Internet Routing

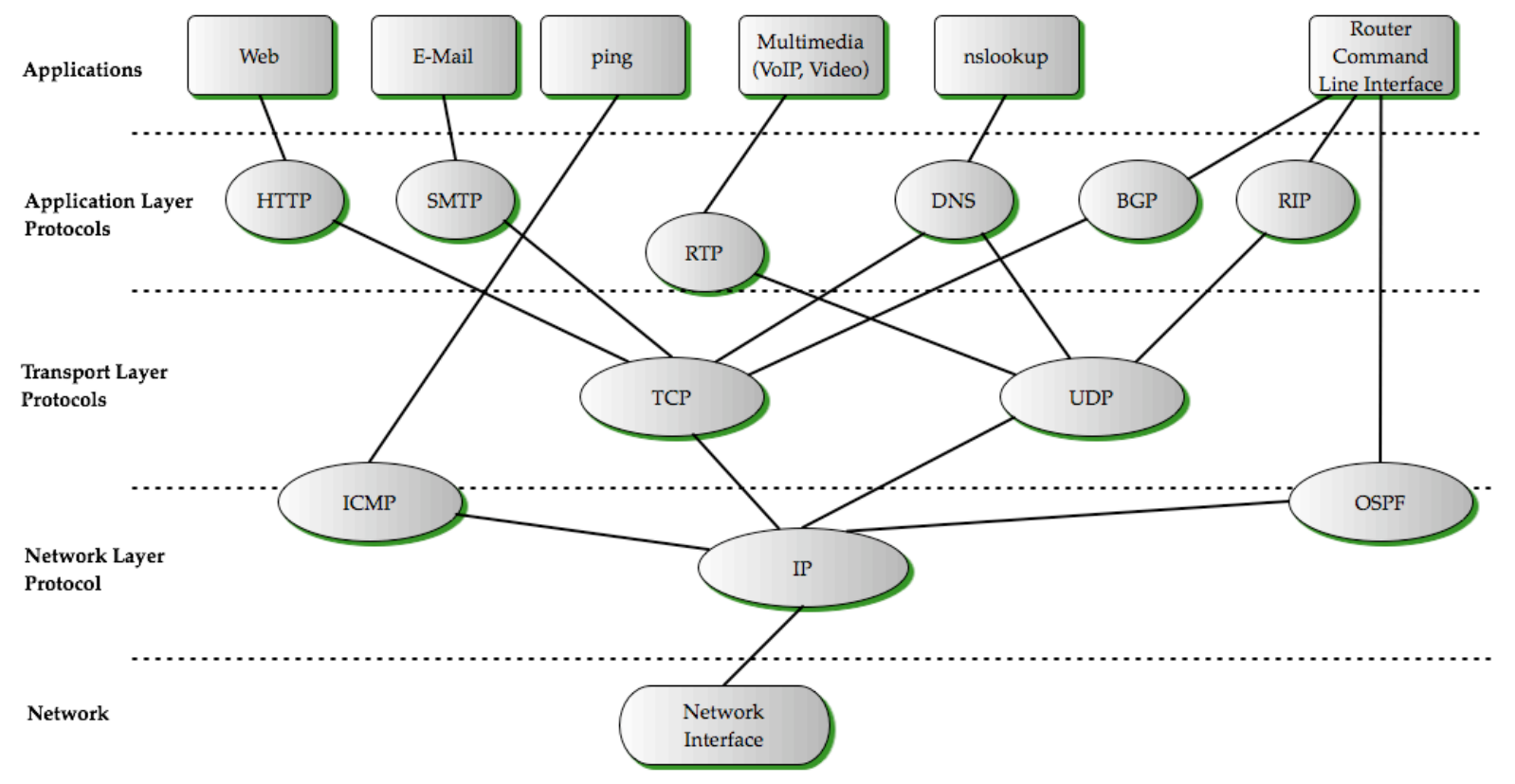
Deep Medhi
Computer Science & Electrical Engineering Department
University of Missouri-Kansas City, USA
http://www.csee.umkc.edu/~dmedhi
dmedhi@umkc.edu
SBRC'2012 Tutorial, April 2012

UMKC

Outline

- A bit about myself
- Internet packet delivery: overview
- IP Addressing: refresher
- Routing within a domain
- Address block assignment
- Intra-domain routing protocols
- IP traffic engineering
- Autonomous Systems
- AS and ISPs
  - Tiering and Peering
- Policy-based Routing
- Multi-homing example
- BGP table growth/issues
- PSTN-IP interworking

Protocol Layering in TCP/IP stack (including for routing info exchange)

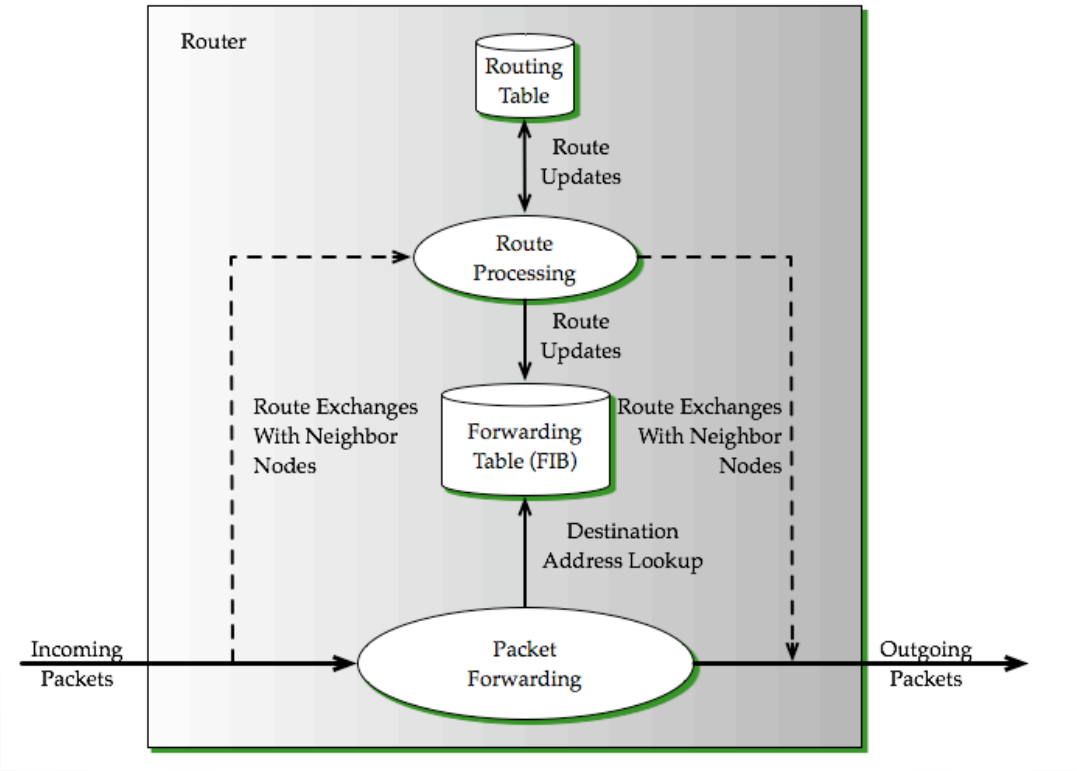


Routing/forwarding: high-level view

- Determine destination address
  - Addressing matters!
- From routing table, determine next hop
- Forward efficiently
- Minimize loss

  - Router implementation: efficiency, lookup
  - Depends also on Network Traffic Engineering

Router: A high-level architecture



IP Address: a quick overview

- Currently: IPv4 (version 4)
  - 32-bit address
  - listed as 4-octet-dotted-decimal

IP address:	192	.	168	.	40	.	3
binary form:	11000000		10101000		00101000		00000011

- Identifies a host/net (<u>not</u> applications)

Netmask Example: IP prefix

11000000 10101000 00101000 00000011	192.168.40.3 (dest. host)
11111111 11111111 11111000 00000000	netmask (/21)
----------	bitwise "AND"
11000000 10101000 00101000 00000000	192.168.40.0 (net-addr)

Network address is: 192.168.40.0/21

in CIDR notation

contiguous address block: **192.168.40.0** to **192.168.47.255**

Address block assignment

- Organizations are assigned *contiguous* address blocks (along with a network mask)
  - UMKC: 134.193.0.0/16, Net-Mask: 16-bits
  - Need to take care of intra-domain routing
  - In case of multiple routers, it is up to the domain to divide its address space into sub-networks "attached" to each router
- In general, block boundary need not be /16 for assignments
  - Classless Interdomain routing (CIDR), i.e., Can be /21 etc
- Address block commonly referred to as IP prefix

On address assignment

- Advantage of assigning contiguous address blocks to an organization:
  - Outside world doesn't need routing table entry for *each and every* subnet (especially hosts) that resides within the organization

Address block assignment

- Five Regional Internet Registries
  - ARIN, RIPE, APNIC, LACNIC, AfriNIC
    (ARIN.net, APNIC.net, RIPE.net, AfriNIC.net, LACNIC.net)
- National Internet Registries: mostly in Asai-Pacific region
- Any "entity" can request IP address block
  - Slow-Start Policy (RFC 2050)
  - There are some restrictions (each has different restrictions)

Allocation Policy: Examples

- ARIN (http://www.arin.net/policy/nrpm.html)" : **4.2.1.5. Minimum allocation**

  "In general, ARIN allocates /20 and larger IP address prefixes to ISPs. If allocations smaller than /20 are needed, ISPs should request address space from their upstream provider. For multihomed ISPs, ARIN allocates /22 and larger IP address prefixes. If allocations smaller than /22 are needed, multihomed ISPs should request address space from their upstream provider."

- LACNIC (http://www.lacnic.net/en/politicas/)
  - 2.3.3.1 : "The minimum initial allocation size applicable to Internet Service Providers established within LACNIC's service region is a /22.."

Intra-domain/Inter-domain

- Inter-domain Routing
  - Protocol:
    - Usually, Border Gateway Protocol (especially at the "core")
    - Other means possible (e.g., route redistribution)
- Intra-domain Routing
  - Protocols:
    - IS-IS, OSPF, IGRP, EIGRP, RIP
      - (all provide also support for "inter-domain" routing for route redistribution)

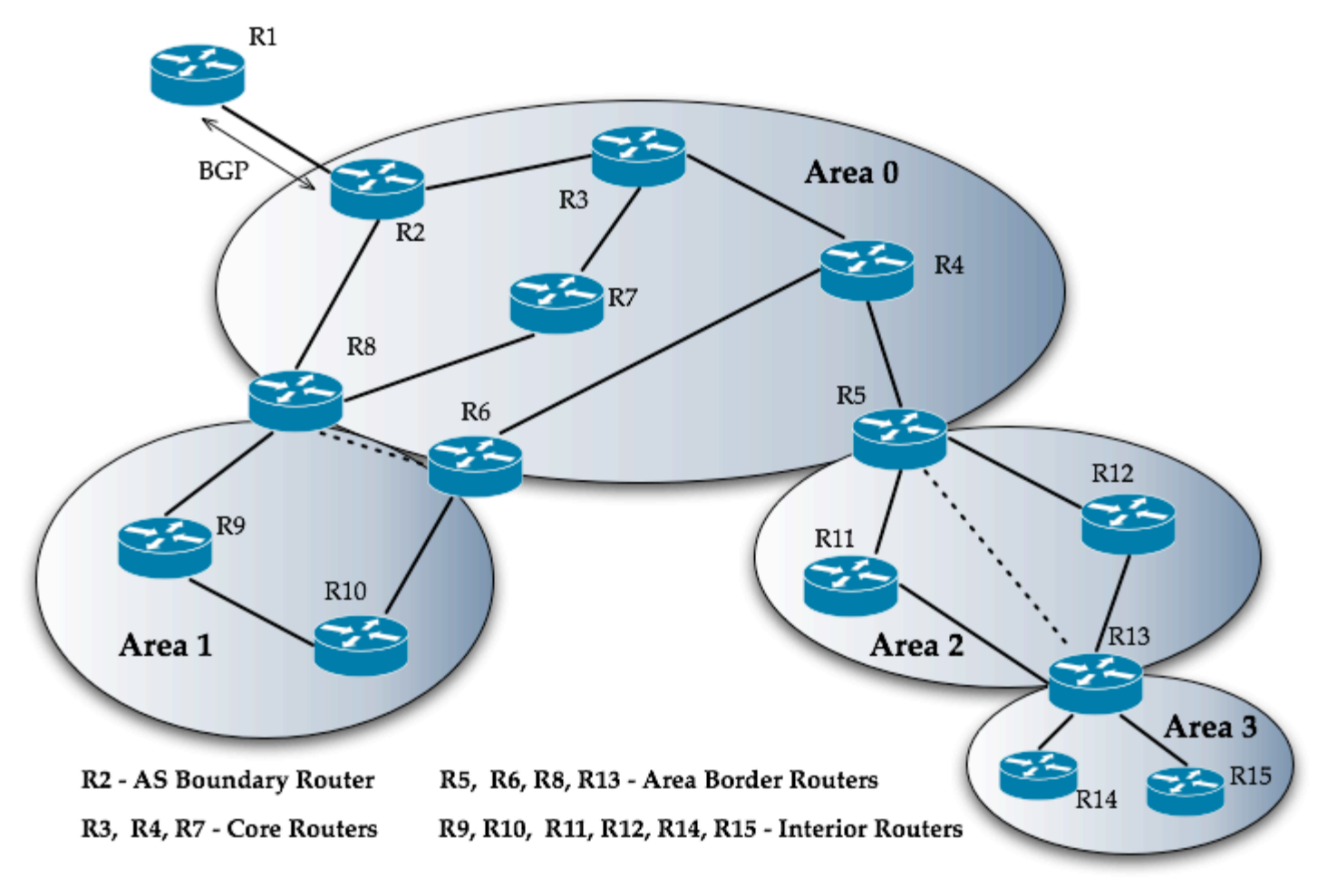
RIP, IGRP, EIGRP

- RIP: old protocol (distance vector protocol)
  - Use limited to small networks
- IGRP: still a distance vector protocol
  - Cisco's extension of RIP
- EIGRP: a loop-free (enhanced) distance vector protocol
  - Based on diffusing computation
- All three usually for small networks

OSPF

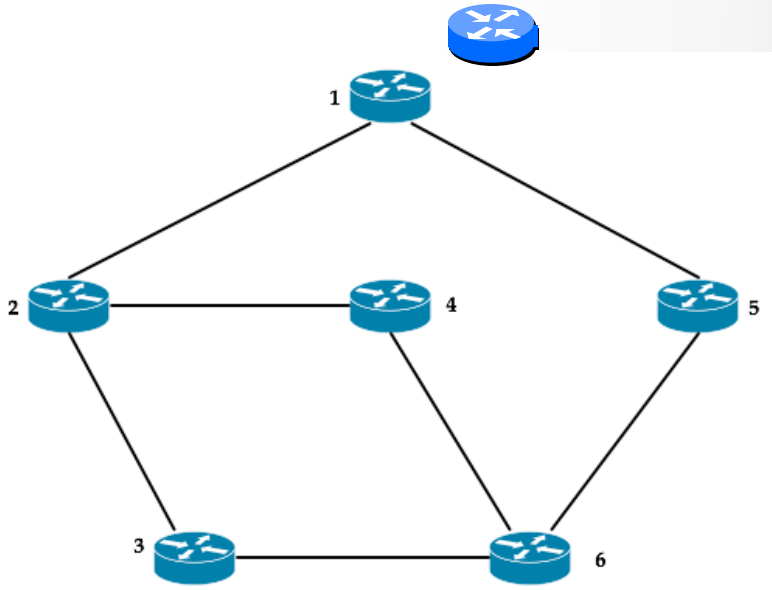
- Popularly deployed in large ISPs
- Uses link state protocol
  - For flooding link cost information
    - Link state advertisement, Link state update
- Dijkstra's shortest path first algorithm used for path determination (next-hop stored)
- Allows: Equal-cost multi-path
- Multiple "link" types defined
  - E.g., "LAN" link
- Link cost: through metric values
  - OSPF: 1 to 2^{16}-1

OSPF area


R1
BGP
R2
R3
Area 0
R4
R7
R8
R6
R5
R9
R10
Area 1
R11
R12
Area 2
R13
Area 3
R14
R15
R2 - AS Boundary Router
R5, R6, R8, R13 - Area Border Routers
R3, R4, R7 - Core Routers
R9, R10, R11, R12, R14, R15 - Interior Routers

Equal-Cost Multi-Path (ECMP)

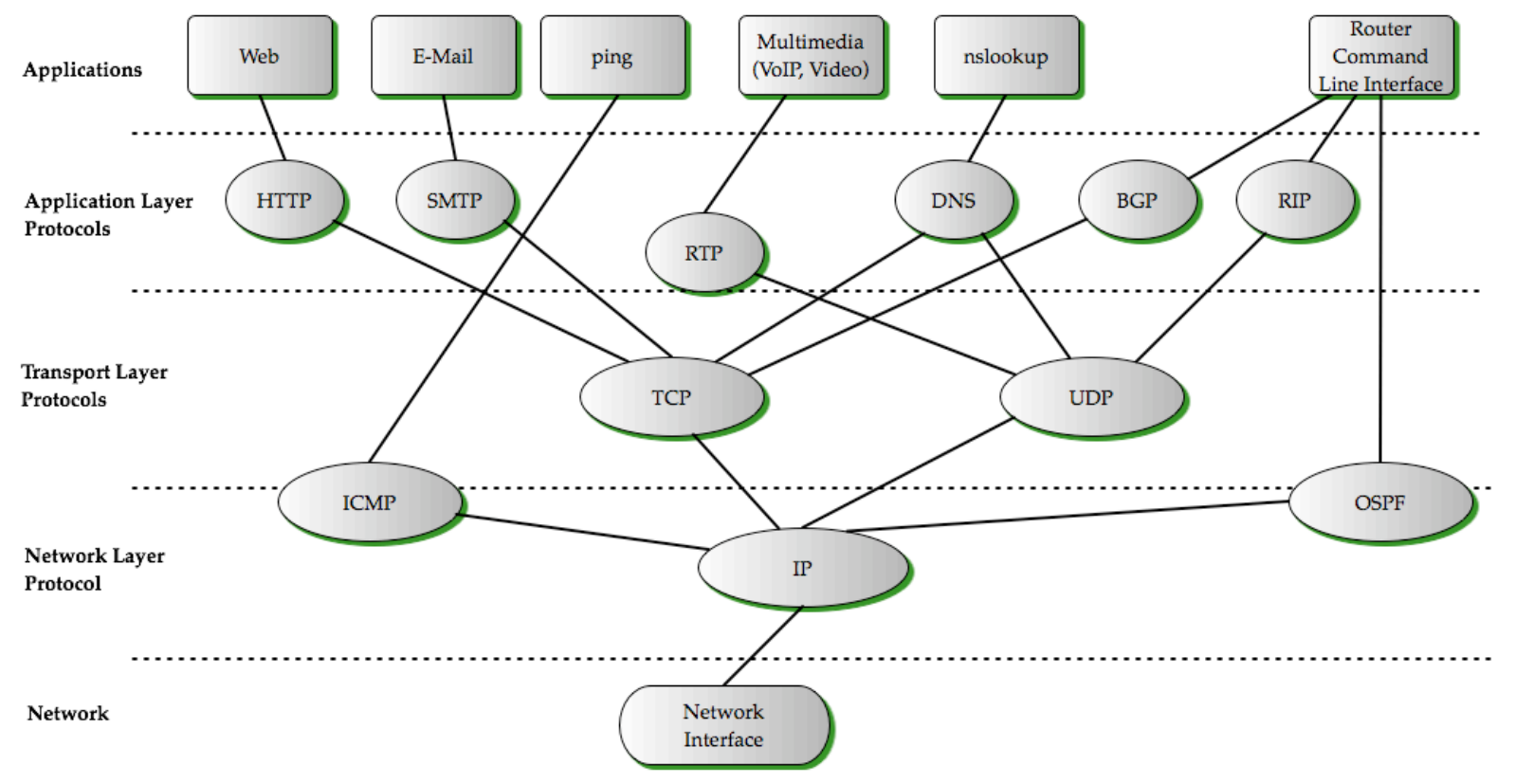
- Split is outgoing link based, NOT path based!
- From 1 to 6:
  - 50% each on 1-2, 1-5
  - 50% split at 2 on 2-3, 2-4
- From 6 to 1:
  - 33% each on links 6-3, 6-4, 6-5
  - At node 2, traffic is combined
- It is approximate
  - Based on microflow



IS-IS

- Came from the OSI world
- It uses the concept of Address Family identifier
  - Thus, IP addressing can be captured as well ("Integrated IS-IS")
- Terminologies are quite different from OSPF
- Many large ISPs deploy IS-IS
  - Stable implementation
- Link cost: through metric values (different than OSPF!)
  - Narrow metric (original): 0 to 2^6-1 (=63)
  - Wide metric (later added): 0 to 2^{24}-1

Routing protocol stack: command line interface



Routing Management encompases

- **Traffic Monitoring and Traffic Engineering System**
- **IP Prefix Management**
- Policy Management

With varying degree depending on the size of the provider and role of the provider (edge or transit or core)

IP traffic engineering

- Issue for large and medium ISPs
  - How to engineer given the traffic volume
- Need to live with the shortest path routing paradigm
  - Link-weight setting problem
    - Requires traffic matrix determination
      - A non-trivial problem in IP world

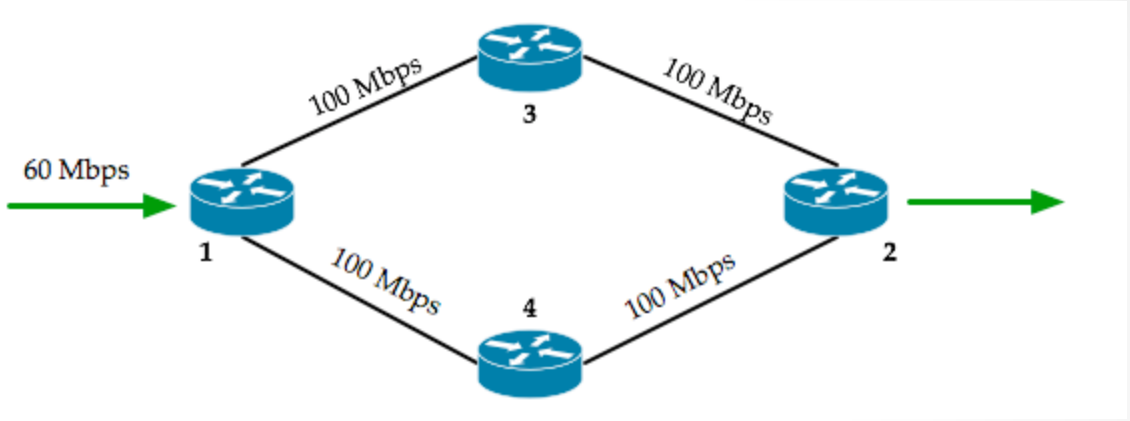
For reliable delivery services:

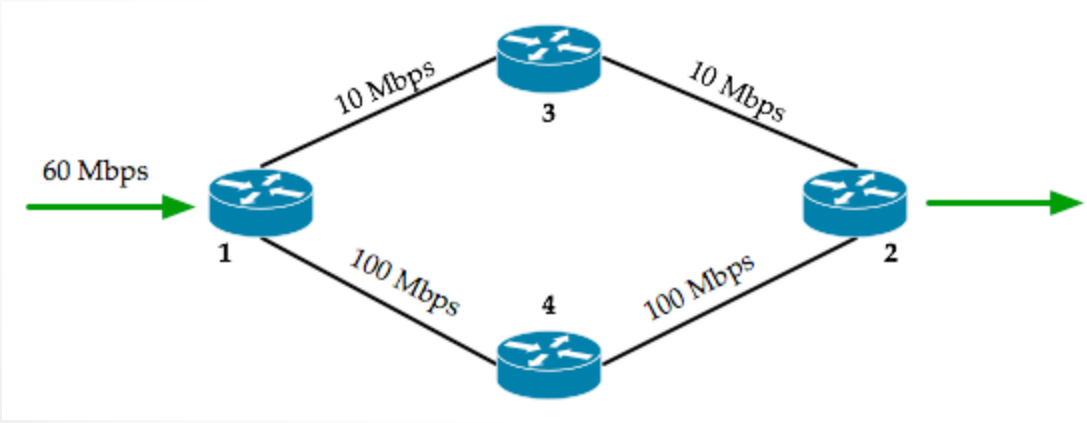
- Minimize loss, minimize round-trip time for Web/email services (which are based on TCP):

$$\text{TCP Throughput} = \frac{1.22\,S}{RTT \times \sqrt{q}}.$$

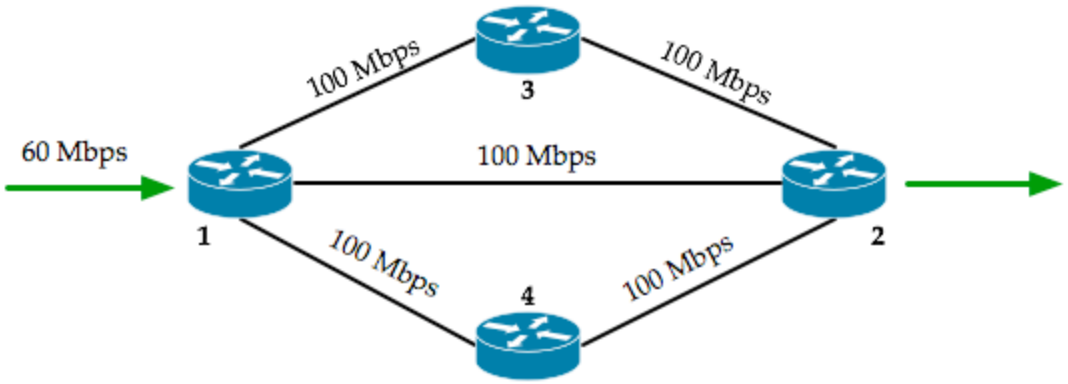
  - RTT := Round-trip time
  - q := packet loss probability
  - S := Maximum segment size (e.g., Ethernet frame limited)

IP Traffic Engineering: Link weight setting problem

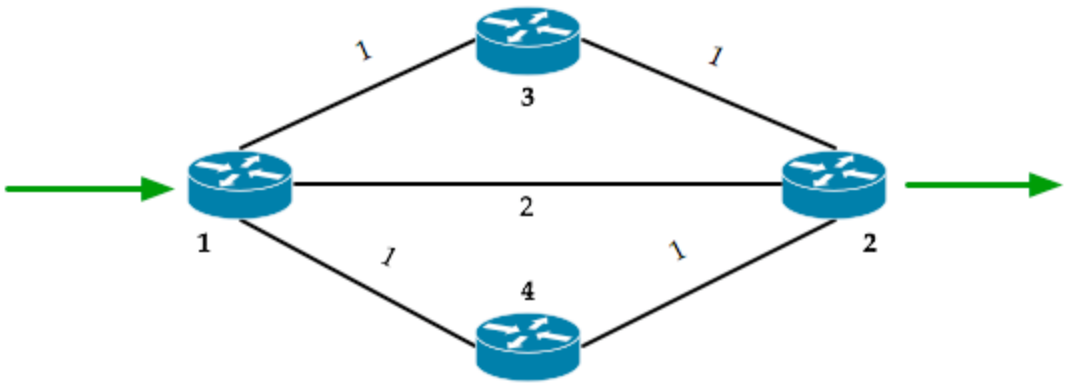




Add capacity (link) …

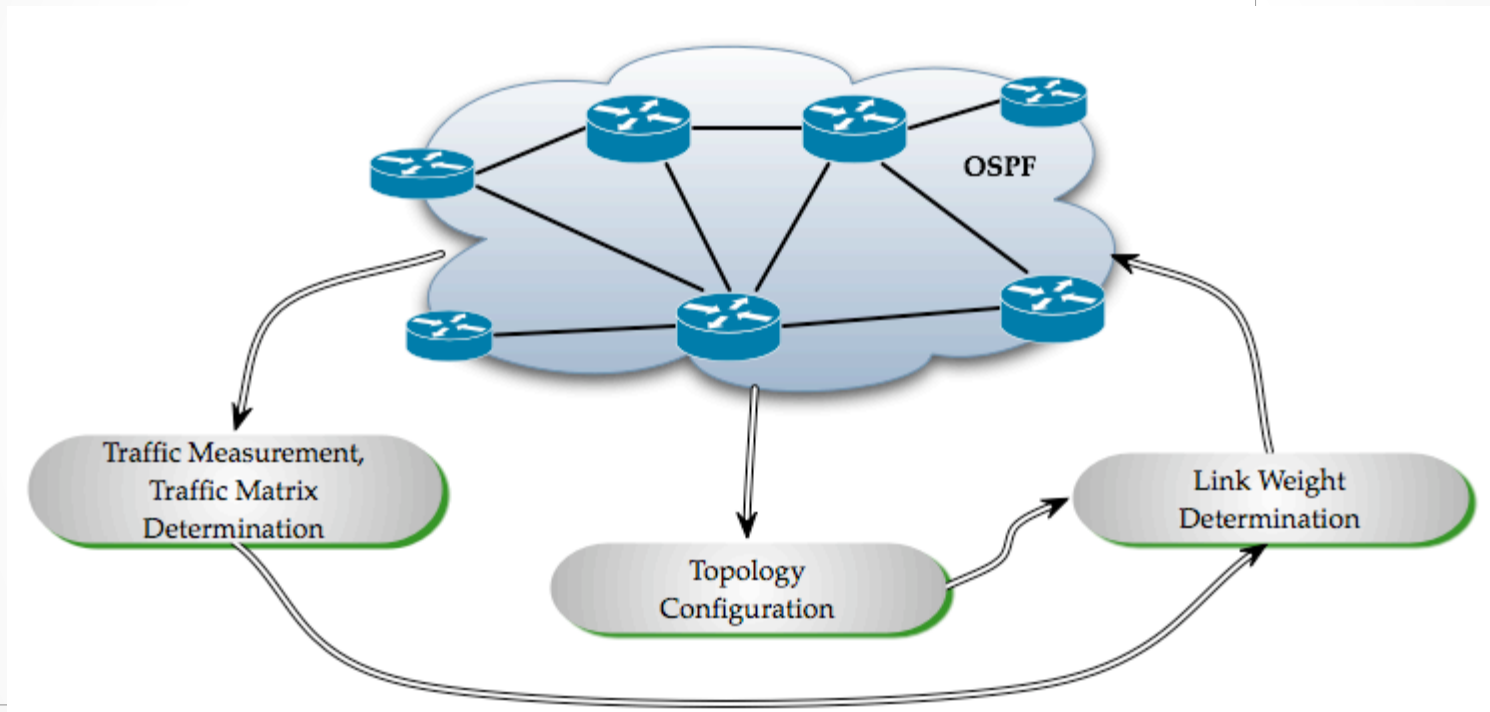


Link Weights …



- By picking weight carefully, ECMP can be invoked

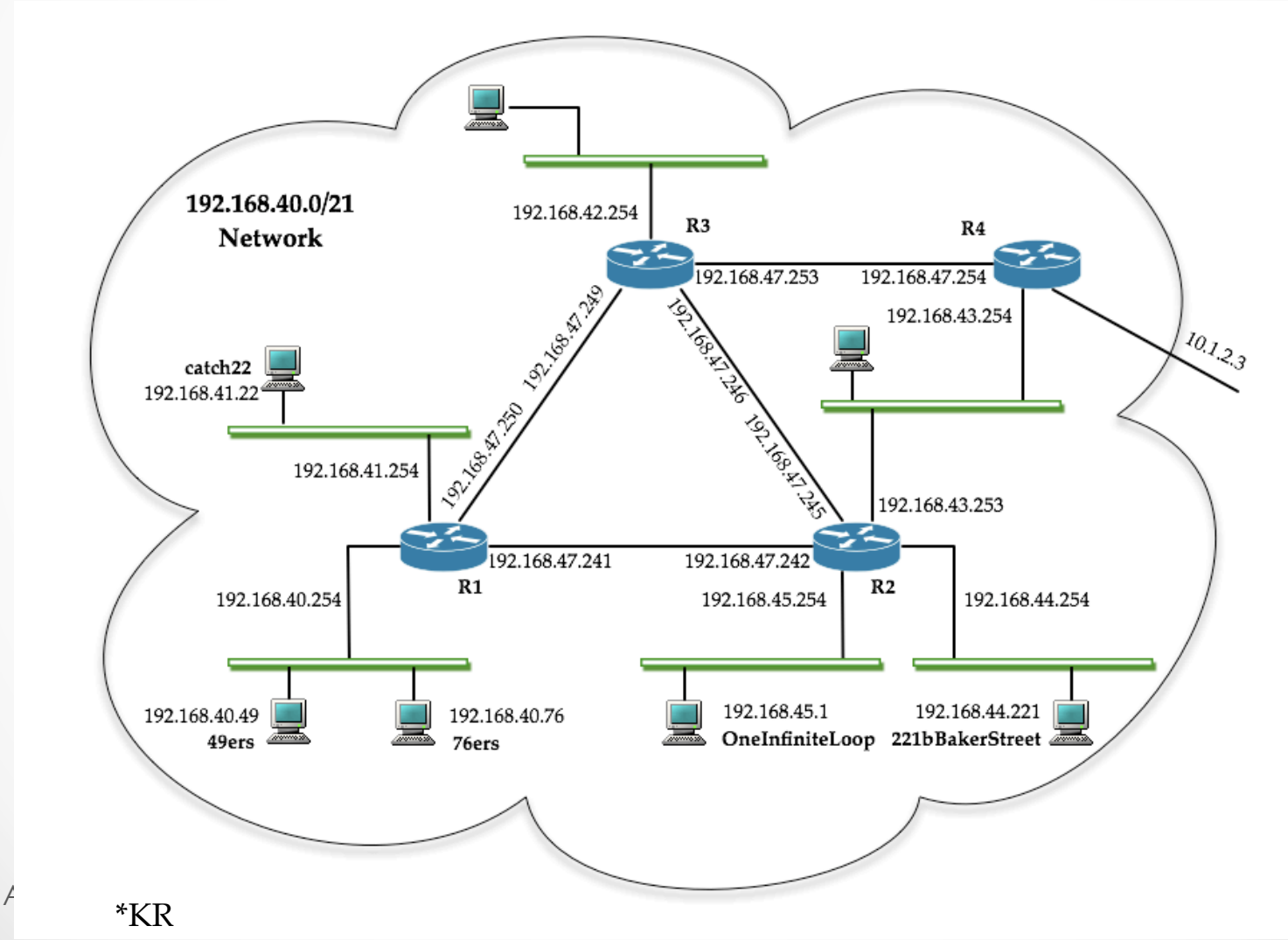
IP traffic engineering: operational view



IP/MPLS environment for ISPs

- Multi-Protocol Label Switching
  - A 2.5 layer solution
- "Virtual" links(Label-switched paths, LSP) can be set up to control behavior of traffic for specific customers
  - Need to address Path optimization problem for traffic engineering, minimize number of tunnels (from management point of view)

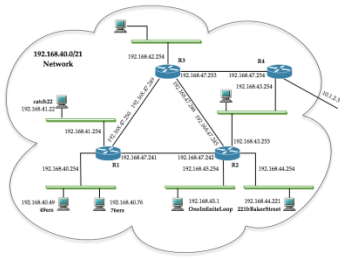
Routing within a domain: Network 192.168.40.0/21



Routing Table/forwarding info:

R1:

Net	mask	NextHop	Interface
192.168.40.0	255.255.255.0	direct	en0
192.168.41.0	255.255.255.0	direct	en1
192.168.42.0	255.255.255.0	192.168.47.249	sl2
192.168.43.0	255.255.255.0	192.168.47.242	sl1
192.168.44.0	255.255.255.0	192.168.47.242	sl1
192.168.45.0	255.255.255.0	192.168.47.242	sl1
192.168.47.240	255.255.255.252	direct	sl1
192.168.47.248	255.255.255.252	direct	sl2
0.0.0.0	0.0.0.0	192.168.47.242	sl1

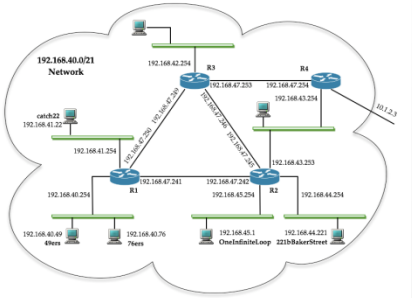


Border Router

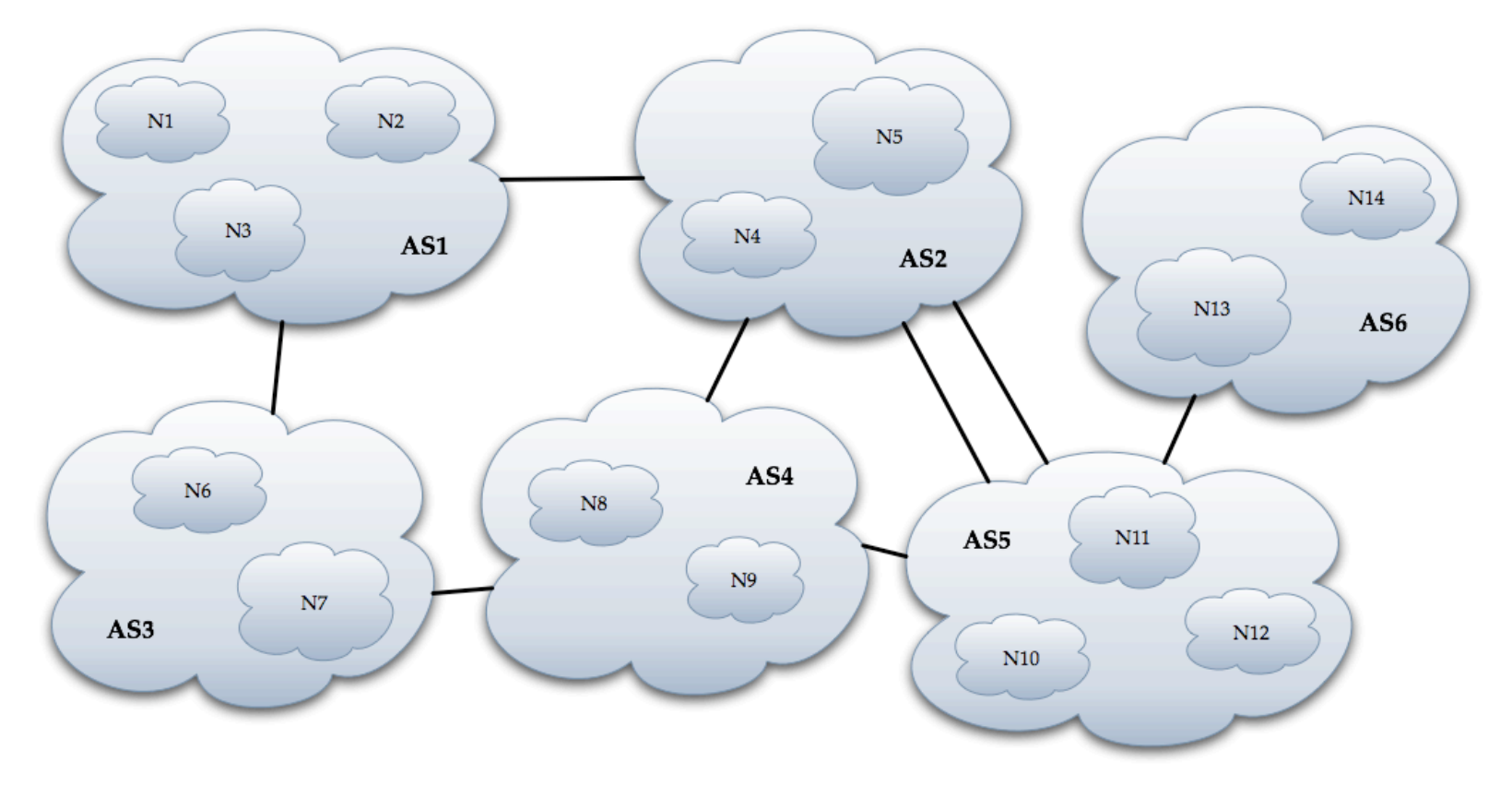
R4:

Net	mask	NextHop	Interface
192.168.40.0	255.255.255.0	192.168.43.253	en0
192.168.41.0	255.255.255.0	192.168.43.253	en0
192.168.42.0	255.255.255.0	192.168.47.253	sl0
192.168.43.0	255.255.255.0	direct	en0
192.168.44.0	255.255.255.0	192.168.43.253	en0
192.168.45.0	255.255.255.0	192.168.43.253	en0
192.168.47.252	255.255.255.252	direct	sl0
0.0.0.0	0.0.0.0	10.1.2.3	sl1

uRPF: protection for spoofing attacks (not shown)



Interconnected Autonomous Systems



N_i : IP prefix defined networks (e.g. 134.193.0.0/16)

BGP: basic operation

- Once session is set up (over TCP), two BGP speakers exchanges the following types of messages:
  - OPEN
    - ‘hello’; announce hold time
  - UPDATE
    - Exchange info about IP prefixes (“main function”)
  - KEEPALIVE
    - Periodic; set to about one third the value of hold time
      - But no more than once every second
  - NOTIFICATION
    - Close a session gracefully
  - ROUTE-REFRESH
    - A relatively newer message type: this is used for pulling information from your neighboring BGP speaker

Autonomous Systems

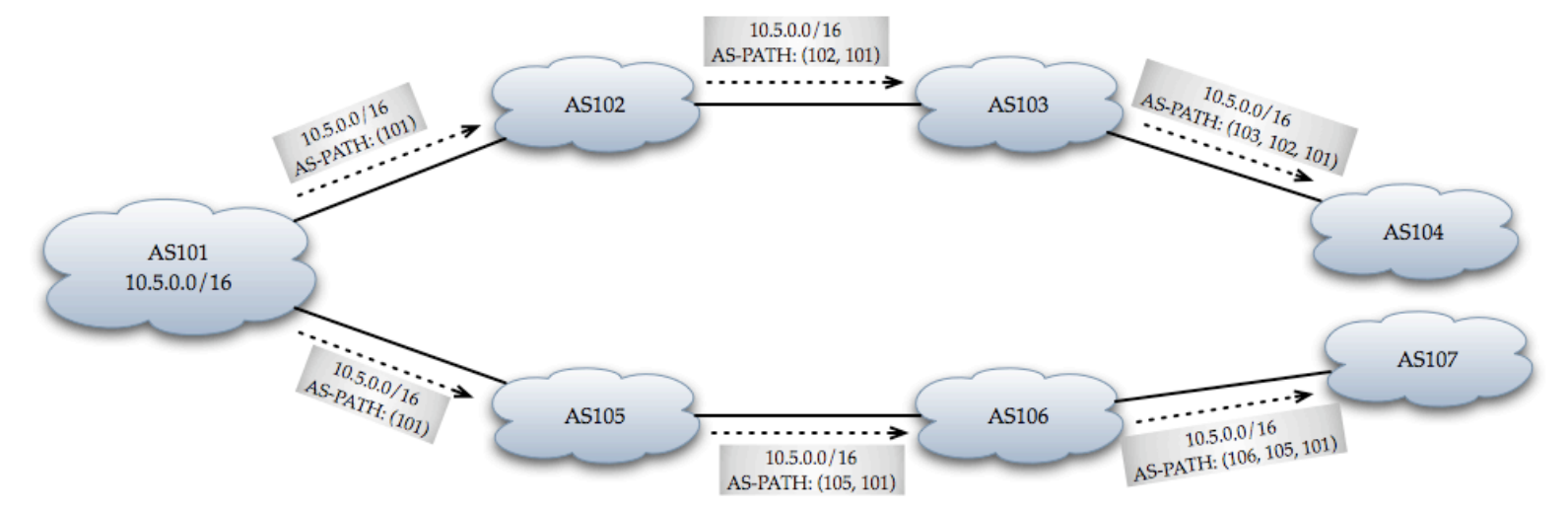
- Each Autonomous system has a unique 16-bit AS number (new extension to 32-bit)
  - UMKC: 3390
  - MOREnet: 2572
  - Sprint: 1239
- Think about Autonomous System numbers like the first two digits in postal ZIP code (66xxx, 64xxx)
- Every valid public address block (IP prefix) must be a member of just one AS at any time instance.

Goal: Tell rest of the world of your presence (i.e., address block/IP prefix)

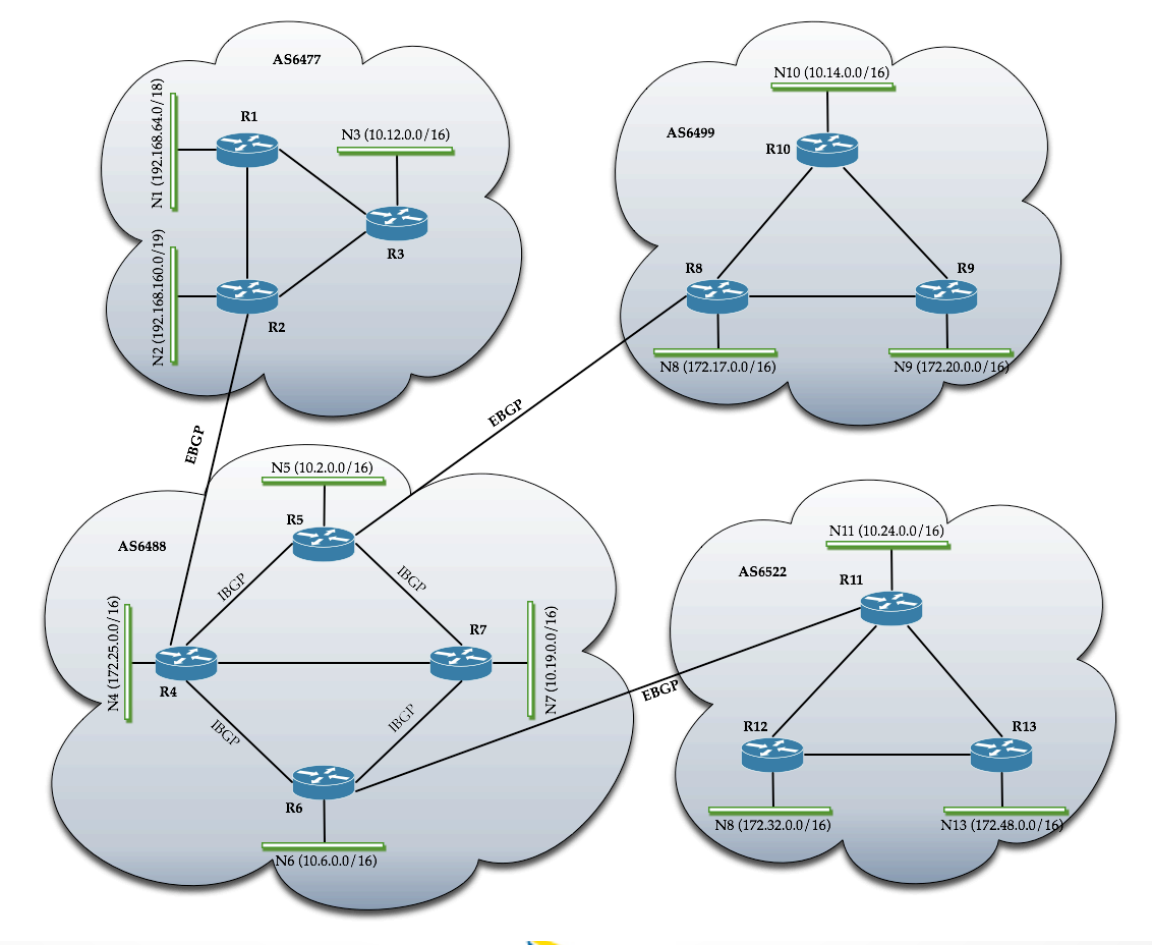
- You advertise your address block
  - My block is 134.193.0.0/16 and I am in AS # 3390
  - Next AS receives it, prepends its AS #, and announces downstream
    - 134.193.0.0/16: (2572, 3390) (MOREnet, UMKC)
  - Repeated:
    - 134.193.0.0/16: (1239, 2572, 3390) (Sprint, MOREnet, UMKC)
- Once the rest of the world knows, it knows how to reach you since each AS knows who to forward to
- Since AS may hear from two different ASes about a particular IP prefix, need to determine shortest AS path.
- BGP: Border Gateway Protocol
  - Reachability information
  - Withdrawal
- Border routers in ASes are called BGP speakers

Pre-pending AS-PATH

- Avoid loops!



Reachability, Interior-BGP, Exterior-BGP

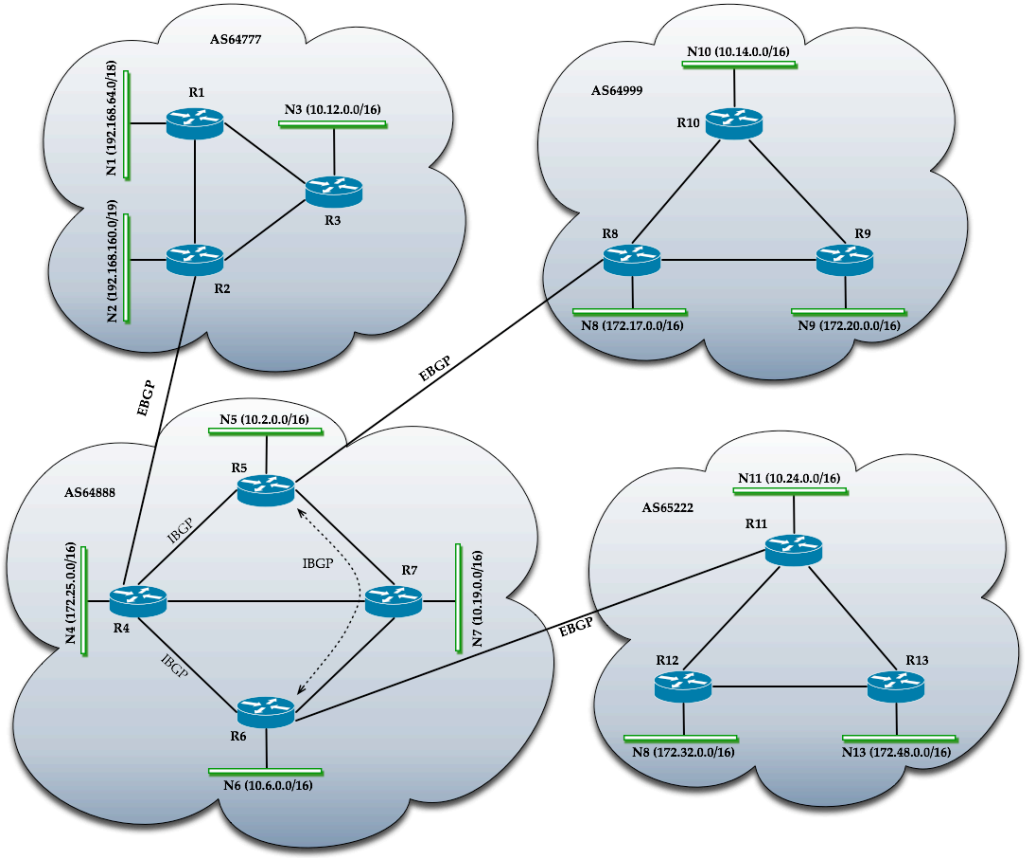


BGP: I-BGP and E-BGP

- For inter-BGP, it's E-BGP
- For intra-AS, do interior BGP (I-BGP)
- For inter-AS, do exterior BGP (E-BGP)

- Why I-BGP?
  - Inside an AS, need a way to communicate BGP information among different speakers.
- Two important rules:
  - Rule 1: A BGP speaker can advertise IP prefixes it has learned from an EBGP speaker to a neighboring IBGP speaker; similarly, a BGP speaker can advertise IP prefixes it has learned from an IBGP speaker to EBGP speaker
  - Rule 2: An IBGP speaker cannot advertise IP prefixes it has learned from an IBGP speaker to another neighboring IBGP speaker
    - (needed since looping is possible

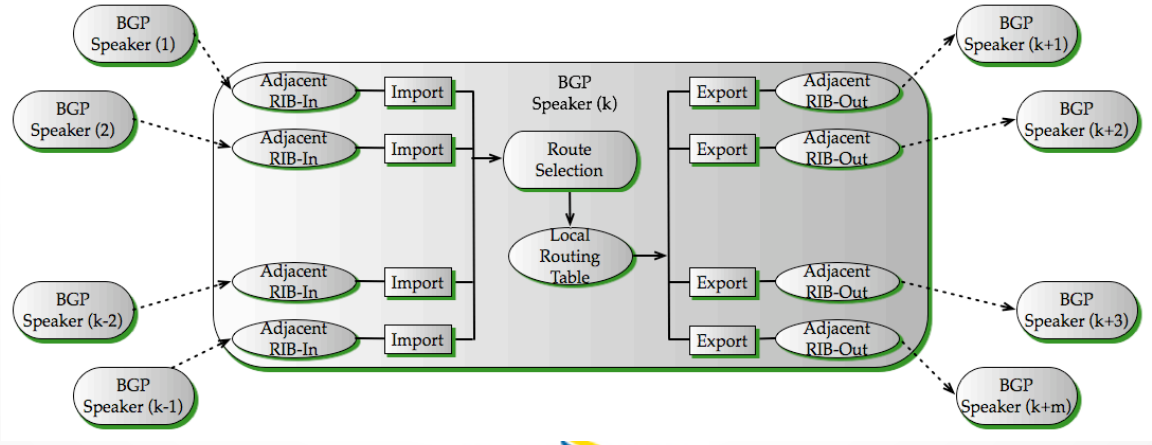
E-BGP and I-BGP



- R4 can learn about N11, N12, N13 from AS65222 from R6 and pass on to R2
- But R4 cannot inform R5 abou N11, N12, N13

BGP Path Selection Process

- Preferred order (key ones)
  - Local-pref, AS-number, Multi-Exit Discriminator
- Policy based
  - Input filter
  - Output filter



Policy Specification at a Border (BGP speaker) & Management

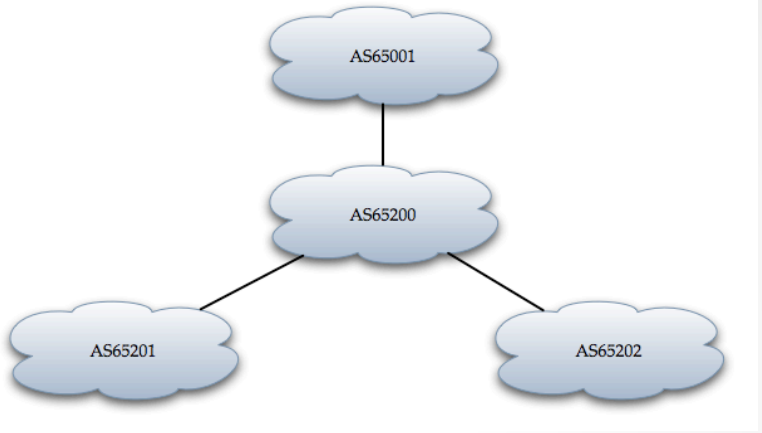
Example

Import Policy	Export Policy
– Do not accept default 0.0.0.0/0 from AS64617. – Assign 192.168.1.0/24 coming from AS64617 preference to receiving it from AS64816. – Accept all other IP prefixes.	– Do not propagate default route 0.0.0.0/0 except to internal peers. – Do not advertise 192.168.1.0/24 to AS64999. – Assign 172.22.8.0/24 a MED metric of 10 when sent to AS64999.

Policy-based routing at AS level

- RPSL
- Who am I going to allow

to transit through my network:



- Examples:

  import: from AS65001 accept ANY
  import: from AS65201 accept <^AS65201+$>
  import: from AS65202 accept <^AS65202+$>
  export: to AS65201 announce ANY
  export: to AS65202 announce ANY
  export: to AS65001 announce AS65200 AS65201 AS65202

- Can be more specific too, such as

  import: from AS65201 accept {134.193.0.0/16}

Autonomous System and ISPs

- How are they related?
  - Most large ISPs have an assigned Autonomous System number
- Are all ISPs same?
  - No (except for having AS number)
- What about content providers (such as Akamai, Google, Yahoo)? they have their own AS # too!
- In the business side of Internet, there are different sizes and shapes of ISPs
  - So you form business relations depending on which market slice you're targeting your ISP business to be
    - core provider, transit provider, access provider

Do you really need your own AS number to connect to the Internet?

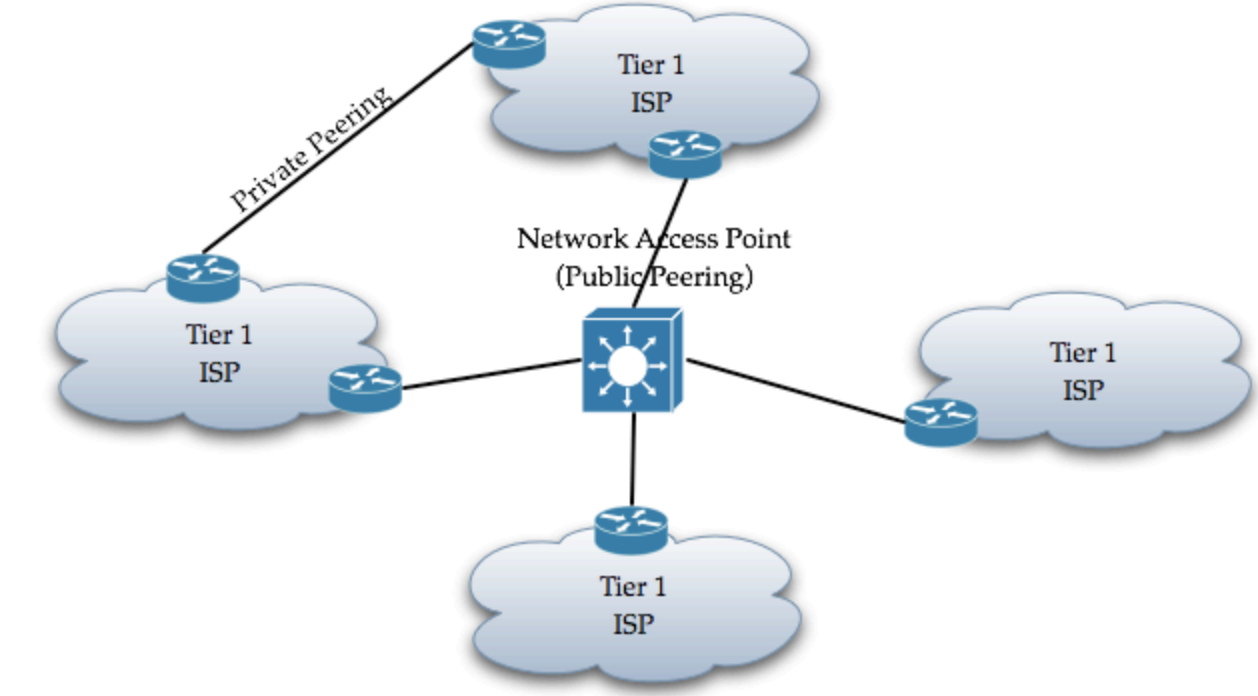
- The answer is no.
- You can have your own IP prefix (address block)
- You can subscribe to an ISP that has an AS number
  - This ISP will announce your IP prefix (along with the rest it has) to the rest of the Internet
  - You will set up a link connectivity with your ISP
    - Data rate dependent pricing
    - You can run your own routing protocol – need not be the same as your ISP
    - Use route redistribution (through RIPv2, IGRP)

ISP Tiers

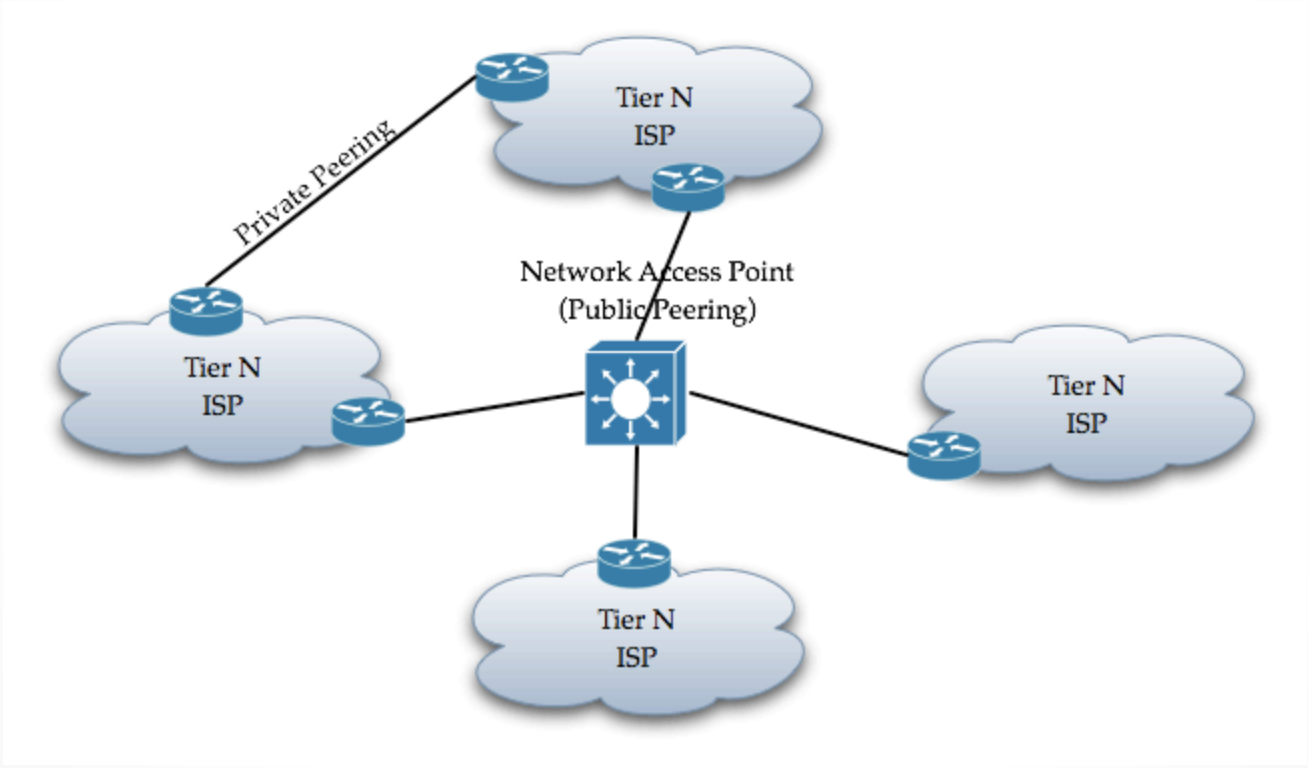
- Tier 1, Tier 2, Tier 3, …

- Tier 1: backbone ISPs (maintains default-free routing tables): Sprint, AT&T, Verizon (UUNet), …
  - They connect to each other at multiple locations ("peering")
- Tier 2: sort of transit providers
- Tier n: access providers for end users (DSL/Cable-Modem ISPs, Universities, Companies)
  - They may change their "provider" ; thus, the AS "home" may change – this needs to be announced then
  - They may multi-home to two or more providers

ISPs form peering relation: Public and/or Private Peering



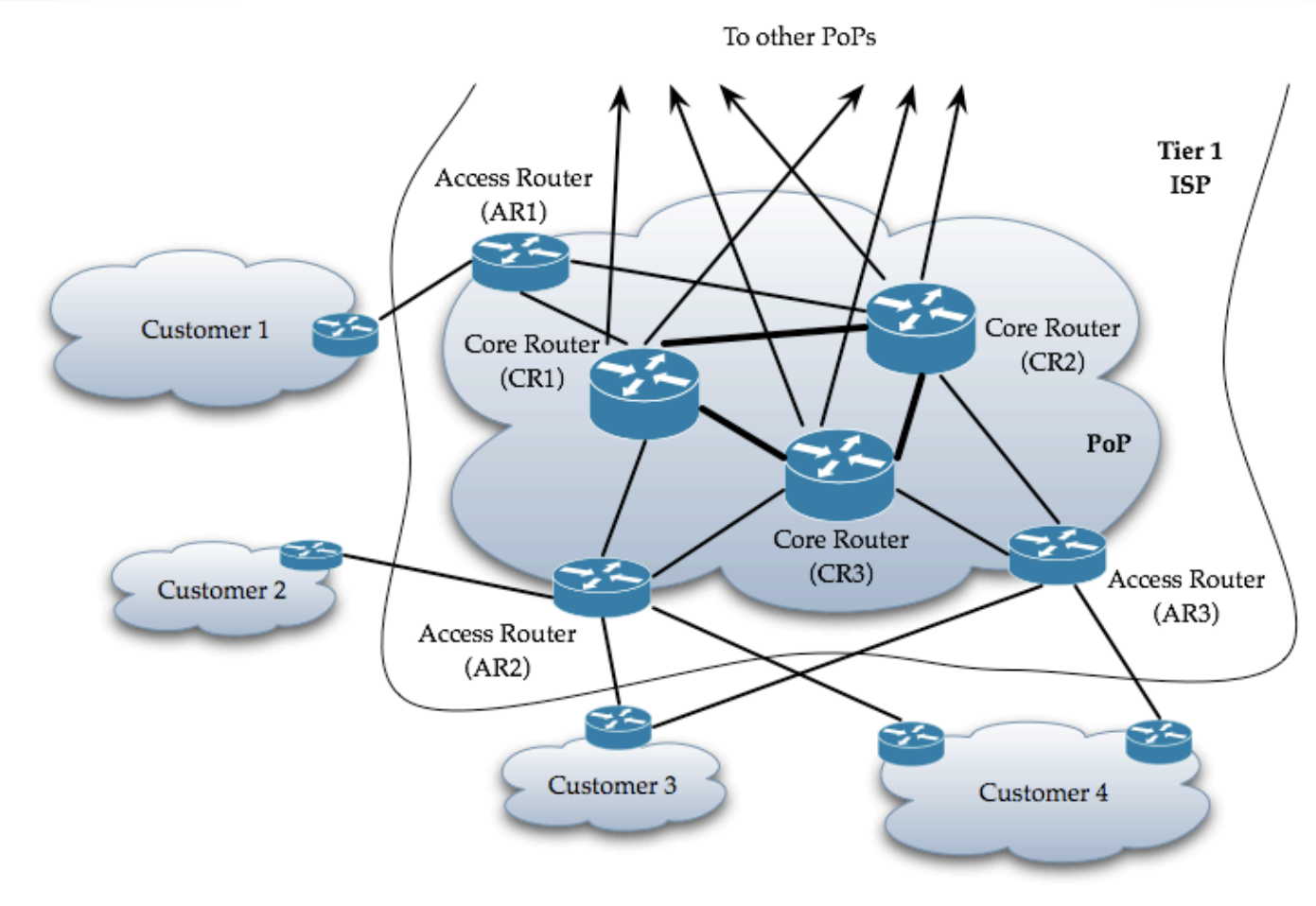
Tier "n" public peering more common these days



Management Issue for Large "core" ISPs

- Customer Management
- SLAs & Policy
- Examples at:
  - http://www.att.com/peering/
  - http://www.verizonbusiness.com/uunet/peering/
  - http://www.sprintlink.net/policy/bgp.html
- Traffic Engineering (early exit routing)

PoP view at a Tier-1 Provider



Traffic Exchange and Payment Relations: Examples

- Multilateral agreement: several ISPs build/use shared facilities and share cost; for example, this agreement can be possible with public exchange points or private exchange points.
- Bilateral agreement: two providers agree to exchange traffic if traffic is almost symmetric, or agree on a price, taking into account the imbalance in traffic swapped; for example, in a private peering setting.
- Unilateral agreement for transit: a customer pays its provider "access" charge for carrying traffic; for example, a tier 4 ISP would pay a charge to tier 3 ISP.
- Sender Keeps All (SKA): ISPs do not track or charge for traffic exchange; this is possible in private peering, and in some public peering.

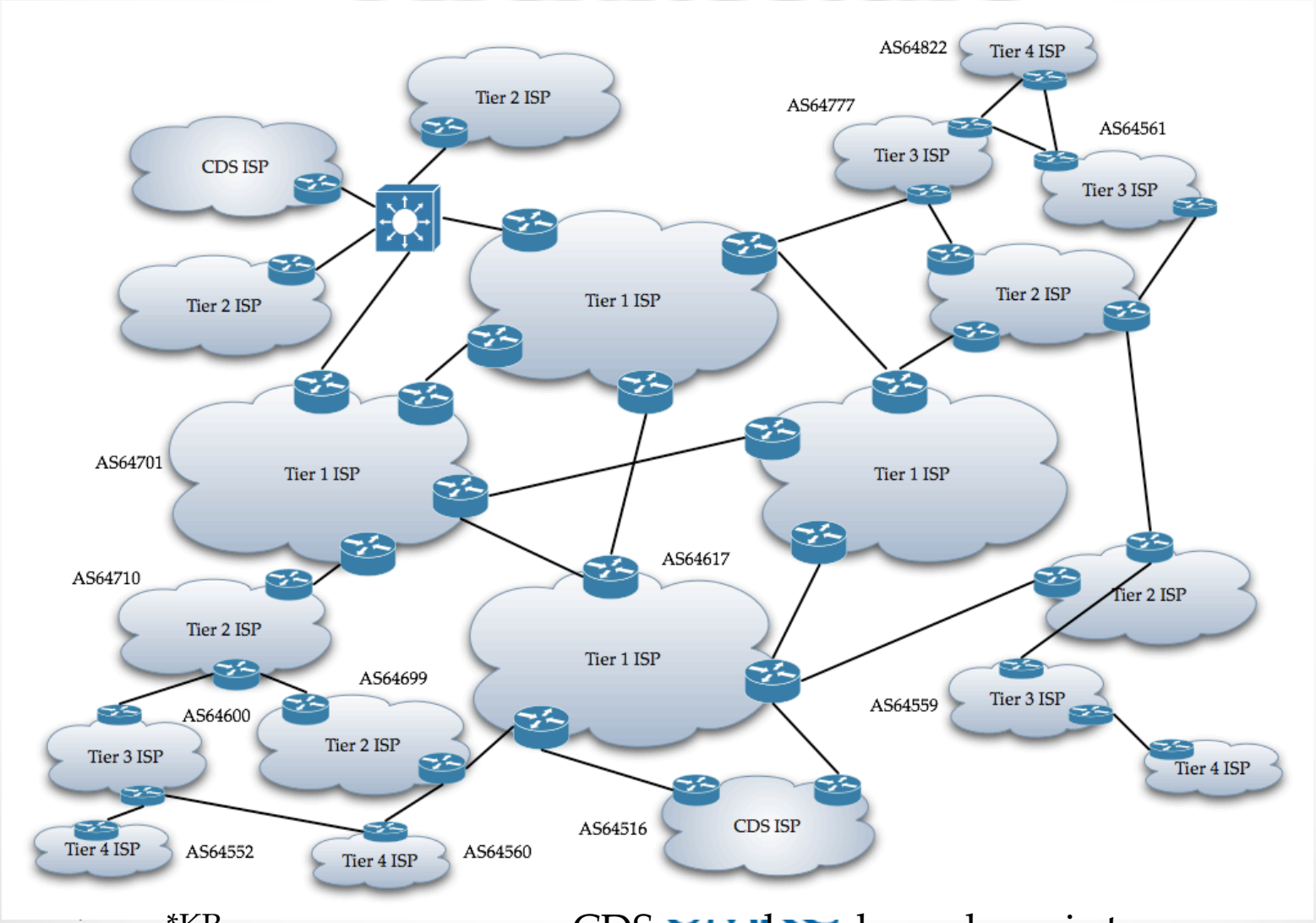
Internet Exchange Points (IXP)

- Largest three
  - AMS-IX: Amsterdam (http://www.ams-ix.net)
    - peak traffic 1558 Gbps (!)
  - London (http://www.linx.net)
    - peak traffic 1297 Gbps
  - Japan (http://www.jpix.ad.jp/en/)
    - over 100 members; peak traffic 173 Gbps
- Brazil: PTT Metro: 16 IXPs, Sao Paulo is the largest (peak: 40 Gbps)
- In US: private peering is quitecommon at tier-1 level than in the rest of the world
  - public peering at other levels

AMS-IX.net

- **Each member may set their own policy on peering**
  - **https://www.ams-ix.net/connected/**
- **AMS-IX policy on private peerings**
  - "AMS-IX poses no limits on peering arrangements between our members; nor do we interfere with private peering connections: in this respect we act purely as a L2 facilitator. "

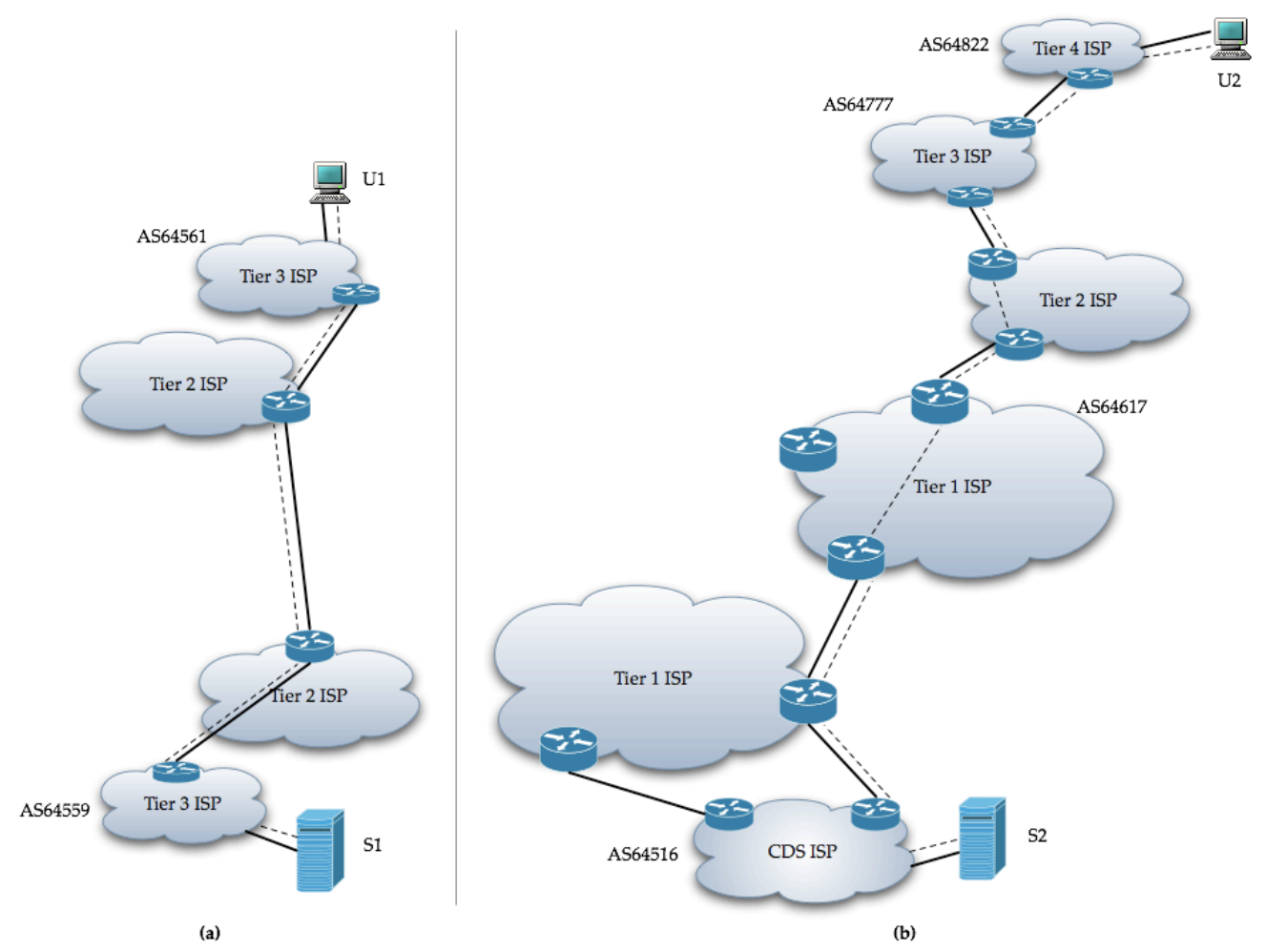
Representative Architecture



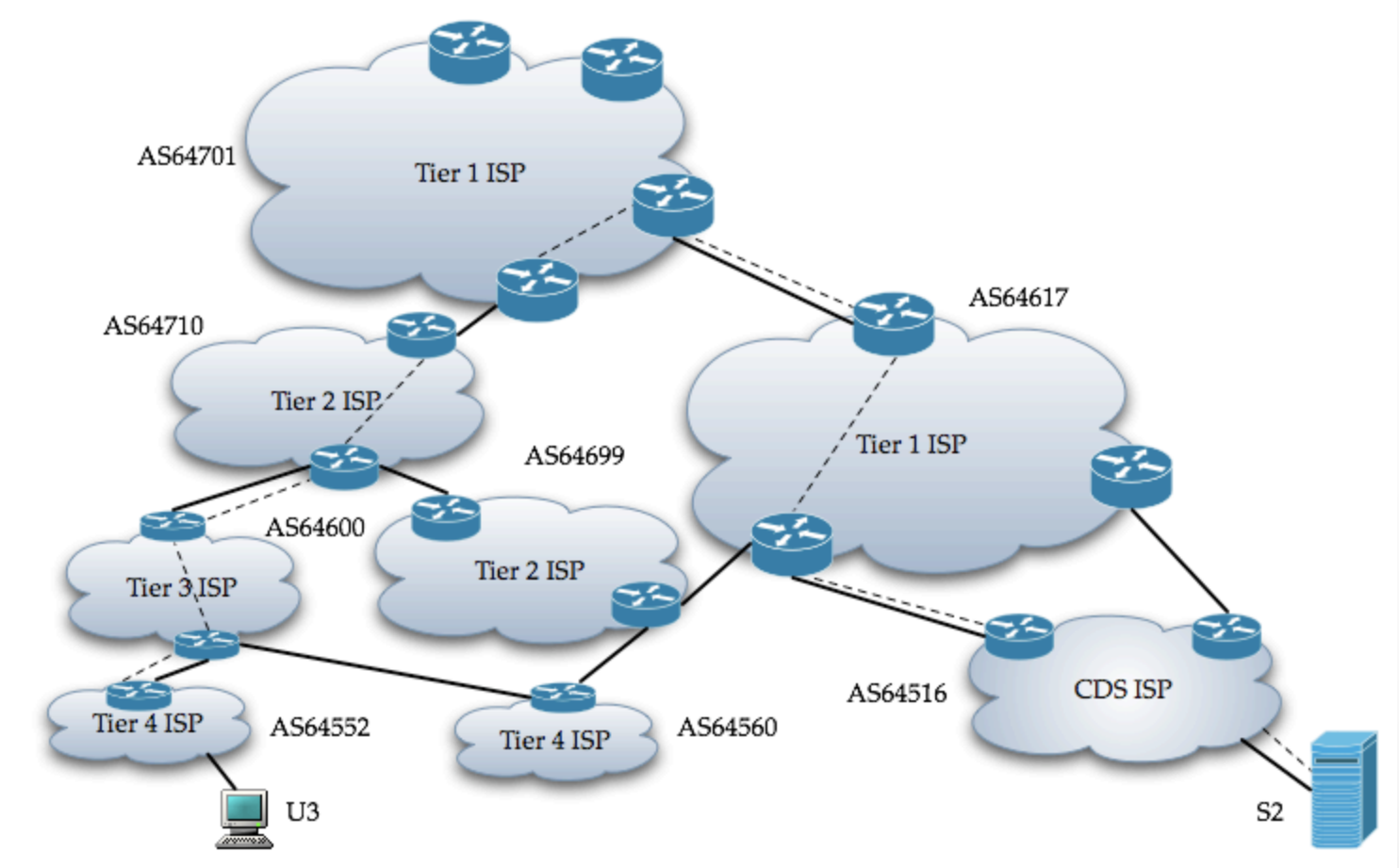
*KR

CDS: google, yahoo, akamai etc

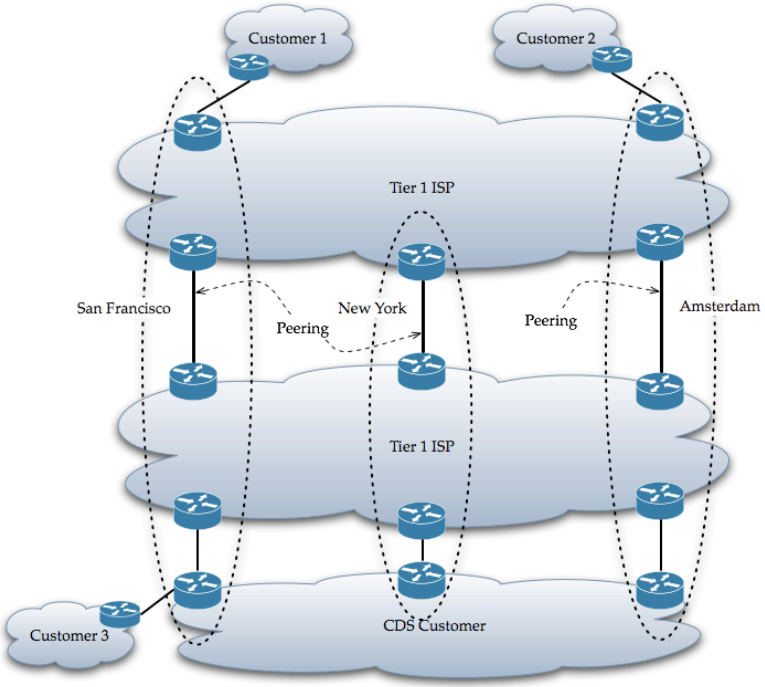
Packet Flow examples



Multi-Path Scenario & Policy Decision



Multi-homing connectivity, early-exit routing



Impact on IP traffic engineering due to early-exit routing

--need to track traffic demand groups a bit differently than mentioned earlier

IP prefix hijacking

Terminology

- Both BGP Hijacking and IP Hijacking are commonly used in the literature
  - Means the same thing
- Technically, it's *IP prefix hijacking*, which is accomplished because of how BGP works
  - Recall IP Prefix means an address block, typically identified as
    - 192.168.40.0/21
    - IP net id with a netmask

BGP threats

- Error due to configuration
- Untrusted Origin
- Untrusted modification (along the path)
- BGP speakers compromise
- BGP packet sniffing

Some solution for BGP threats

- RFC 2385
  - Implement what’s recommended
  - MD5 validation of TCP sessions (between two BGP speakers)
- RFC 2827:
  - Is neighbor announcing their own space?
- Contain problem near the origin, one AS upstream (need to rely on other sources of information)
- BGP Configurations checking

BGP extensions proposed

- Secure BGP (S-BGP)
  - Digitally signed, PKI based, validate BGP speakers
  - Limitations:
    - PKI infrastructure, Increase CPU & memory for processing, "reluctance" by many ISPs
- Secure Origin BGP (SO-BGP)
  - Cisco's solution in response to S-BGP: verifies AS path as announced, security message extension implemented (incremental deployment with requiring PKI)

BGP session hijacking

- Setting up a session between two BGP speakers (which is over TCP)
  - Hard to hijack
  - Usually, you keep a white list of which sources you're expecting to connect with (for example a direct neighbor, unless a 'long' session such as connecting to RouteViews)

Bogons: IP prefix and AS

- An IP prefix or an AS number that should never appear in the Internet
  - Not assigned yet (by RIR) or reserved space (RFC 3330), …
- Source of DDoS attacks
- An organization ‘claims’ that they have obtained an IP prefix from an RIR and “convinces” an ISP to announce it!

- Zombie blocks
  - Address currently not used in the Internet

Solution to Bogons/ Zombies

- Proper filtering can minimize bogon/zombie problems
  - build a black list and apply policy before announcing further a received IP prefix
  - Maintain full list (based on IP prefixes learned) instead of doing default routing [reverse path forwarding]

Relation between BGP and actual packet forwarding

- BGP only provides reachability information for an IP prefix
- BGP does not provide data delivery

- Means you can't trust what you get from BGP!
  - Source of trouble
  - However, not easy to always identify

IP prefix hijacking: unintentional or intentional

- Unintentional
  - An AS incorrectly 'set something' that gives the impression of origination of routes
    - Such errors are typically "large-scale" so it be figured out
  - Some examples:
    - 1997 example (AS 7007 problem)
      - http://merit.edu/mail.archives/nanog/1997-04/msg00380.html
    - December 2004 Example (AS 9121 problem)
    - Recent "YouTube/Pakistan" example (February 2008)
      - See; http://www.narus.com/blog/2008/02/

- Intentional
  - An AS does original an IP prefix
    - But may do only for a few (just one)
      - Feeble signal - Hard to detect!

To make matters worse

- An IP prefix may change its origin at any time legitimately
  - You change your ISP (i.e. would need to change your AS number)
  - UMKC's IP prefix 134.193.0.0/16 now with AS3390; tomorrow with AS 65512
- You may have multiple origin AS (MOAS) for diversity region
  - 134.193.0.0/16 announced as origin by two ASes, AS 3390 and AS 65516
- Also, you have one and just added another one!
- Even if you found that an IP prefix is hijacked, how do you inform them of the problem?
  UMKC's IP prefix is 134.193.0.0/16 and with native AS 3390
  - Now if this prefix is hijacked, to say AS 65512, then someone sends UMKC's sys admin an email will go via AS65512 and then "die" (host unreachable!)

Another issue: Longest prefix matching

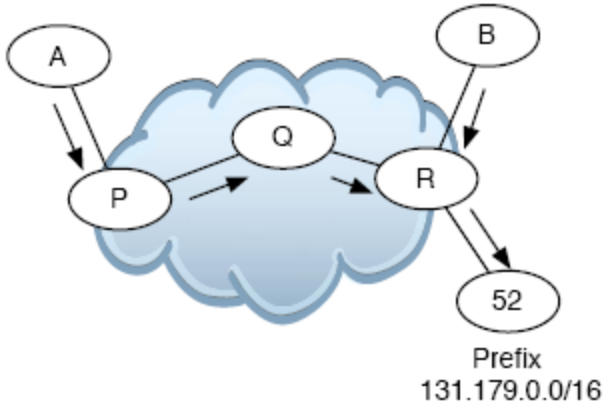
- BGP routing table suppose contains entries for IP prefix 134.193.0.0/16 and 134.193.0.0/24 (second one is more specific)
  - "Punching hole"

- Suppose a packet destined for 134.193.0.128 would go to 134.193.0.0/24; however, a packet destined for 134.193.0.1 would go 134.193.0.0/16

Different ways an AS can hijack an IP prefix

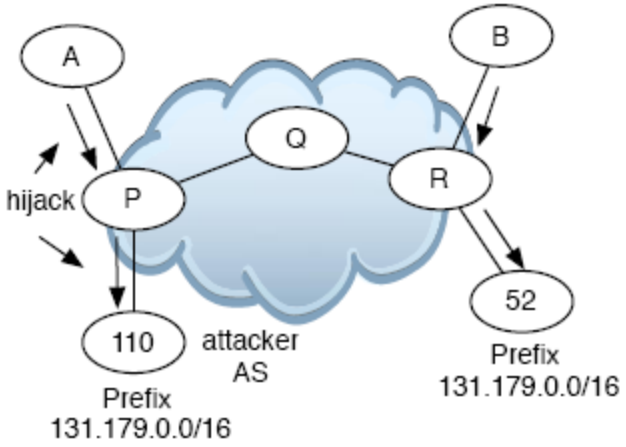
- Advertise falsely that it's the origin of the IP prefix
- Modify part of the AS path during "transit"
- Advertise falsely a more specific IP prefix

- Normal 131.179.0.0/16 with AS 52



a. True origin AS 52 announces prefix 131.179.0.0/16

- 1391.179.0.0/16 is hijacked by AS 110
  - B's traffic will go normally to AS 52
  - A's will falsely go to AS 110



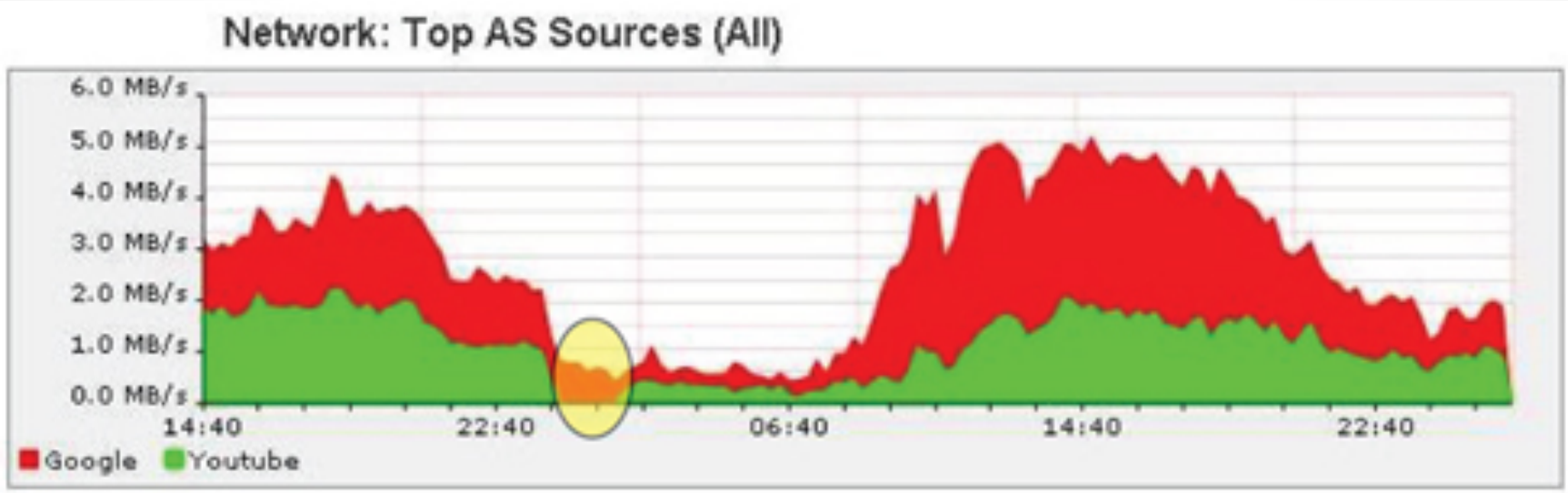
b. False origin AS 110 announces prefix 131.179.0.0/16 and hijacks A's route

Figure source: PHAS paper

YouTube/Pakistan Example (February 2008)

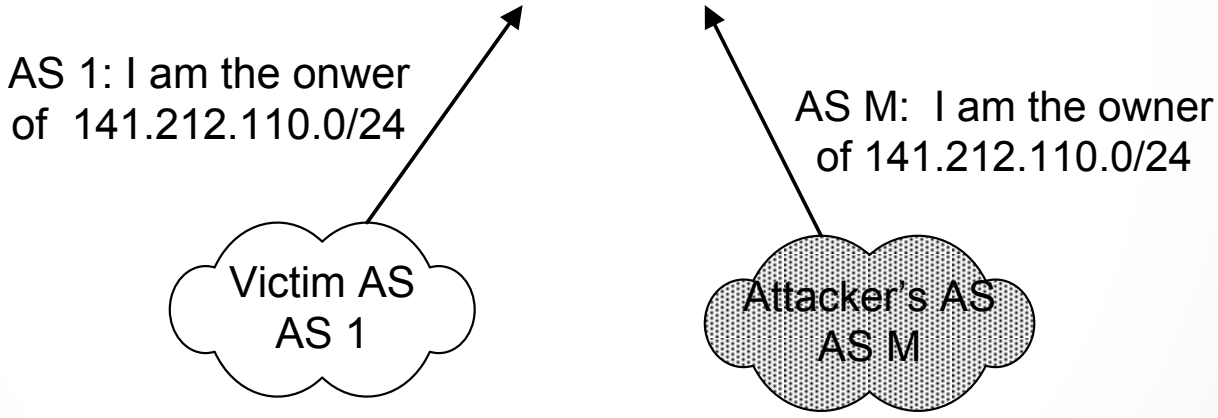
- BGP data intended to block access to YouTube within Pakistan was accidentally broadcast to other service providers, causing a widespread YouTube outage.
- The chain of events that led to YouTube's partial black-out was kicked off Friday when the
- Pakistan Telecommunication Authority (PTA) ordered the country's ISPs to block access to YouTube
- ISPs in Pakistan followed the order for blocking YouTube by creating BGP data that redirected routers looking for YouTube.com's servers to nonexistent IP prefix (destination).
- It shared this data, unfortunately, upstream with Hong Kong's PCCW
- PCCW, in turn shared it with other ISPs throughout the internet.
- Because Pakistan's BGP traffic was offering very precise routes to what it claimed were YouTube's internet servers, routers took it to be more accurate than YouTube's own information about itself.

Drop in traffic to YouTube



Source: Supranamaya Ranjan, *YouTube Prefix Hijacking*, February 28th, 2008, http://www.narus.com/blog/2008/02/

- Attacker announces the prefix belonging to other ASes using his *own* AS number.



- Leading to MOAS (Multiple Origin AS) conflicts

Having a sounding board

- BGP Monitoring Sites:
  - Oregon RouteViews & RIPE RRC
  - These sites collect data directly from "signed up" BGP speakers throughout the world (i.e., ISPs, mostly tier-1 ISPs)
    - However, IP prefix owners aren't signed up for this service, neither can they [because this is a BGP only service under standard BGP protocol]
- Why helpful
  - Providers can check themselves like a 'mirror'
    - Is it showing where I'm supposed to be

Goal

- Developing an Prefix Hijacking Alert System (see PHAS reference)

- Main requirement: Alert accurately

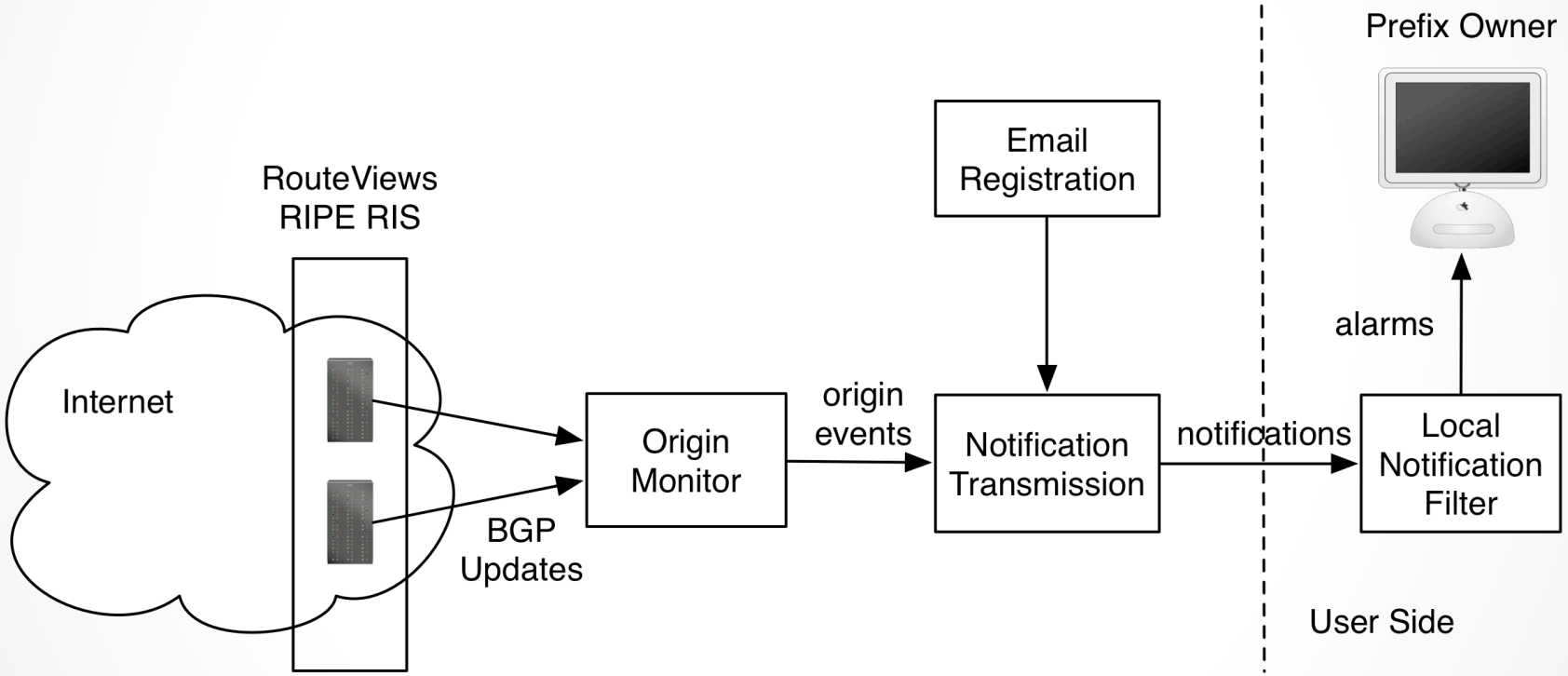
  Remember the issues discussed earlier!

RouteViews based Alert System

- Step 1: Monitor RouteViews BGP tables and updates in (near) Real-Time
- Step 2: Keep a database of Origins used to reach each Prefix
- Step 3: Report any change in Origins used to reach the Prefix
- Step 4: Owner applies local filter rules to determine significance

*From PHAS presentation

Architecture of an Alert system



*From PHAS presentation

- Email registration:
  - More than at your domain name (e.g., have one with gmail or yahoo mail also)

Some notations:

- O_{SET} (P, t) := Origin set of AS for Prefix P at time t
  - Could be a singleton set (in most cases), unless it's MOAS, or hijacked
  - Example, UMKC normally
    - O_{SET} (134.193.0.0/16, t) = {3390}
    - If hijacked at time t1 by AS 65512, then
      - O_{SET} (134.193.0.0/16, t) = {3390, 65512}
- $M_1, M_2, \ldots, M_N$:= the N BGP speakers providing data to sites such as RouteViews or RRC
- *Origin*(M_i, P, t) := Origin AS set for prefix P as known to BGP speaker M_i at time t
- Now if
  - *Origin*(M_i, P, t) ≠ *Origin*(M_j,P,t),
  - then there is something wrong

Basic Algorithm

Algorithm 1: Instantaneous Origin Change

Initialize $origin(M_i, P, t_0)$ using the initial routing table of M_i at time t_0;
$O_{SET}(P, t_0) = \cup_{i=1}^{N} origin(M_i, P, t_0)$;
if *update for prefix* P *at time* t *from router* M_i *is an announcement* **then**
    $origin(M_i, P, t)$ = the last AS in the announced path;
else
    $origin(M_i, P, t) = \{\}$;
$O_{SET}(P, t) = \cup_{i=1}^{N} origin(M_i, P, t)$;
if $O_{SET}(P, t) \neq O_{SET}(P, t-1)$ **then**
    $origin\_gain = O_{SET}(P, t) - O_{SET}(P, t-1)$;
    $origin\_loss = O_{SET}(P, t-1) - O_{SET}(P, t)$;
    send $[O_{SET}(P, t), origin\_gain, origin\_loss]$ to prefix owner;

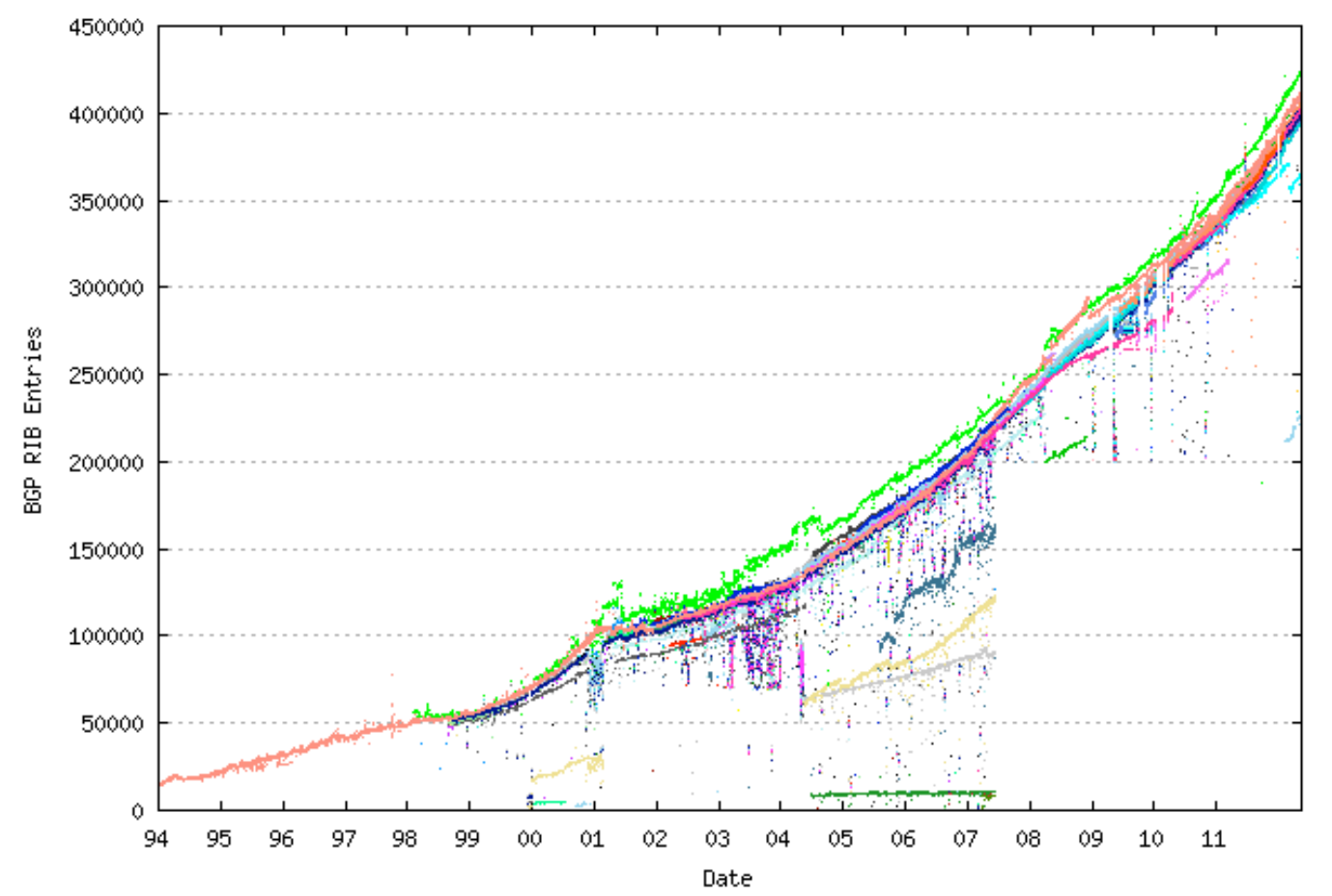
Drawback with basic algorithm

- From measurement data
  - Some prefixes have large number of origin AS events, which are legitimate
    - Don't want to generate a notification each time ("false notification")
  - Some prefixes have unstable connection to the Internet
    - In BGP it shows up withdrawal and announcement: repeated oscillation
      - Shouldn't be notified
- Variation of the basic algorithm:
  - Introduce a hold-down time window with adaptive adjustment (sort of like route damping)[see Lad et all paper]

Routing Table issues at AS boundary

- At border routers (BGP speaker)
  - Need to maintain entry for each address block (IP prefix), although some aggregation possible; why
    - Address block "flat"
      - E.g. 134.193.0.0/16 UMKC, "next" one: 134.194.0.0/16 is DoD NIC in Columbus Ohio
      - Need separate entries, aggregation not possible
  - Current size of IP prefixes: 230,000 entries
    - For each IP prefix, a next hop entry is needed
  - Can future routers handles lookup efficiently if this table keeps growing
    - Address lookup research, hardware research etc

BGP routing table growth



Data source: courtesy Geoff Huston

BGP issues

- Announcement/Withdrawal
  - Can be very long time for convergence
    - Especially, if multihomed
  - BGP Stable Paths Problem
- Actual failure
- Routing table growth, currently at 400,000 (up from about 250,000 a few years ago): Can core routers handle it efficiently?
- Locator and Identifier separation: and reduce the size of routing table

LISP

- LISP separates the current single numbering space into Endpoint IDentifiers' (Non-Routable) and Routing LOcators (Routable).

- LISP Terminologies

- Ingress Tunnel Router (ITR) encapsulates the IP packet (from EID's) with LISP header and forwards the packet to appropriate ETR (Outer header IP address).
- Egress Tunnel Router (ETR) decapsulates the outer LISP header and forwards the packet to appropriate EID (Inner header IP address).

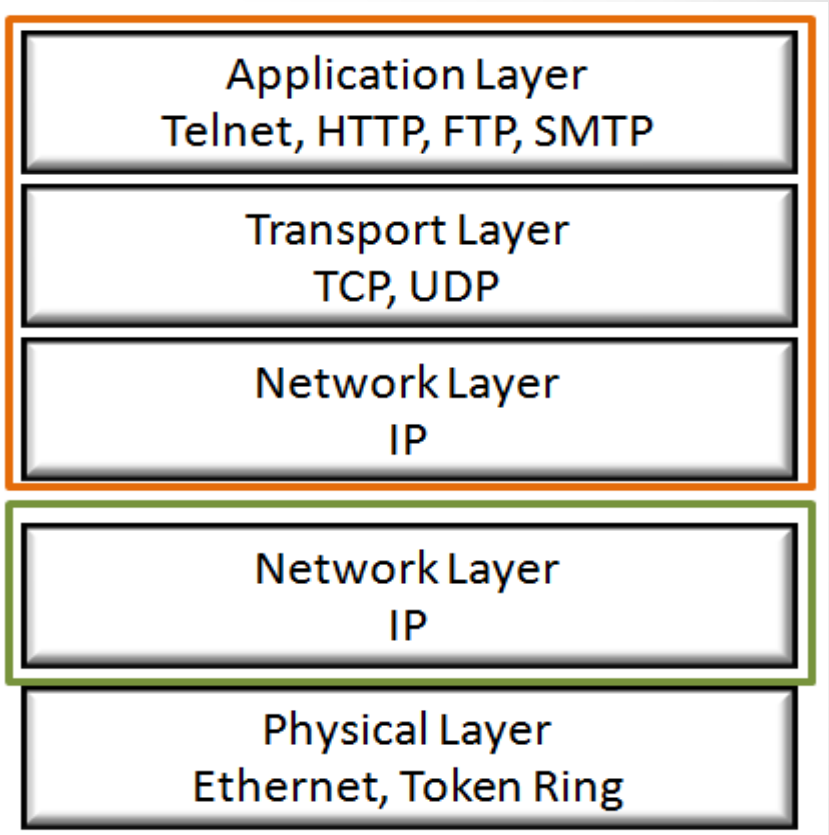
EID and RLOC

- EID's are stored in the inner IP header.
- RLOC's are stored in the outer LISP header.
- It is based on a simple IP-in-IP tunneling approach.

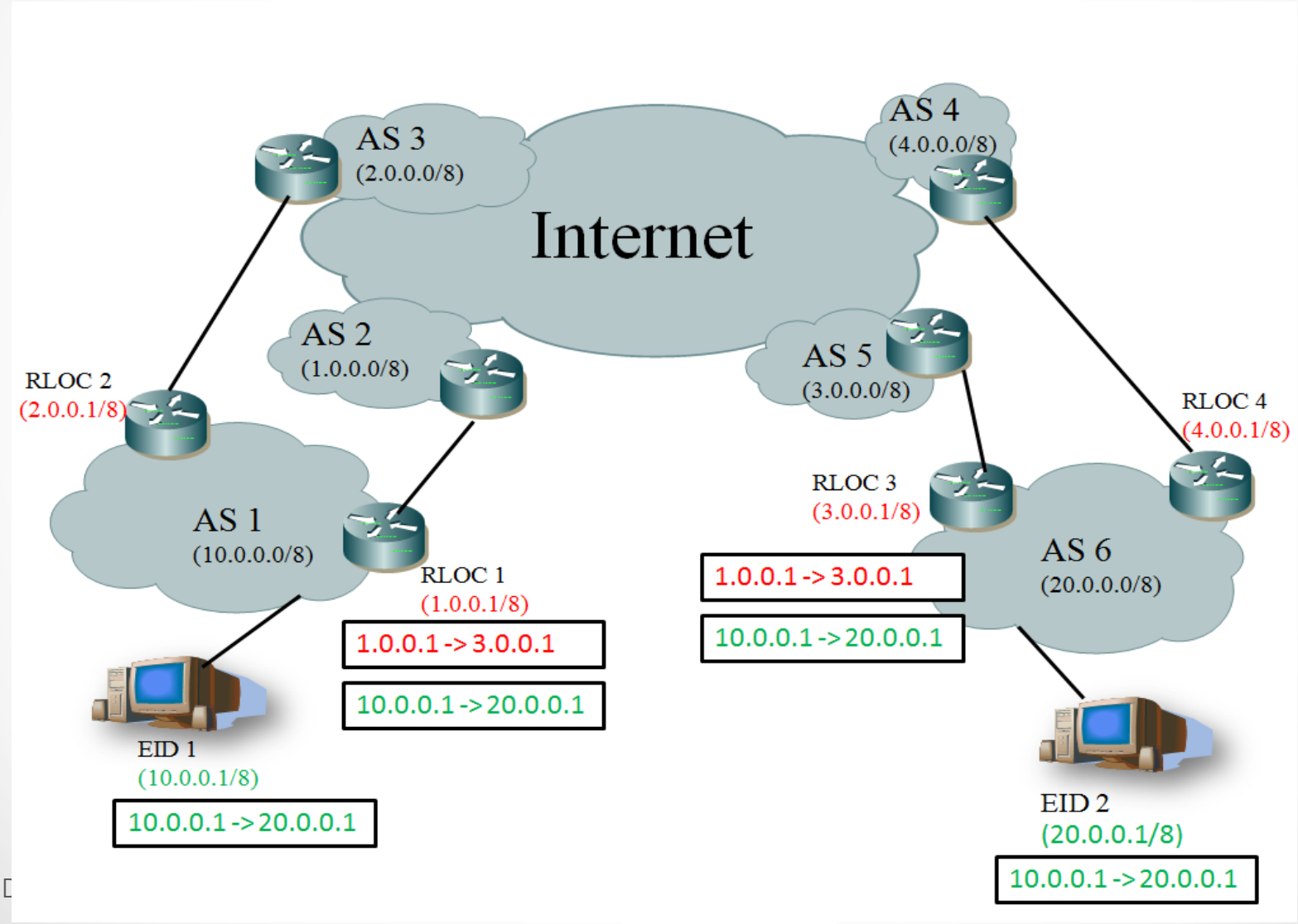
➢ LISP Architecture

Host Stack supplies IPv4/IPv6 EID's

LISP Routers supplies IPv4/ IPv6 RLOC's



LISP operation



Summary

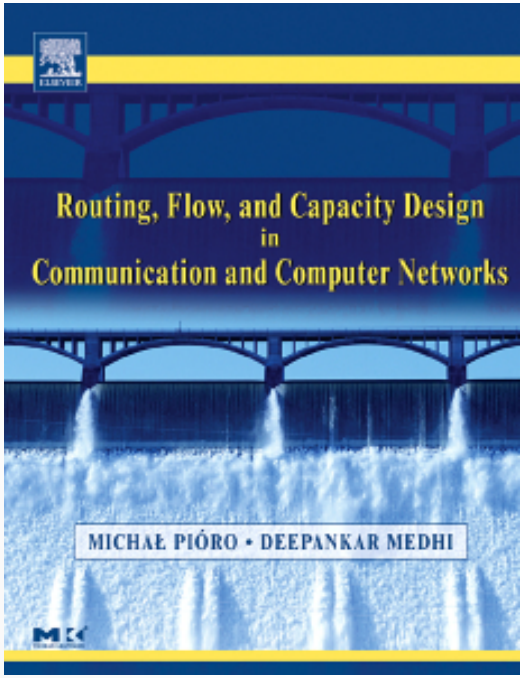
- An address block (IP prefix) is "home" to an autonomous system/ISP
  - They may move to another ISP
- The network of ASes form the Internet at core
- An AS can run different routing protocols – OSPF, IS-IS
  - An AS can have multiple providers "internally"
  - IP traffic engineering
- IP prefixes are announced/withdrawn using BGP protocol
- Border routers (BGP speakers) computed shortest AS-path subject to policy constraints
  - For each IP prefix, a next-hop entry is created at border routers
  - IP prefix is over 200,000; routing table growth issue
- ISPs can be of different tiers
  - They form business relations: core, transit, access
  - Public/Private peering, or transit
- Internet Exchange Points (almost at every tier)
- IP Prefix hijacking
- LISP

Acknowledgment

- Many figures by Karthik Ramasamy! (*KR)
- Many other figures borrowed from a number of sources
- Geoff Huston for BGP data

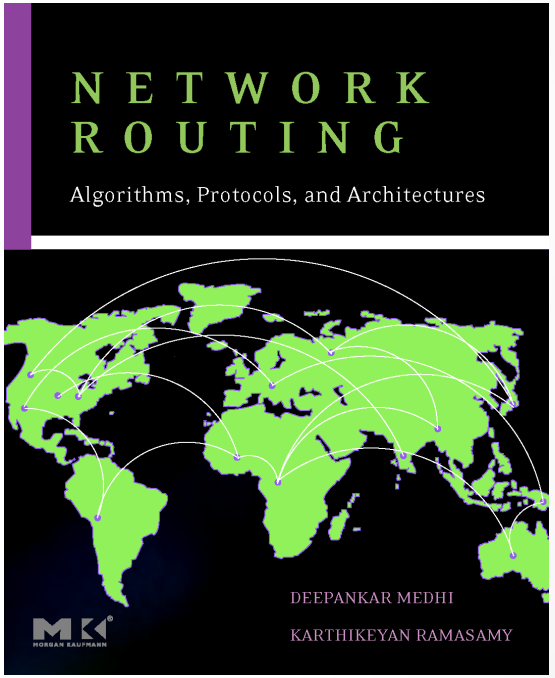
Two books

- Publisher: Morgan Kaufmann/Elsevier



July 2004

ISBN 978-0-12-557189-0



March 2007

ISBN 978-0-12-088588-6

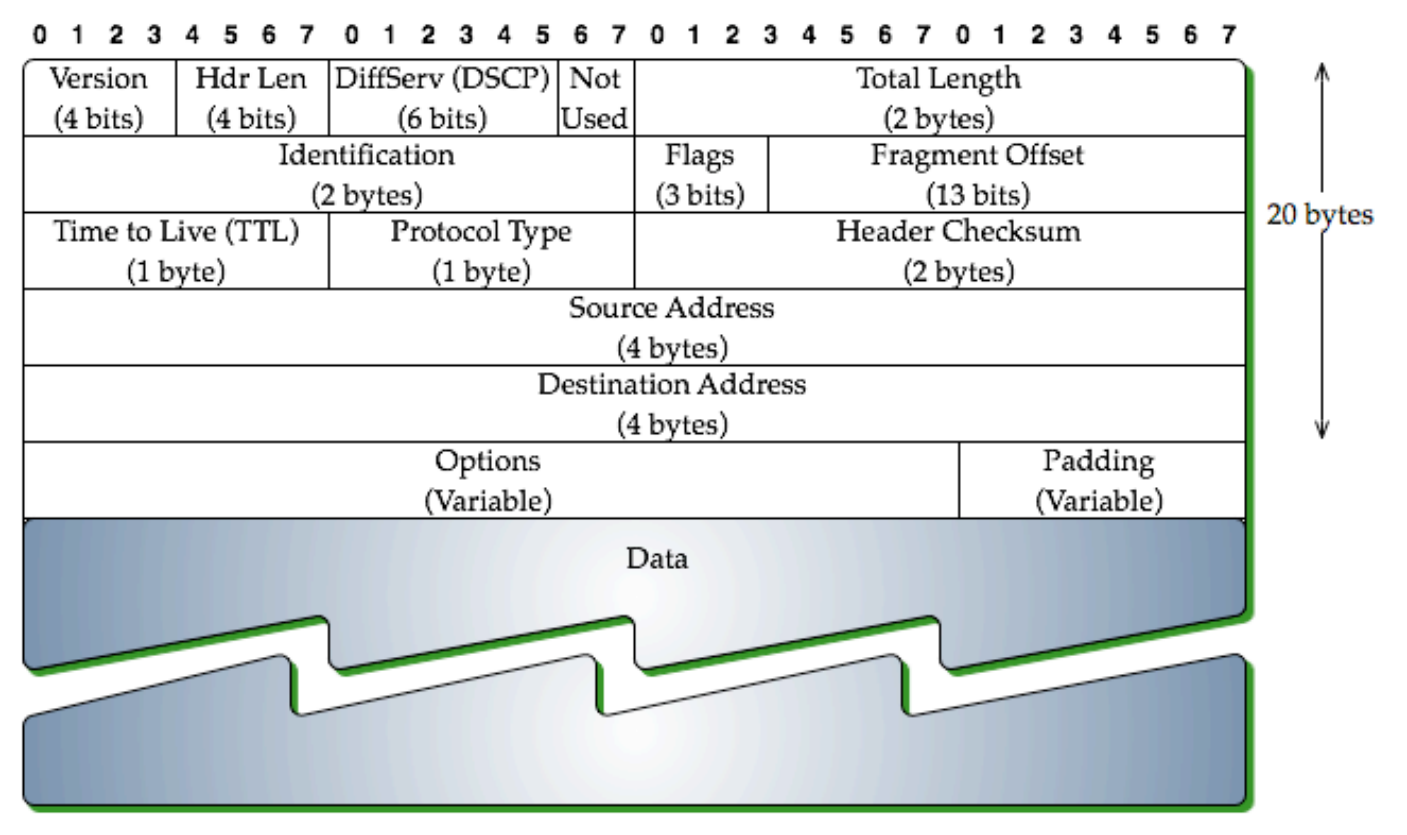
Acronym List

- IP: Internet Protocol
- TCP: Transmission Control Protocol
- UDP: User Datagram Protocol
- RIP: Routing Information Protocol
  -- Distance Vector Protocol
- IGRP: Interior Gateway Routing Protocol
  -- Distance Vector Protocol
- EIGRP: Enhanced Interior Gateway Routing Protocol
  -- Enhanced loop-free Distance Vector Protocol
- OSPF: Open Shortest Path First Protocol
  -- Link state protocol
- IS-IS: Intermediate System-to-Intermediate System
  -- Link state protocol
- BGP: Border Gateway Protocol
  -- Path Vector Protocol
- RPSL: Routing Policy Specification Language, RFC 2622

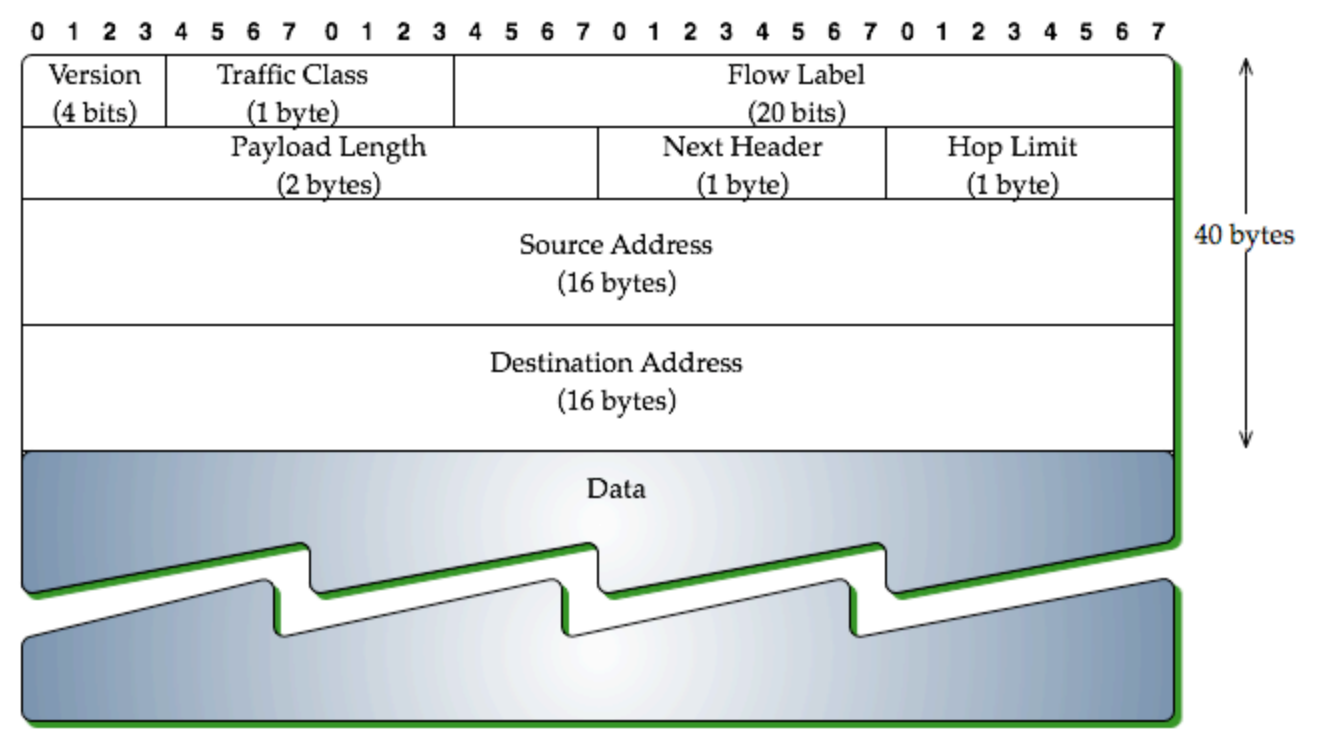
Acronym List (cont'd)

- ICMP: Internet Control Message Protocol
- CIDR: Classless Inter-Domain Routing
- IP Prefix: a contiguous IP address block such as 134.193.0.0/16
  means 134.193.0.0 - 133.193.255.255
- LSP: Label Switched Path
- MPLS: Multiprotocol Label Switching
- AS: Autonomous System
- MED: Multi-Exit Discriminator
- PoP: Point-of-Presence
- SLA: Service-Level Agreement
- RIR: Regional Internet Registries:
  they are ARIN.net, APNIC.net, RIPE.net, AfriNIC.net, LACNIC.net

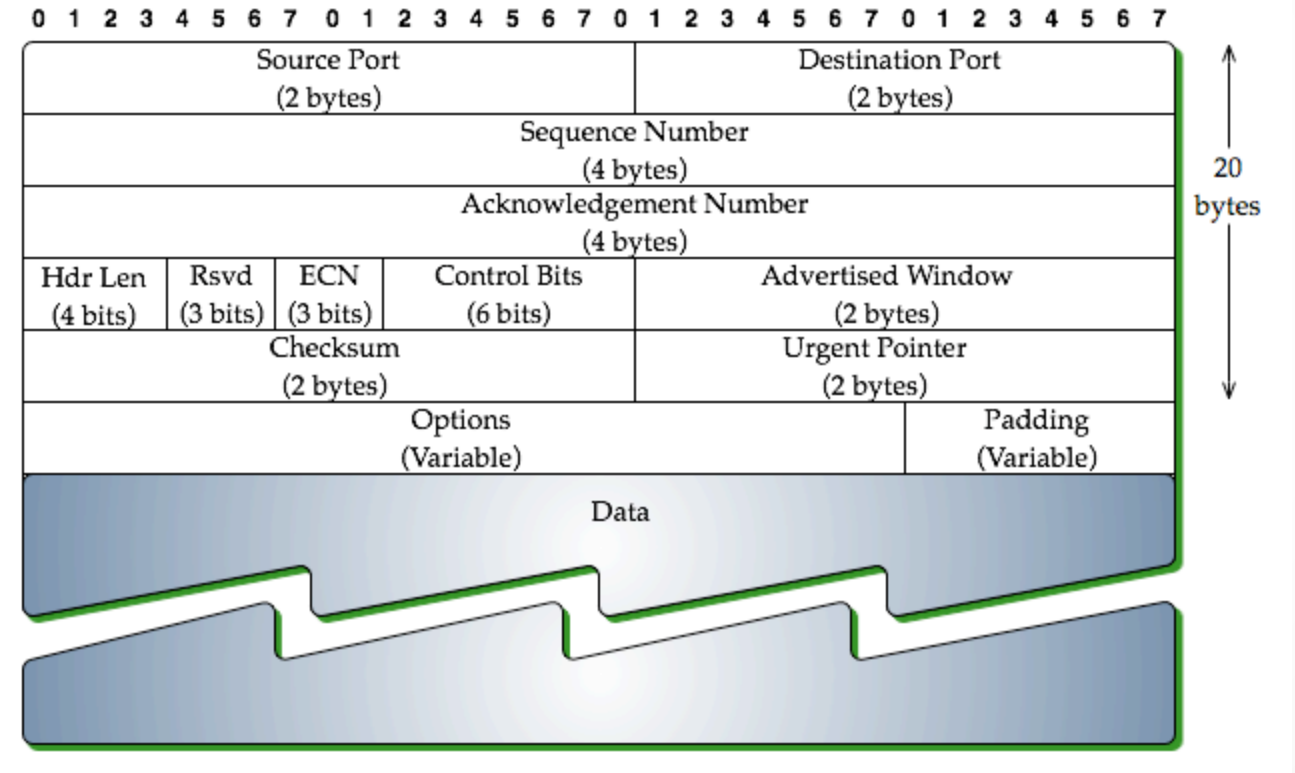
IPv4 packet (datagram) format



IPv6 packet format



TCP packet format



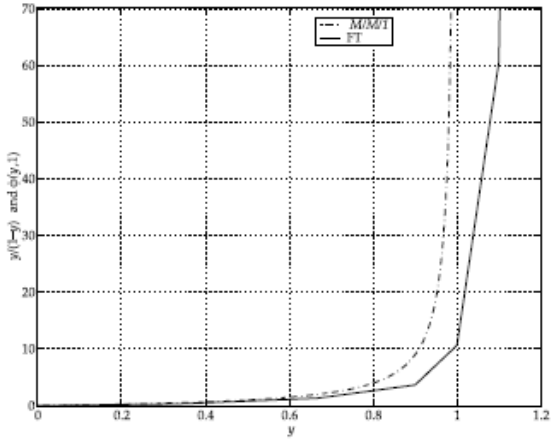
IP Traffic Engineering: Formal Approach: Multi-Commodity Network Flow Models

- Notations:

Notation	Explanation
K	Number of demand pairs with positive demand volume
L	Number of links
h_k	Demand volume of demand index $k = 1, 2, ..., K$
c_ℓ	Capacity of link $\ell = 1, 2, ..., L$
P_k	Number of candidate paths for demand k, $k = 1, 2, ..., K$
$\delta_{kp\ell}$	Link-path indicator, set to 1 if path p for demand pair k uses the link ℓ; 0, otherwise
ξ_{kp}	Unit cost of flow on path p for demand k
$\hat{\xi}_\ell$	Unit cost of flow on link ℓ
w_ℓ	Link weight for link $\ell = 1, 2, ..., L$
$x_{kp}(w)$	Flow amount on path p for demand k for given link weight system w
x_{kp}	Flow amount on path p for demand k
y_ℓ	Link flow variable for link ℓ
r	maximum link utilization variable
*	Use as a superscript with a variable to indicate optimal solution, e.g., x^*_{kp}

Objectives: commonly used

- Minimize Average Delay
  - Non-linear! Can be transformed to a piece-wise linear convex function, which can be replaced with a linear object with additional constraints
- Minimize Maximum Link Utilization
  - Can be transformed to a linear objective with constraints



(based on min-max link utilization)

$$
\begin{aligned}
&\textit{minimize}_{\{w,r\}} && F = r \\
&\textit{subject to} && \sum_{p=1}^{P_k} x_{kp}(w) = h_k, && k = 1, 2, ..., K \\
& && \sum_{k=1}^{K} \sum_{p=1}^{P_k} \delta_{kp\ell}\, x_{kp}(w) = y_\ell, && \ell = 1, 2, ..., L \\
& && y_\ell \le c_\ell\, r, && \ell = 1, 2, ..., L \\
& && w_1, w_2, ..., w_L \in \mathcal{W} \\
& && x_{kp}(w) \ge 0, && p = 1, 2, ..., P_k, \quad k = 1, 2, ..., K \\
& && y_\ell \ge 0, && \ell = 1, 2, ..., L \\
& && r \ge 0.
\end{aligned}
$$

Approach:

$$
\begin{aligned}
&\textit{minimize}_{\{x,y,r\}} && F = r && \\
&\textit{subject to} && \sum_{p=1}^{P_k} x_{kp} = h_k, && k = 1, 2, \ldots, K \\
& && \sum_{k=1}^{K} \sum_{p=1}^{P_k} \delta_{kp\ell}\, x_{kp} = y_\ell, && \ell = 1, 2, \ldots, L \\
& && y_\ell \le c_\ell\, r, && \ell = 1, 2, \ldots, L \\
& && x_{kp} \ge 0, && p = 1, 2, \ldots, P_k, \quad k = 1, 2, \ldots, K \\
& && y_\ell \ge 0, && \ell = 1, 2, \ldots, L \\
& && r \ge 0. &&
\end{aligned}
$$

- “
- C
  - Dual solution is related to the weights

- Summarizing
  - IP traffic engineering is an important research area
  - Need traffic measurement and traffic volume estimation
  - Specialized Algorithm for Large-scale networks
    - Heuristic based
    - Dual-based weights
    - Popular objective function: 1) Minimize average delay, 2) Minimize maximum link utilization (load balancing)
    - You want to watch for how different factors are affected:
      - Link utilization, length of shortest paths, delay estimate, variation due to load perturbation