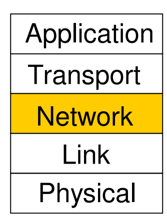
CSE 461 – Interdomain routing

Interdomain routing

- Focus:
  - Routing across internetworks made up of different parties

- Route scaling
- Route policy

- The protocol: BGP



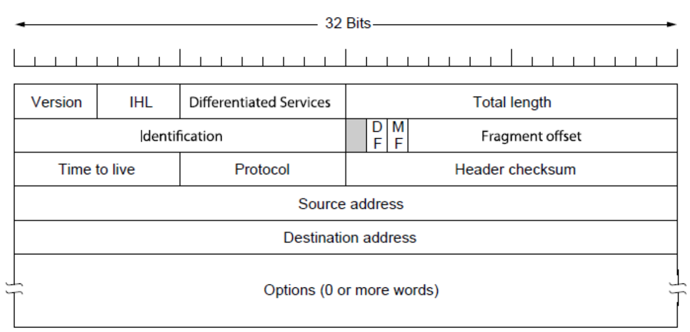
IPv4 or IPv6?

- We’re at the cusp of a multi-decade transition from IPv4 to IPv6
- What’s the big rush?

3

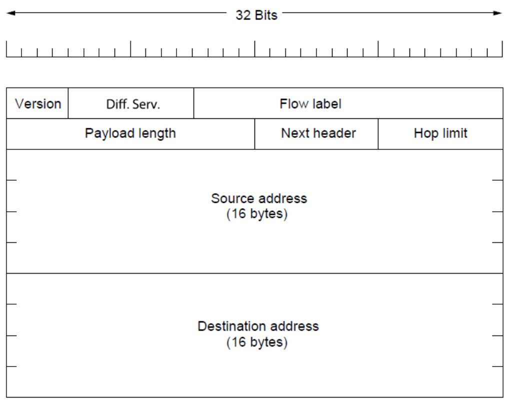
IPv4 (1981): The Problem(s)

- Version is 4; addresses are 32 bit addresses
  - TTL + header checksum → header modification on each hop



4

IPv6 Header



IP Version 6

Additions:
- Longer addresses (128 bits)
- Flow label is added (grouping hint to network)

Simplifications:
- Header checksum is gone
- Weird stuff moved to optional extensions (e.g., fragments and identification)
- (Upper) Protocol combined with Next Header
- Header length is now fixed
- TTL renamed “Hop Limit”

IPv6 Specification

```
Network Working Group                                        S. Deering
Request for Comments: 2460                                        Cisco
Obsoletes: 1883                                               R. Hinden
Category: Standards Track                                         Nokia
                                                          December 1998

                Internet Protocol, Version 6 (IPv6)
                           Specification

1.  Introduction

   IP version 6 (IPv6) is a new version of the Internet Protocol,
   designed as the successor to IP version 4 (IPv4) [RFC-791].  The
   changes from IPv4 to IPv6 fall primarily into the following
   categories:

      o  Expanded Addressing Capabilities

         IPv6 increases the IP address size from 32 bits to 128 bits, to
         support more levels of addressing hierarchy, a much greater
         number of addressable nodes, and simpler auto-configuration of
         addresses.  The scalability of multicast routing is improved by
         adding a "scope" field to multicast addresses.  And a new type
         of address called an "anycast address" is defined, used to send
         a packet to any one of a group of nodes.
```


IPv6 Rollout (1999)

```
From: IANA [iana@ISI.EDU]
Sent: Wednesday, July 14, 1999 12:32 PM
To: iana-announce@ISI.EDU
Cc: 'iana'
Subject: Delegation of IPv6 address space

Internet Community,

After much discussion concerning the policy guidelines for the deployment of
IPv6 addresses, in addition to the years of technical development done
throughout the Internet community, the IANA has delegated the initial IPv6
address space to the regional registries in order to begin immediate
worldwide deployment of IPv6 addresses.

We would like to thank the current Regional Internet Registries (RIR) for
their invaluable work in the construction of the policy guidelines, which
seem to have general consensus from the Internet community.  We would also
like to thank the efforts of the IETF community and the support of the IAB
in making this effort a reality.

This is an historic moment in the continued development of the Internet.

Thank you for your valuable support and participation in the Internet
community.
```

NATs Live On

Network Working Group — P. Srisuresh
Request for Comments: 5128 — Kazeon Systems
Category: Informational — B. Ford
M.I.T.
D. Kegel
kegel.com
March 2008

State of Peer-to-Peer (P2P) Communication across Network Address Translators (NATs)

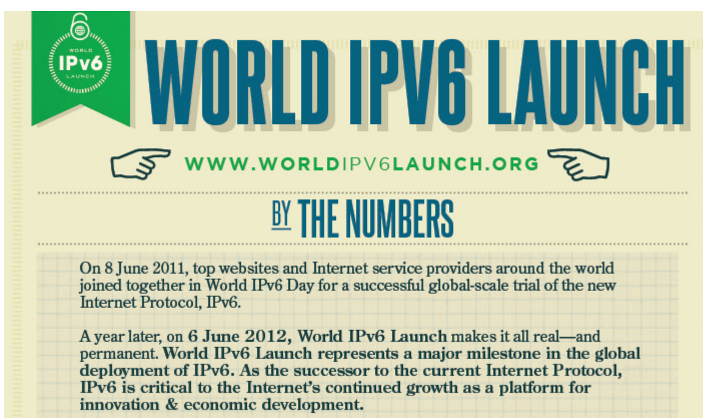
Status of This Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

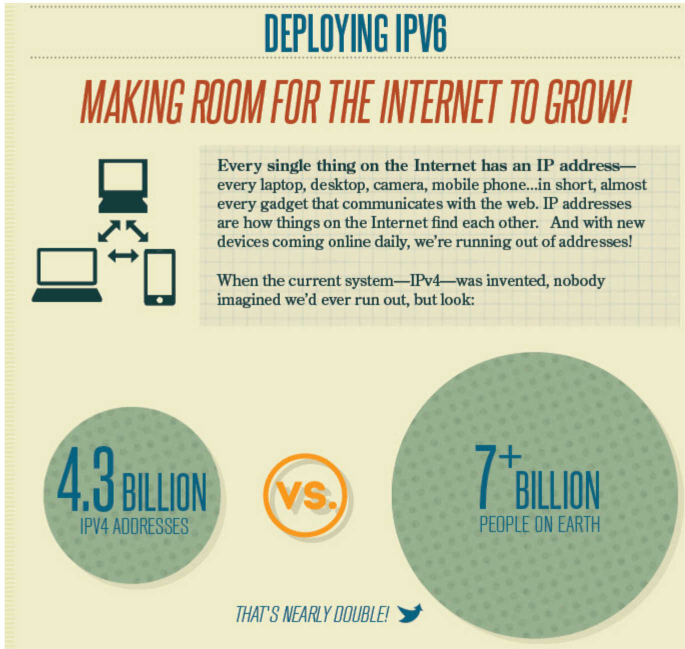
Abstract

This memo documents the various methods known to be in use by applications to establish direct communication in the presence of Network Address Translators (NATs) at the current time. Although this memo is intended to be mainly descriptive, the Security Considerations section makes some purely advisory recommendations about how to deal with security vulnerabilities the applications could inadvertently create when using the methods described. This memo covers NAT traversal approaches used by both TCP- and UDP-based applications. This memo is not an endorsement of the methods described, but merely an attempt to capture them in a document.

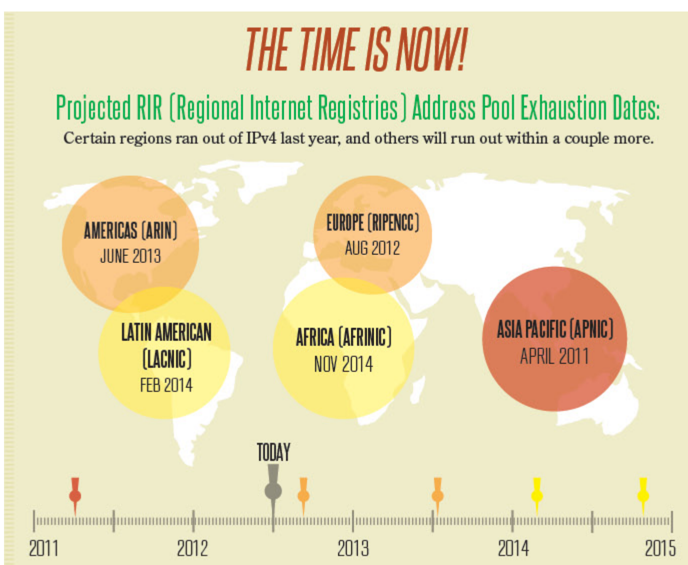
9



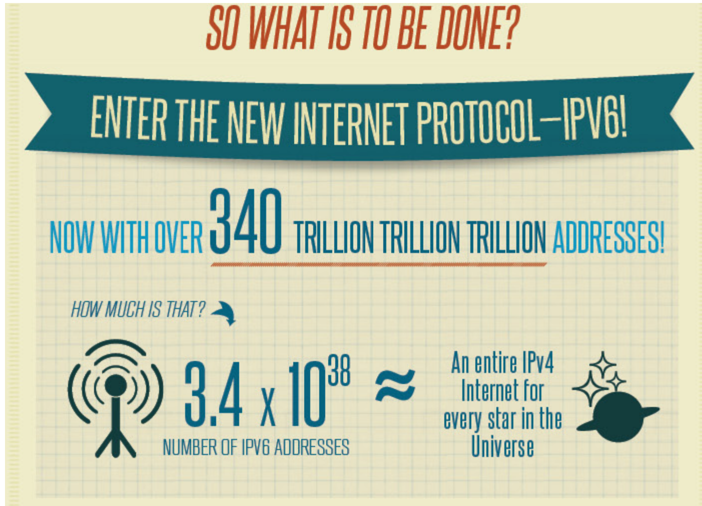
10


DEPLOYING IPV6
MAKING ROOM FOR THE INTERNET TO GROW!
Every single thing on the Internet has an IP address—every laptop, desktop, camera, mobile phone...in short, almost every gadget that communicates with the web. IP addresses are how things on the Internet find each other. And with new devices coming online daily, we're running out of addresses!
When the current system—IPv4—was invented, nobody imagined we'd ever run out, but look:
4.3 BILLION
IPV4 ADDRESSES
VS.
7+ BILLION
PEOPLE ON EARTH
THAT'S NEARLY DOUBLE!

11


THE TIME IS NOW!
Projected RIR (Regional Internet Registries) Address Pool Exhaustion Dates:
Certain regions ran out of IPv4 last year, and others will run out within a couple more.
AMERICAS (ARIN)
JUNE 2013
EUROPE (RIPENCC)
AUG 2012
LATIN AMERICAN (LACNIC)
FEB 2014
AFRICA (AFRINIC)
NOV 2014
ASIA PACIFIC (APNIC)
APRIL 2011
TODAY
2011
2012
2013
2014
2015

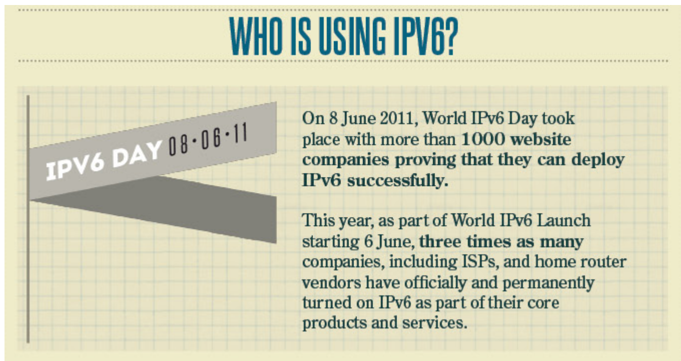
12


SO WHAT IS TO BE DONE?
ENTER THE NEW INTERNET PROTOCOL—IPV6!
NOW WITH OVER 340 TRILLION TRILLION TRILLION ADDRESSES!
HOW MUCH IS THAT?
3.4 x 10^38
NUMBER OF IPV6 ADDRESSES
≈
An entire IPv4 Internet for every star in the Universe

13


BUT THERE ARE CHALLENGES!
IPV4
IPV6
IPv4 is not forward compatible, so companies —ISPs, websites, home router vendors—have to actively enable IPv6 on their products and services.

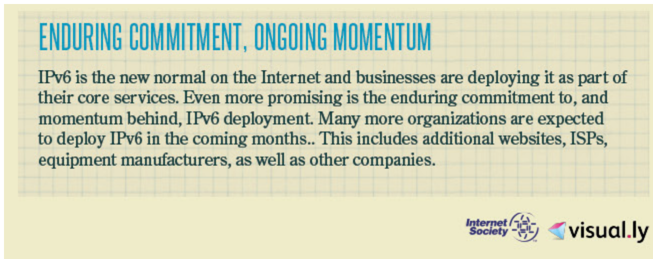
14


WHO IS USING IPV6?
IPV6 DAY 08·06·11
On 8 June 2011, World IPv6 Day took place with more than 1000 website companies proving that they can deploy IPv6 successfully.
This year, as part of World IPv6 Launch starting 6 June, three times as many companies, including ISPs, and home router vendors have officially and permanently turned on IPv6 as part of their core products and services.

15


WORLD IPV6 LAUNCH
PARTICIPATING COMPANIES
06·06·12
3,000+ WEBSITE OPERATORS
AOL, BING, CISCO, FACEBOOK, GOOGLE, MOZILLA, NASA, NETFLIX, WIKIPEDIA, YAHOO, YOUTUBE...
65 NETWORK OPERATORS
AT&T, COMCAST, FREE, KDDI, TIME WARNER CABLE, VERIZON WIRELESS, XS4ALL...
5 HOME ROUTER VENDORS
CISCO, D-LINK, NEC ACCESS TECHNICA, YAMAHA CORPORATION, ZYXEL...
AND MORE
ALL OVER THE WORLD!
WWW.WORLDIPV6LAUNCH.ORG/PARTICIPANTS

16



Google Measurements (2010)

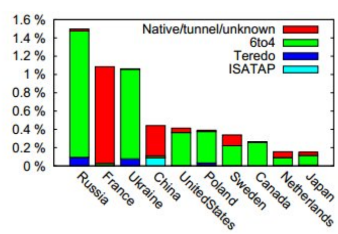
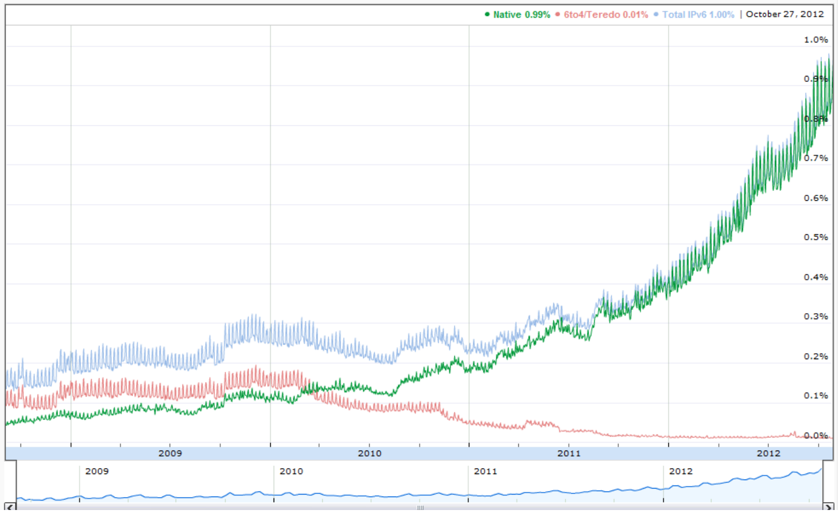


Fig. 6. Working IPv6 ratio for top-10 countries by connectivity type.

Google Measurements (2012)



19

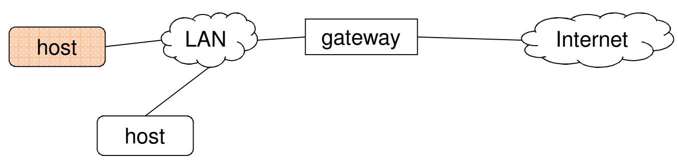
Preliminaries: Host View of IP (IPv4)

- Machine boots and looks around
  - It finds its network interface(s)
    - NIC knows its own MAC address and can start using the network at the link layer
- System needs an IP presence
- What happens?

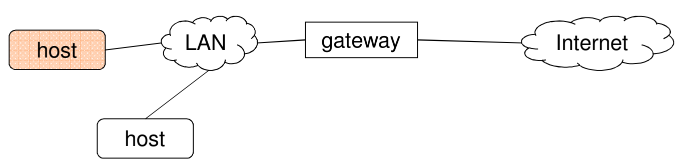
20

DHCP

- Host tries to find a DHCP server
  - *Discovery* via link-layer broadcast
- DHCP server provides host with:
  - an IP address *lease*
    - Why a lease?
  - The hosts *hostname*
  - The IP address of a *gateway*

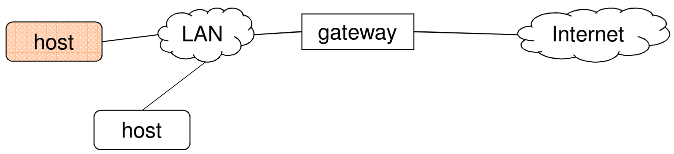


ARP



- Host knows it's at 192.168.0.17, wants to talk with 192.168.0.22
- Same "IP network" means needs to communicate directly
  - Link layer delivery for final hop
- How does it get the MAC address for 192.168.0.22?
  - ARP
- Similarly, if gateway receives an incoming packet for 192.168.0.17, needs to find new host's MAC address

NAT



- There are most hosts than available 32-bit IPv4 addresses
  - Make some of them private (non-routable)
- Gateway maps (hostIP, port, destIP, port) to (gwIP, gwPort, destIP, port)
  - Maps (destIP, port, gwIP, gwPort) to (destIP, port, gwIP, gwPort)
- Pros:
  - Saves global IP address space
  - Crude firewall
- Cons: crude firewall

23

24

Implications of NAT

- Your home machine can connect to any CSE machine, but your home machine (probably) can't be connected to from any home machine
  - Your phone can..
- Peer-to-peer (P2P) is difficult
  - E.g., Skype
  - Approaches:
    - "Punch holes" in your NAT
    - Use an intermediary to coordinate simultaneous connection
    - Use an intermediary to forward your traffic

NAT Implications

DHCP/ARP/NAT Summary

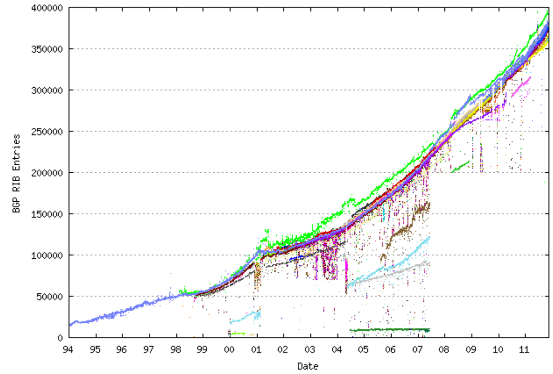
- Wireshark
  - http://www.wireshark.org

Back to the Internet (IPv4): Two key problems

- Scale
  - Size of routing tables, computation, messages
  - All grow with the size of the network
- Policy
  - Different parties with different goals make different decisions
  - ISPs are out to make money (locally good paths), not save the world (global shortest path)

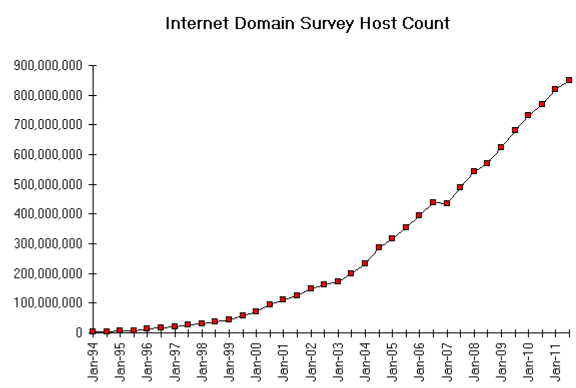
Problem: Core BGP Table Growth 1994-2010

- Growth of the BGP routing table kept at ISP routers
- Size roughly indicates routing/forwarding workload



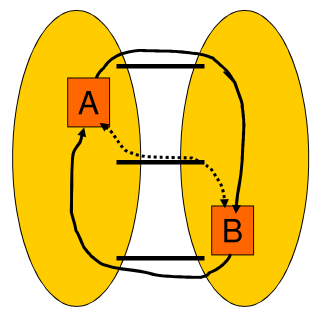
www.cidr-report.org

For context: reachable Internet hosts



Problem: Independent decisions

- Multiple parties can greatly influence the routes chosen
- Example: Early Exit / Hot Potato
  - “if it’s not for you, get rid of it”
  - Combination of best local policies is not globally best
- Side-effect: route asymmetry



31

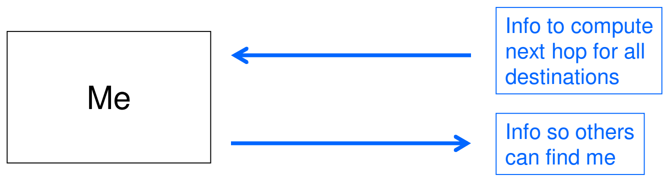
Solutions?

- Scale solution
  - Standard approach of information hiding
    - In the forms of IP prefixes and ASes
- Policy solution
  - No great solutions here!
  - Let everyone make their own decisions to the extent possible
  - Economic model gives rise to common commercial policies,
    - e.g, transit vs peering

32

Preliminaries

- Basic issue is how much information is required to effect routing
  - To scale, we want to be able to control it, at the least

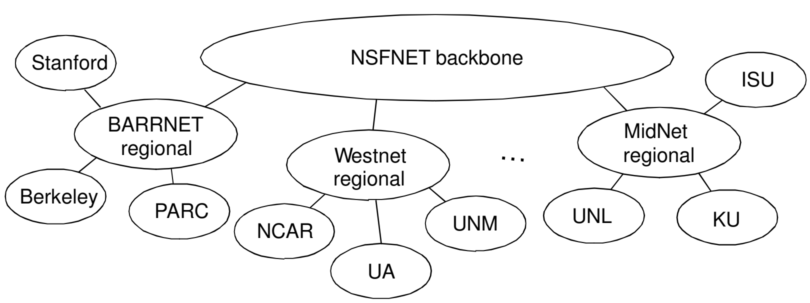


  - Aggregation: reduce amount others need to know
  - Hierarchy: reduce the amount that I need to know

m9.33

Original Structure of the Internet

- No longer quite right, but…

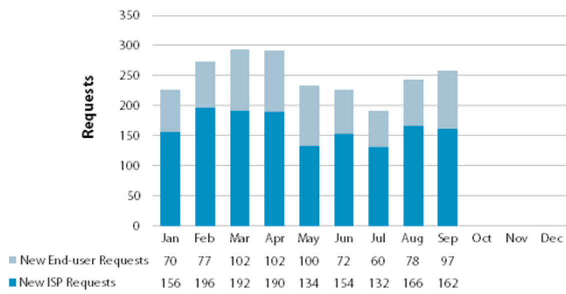


- Hierarchy lets us aggregate destination addresses
  - Don't need to know every host IP at Berkeley, just which direction all Berkeley hosts are

m9.34

IP address assignment is hierarchical

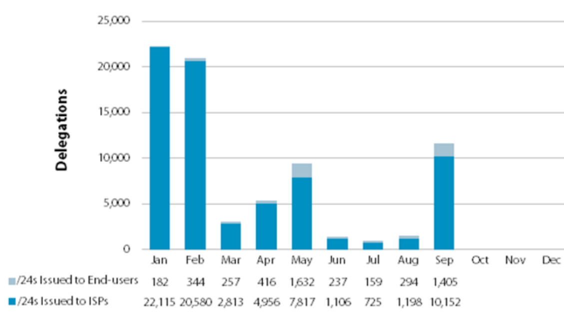
IANA owns everything, assigns blocks to Regional Internet Registries (RIR), who assign to ISPs/users
e.g., ARIN = American Registry for Internet Numbers)



http://www.arin.net/statistics/index.html

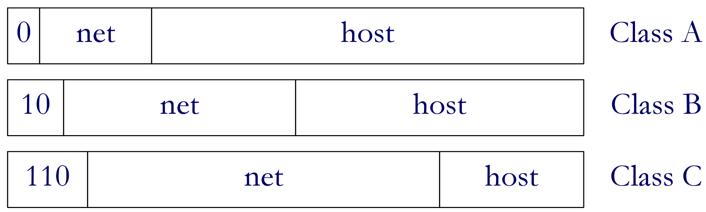
m9.35

Example (cont.)



m9.36

Old-style IP Address Classes

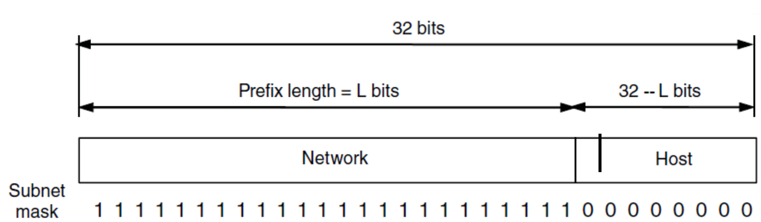


- Network mask defined as part of the address
  - Three sizes, class A, B and C, for different size networks.
- It’s easy to extract network number given the full IP
- Look up network number in routing table

Linternet.37

Scaling with IP prefixes - CIDR

- Route to blocks of addresses called “prefixes”
  - Written as IP prefix “x.x.x.x/length” for $2^{(32\text{-length})}$ addresses
  - Replaces old fixed blocks of lengths 8, 16 and 24
  - Only store one entry for a prefix in the routing table



38

IP Forwarding -- Longest Matching Prefix

Destination	Gateway
Default (0/0)	192.168.1.1
192.168.1.0/24	Link #4

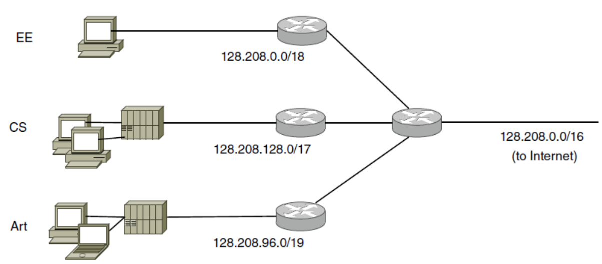
My PC's routing table (netstat –r)

- Can't tell from an address which prefix it belongs to, so match on the longest prefix for forwarding
  - Routers in the Internet may have 100s of 1000s of prefixes
- Example:
  - Send a packet to my printer (192.168.1.254)
  - Send a packet to cnn.com (157.166.224.25)

Subnetting

- Can internally divide a prefix
  - Better manageability and efficient allocation

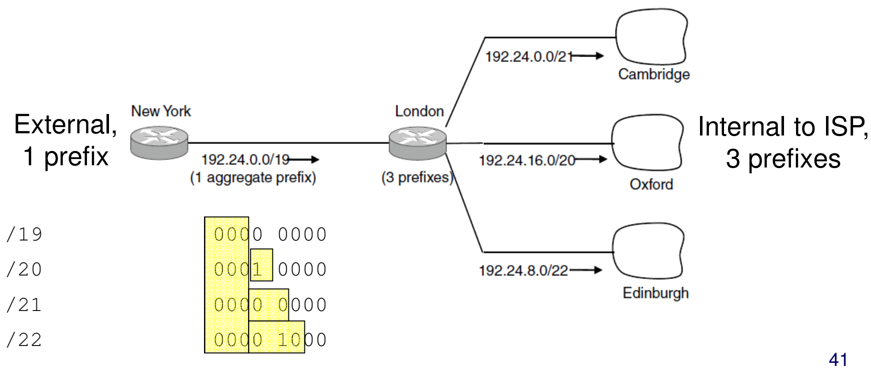
Internal, 3 prefixes



External, 1 prefix

Aggregation (CIDR, supernetting)

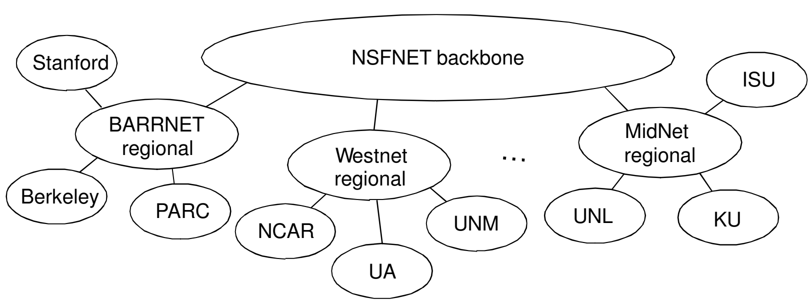
- Can externally combine prefixes
  - Same mechanism, different goal -- smaller routing tables
  - Would reduce table size by up to 40% if use was widespread!



41

Original Structure of the Internet

•Like address assignment: hierarchical

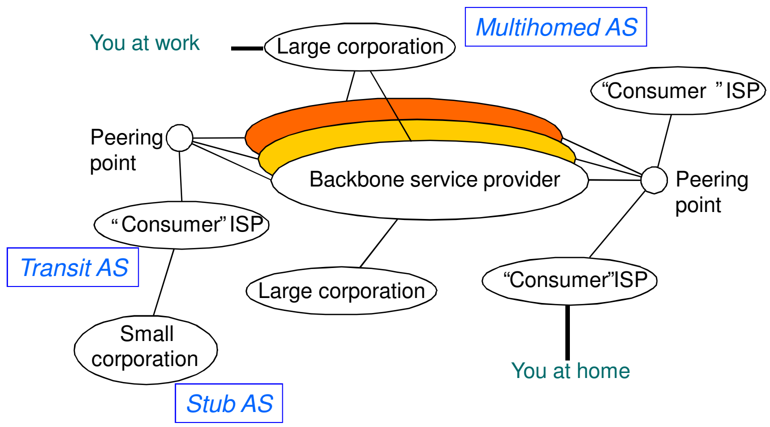


•What's "wrong" with this?

m9.42

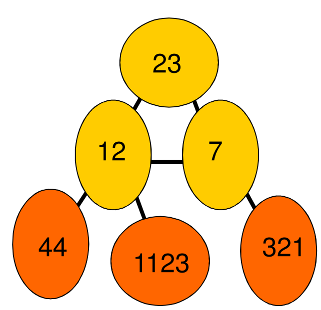
Current Structure

•Inter-domain versus intra-domain routing



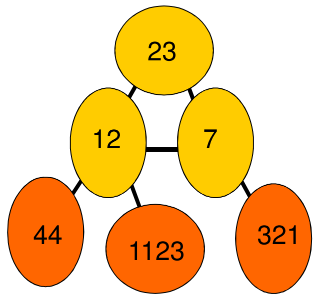
Scaling with ASes

- Network comprised of many Autonomous Systems (ASes) or domains
- To scale, use hierarchy to separate inter-domain (BGP) and intra-domain (OSPF) routing



Path Vectors

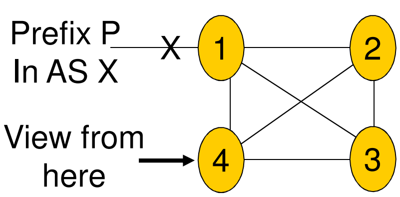
- Similar to distance vector, except send entire paths
  - e.g. 321 hears [7,12,44]
  - stronger avoidance of loops
  - supports policies (later)
- Modulo policy, shorter paths are chosen in preference to longer ones
- Reachability only – no metrics



Linterdomain.45

An Ironic Twist on Convergence

- Recently, it was realized that BGP convergence can undergo a process analogous to count-to-infinity!

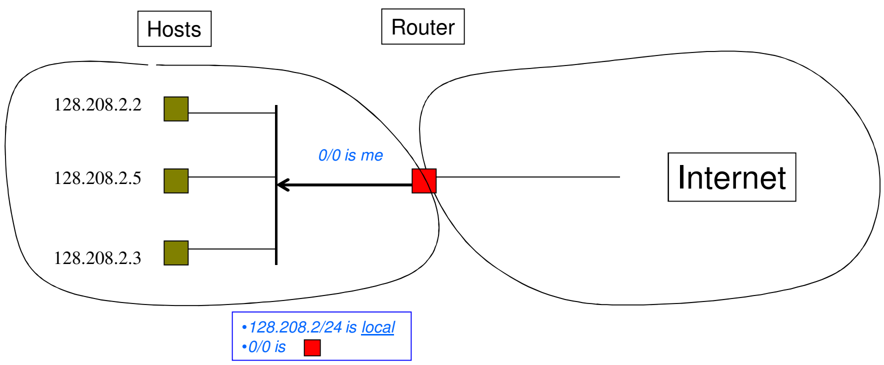


- AS 4 uses path 4 1 X. A link fails and 1 withdraws 4 1 X.
- So 4 uses 4 2 1 X, which is soon withdrawn, then 4 3 2 1 X, …
- Result is many invalid paths can be explored before convergence

Linterdomain.46

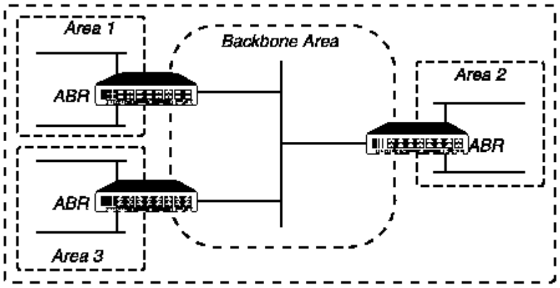
Applying Hierarchy to "ASes"

•We've already seen an example: host gateways



m9.47

Generalizing: Routing Areas



- Routers within an area (only) exchange full link state information
  - Limit cost of link state traffic / computation
  - (Different areas could have different cost metrics)
- Area border routers (ABRs) summarize area to other ABRs
- ABRs summarize rest of world to an area
- (Areas can have more than one ABR.)

m9.48

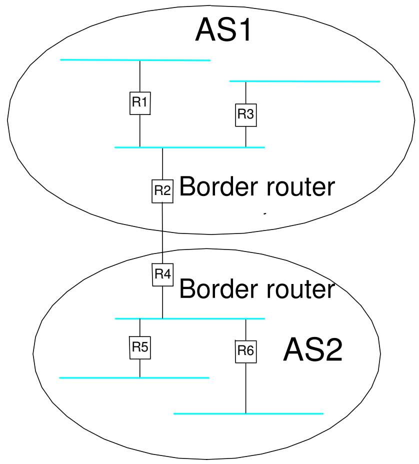
Inter-Domain Routing

• Border routers summarize and advertise internal routes to external neighbors and vice-versa

• Border routers apply <u>policy</u>

• Internal routers can use notion of default routes

• Core is "default-free"; routers must have a route to all networks in the world



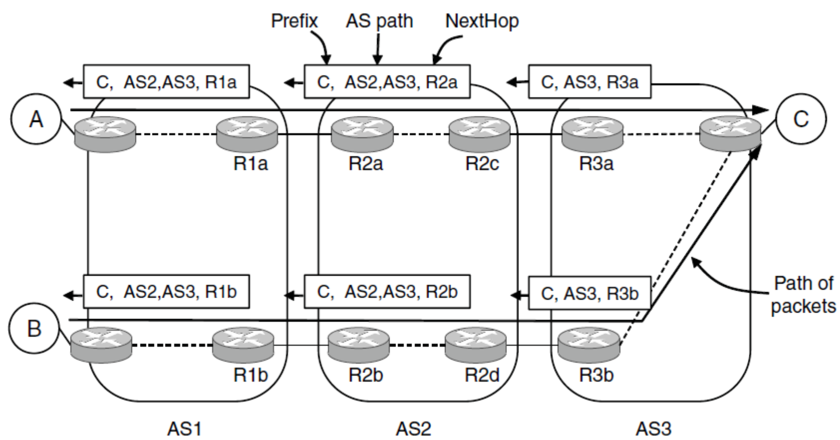
m9.49

BGP

- Interdomain routing protocol of the Internet
- Each AS tells other ASes the paths it is offering
  - Paths are summaries to prefixes via the sequence of ASes
  - No detailed paths of cost metrics to particular IPs
  - This happens at each border router of the AS
- Each AS picks the paths it wants to use to send traffic
  - Default rule: prefer shortest AS path, then shortest internal path
  - But selection heavily customized by ISPs
  - This happens at each border router of the AS

50

BGP



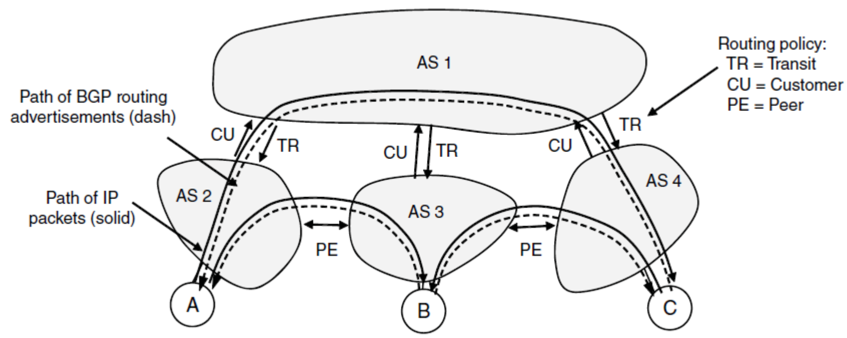
51

Policies

- Each ISP decides which routes to advertise, which to use
  - Choice of routes may depend on owner, cost, AUP, …
- Example: providers sell Transit to their customers
  - Customer announces their prefixes to provider for the rest of the Internet to reach them; Provider announces all other prefixes to customer for them to reach the rest of the Internet
- Example: parties Peer for mutual benefit
  - Peers announce path to their customer's prefixes to each other but do not propagate announcements further

52

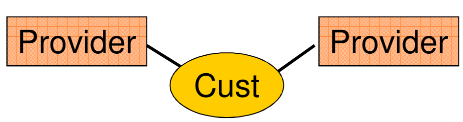
Policies



- Q: What routing do A, B, and C need to do?

Multi-Homing

- Connect to multiple providers for reliability, load sharing



- Choose the best outgoing path to P out of any of the announcements to P that we hear from our providers
  - Easy to control outgoing traffic, e.g, for load balancing
- Advertise the possible routes to P to our providers
  - Less control over what paths other parties will use to reach us

Brief Foray Into Security Issues

- Movie break:
  - http://opennet.net/youtube-censored-a-recent-history

Linterdomain.55

Prefix Hijacking

http://arstechnica.com/old/content/2008/02/insecure-routing-redirects-youtube-to-pakistan.ars

Insecure routing redirects YouTube to Pakistan
By Iljitsch van Beijnum | Last updated February 25, 2008 3:31 AM

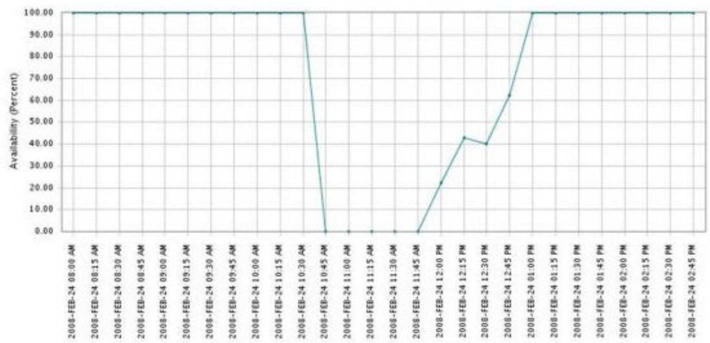
On Sunday, YouTube became unreachable from most, if not all, of the Internet. No "sorry we're down" or cutesy kitten-with-screwdriver page, nothing. What happened was that packets sent to YouTube were flowing to Pakistan. Which was curious, because the Pakistan government had just instituted a ban on the popular video sharing site. What apparently happened is that Pakistan Telecom routed the address block that YouTube's servers are into a "black hole" as a simple measure to filter access to the service. However, this routing information escaped from Pakistan Telecom to its ISP PCCW in Hong Kong, which propagated the route to the rest of the world

In the case of YouTube and Pakistan Telecom, YouTube injected the address block 208.65.152.0/22 in the Internet's routing tables, while Pakistan Telecom advertised the 208.65.153.0/24 block. So even though YouTube's routing information was still there, packets would flow towards Pakistan Telecom because of the longest match first rule.

m9.56

Prefix Hijacking

http://news.cnet.com/8301-10784_3-9878655-7.html



This graph that network-monitoring firm Keynote Systems provided to us shows the worldwide availability of YouTube.com dropping dramatically from 100 percent to 0 percent for over an hour. It didn't recover completely until two hours had elapsed.
(Credit: Keynote Systems)

m9.57

Another security issue

Web surfing break:

http://www.iana.org/abuse

m9.58

Bogons

Possible Bogus Routes and AS Announcements

Possible Bogus Routes

Prefix	Origin AS	AS Description	Unallocated block
1.0.0.0/8	AS237	MERIT-AS-14 - Merit Network Inc.	1.0.1.0 - 1.1.0.255
2.0.0.0/16	AS12654	RIPE-NCC-RIS-AS RIPE NCC RIS project	2.0.0.0 - 2.255.255.255
2.1.0.0/21	AS12654	RIPE-NCC-RIS-AS RIPE NCC RIS project	2.0.0.0 - 2.255.255.255
2.1.24.0/24	AS12654	RIPE-NCC-RIS-AS RIPE NCC RIS project	2.0.0.0 - 2.255.255.255
2.2.2.0/24	AS12654	RIPE-NCC-RIS-AS RIPE NCC RIS project	2.0.0.0 - 2.255.255.255
41.77.236.0/22	AS5.8		41.77.232.0 - 41.77.239.255
41.190.64.0/22	AS28683	OPT-NTIC-AS Office des Postes et telecommunications du Benin	41.190.64.0 - 41.190.67.255
41.190.66.0/24	AS37039		41.190.64.0 - 41.190.67.255
41.202.96.0/19	AS29571	CITelecom-AS	41.202.96.0 - 41.202.127.255
41.216.32.0/19	AS28683	OPT-NTIC-AS Office des Postes et telecommunications du Benin	41.216.32.0 - 41.216.63.255
41.220.144.0/20	AS36918	OTAVSAT-AS ORASCOM TELECOM ALGERIE VSAT	41.220.144.0 - 41.220.159.255
41.220.159.0/24	AS36918	OTAVSAT-AS ORASCOM TELECOM ALGERIE VSAT	41.220.144.0 - 41.220.159.255
41.222.79.0/24	AS36938	AMSCOTELECOMS Amsco Telecommunications Nigeria Limited	41.222.72.0 - 41.222.79.255
41.223.24.0/22	AS25747	VSC-SATELLITE-CO - VSC Satellite Co.	41.223.24.0 - 41.223.27.255
41.223.92.0/22	AS36936	CELTEL-GABON Celtel Gabon Internet Service	41.223.92.0 - 41.223.99.255
41.223.188.0/24	AS22351	INTELSAT Intelsat Global BGP Routing Policy	41.223.188.0 - 41.223.199.255
41.223.189.0/24	AS26452	BRING-AS - BringCom, Inc.	41.223.188.0 - 41.223.199.255
41.223.196.0/24	AS36990		41.223.188.0 - 41.223.199.255
41.223.197.0/24	AS36990		41.223.188.0 - 41.223.199.255
41.223.198.0/24	AS36990		41.223.188.0 - 41.223.199.255
41.223.199.0/24	AS36990		41.223.188.0 - 41.223.199.255
46.0.0.0/16	AS12654	RIPE-NCC-RIS-AS RIPE NCC RIS project	46.0.0.0 - 46.255.255.255
46.1.0.0/21	AS12654	RIPE-NCC-RIS-AS RIPE NCC RIS project	46.0.0.0 - 46.255.255.255
46.1.24.0/24	AS12654	RIPE-NCC-RIS-AS RIPE NCC RIS project	46.0.0.0 - 46.255.255.255

m9.59