Distributed Computer Systems Engineering

CIS 508: Lecture 10
Networking Protocols II

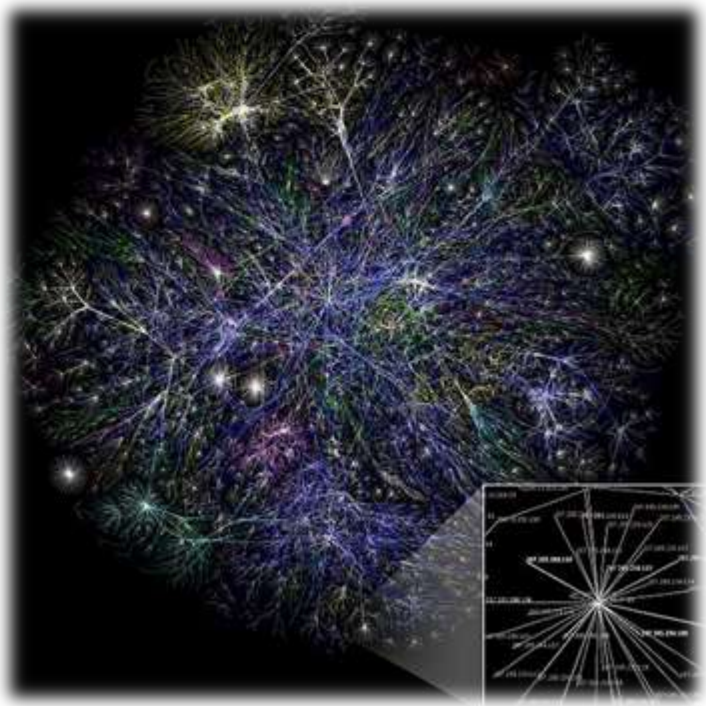
Lecturer: Sid C.K. Chau
Email: ckchau@masdar.ac.ae

Distributed Decision: **Routing**

- A collection of systems need to reach a decision for a specific task
- Information is exchanged among the systems
- No single system needs to keep track of the whole information of all systems
- Classical example: *routing*
    - Deciding the best path from source to destination
    - Each forwarding node only keeps track of local (neighbor) link states and next-hop forwarding decisions
    - E.g. Bellman-Ford algorithm, dynamic programming

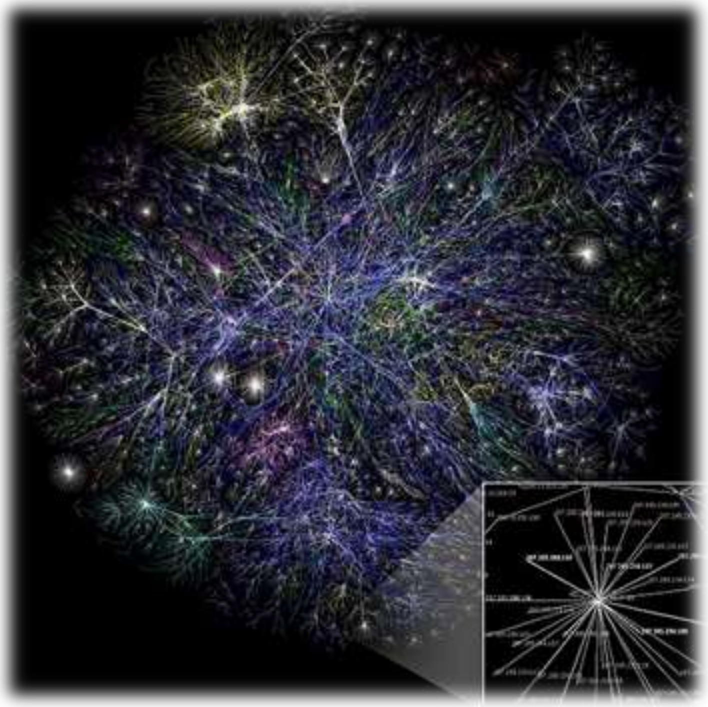
How to route on the Internet?

- Shortest-path routing paradigm
  - Finding a path between two nodes in a network such that the sum of the weighted edges is minimized
  - Open Shortest Path First (OSPF)
  - Does routing on the Internet always follow shortest-path routing?
  - ***No, far from it!***

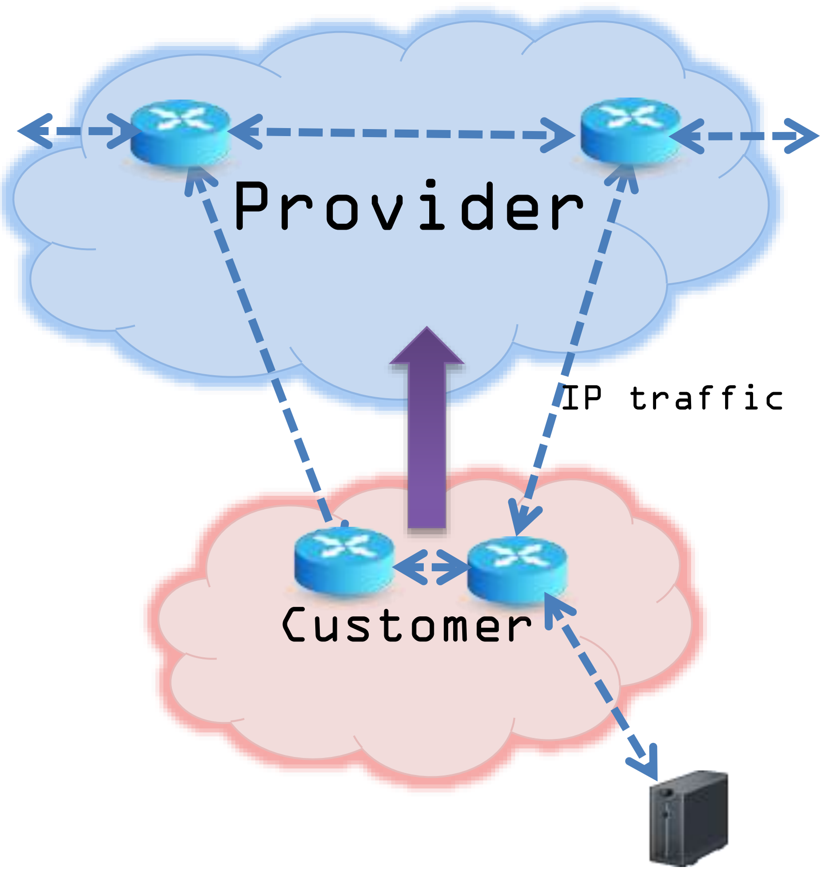


Forming the Internet

- The Internet comprises of billions of routers, switches, hubs
- The network facilities belong to many different organizations
  - Most of organizations are not sharing the facilities selflessly
  - Complicated business, governmental, institutional polices affecting the formation of the Internet
- Information of the networks belonging to different organizations is not always public
- Routing in the Internet is never the same as routing within an organization

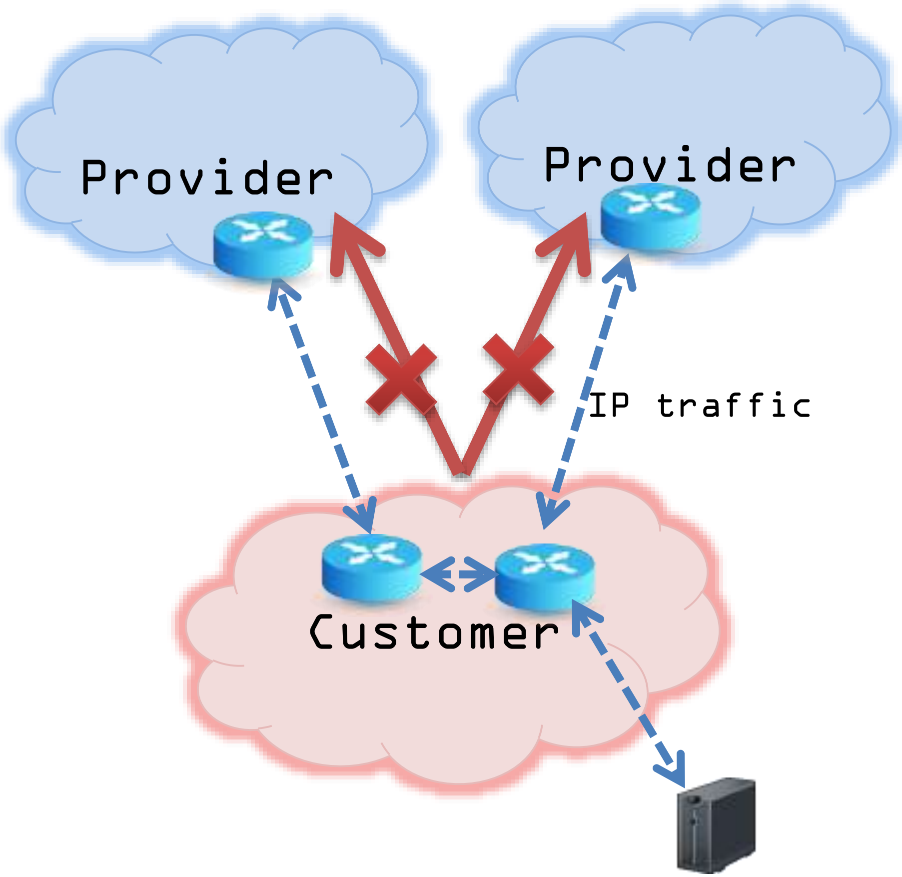


Customer-Provider Relationship



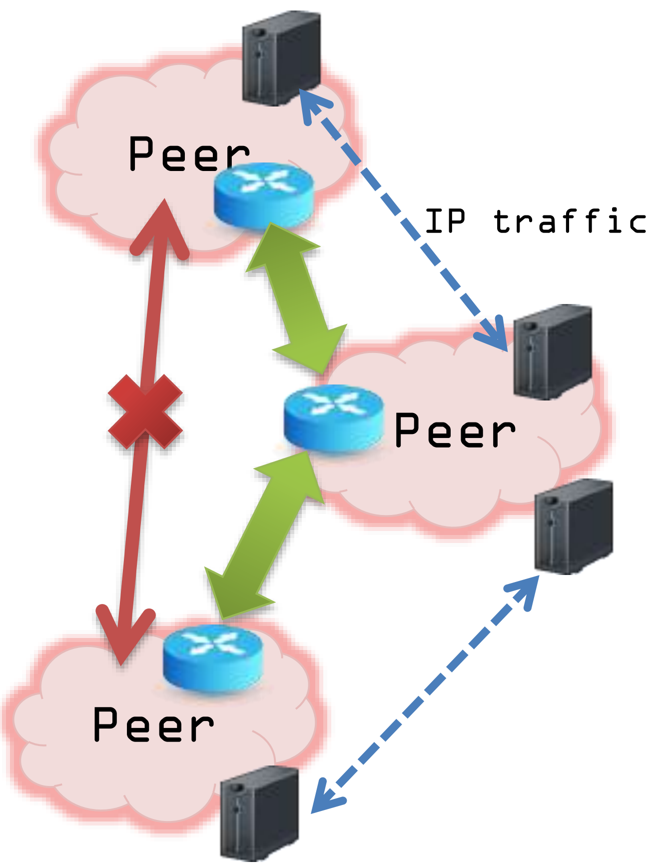
- Provider
  - Superior network infrastructure
  - Well-connected network connectivity
  - Presence in wider geography
- Customer
  - Inferior network infrastructure
  - Constrained network connectivity
  - Presence in local area
- Customer needs to ***pay*** provider for transit service to other networks

Customer-Provider Relationship



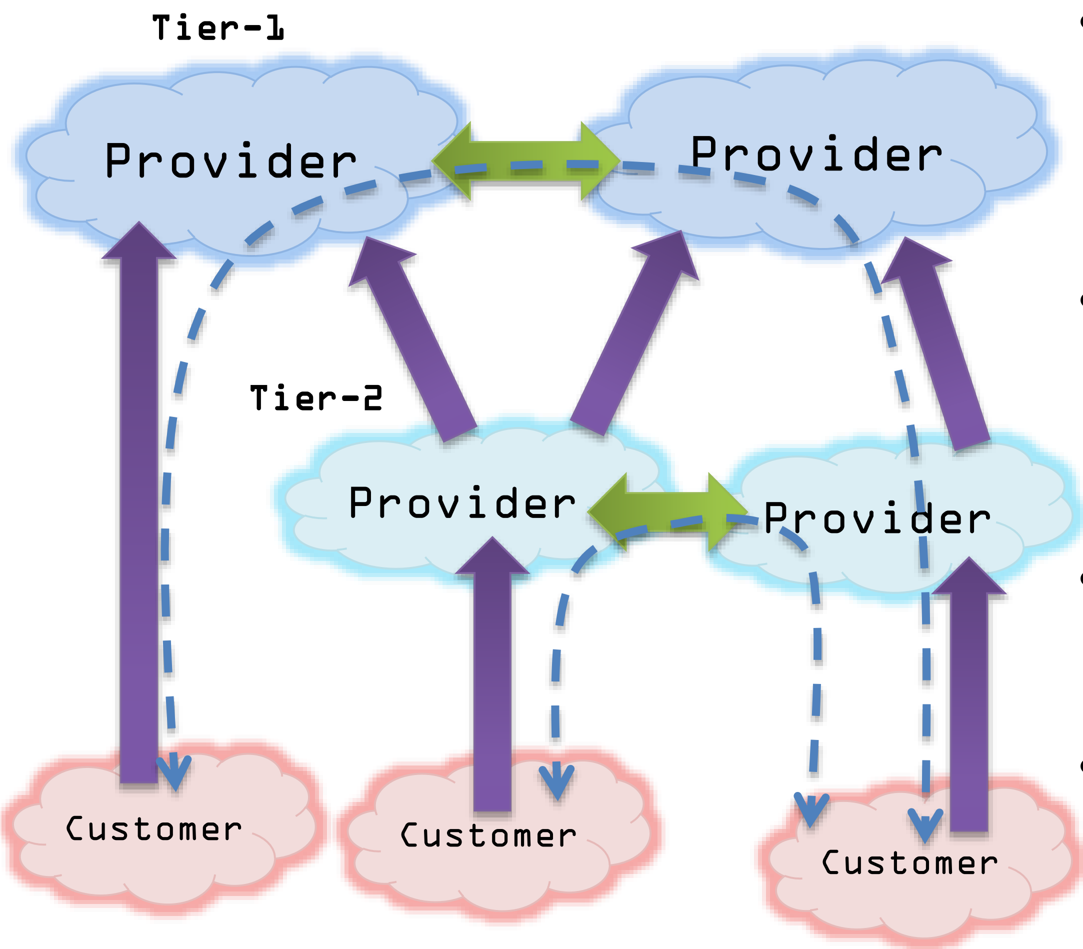
- Customer needs to ***pay*** provider for transit service to other networks
- Transit service
  - One party provides guaranteed connectivity to another party
- Provider *cannot* use customer for transit service
- Uni-directional connectivity

Peering Relationship



- Peering
  - Agreement among networks to directly exchange data traffic
  - Bilateral agreement
  - Not traffic transit through peer to other network
- Peer need not pay each other for mutual connectivity

General Relationship

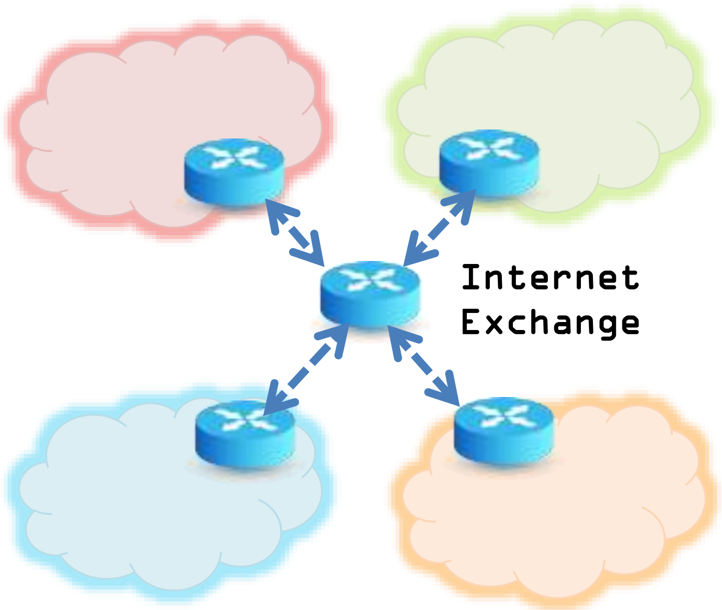


- Tier-1 Providers
  - Global network connectivity
  - E.g. AT&T
- Tier-2 Providers
  - Local network connectivity
  - E.g. Etisalat
- Customers
  - E.g. Masdar
- Peering allows connectivity between the customers of Tier 2 providers
- Not always use Tier 1 providers

Peering vs. Not Peering

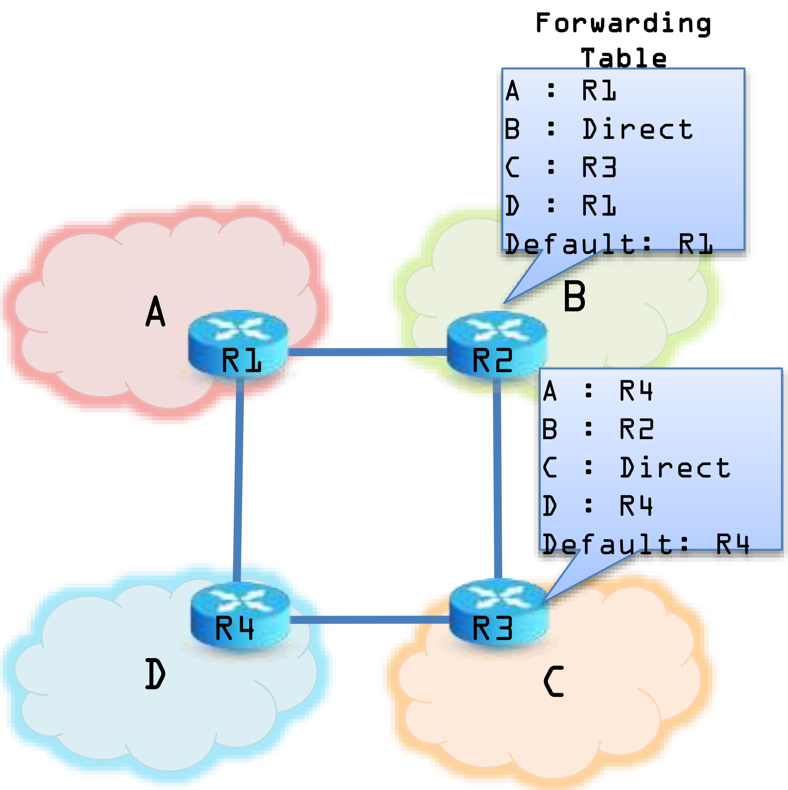
- *Peering*
  - Reduces upstream transit costs
  - Can increase end-to-end performance
  - May be the only way to connect your customers to some part of the Internet (for “Tier 1” providers)
- *Not Peering*
  - Acquiring new customers makes profit
  - Peers are usually your competitors
  - Peering relationships may require periodic renegotiation
- Peering agreements are often confidential
- Decision to peer or not peer is a complicated business decision

Internet Exchange Point (IXP)



- Internet Exchange Point (IXP)
  - A third-party facility that enables peering among multiple parties
  - Can be not-for-profit or for-profit
  - E.g. LINX (London Internet Exchange), Ankabut
  - Separate peering can be established among participating parties
  - Participating in IXP *not implying* accepting data traffic from all other participants

Basic Ideas: Routing vs. Forwarding

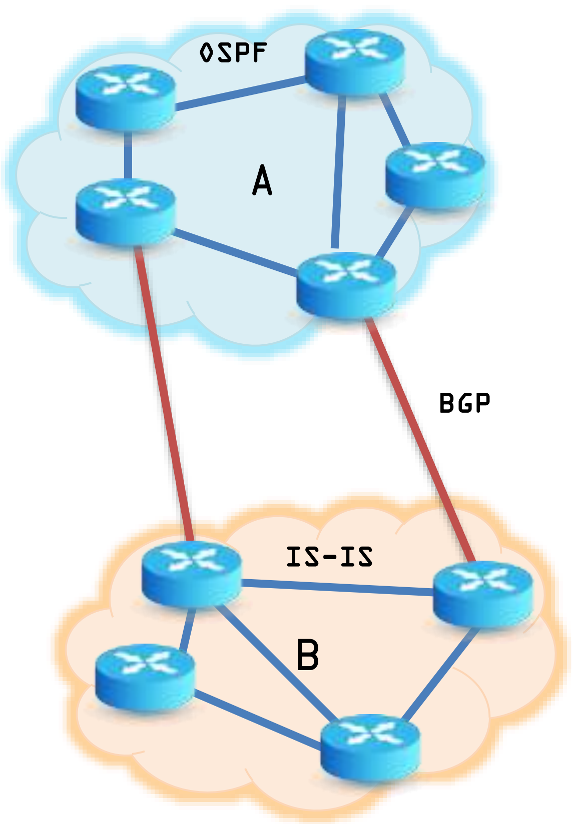


- Forwarding
  - Each network controls its forwarding decisions
  - Decide next hop for its traffic
  - Forwarding table stores information & configuration
- Routing
  - Establish end-to-end paths
  - Forwarding table may be default
  - Routing protocol can be used to determine the forwarding tables

Forwarding Tables

- *Static approach*
  - Administrator manually configures forwarding table entries
  - Advantages: More control; not restricted to destination-based forwarding
  - Disadvantages: Do not scale; slow to adapt to network failures
- *Dynamic approach*
  - Routers exchange network reachability information using routing protocols; Routers use this to compute best routes
  - Advantages: Rapidly adapt to topology changes; scalable
  - Disadvantages: Can be made to scale well; Complex distributed algorithms; Consume CPU, Bandwidth, Memory
  - Debugging can be difficult
- In practice : a mix of these. Static mostly at the “edge”

Routing Protocols & Architecture



- Interior Gateway Protocol (IGP): routing protocol within an organization
  - Metric based: OSPF, IS-IS, RIP, EIGRP (Cisco)
  - Based on minimum weighted path
- Exterior Gateway Protocol (EGP): routing protocol across different organizations
  - Policy based: BGP (Border Gateway Protocol)
  - Routing scope of BGP is the entire Internet

Approaches of Distributed Routing

- *Link State* (e.g. OSPF, IS-IS)
  - Topology information is flooded within the routing domain
  - Best end-to-end paths are computed locally at each router.
  - Best end-to-end paths determine next-hops.
  - Based on minimizing some notion of distance
  - Works only if policy is shared and uniform
- *Vectoring* (e.g. RIP, BGP)
  - Each router knows little about network topology
  - Only best next-hops chosen by each router for each destination
  - Best end-to-end paths result from composition of all next-hop choices
  - Not require any notion of distance
  - Not require uniform policies at all routers

Classification

	Link State	Vectoring
IGP	OSPF IS-IS	RIP
EGP		BGP

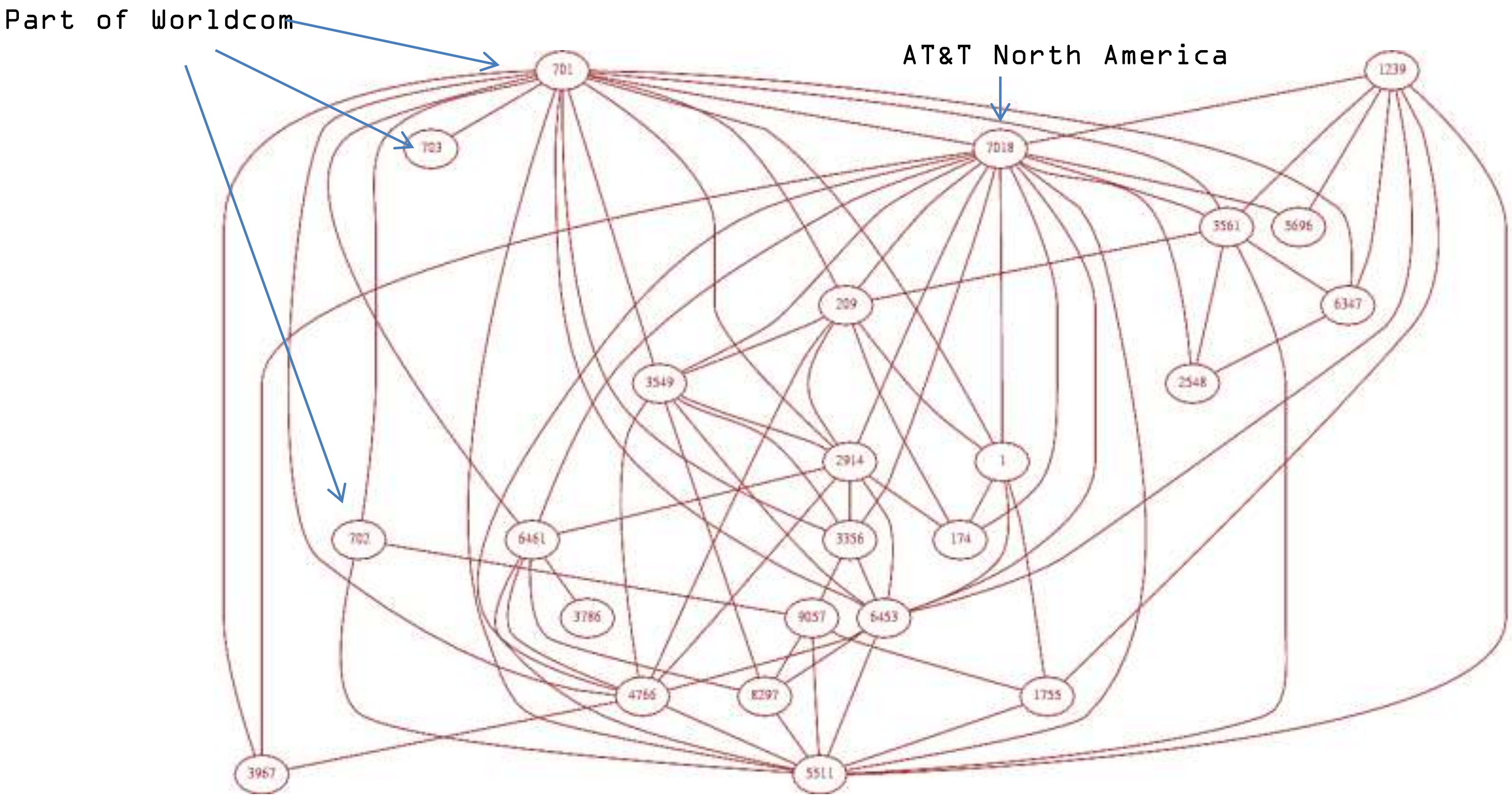
Autonomous Systems

- A collection of physical networks grouped together using IP
  - Having a unified administrative routing policy (e.g. a single organization)
  - Can hide its internal network topology (encapsulation)
  - Examples
    - Campus networks, Corporate networks
    - ISP Internal networks
- An autonomous system is an autonomous routing domain that has been assigned an *Autonomous System Number* (*ASN*)
  - *"… the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it"*
    RFC 1930: Guidelines for creation, selection, and registration of an Autonomous System

AS Numbers (ASNs)

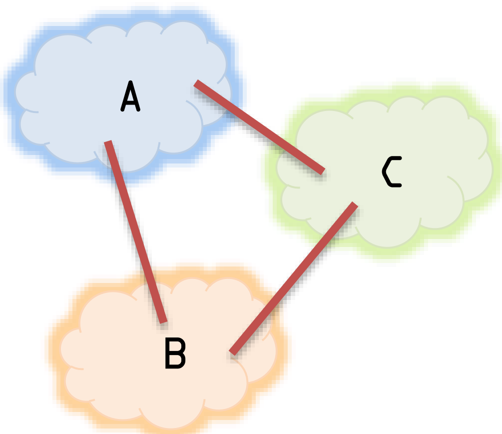
- ASNs are 16 bit values (64512 through 65535 are “private”)
- Currently over 11,000 in use
- Examples
  - Genuity (formerly known as BBN): 1
  - MIT: 3
  - Harvard: 11
  - UC San Diego: 7377
  - AT&T: 7018, 6341, 5074, …
  - UUNET: 701, 702, 284, 12199, …
  - Sprint: 1239, 1240, 6211, 6242, …
- ASNs represent units of routing policy

AS Graph: Example

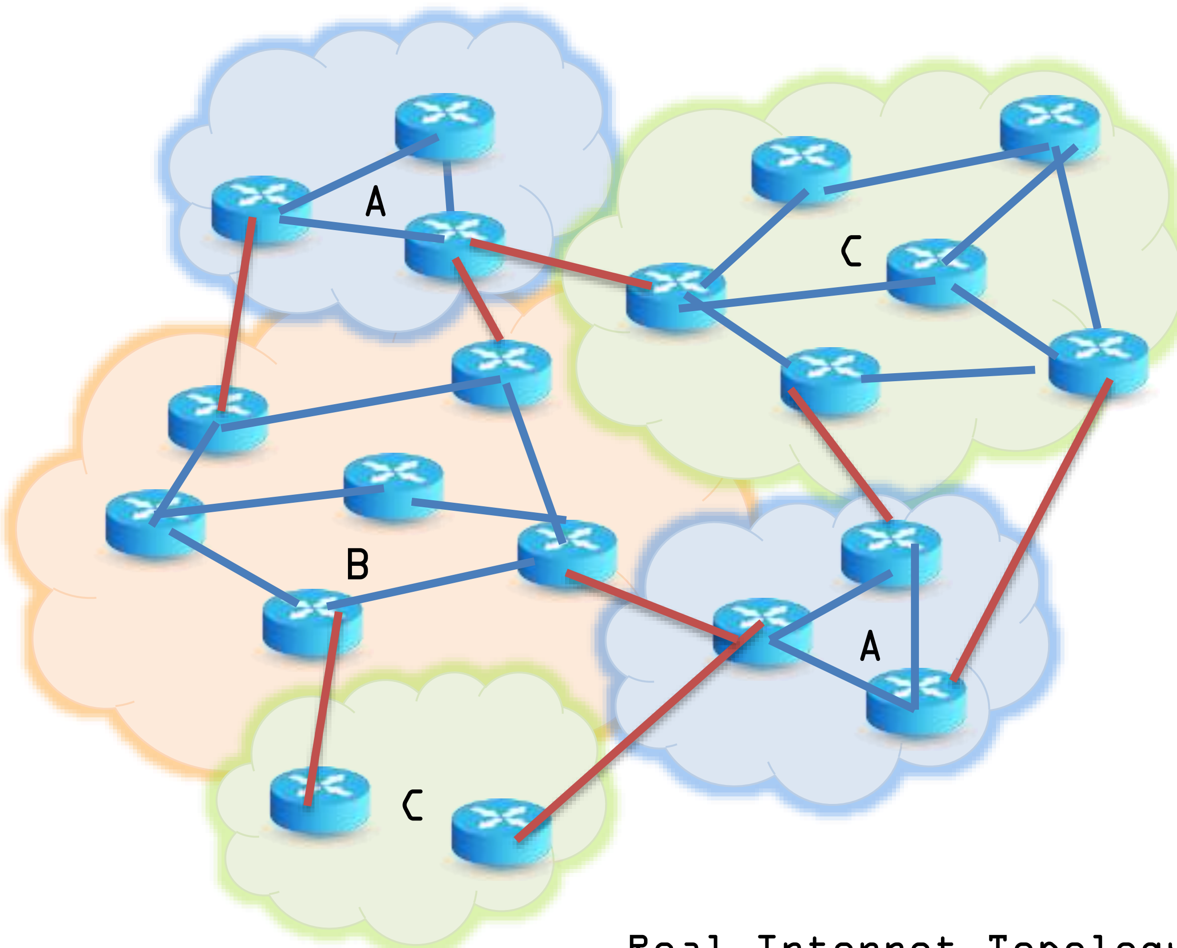


- The subgraph showing all ASes that have more than 100 neighbors in full graph of 11,158 nodes. July 6, 2001. Point of view: AT&T route-server

AS Graph ≠ Internet Topology



AS Graph

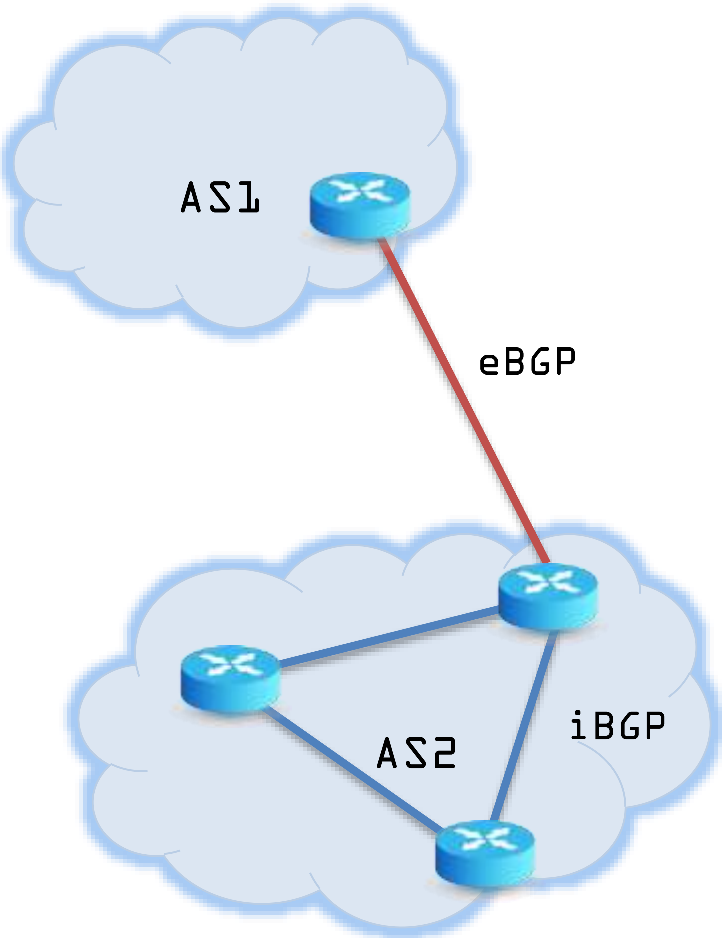


Real Internet Topology

BGP-4

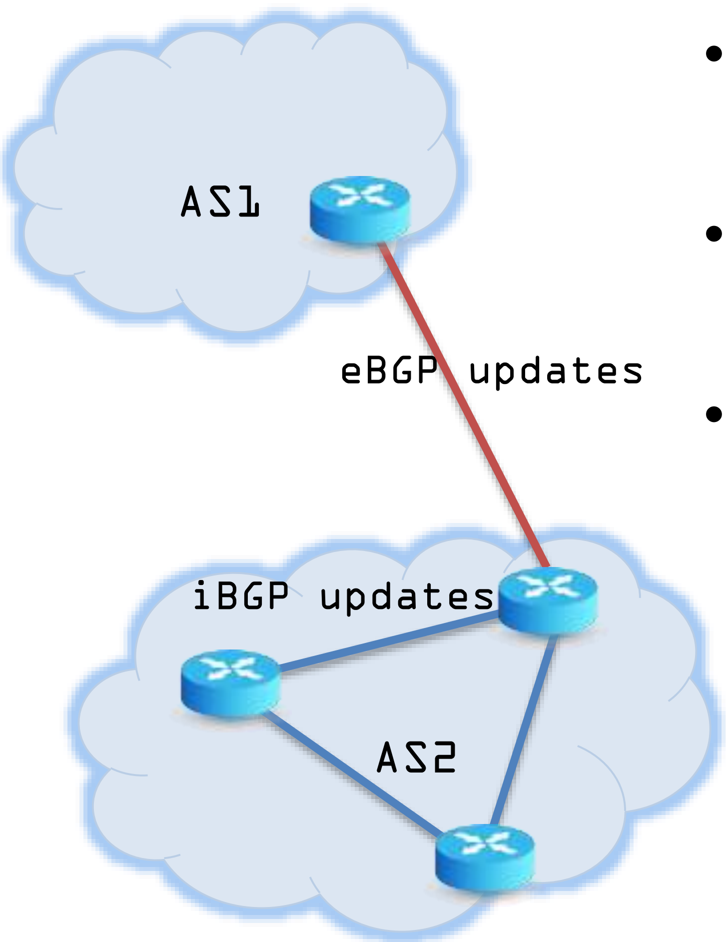
- BGP-4: current operational version of Border Gateway Protocol
- Operate on AS level graph
- Policy-based routing protocol
- The de facto EGP of today's global Internet
- Relatively simple protocol, but configuration is complex and the entire world can see, and be affects by misconfigurations
  - 1989 : BGP-1 [RFC 1105]
  - Replacement for EGP (1984, RFC 904)
  - 1990 : BGP-2 [RFC 1163]
  - 1991 : BGP-3 [RFC 1267]
  - 1995 : BGP-4 [RFC 1771]

BGP Neighbor Relationships



- BGP is divided into two parts
  1. *eBGP*: protocol with external neighbor in a different Autonomous System
     - Functions: announce and withdraw forwarding decisions in AS level
  2. *iBGP*: protocol with internal neighbor in the same Autonomous System
     - Functions: communicate with internal neighbor to consolidate routing operations

BGP Messages



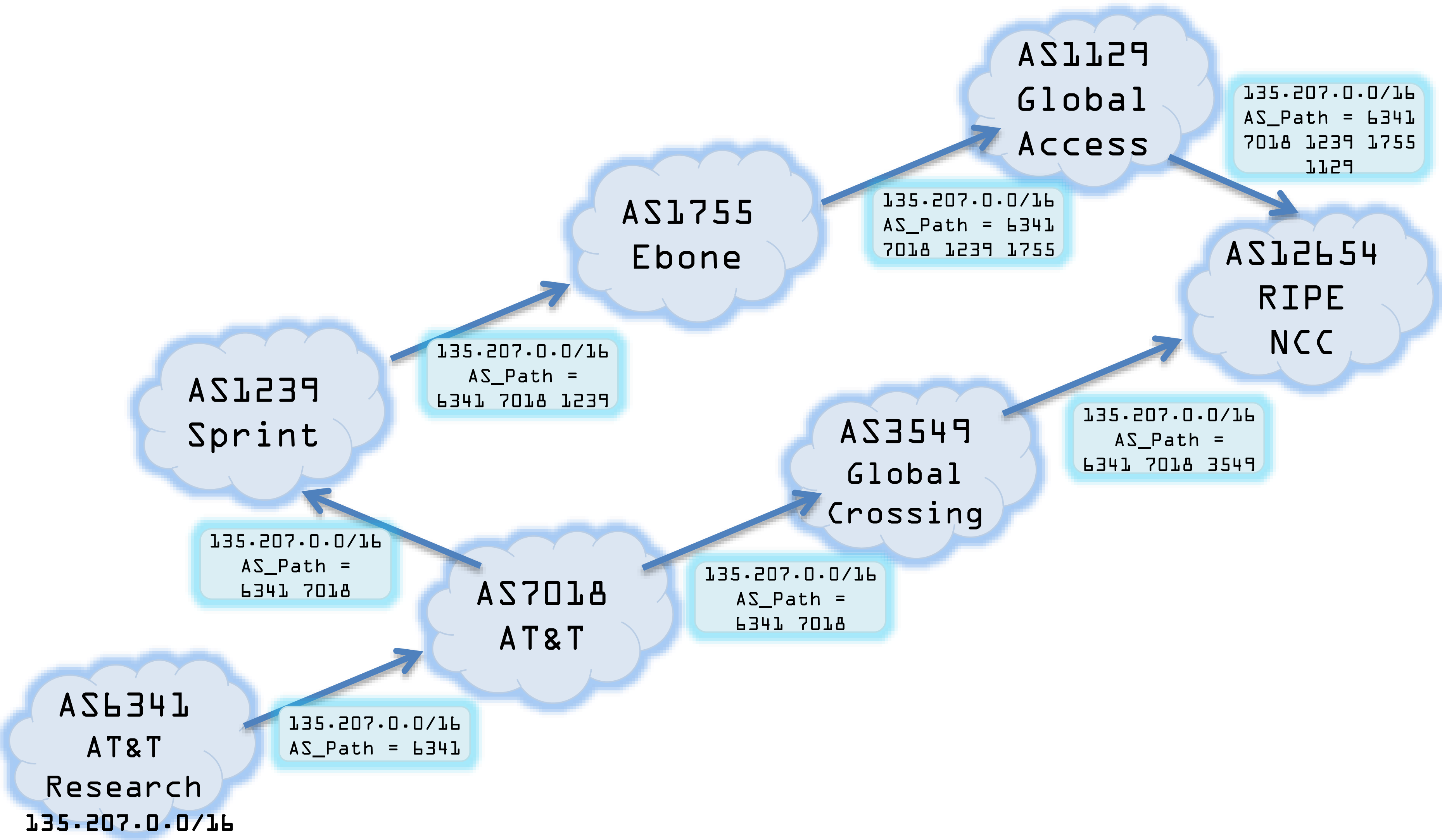
- `Open` : Establish a peering session
- `Keep Alive` : Handshake at regular intervals
- `Notification` : Shuts down a peering session
- `Update` : Announcing new routes or withdrawing previously announced routes

`announcement = IP prefix + attribute values`

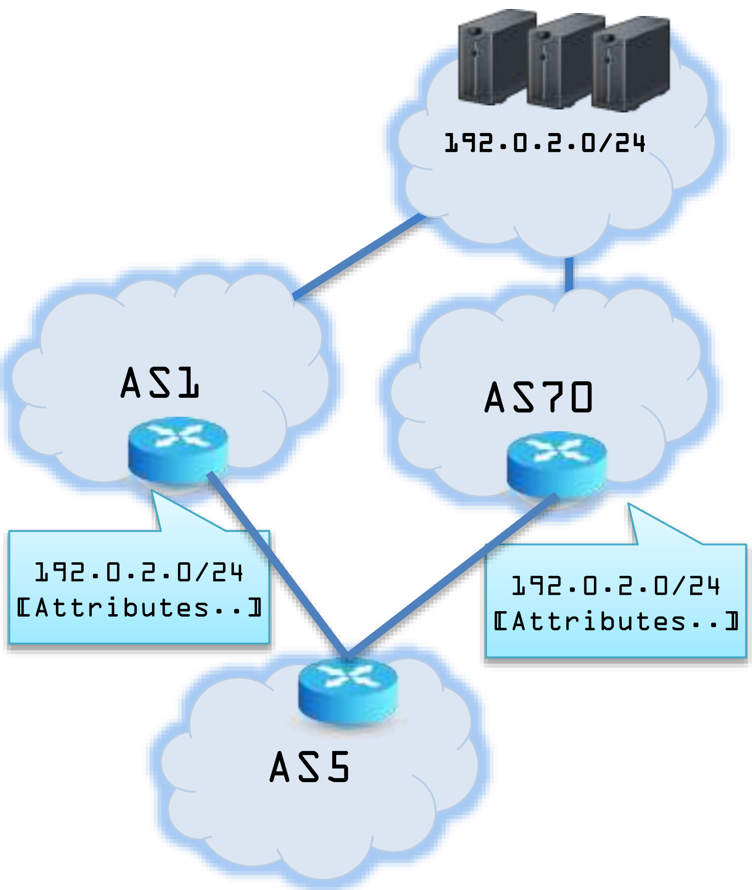
BGP Attributes

Value	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MED	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	
12	ADVERTISER	[RFC1863]
13	RCID_PATH / CLUSTER_ID	[RFC1863]
14	MP_REACH_NLRI	[RFC2283]
15	MP_UNREACH_NLRI	[RFC2283]
16	EXTENDED COMMUNITIES	
...		
255	reserved for development	

Example: AS_Path Attribute



How to Select Best Routes

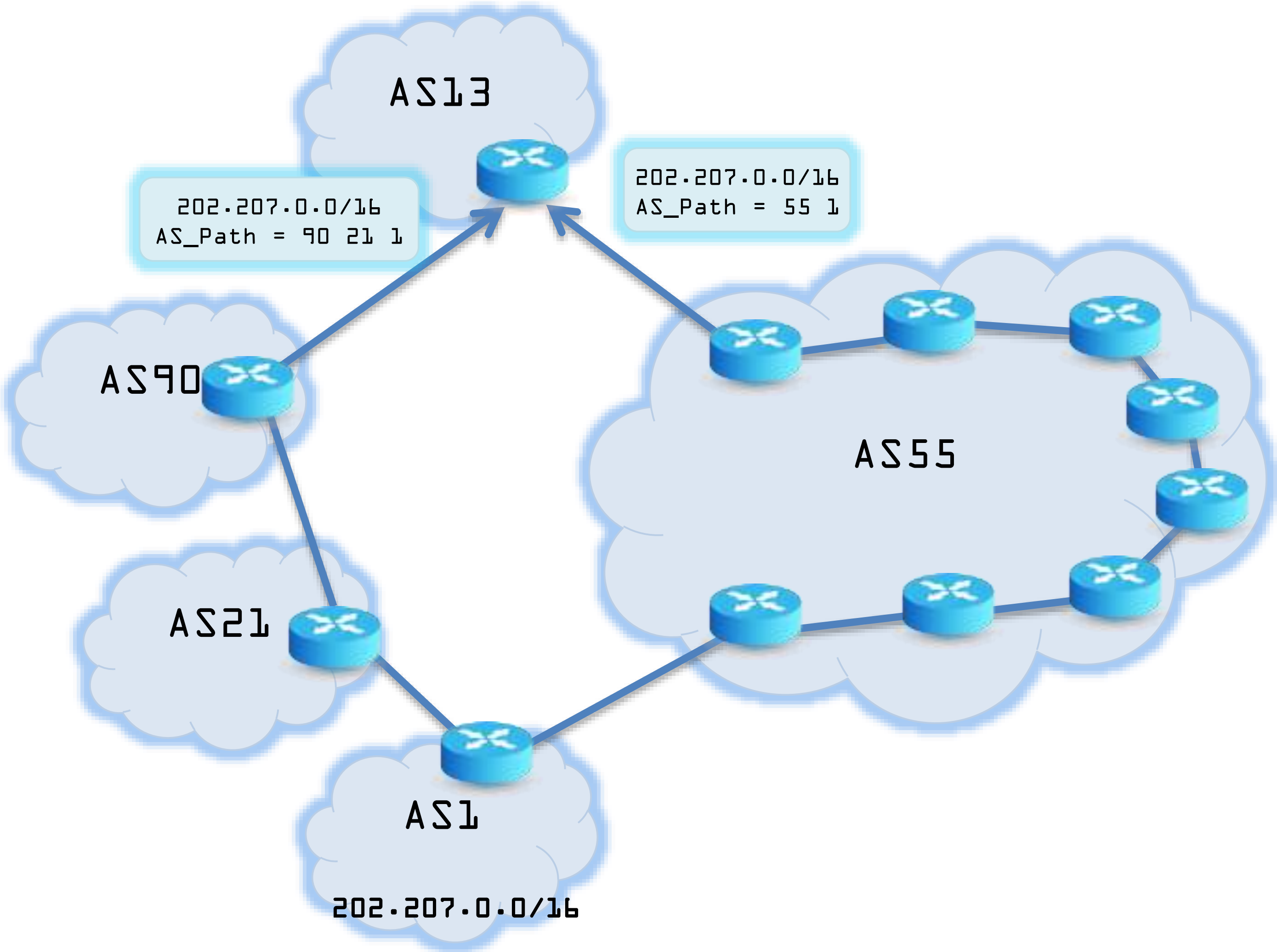


- A BGP router can receive multiple announcements of routes the same destination (IP prefix)
- At most a single route is selected, based on
  - Attributes in announcements
  - Internal route selection policy
  - There is no uniform routing policy across different Autonomous Systems

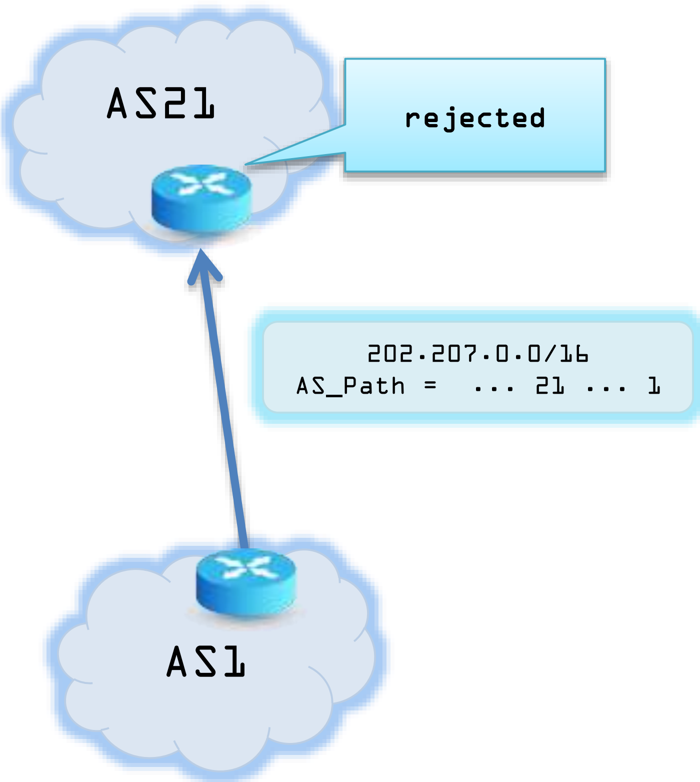
Possible Route Selection Policy

Highest internal preference	**Enforce relationships**
Shortest `AS_Path` Lowest `MED` i-BGP < e-BGP Lowest IGP cost to BGP egress	**Traffic engineering**
Lowest router ID	**Default tie-breaking**

Shorter AS_Path ≠ Shortest Path

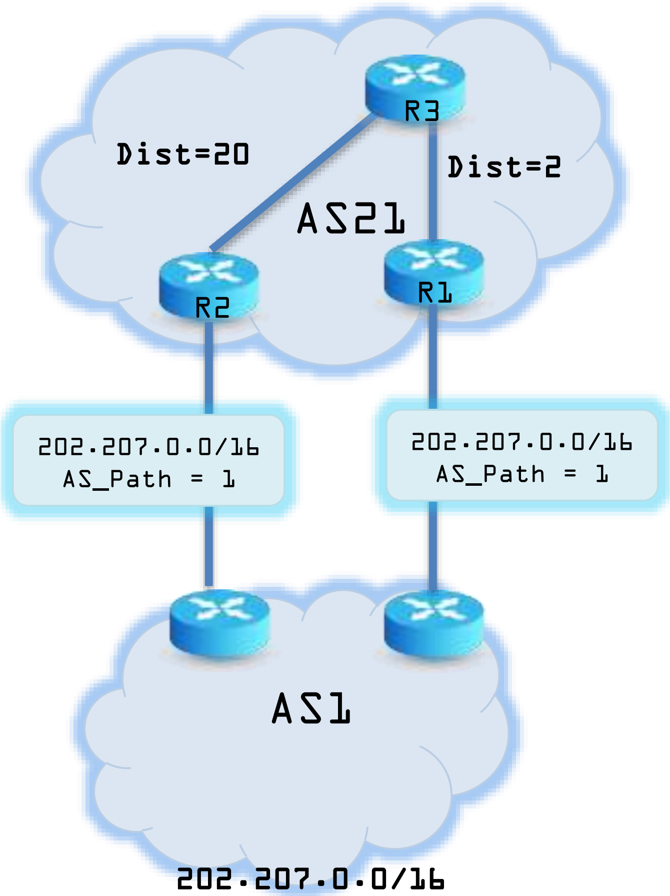


Loop Prevention



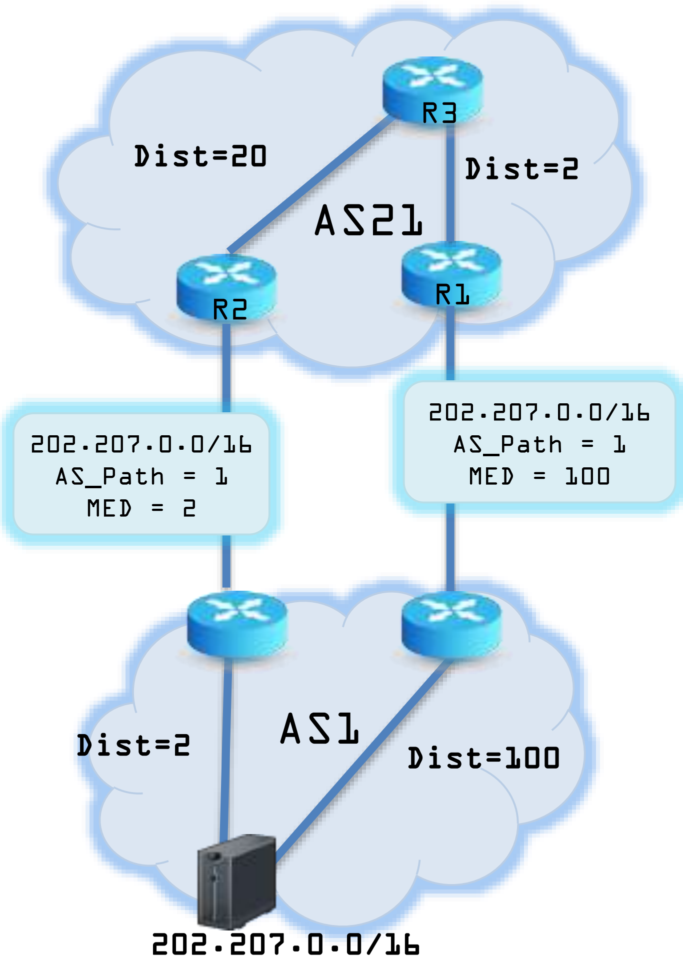
- A loop occurs when Autonomous System accepts route that traverses itself
- To prevent loop, Autonomous System rejects route announcement that contains its AS number

Hot Potato Routing



- There can be multiple route through an Autonomous System to reach the same destination
- Which path is selected?
- *Hot potato routing*: get traffic off of your network as soon as possible
- Minimize the distance traversed within the Autonomous System
- Example: R3 should selects path to R1

Hot Potato Routing

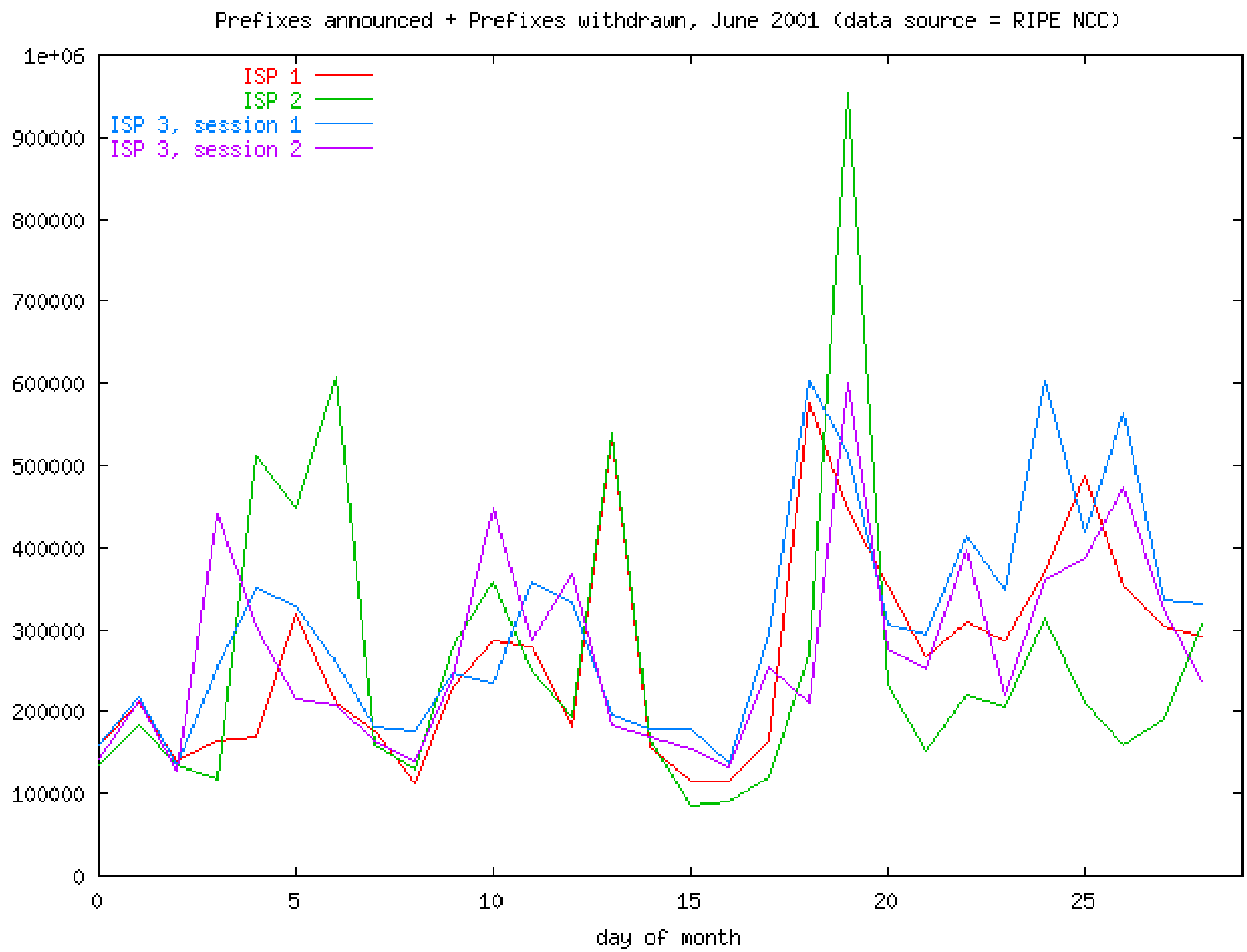


- Hot potato routing cannot optimize the path selection through other AS
- Some Autonomous Systems may be willingly to declare its internal path distance, if there are multiple paths through it to the same destination
- Multi exit description (MED) can convey the internal information to the next Autonomous Systems
- Multi exit description may not reveal the true internal path distance

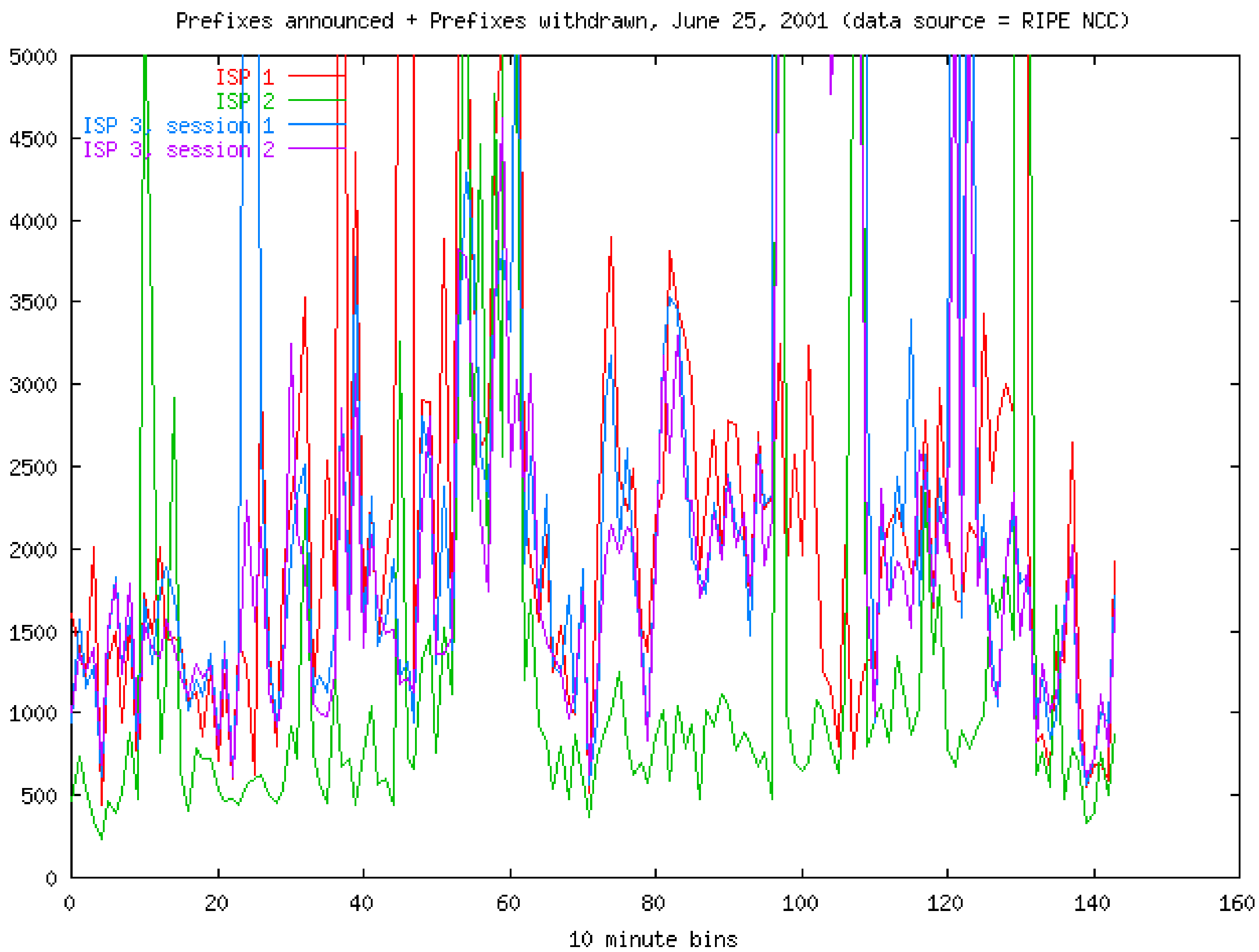
BGP Dynamics

- Why are BGP updates?
  - Misconfiguration (top reason)
  - Traffic engineering
  - BGP exploring many alternate paths
  - Software bugs in implementation of routing protocols
  - BGP session resets due to congestion or lack of interoperability
  - IGP instability exported by use of MEDs or IGP tie breaker
  - Sub-optimal vendor implementation choices
  - Bad policy
- How many updates are flying around the Internet?
- How long does it take routes to change?

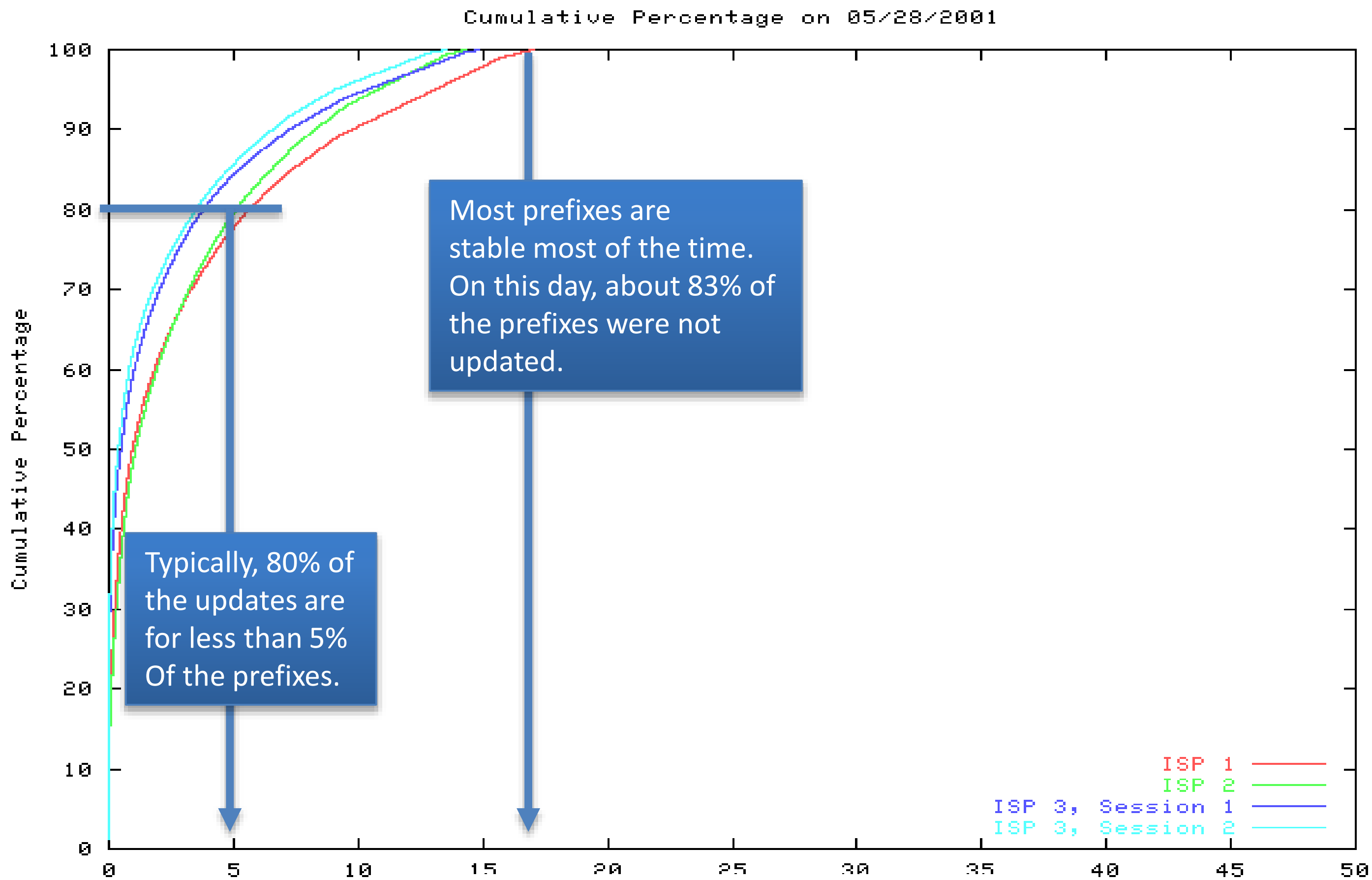
Daily Update Count



Route Flapping



Route Flapping



Percent of BGP table prefixes

Data source: RIPE NCC

Suppress Updates

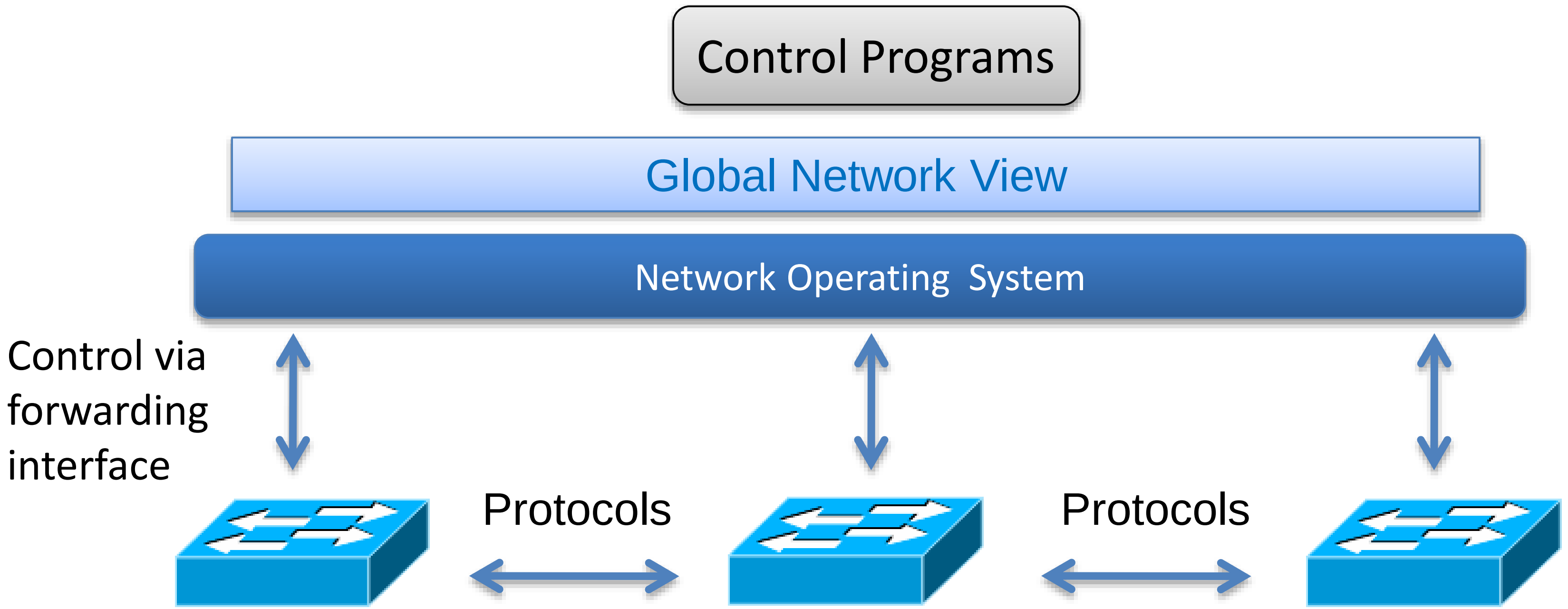
- Rate limiting on sending updates
  - Send batch of updates every Minimal Route Advertisement Interval seconds)
  - Default value is 30 seconds
  - A router can change its mind about best routes many times within this interval without telling neighbors
  - Effective in dampening oscillations inherent in the vectoring approach
- Route Flap Dampening
  - Punish routes for “misbehaving”
  - Must be turned on with configuration

Software-Defined Networking

- Today’s router and switches run on proprietary OSes
  - Cisco IOS, Juniper JunOS, Alcatel TimOS
  - Need to manage device-by-device locally and differently
- Software-defined Networking
  - Abstracting the control plane (routing policy, protocols) by a common API across devices and manufacturers
  - Enable global logically-centralized controller across the network
  - Increase interoperability and compatibility
  - Towards an OS for Network

Idea: An OS for Networks

Software-Defined Networking (SDN)

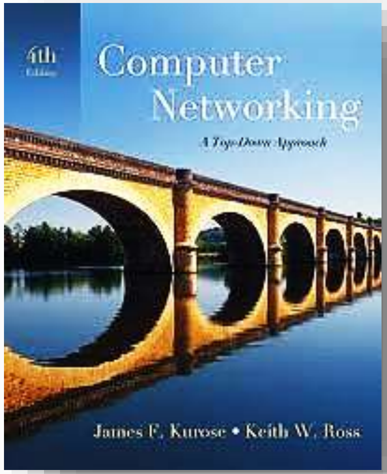


Software-Defined Networking

- Increasingly being supported by the industry
  - Juniper, NEC, HP, Netgear, …
  - Used for enterprise networks
  - Adopted by large data center: Google
  - Large-scale adoption in near future
- Will SDN replace BGP?
  - Possibly
  - However, ISPs will retain the control and visibility of their policies

References

- Computer Networking
  James F. Kurose and Keith W. Ross; Pearson Addison-Wesley
  - Chapter 4.6



- BGP Tutorial , *Tim Griffin*
  - http://www.cl.cam.ac.uk/~tgg22/talks/BGP_TUTORIAL_ICNP_2002.ppt
- Software-defined Networking
  - https://www.coursera.org/course/sdn