Lecture 20: Internet Routing

COMP 332, Spring 2024

Victoria Manfredi

WESLEYAN UNIVERSITY

Acknowledgements: materials adapted from Computer Networking: A Top Down Approach 7th edition: ©1996-2016, J.F Kurose and K.W. Ross, All Rights Reserved as well as from slides by Abraham Matta at Boston University, and some material from Computer Networks by Tannenbaum and Wetherall.

Today

1. Announcements
   - Homework 7 due next Wed.

2. Internet routing
   - overview
   - intra-AS routing
   - inter-AS routing

3. Internet addressing
   - IPv6
   - Dynamic Host Configuration Protocol (DHCP)
   - Network Address Translation (NAT)

Internet Routing

OVERVIEW

From graph algorithms to routing protocols

Need to address Internet reality

1. Internet is network of networks
   - hierarchical structure
   - routers not all identical
     - some routers connect different networks together
   - each network admin may want to control routing in its own network

2. Scalability with billions of destinations
   - don't all fit in one routing table
   - can't exchange routing tables this big
     - would use all link capacity

Scalable routing on the Internet

Aggregate routers into regions called Autonomous Systems

Autonomous Systems (AS)

- aka domain
- network under single administrative control
  - company, university, ISP, …
- 30,000+ ASes: AT&T, IBM, Wesleyan …
- each AS has a unique 16-bit AS #
  - Wesleyan: AS167
  - BBN: used to be AS1: was first org to get AS # then L3 later acquired

```
AS160   U-CHICAGO-AS - University of Chicago, US
AS161   TI-AS - Texas Instruments, Inc., US
AS162   DNIC-AS-00162 - Navy Network Information Center (NNIC), US
AS163   IBM-RESEARCH-AS - International Business Machines Corporation,
AS164   DNIC-AS-00164 - DoD Network Information Center, US
AS165   DNIC-AS-00165 - DoD Network Information Center, US
AS166   IDA-AS - Institute for Defense Analyses, US
AS167   WESLEYAN-AS - Wesleyan University, US
AS168   UMASS-AMHERST - University of Massachusetts, US
AS169   HANSCOM-NET-AS - Air Force Systems Networking, US
```

Hierarchical routing

2-level route propagation hierarchy

1. intra AS routing protocol between routers in same AS
   - aka intra domain routing protocol
   - aka interior gateway protocol
   - each AS selects its own

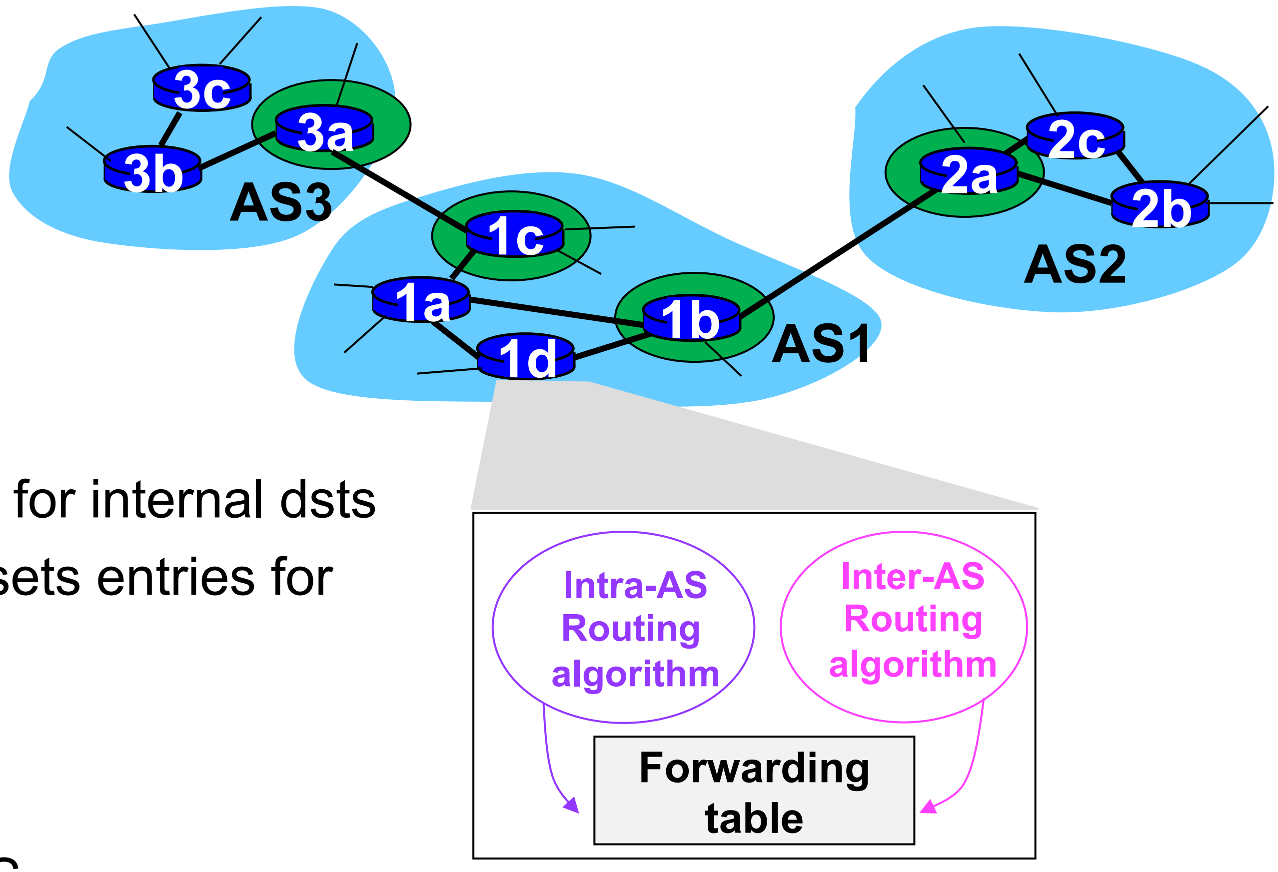
Focus is performance

2. inter AS routing protocol between gateway routers in different ASes
   - aka inter domain routing protocol
   - aka exterior gateway protocol
   - Internet-wide standard

Policy may dominate performance

Q: Can routers in different ASes run different intra AS routing protocol?

Hierarchical routing



Forwarding table

- intra-AS sets entries for internal dsts
- inter-AS & intra-AS sets entries for external dsts

Gateway router

- at edge of its own AS
- direct link to router in another AS
- perform inter-AS as well as intra-AS routing
- distributes results of inter-AS routing to other routers in AS

Example: set forwarding table in router 1d

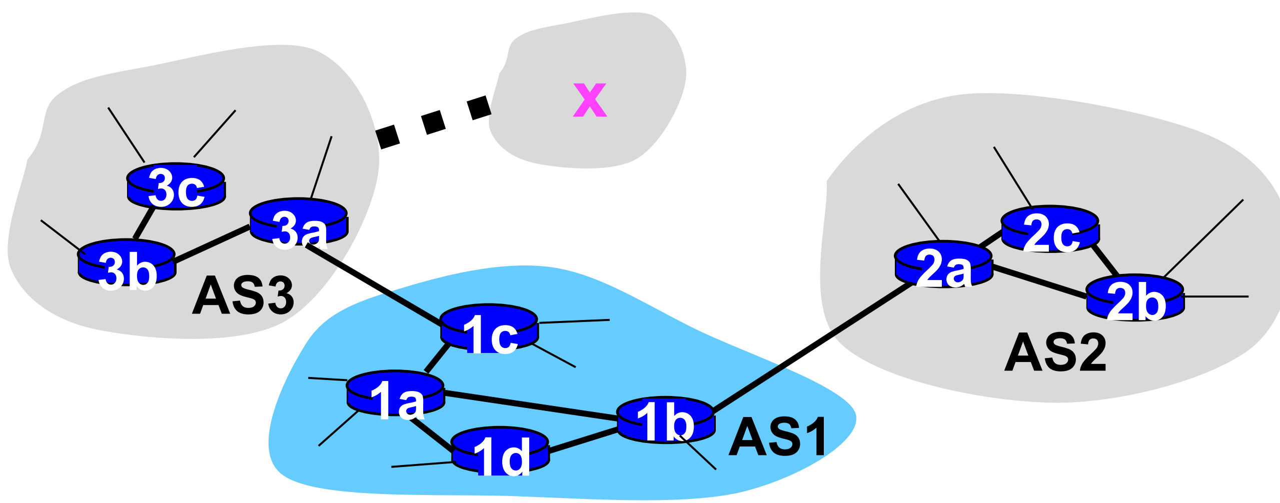
AS1 learns (from inter-AS protocol)

– subnet **x** is reachable via AS3 (gateway 1c) but not via AS2

Router 1d learns (from intra-AS protocol)

– that its interface **y** is on least cost path to 1c.

– installs forwarding table entry **(x,y)**



Q: What if multiple ASes can be used to reach x?

Internet ROUTING

INTRA-AS ROUTING

Most common intra-AS routing protocols

RIP

- Routing Information Protocol
- distance vector protocol

(E)IGRP

- (Enhanced) Interior Gateway Routing Protocol
- Cisco proprietary for decades, until 2016
- distance vector protocol

IS-IS

- Intermediate System to Intermediate System
- link state protocol

OSPF

- Open Shortest Path First
- link state protocol

Open Shortest Path First (OSPF)

Open

- i.e., publicly available

Link-state algorithm

1. Each router floods its link state to all other routers in AS
   - msgs carried directly over IP, authentication possible
   - supports unicast (1src –1dst) and multicast (1src - multiple dst)
2. Each router builds topology map
3. Route computation using Dijkstra's
   - can have multiple paths with same cost
     - traffic can go over different paths
   - can have different costs per link depending on type of service
     - e.g., satellite link cost: low for best effort, high for real time

Internet ROUTING

INTER-AS ROUTING

Inter-AS routing

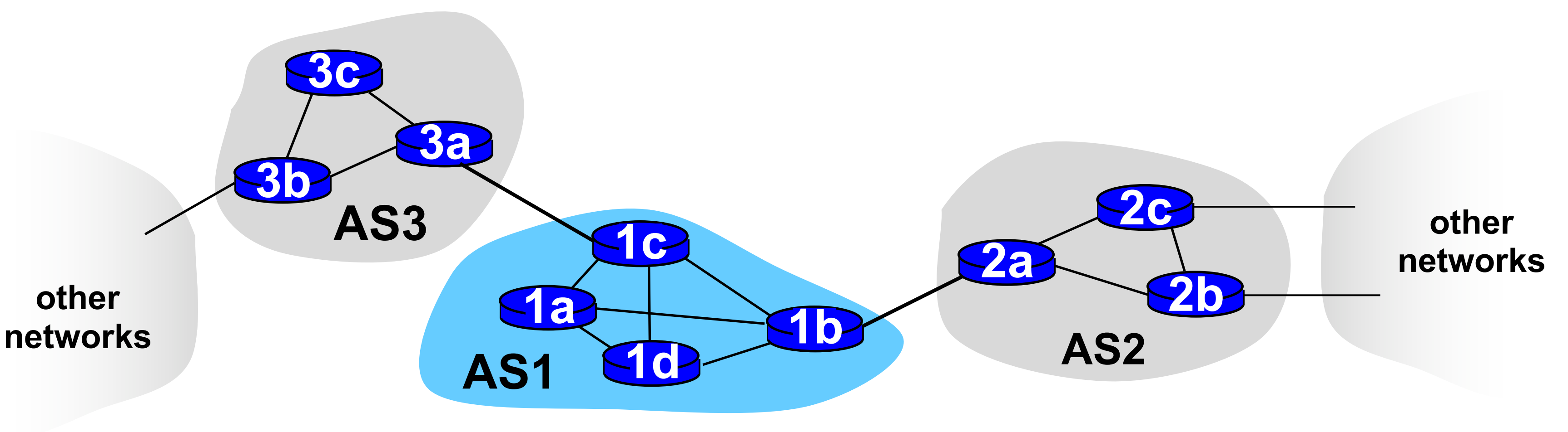
Router in AS1 receives pkt destined outside of AS1

– router forwards pkt to gateway router, but which one?

AS1 must learn which dsts reachable through neighbor ASes

– propagate this reachability info to all routers in AS1

⇒ job of inter-AS routing!



Border Gateway Protocol (BGP)

Defacto inter-domain routing protocol

- allows subnet to advertise its existence to rest of Internet
- path vector protocol

BGP provides way to find good routes to other networks

- based on reachability info and policy

Q: why must all ASes use same inter-AS protocol

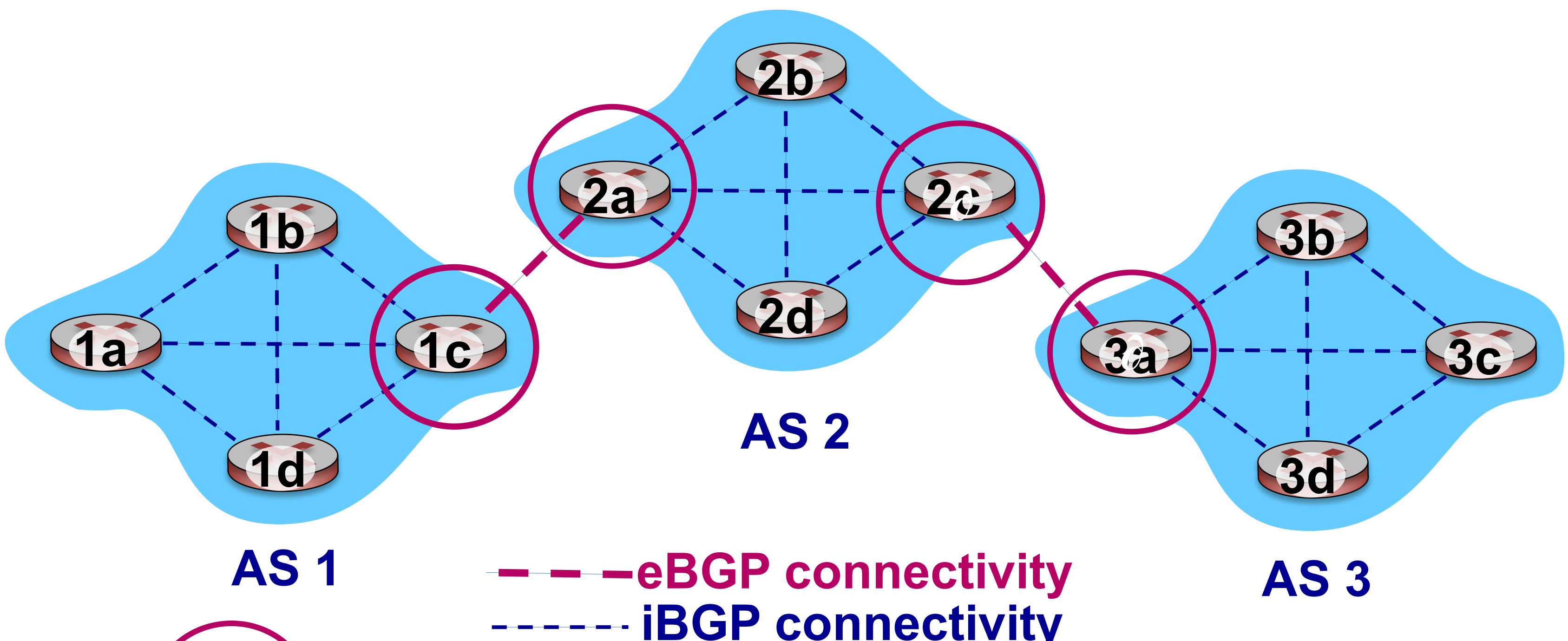
eBGP vs. iBGP connections

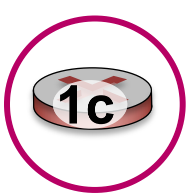
eBGP: external

obtain subnet reachability info (routes) from neighboring Ases

iBGP: internal

propagate externally learned reachability info (routes) to all routers in AS similar to intra-AS routing protocols but more scalable



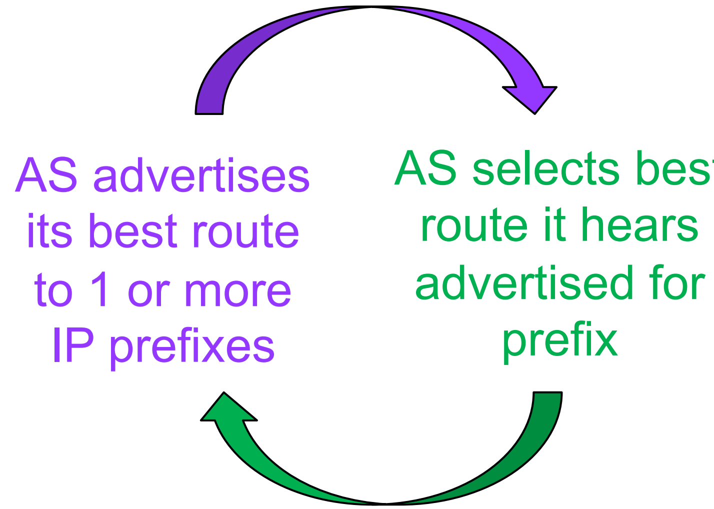


gateway routers run both eBGP and iBGP protocols

How BGP works

Similarities with distance vector

- per dst route info advertised
- no global sharing of network topology
- iterative distributed convergence



Differences from distance vector

- selects best route based on policy not min cost
- path vector routing
  - advertises entire path for each dst rather than cost
    - allows policies based on full path
    - avoids loop: if your AS is in path then discard
  - selective route advertisements
    - choose not to advertise route to dst for policy reasons
    - aggregate routes for scalability: e.g., a.b.*.* and a.c.*.* become a.*.*.*

Policy-shaped route selection

Political, economic, security considerations

Shaped by business relationships between ASes

- AS1 is customer of AS2 (AS 1 pays AS2)
- AS1 is provider of AS 2
- AS1 is peer of AS 2 (peers don't pay each other to exchange traffic)

E.g.,

- don't want to carry commercial traffic on university network
- traffic to apple shouldn't transit through google
- pentagon traffic shouldn't transit through Iraq

Why BGP is so complicated!

Why different intra- vs. inter-AS routing?

Policy

- inter-AS
  - admin wants control over how its traffic routed, who routes through its net
- intra-AS
  - single admin, so no policy decisions needed

Scale

- hierarchical routing saves table size, reduced update traffic

Performance

- inter-AS
  - policy may dominate over performance
- intra-AS
  - can focus on performance

Routing blackholes

The Register®
Biting the hand that feeds IT

CENTER SOFTWARE SECURITY DEVOPS BUSINESS PERSONAL TECH SCIENCE

Data Center ▸ Networks

Google routing blunder sent Japan's Internet dark on Friday

Another big BGP blunder

By Richard Chirgwin 27 Aug 2017 at 22:35 40 SHARE ▼

Last Friday, someone in Google fat-thumbed a border gateway protocol (BGP) advertisement and sent Japanese Internet traffic into a black hole.

The trouble began when The Chocolate Factory "leaked" a big route table to Verizon, the result of which was traffic from Japanese giants like NTT and KDDI was sent to Google on the expectation it would be treated as transit.

Since Google doesn't provide transit services, as BGP Mon explains, that traffic either filled a link beyond its capacity, or hit an access control list, and disappeared.

The outage in Japan only lasted a couple of hours, but was so severe that Japan Times reports the country's Internal Affairs and Communications ministries want carriers to report on what went wrong.

BGP Mon dissects what went wrong here, reporting that more than 135,000 prefixes on the Google-Verizon path were announced when they shouldn't have been.

The Register®
Biting the hand that feeds IT

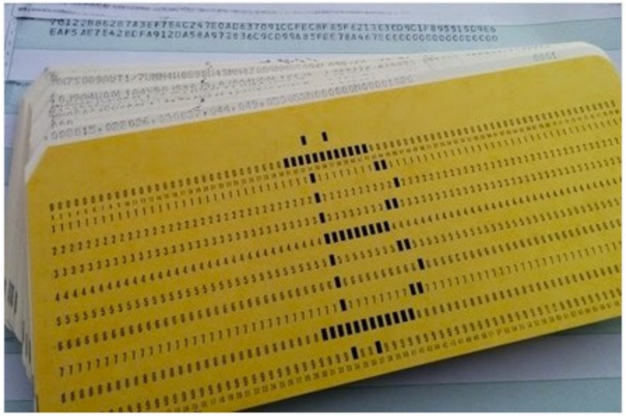
CENTER SOFTWARE SECURITY DEVOPS BUSINESS PERSONAL TECH SCIENCE

Security

Evil ISPs could disrupt Bitcoin's blockchain

Boffins say BGP is a threat to the crypto-currency

By Richard Chirgwin 11 Apr 2017 at 03:03 11 SHARE ▼



Attacks on Bitcoin just keep coming: ETH Zurich boffins have worked with Aviv Zohar of The Hebrew University in Israel to show off how to attack the crypto-currency via the Internet's routing infrastructure.

That's problematic for Bitcoin's developers, because they don't control the attack vector, the venerable Border Gateway Protocol (BGP) that defines how packets are routed around the Internet.

BGP's problems are well-known: conceived in a simpler era, it's designed to trust the information it receives. If a careless or malicious admin in a carrier or ISP network sends incorrect BGP route information to the Internet, they can black-hole significant chunks of 'net traffic.

In this paper at arXiv, explained at this ETH Website, Zohar and his collaborators from ETH, Maria Apostolaki and Laurent Vanbever, show off two ways BGP can attack Bitcoin: a partition attack, and a delay attack.

BGP hijacking

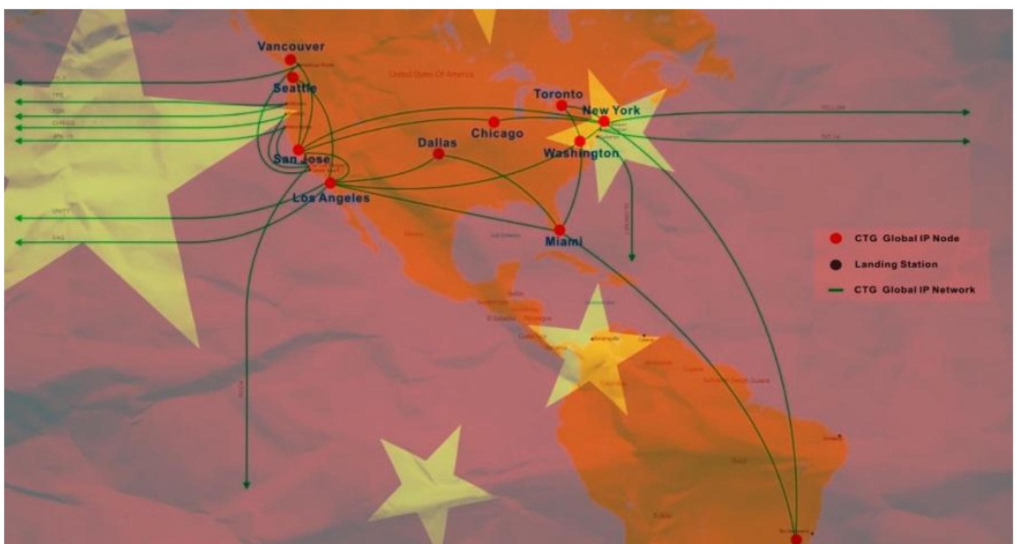
https://www.zdnet.com/article/china-has-been-hijacking-the-vital-internet-backbone-of-western-countries/

China has been 'hijacking the vital internet backbone of western countries'

Chinese government turned to local ISP for intelligence gathering after it signed the Obama-Xi cyber pact in late 2015, researchers say.

By Catalin Cimpanu for Zero Day | October 26, 2018 -- 12:39 GMT (05:39 PDT) | Topic: Security

2



MORE FROM CATALIN CIMPANU

Security
Many CMS plugins are disabling TLS certificate validation... and that's very bad

Security
Google launches reCAPTCHA v3 that detects bad traffic without user interaction

Security
US bans exports to Chinese DRAM maker citing national security risk

Security
Pakistani bank denies losing $6 million in country's 'biggest cyber attack'

Takeaways

Internet is divided into autonomous systems (AS)

- routing is done between (inter) and within (intra) autonomous systems: hierarchical routing
- one AS contains many subnets and many routers

Traffic is still routed to subnets

- but how those routes are set depend on intra-AS and inter-AS routing protocols.

Intra-AS routing

- performance focused
- RIP (Distance-Vector), OSPF (Link-State)

Inter-AS routing

- policy focused
- BGP (Path-Vector)

Internet Addressing

IPV6 ADDRESSES

IPv6 motivation

Initial motivation

- 32-bit address space soon to be completely allocated
- 128-bit IPv6 address

Additional motivation

- header format helps speed processing/forwarding
- header changes to facilitate QoS

IPv6 packet format

- fixed-length 40 byte header
- no fragmentation allowed

Dig www.google.com

```
>   dig AAAA www.google.com

; <<>> DiG 9.10.6 <<>> AAAA www.google.com
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 59860
;; flags: qr rd ra; QUERY: 1, ANSWER: 1, AUTHORITY: 0, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 512
;; QUESTION SECTION:
;www.google.com.                        IN      AAAA

;; ANSWER SECTION:
www.google.com.         172     IN      AAAA    2607:f8b0:4006:821::2004

;; Query time: 21 msec
;; SERVER: 2001:558:feed::1#53(2001:558:feed::1)
;; WHEN: Thu Apr 20 12:02:23 EDT 2023
;; MSG SIZE  rcvd: 71
```

AAAA is an IPv6 record

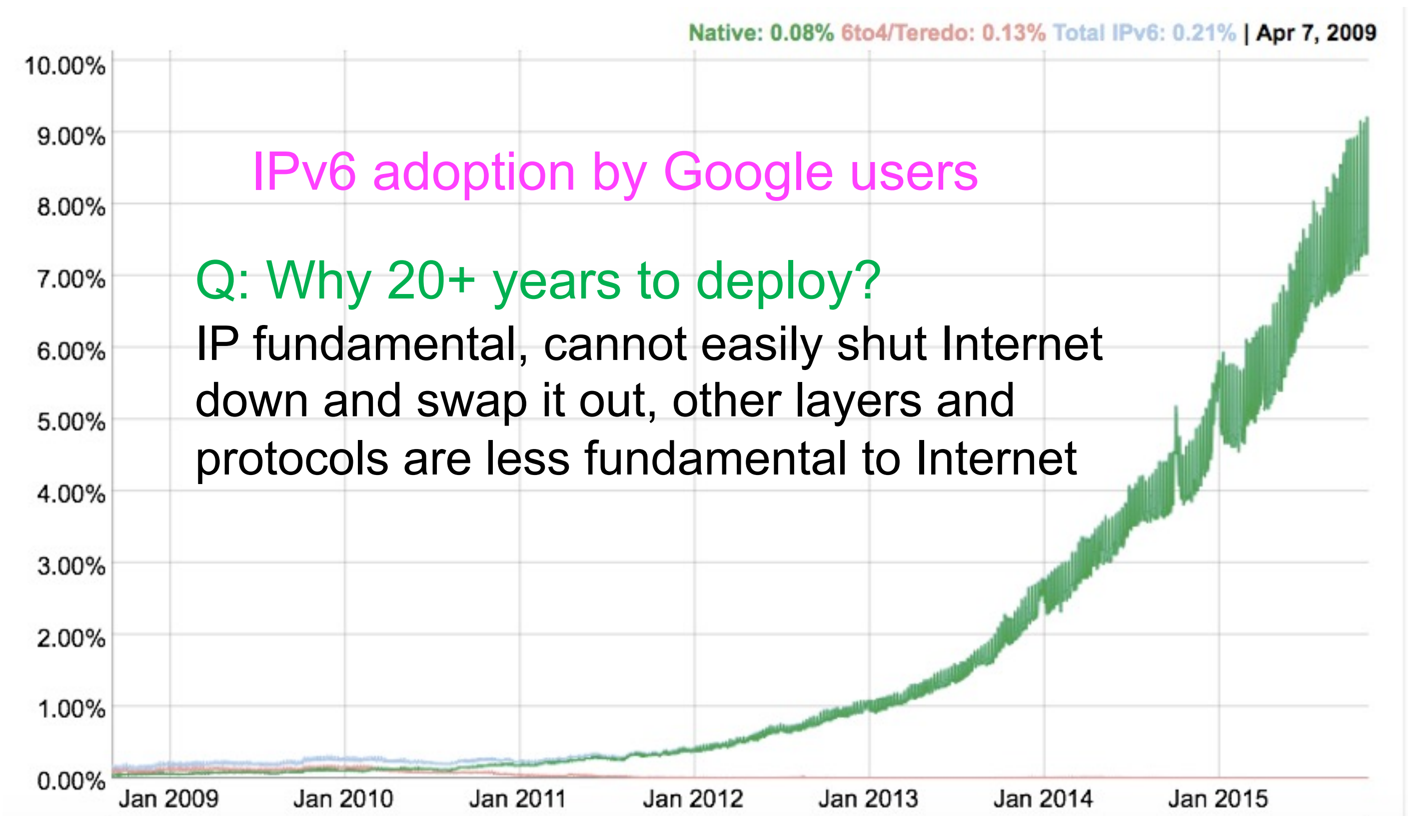
Ifconfig example

```
> ifconfig
lo0: flags=8049<UP,LOOPBACK,RUNNING,MULTICAST> mtu 16384
        options=1203<RXCSUM,TXCSUM,TXSTATUS,SW_TIMESTAMP>
        inet 127.0.0.1 netmask 0xff000000
        inet6 ::1 prefixlen 128
        inet6 fe80::1%lo0 prefixlen 64 scopeid 0x1
        nd6 options=201<PERFORMNUD,DAD>
gif0: flags=8010<POINTOPOINT,MULTICAST> mtu 1280
stf0: flags=0<> mtu 1280
en0: flags=8863<UP,BROADCAST,SMART,RUNNING,SIMPLEX,MULTICAST> mtu 1500
        ether 78:4f:43:73:43:26
        inet6 fe80::1c8d:4bcb:b52d:9d1d%en0 prefixlen 64 secured scopeid 0x5
        inet 10.66.104.246 netmask 0xfffffc00 broadcast 10.66.107.255
        nd6 options=201<PERFORMNUD,DAD>
        media: autoselect
        status: active
```

IPv6 deployment

Standardized ~1998

- 2008: IPv6 < 1% of Internet traffic
- 2011: IPv6 increasingly implemented in OS, mandated by governments and cell providers for new network devices, ….
- as recently as last year, Wesleyan did not support IPv6



IPv6 adoption by Google users

Q: Why 20+ years to deploy?

IP fundamental, cannot easily shut Internet down and swap it out, other layers and protocols are less fundamental to Internet

Internet Addressing

DYNAMIC HOST CONFIGURATION PROTOCOL (DHCP)

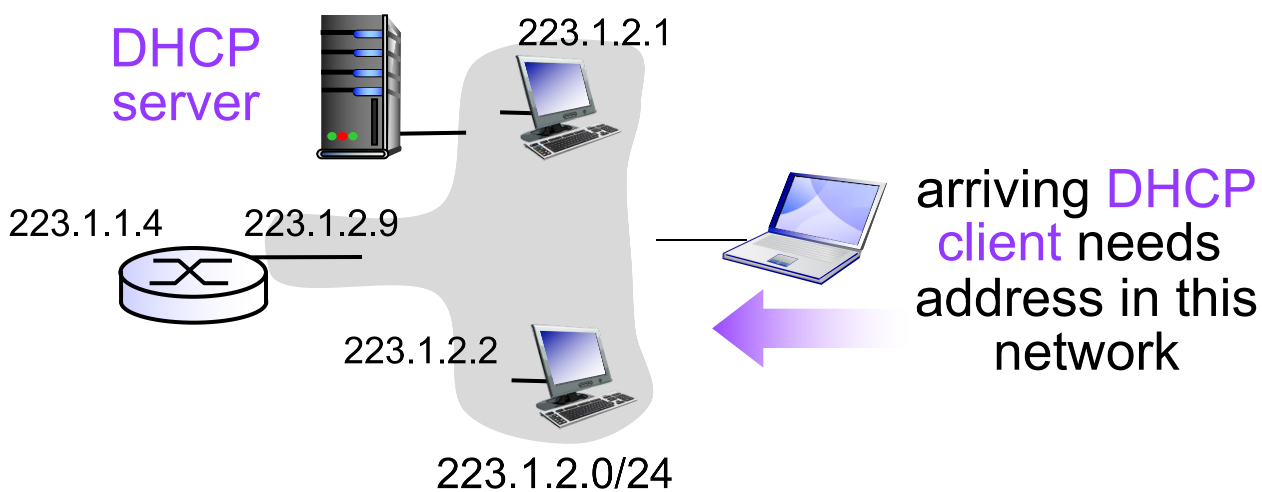
DHCP: Dynamic Host Configuration Protocol

Goal

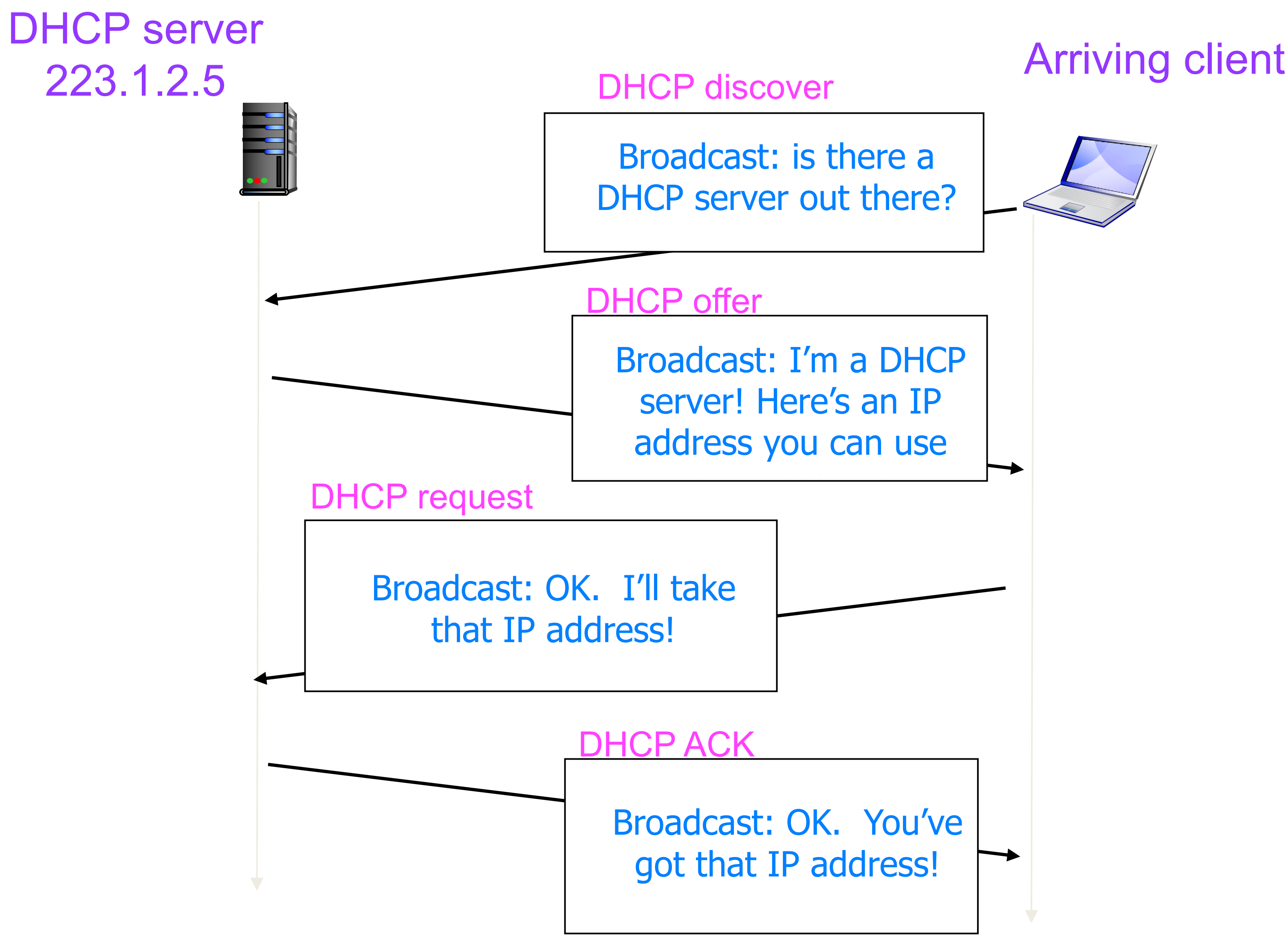
- let host dynamically obtain IP addr from server when it joins network

Benefits

- reuse of addresses by different hosts
  - only hold address while connected to network
  - host can renew its lease on address in use
- support for mobile users who want to join network



Client-server scenario



Q: What layer is DHCP in?
Q: What transport layer protocol does DHCP run over?

No.	Time	Source	Destination	Proto	Length	Info
1163	6.261619	0.0.0.0	255.255.255.255	DHCP	342	DHCP Discover - Transaction ID 0xecc8a20d
1199	6.565966	0.0.0.0	255.255.255.255	DHCP	342	DHCP Discover - Transaction ID 0xecc8a20e
1201	6.570664	129.133.176.5	vmanfredismbp2.wi…	DHCP	342	DHCP Offer - Transaction ID 0xecc8a20e
1205	7.573840	0.0.0.0	255.255.255.255	DHCP	342	DHCP Request - Transaction ID 0xecc8a20e
1206	7.581751	129.133.176.6	vmanfredismbp2.wi…	DHCP	342	DHCP ACK - Transaction ID 0xecc8a20e
1208	7.597775	129.133.176.5	vmanfredismbp2.wi…	DHCP	342	DHCP ACK - Transaction ID 0xecc8a20e

```
▶ Frame 1205: 342 bytes on wire (2736 bits), 342 bytes captured (2736 bits) on interface 0
▶ Ethernet II, Src: 78:4f:43:73:43:26 (78:4f:43:73:43:26), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
▶ Internet Protocol Version 4, Src: 0.0.0.0 (0.0.0.0), Dst: 255.255.255.255 (255.255.255.255)
▶ User Datagram Protocol, Src Port: 68 (68), Dst Port: 67 (67)
▼ Bootstrap Protocol (Request)
    Message type: Boot Request (1)
    Hardware type: Ethernet (0x01)
    Hardware address length: 6
    Hops: 0
    Transaction ID: 0xecc8a20e
    Seconds elapsed: 1
  ▶ Bootp flags: 0x0000 (Unicast)
    Client IP address: 0.0.0.0 (0.0.0.0)
    Your (client) IP address: 0.0.0.0 (0.0.0.0)
    Next server IP address: 0.0.0.0 (0.0.0.0)
    Relay agent IP address: 0.0.0.0 (0.0.0.0)
    Client MAC address: 78:4f:43:73:43:26 (78:4f:43:73:43:26)
    Client hardware address padding: 00000000000000000000
    Server host name not given
    Boot file name not given
    Magic cookie: DHCP
  ▶ Option: (53) DHCP Message Type (Request)
  ▶ Option: (55) Parameter Request List
  ▶ Option: (57) Maximum DHCP Message Size
  ▶ Option: (61) Client identifier
  ▶ Option: (50) Requested IP Address
  ▶ Option: (54) DHCP Server Identifier
  ▶ Option: (12) Host Name
  ▶ Option: (255) End
    Padding: 000000
```

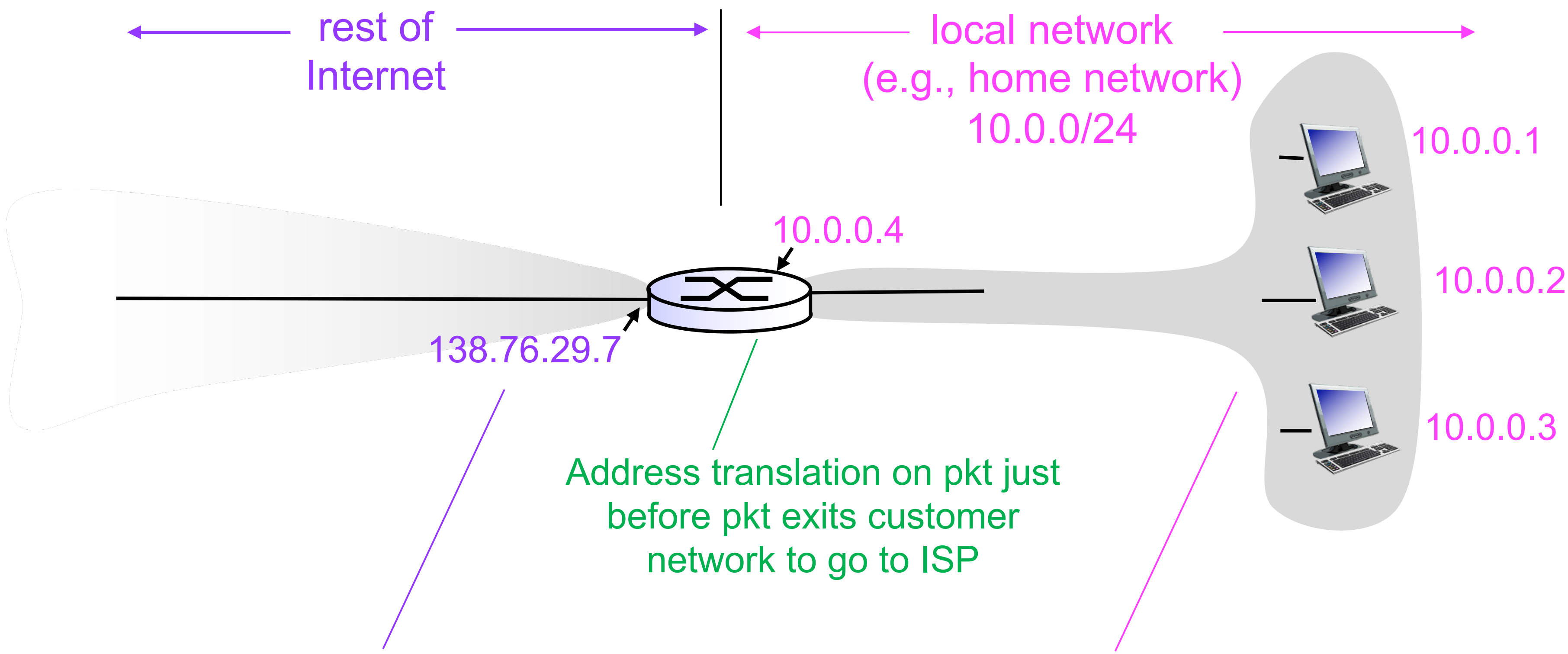
Internet Addressing

NETWORK ADDRESS TRANSLATION (NAT)

Network Address Translation (NAT)

Motivation

– local network uses 1 IP address as far as outside world is concerned



Externally: all packets leaving local network have same single source NAT IP address: 138.76.29.7, different source port #s

Internally: each host gets unique address from set of private subnet addresses, 10.0.0/24

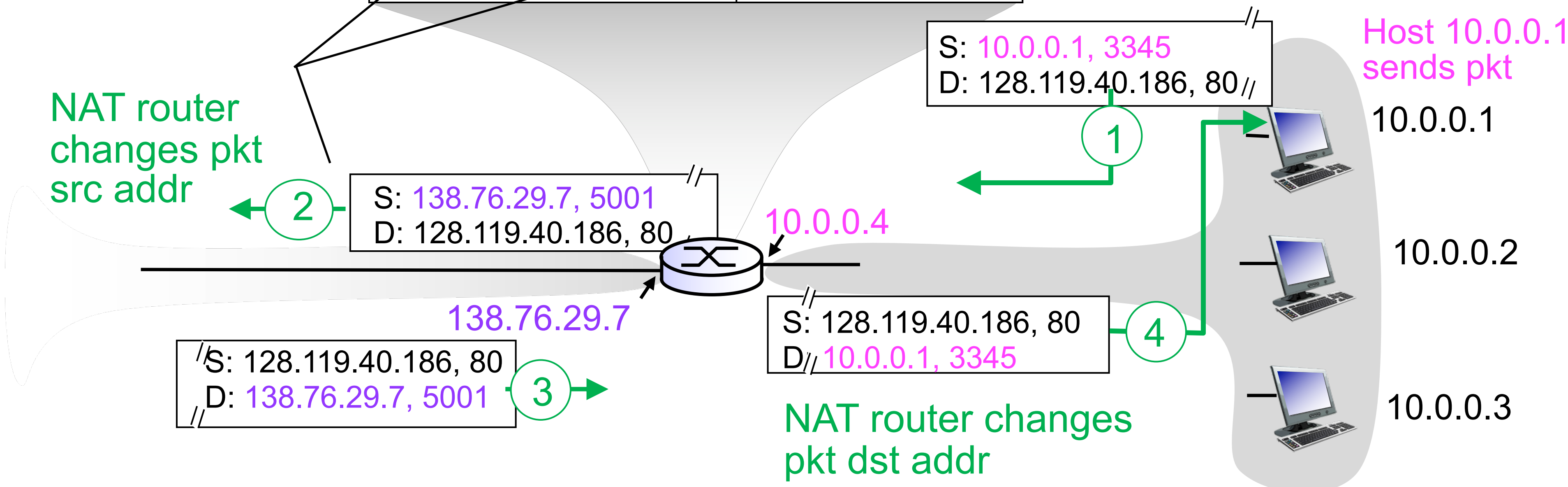
NAT implementation on router

Outgoing packets

Replace (src IP addr, port #) to (NAT IP addr, new port #)

NAT translation table

WAN side addr	LAN side addr
138.76.29.7, 5001	10.0.0.1, 3345
…	…



Incoming packets

Replace (NAT IP addr, new port #) in dst fields with corresponding (src IP addr, port #) in NAT table

Q: # of connections supported with 16-bit port #?

Q: Why was NAT was designed this way? Can ICMP traffic reach host behind NAT router?

Most traffic is TCP or UDP

NAT pros and cons

Pros

- don't need range of addresses from ISP
  - just one public IP address for all devices
- change private addresses of devices
  - without notifying outside world
- change ISP
  - without changing addresses of devices in local network
- security
  - devices inside local network not explicitly addressable or visible

Cons: NAT is controversial!

- routers should only process up to network layer
- address shortage should be solved by IPv6
- violates e2e argument
  - app designers (e.g., p2p) must account for NAT usage
- creates a strange kind of connection-oriented network
- NAT traversal
  - how to connect to server behind NAT? Problems for VOIP, FTP, …

Recall RFC 1958 architectural principles

1. Make sure it works: don't finalize standard before implementing
2. Keep it simple: Occam's razor
3. Make clear choices: choose one way to do it
4. Exploit modularity: e.g., protocol stack
5. Expect heterogeneity: different hardware, links, applications
6. Avoid static options and parameters: better to negotiate
7. Look for a good not necessarily perfect design: onus is on the designers with the outliers to work around design
8. Be strict when sending and tolerant when receiving
9. Think about scalability: no centralized databases, load evenly spread over resources
10. Consider performance and cost: if bad, no one will use network