
- 5 192.33.87.0/24 10 192.33.110.0/24

### Internet Protocol and Forwarding



IP addresses use, structure, allocation

2 IP forwarding longest prefix match rule

> IP header IPv4 and IPv6, wire format

What's inside an IP router?







# Routers maintain forwarding entries for each Internet prefix

Provider 2's Forwarding table





#### Provider 2's Forwarding table

# Let's say a packet for 129.0.1.1 arrives at Provider 2

IP prefix	Output
129.0.0/8	IF#2
129.132.1.0/24	IF#2
129.132.2.0/24	IF#2
129.133.0.0/16	IF#3



When a router receives an IP packet, it performs an IP lookup to find the matching prefix

Provider 2's Forwarding table

IF#3

129.133.0.0/16

Let's say a packet for 129.0.1.1	IP prefix	Output
arrives at Provider 2	129.0.0/8	IF#2
> Provider 2 femuerde it to 15#2	129.132.1.0/24	IF#2
> Provider 2 Torwards It to IF#2	129.132.2.0/24	IF#2



CIDR makes forwarding harder though, as one packet can match many IP prefixes

Provider 2's Forwarding table

# Let's say a packet for 129.133.0.1 arrives at Provider 2

IP prefix	Output
129.0.0.0/8	IF#2
129.132.1.0/24	IF#2
129.132.2.0/24	IF#2
129.133.0.0/16	IF#3



Provider 2's Forwarding table

Let's say a packet for 129.133.0.1	IP prefix	Output
arrives at Provider 2	129.0.0/8	IF#2
M/a have two matched	129.132.1.0/24	IF#2
we have two matches!	129.132.2.0/24	IF#2
	129.133.0.0/16	IF#3



To resolve ambiguity, forwarding is done along the *most specific* prefix (*i.e.*, the longer one)

Provider 2's Forwarding table

IF#3

129.133.0.0/16

Let's say a packet for 129.133.0.1	IP prefix	Output
arrives at Provider 2	129.0.0.0/8	IF#2
> Drovidor 2 forwards it to 15#2	129.132.1.0/24	IF#2
> Provider 2 Torwards It to IF#5	129.132.2.0/24	IF#2



Could we do something better than maintaining one entry per prefix? *Yep!* 

## A child prefix can be filtered from the table whenever it shares the same output interface as its parent

**Routing Table** 



#### Routing Table



#### Routing Table



#### **Exactly the same forwarding as before**

Check out www.route-aggregation.net, to see how filtering can be done automatically

### Internet Protocol and Forwarding



IP addresses use, structure, allocation

IP forwarding longest prefix match rule

3 IP header IPv4 and IPv6, wire format

## Here is what an IPv4 packet look like on a wire

32	hits
J <b>Z</b>	DICS

4	4	8	16	
version	header length	Type of Service	Total Length	
	ldentif	ication	cation Flags Fragment offset 13	
Time <sup>-</sup>	To Live	Protocol	Header checksum	
Source IP address				
Destination IP address				
Options (if any)				
Payload				
### The version number tells us what other fields to expect, typically it is set to "4" for IPv4, or "6" for IPv6

version	header length	Type of Service	Total Length			
Identification			Flags 3	Fragment offset 13		
Time <sup>-</sup>	Time To Live Protocol		Header checksum			
	Source IP address					
	Destination IP address					
	Options (if any)					
Payload						

The header length denotes the number of 32-bits word in the header, typically set to 5 (20 bytes header)

version	<mark>header</mark> length	Type of Service	Total Length		
Identification			Flags 3	Fragment offset 13	
Time <sup>-</sup>	Fo Live	Protocol	Header checksum		
	Source IP address				
		Destination	IP add	ress	
		Options	(if any)	)	
Payload					

The ToS allows different packets to be treated differently, e.g., low delay for voice, high bandwidth for video

version	header length	Type of Service	Total Length		
Identification			Flags 3	Fragment offset 13	
Time <sup>-</sup>	Fo Live	Protocol	Header checksum		
	Source IP address				
	Destination IP address				
	Options (if any)				
Payload					

# The total length denotes the # of bytes in the entire packet, with a maximum of 65 535 bytes

version	header length	Type of Service	Total Length		
Identification			Flags 3	Fragment offset 13	
Time <sup>-</sup>	To Live	Protocol	Header checksum		
	Source IP address				
	Destination IP address				
	Options (if any)				
Payload					

# The next three fields are used when packets get fragmented

version	header length	Type of Service	Total Length			
Identification			Flags 3	Fragment offset 13		
Time <sup>-</sup>	To Live	Protocol	Header checksum			
	Source IP address					
	Destination IP address					
	Options (if any)					
Payload						

#### Every link in the Internet has a Maximum Transmission Unit (MTU)

MTU is the max. # of bytes a link can carry as one unit *e.g.*, 1500 bytes for normal Ethernet

A router can fragment a packet if the outgoing link MTU is smaller than the total packet size

Fragmented packets are recomposed at the destination why not in the network?

Assume Alice is sending 4000B packets to Bob, who is connected to a 1500B MTU link



Because the packet is larger than the MTU, router B will split the packet into fragments





# The Identification header uniquely identify the fragments of a particular packet

version	header length	Type of Service	Total Length			
Identification			FlagsFragment offset313			
Time <sup>-</sup>	To Live	Protocol	Header checksum			
	Source IP address					
	Destination IP address					
	Options (if any)					
Payload						

#### The fragment offset is used to put back the fragments in the right order in case of reordering

version	header length	Type of Service	Total Length			
Identification			Flags 3	Fragment offset 13		
Time <sup>-</sup>	Fo Live	Protocol	Header checksum			
	Source IP address					
	Destination IP address					
	Options (if any)					
Payload						

#### The flags is used to tell whether there are more fragments coming or not

version	header length	Type of Service	Total Length		
Identification			Flags 3	Fragment offset 13	
Time <sup>-</sup>	Fo Live	Protocol	Header checksum		
	Source IP address				
		Destination	n IP add	ress	
	Options (if any)				
Payload					

### The TTL is used to identify packets trapped in a loop, and eventually discard them

header length	Type of Service	Total Length			
Identification			Fragment offset 13		
Time To Live Protocol			Header checksum		
Source IP address					
Destination IP address					
Options (if any)					
Payload					
	header length Identif	header Type of Service   Identification   To Live Protocol   Source IP   Destination   Options   Payl	header Type of Service   Identification Flags   3 3   To Live Protocol Source IP address Destination IP add Options (if any) Payload		

TTL is decremented by 1 at each router, the packet is discarded if it reaches 0



default TTL values

\*nix (Linux/Mac) 64Windows 128

(used for OS fingerprinting)

### The protocol field identifies the higher level protocol carried in the packet, "6" for TCP, "17" for UDP

version	header	Type of Service	Total Length			
Identification			Flags 3	Fragment offset 13		
Time	Fo Live	Protocol	Header checksum			
	Source IP address					
	Destination IP address					
	Options (if any)					
Payload						

# The checksum is the sum of all the 16 bits words in the header (does not protect the payload)

version	header length	Type of Service	Total Length		
Identification			Flags 3	Fragment offset 13	
Time <sup>-</sup>	Fo Live	Protocol	Header checksum		
	Source IP address				
	Destination IP address				
	Options (if any)				
Payload					

### The source and destination IP uniquely identifies the source and destination host

version	header length	Type of Service	Total Length			
Identification		Flags 3	Fragment offset 13			
Time <sup>-</sup>	Fo Live	Protocol	Header checksum			
	Source IP address					
	Destination IP address					
	Options (if any)					
Payload						

#### Options were initially put to provide additional flexibility. For security reasons, there are often deactivated.

version	header length	Type of Service	Total Length						
	Identif	ication	FlagsFragment offset313						
Time	Fo Live	Protocol	Header checksum						
Source IP address									
Destination IP address									
Options (if any)									
Payload									

IP options Record route

Strict source route

Loose source route

Timestamp

Traceroute

Router alert

. . .

see http://www.networksorcery.com/enp/protocol/ip.htm#Options for a full list

#### While there are no new IPv4 available, IPv4 still accounts for more than 98% of all traffic



according to https://ams-ix.net/technical/statistics/sflow-stats/ipv6-traffic and https://ams-ix.net/technical/statistics

### With respect to IPv4, IPv6 is simpler

IPv6 was motivated by address exhaustion IPv6 addresses are 128 bits long, that's plenty!

IPv6 got rid of anything that wasn't necessary spring cleaning

Result is an elegant, if unambitious, protocol

# With respect to IPv4, IPv6 is simpler

IPv6

#### removed

reason

- fragmentation
   chocksum
   leave problems
   to the end host
- checksum
- header length simplify handling

added...

new options mechanism
 expanded addresses
 flow label
 flexibility

#### IPv4 vs IPv6

IPv4 Header					IPv6 Header				
Version	IHL	Type of Service	Total Length		Version	Traffic Class	Flow Label		
Identification		Flags	Fragment Offset	Payload Length		Next Header	Hop Limit		
Time to L	e to Live Protocol Header Checksum								
		Source A	ddress			Source Ad	drace		
Destination Address					Source Address				
	Options		Padding						
_egend				Destination Address					
Field's name kept from IPv4 to IPv6									
Field n	ot ke	pt in IPv6							
Name	and p	position cha	nged in IF	Pv6					
New fi	eld in	IPv6							

### IPv6 enables to insert arbitrary options in the packet see RFC 2460



The problem with IPv4 options is that all of them must be processed by each router, which is slow

In IPv6, only one type of optional header must be processed by each router

#### Internet Protocol and Forwarding

![](_page_134_Picture_1.jpeg)

IP addresses use, structure, allocation

IP forwarding longest prefix match rule

IP header IPv4 and IPv6, wire format

### Next week on Communication Networks

### Internet routing!

### Communication Networks Spring 2018

![](_page_136_Picture_1.jpeg)

Laurent Vanbever nsg.ee.ethz.ch

ETH Zürich (D-ITET) March 26 2018

![](_page_136_Picture_4.jpeg)