Interdomain Routing: 101

Ron Bonica

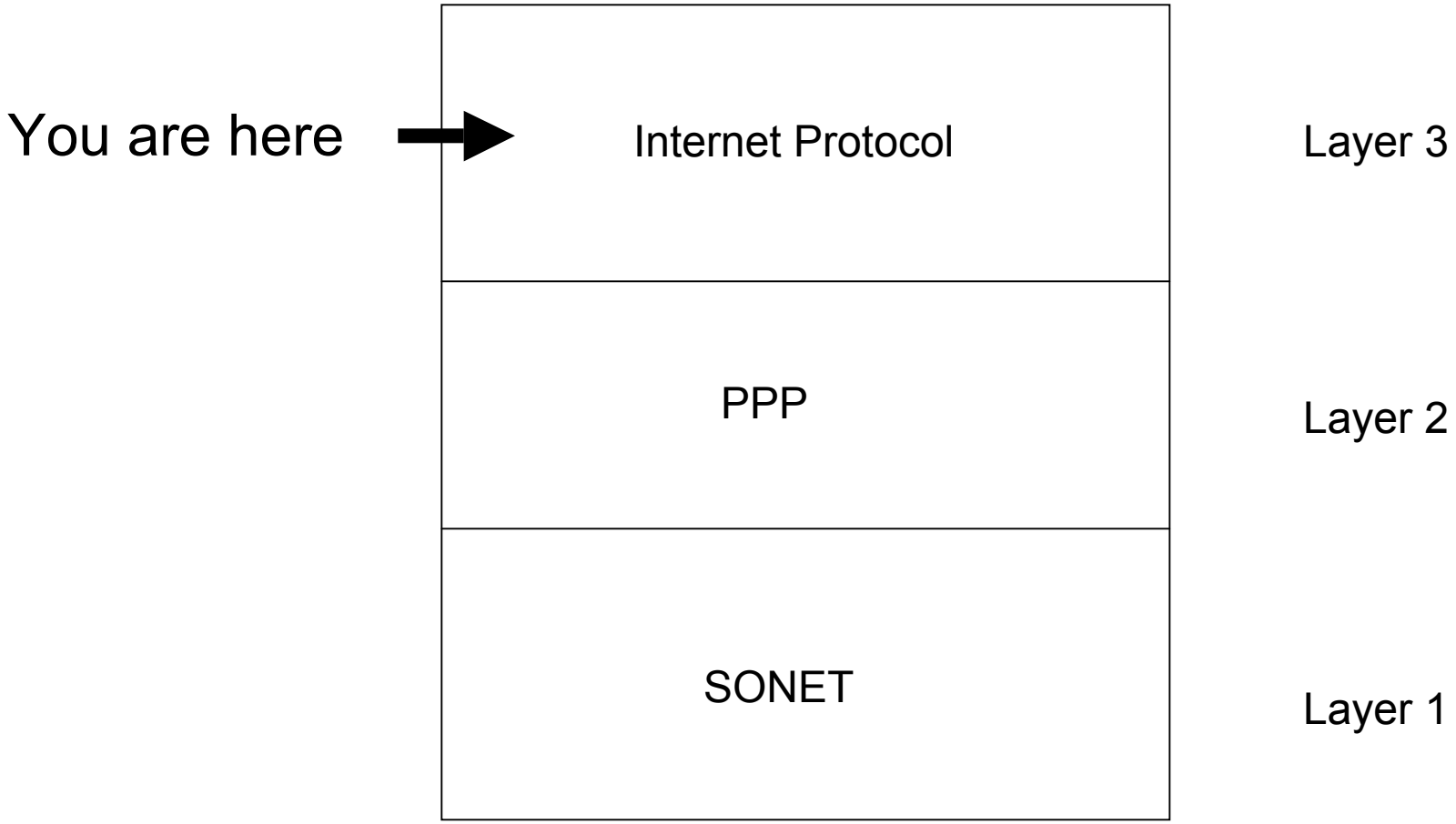
IETF 74

Background Information

Some Essential Terminology

- Network devices maintain ***interfaces*** between one another
- The number of interfaces that a device maintains can vary from one to thousands
- Interface Layer 1, Layer 2 and Layer 3 characteristics

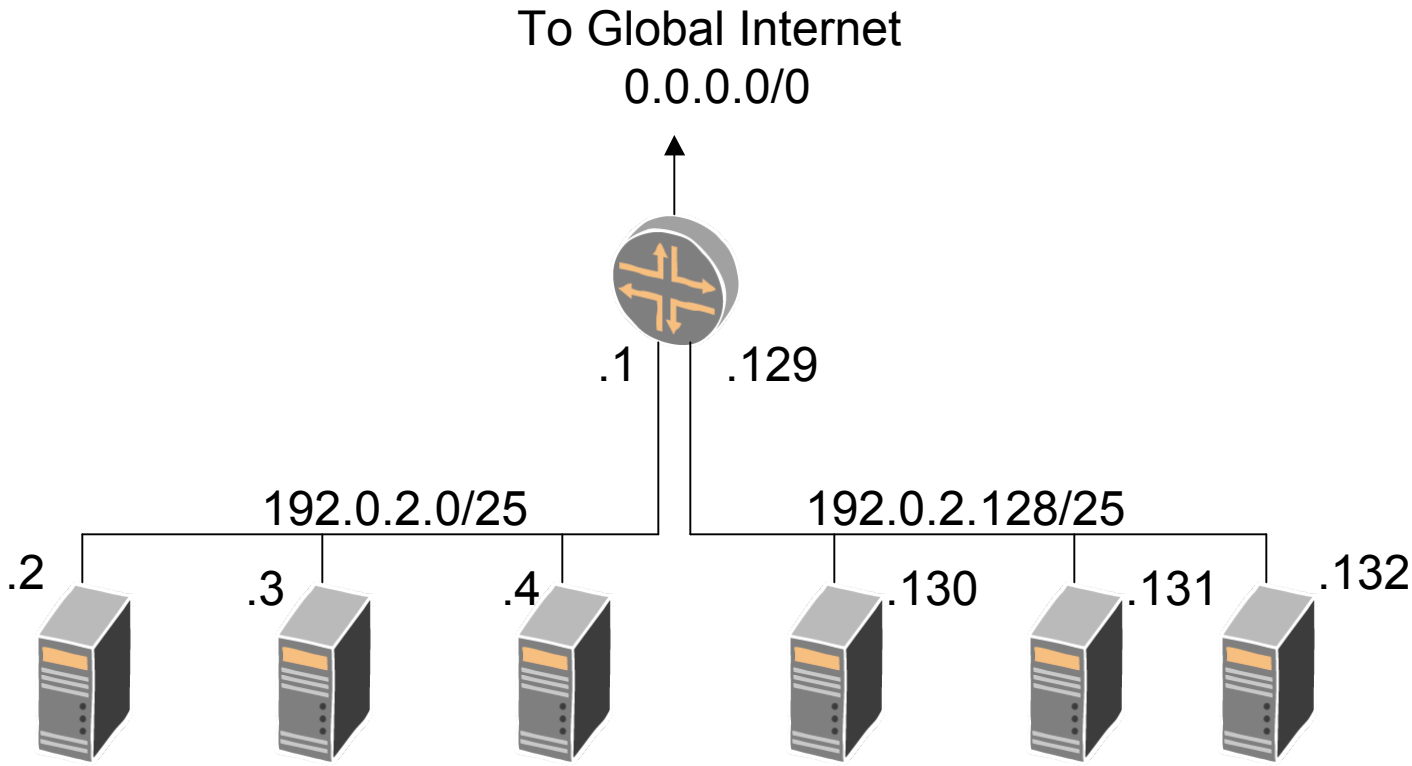
Interface Example



Interface Addressing

- A unique ***IP address*** is assigned to each interface
  - IPv4: 32 bits
  - IPv6: 128 bits
- Globally unique versus various forms of private addressing
- Interfaces can be group together into sub-networks

IP Addressing



IP Forwarding

- An IP datagram arrives on a router interface
- Router identifies ***IP Next-hop***
- Router forwards packet through interface to IP Next-hop
- Forwarding is on a hop-by-hop, packet-by-packet basis

IP Lookup

- Inputs
  - ***IP destination address***
  - ***Forwarding Information Base (FIB)***
- First Order Output
  - IP Next-hop
- Based on longest match

IP Lookup Example (192.0.2.131)

Destination	IP Nexthop
0.0.0.0/0	IP Address A
192.0.2.0/24	IP Address B
192.0.2.128/25	IP Address C
192.0.2.192/26	IP Address D

More About IP Lookups

- Sometimes, vendors get fancy and add additional parameters to the IP lookup
  - QoS-based routing
  - Source-based routing
  - We aren't going to talk about these!
- Once a router has determined the IP next-hop, it may have to perform another lookup to determine what layer 2 framing is required to communicate with the IP next-hop

Routing Loops

- Very possible
  - Logical consequence of FIB sickness
- Very bad
  - Gobble bandwidth
- To some extent, effects are mitigated by IP TTL

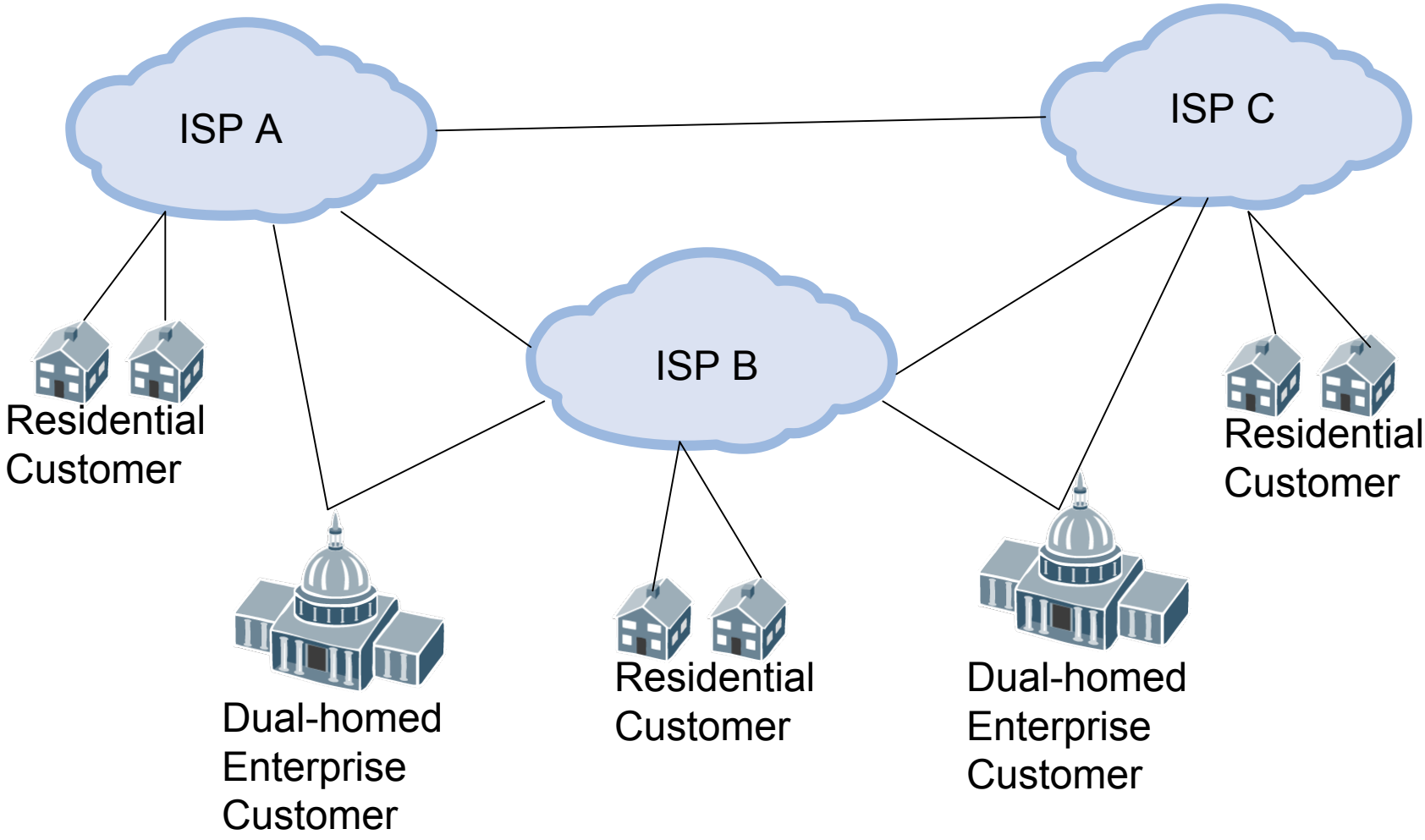
Where Does the FIB Come From

- Simple Case
  - At the edge
- Complex Case
  - An ISP core
  - Today's Topic
- The story is always about autonomy!

The Most Simple Case

- The router in my house
- Static configuration/factory defaults
  - Route to 0.0.0.0/0 through ISP gateway
  - Directly connected interfaces to all of the machines in my house

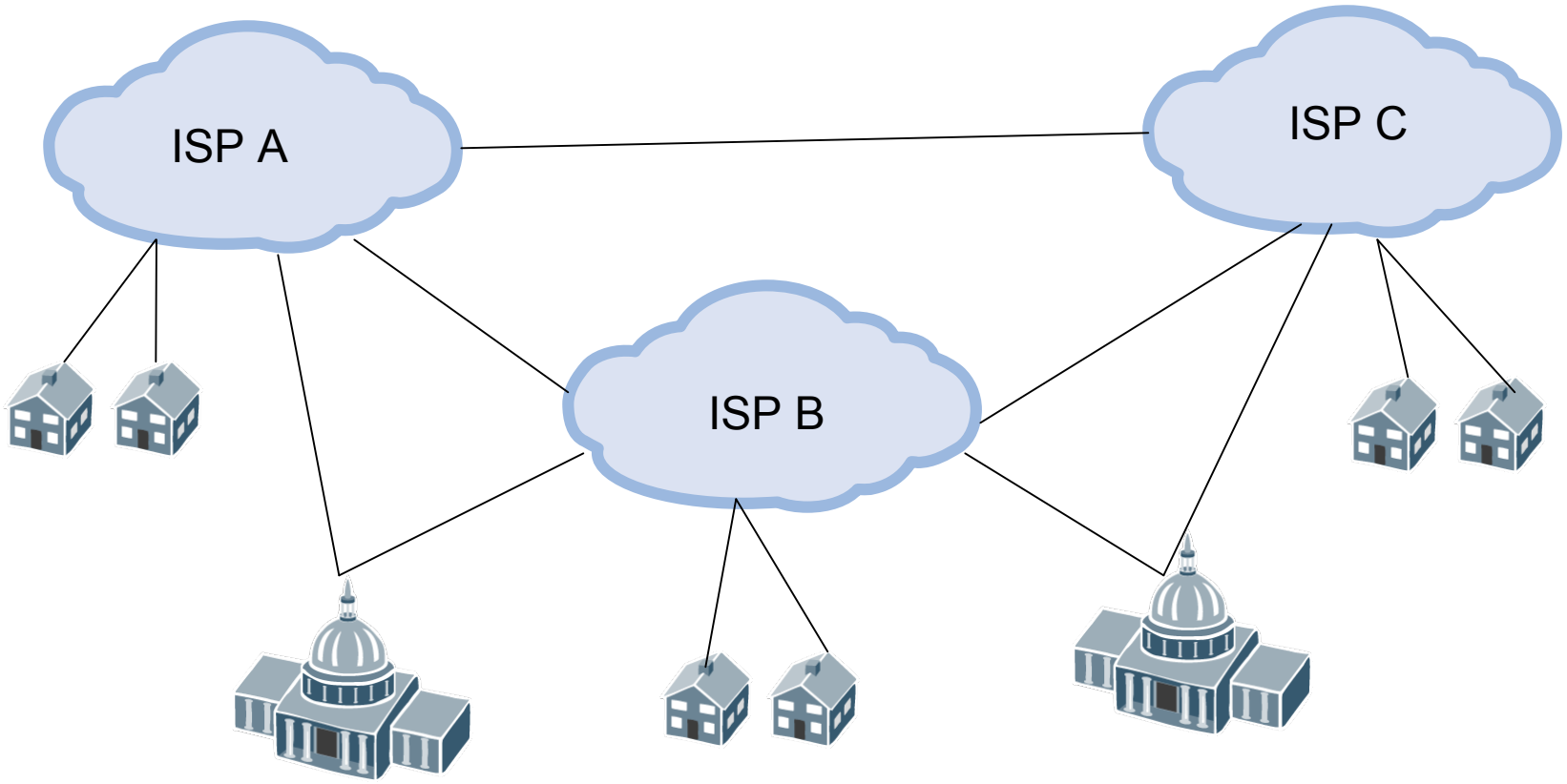
A More Complex Ecosystem



Autonomous Systems

- Each ISP operates an ***Autonomous System***
- Independent Routing Policy
  - Determine who gets connectivity to who
  - Determine how that connectivity can be provided
- Autonomous Systems negotiate connectivity with one another
  - But they don't dictate……

An Example of Routing Policy

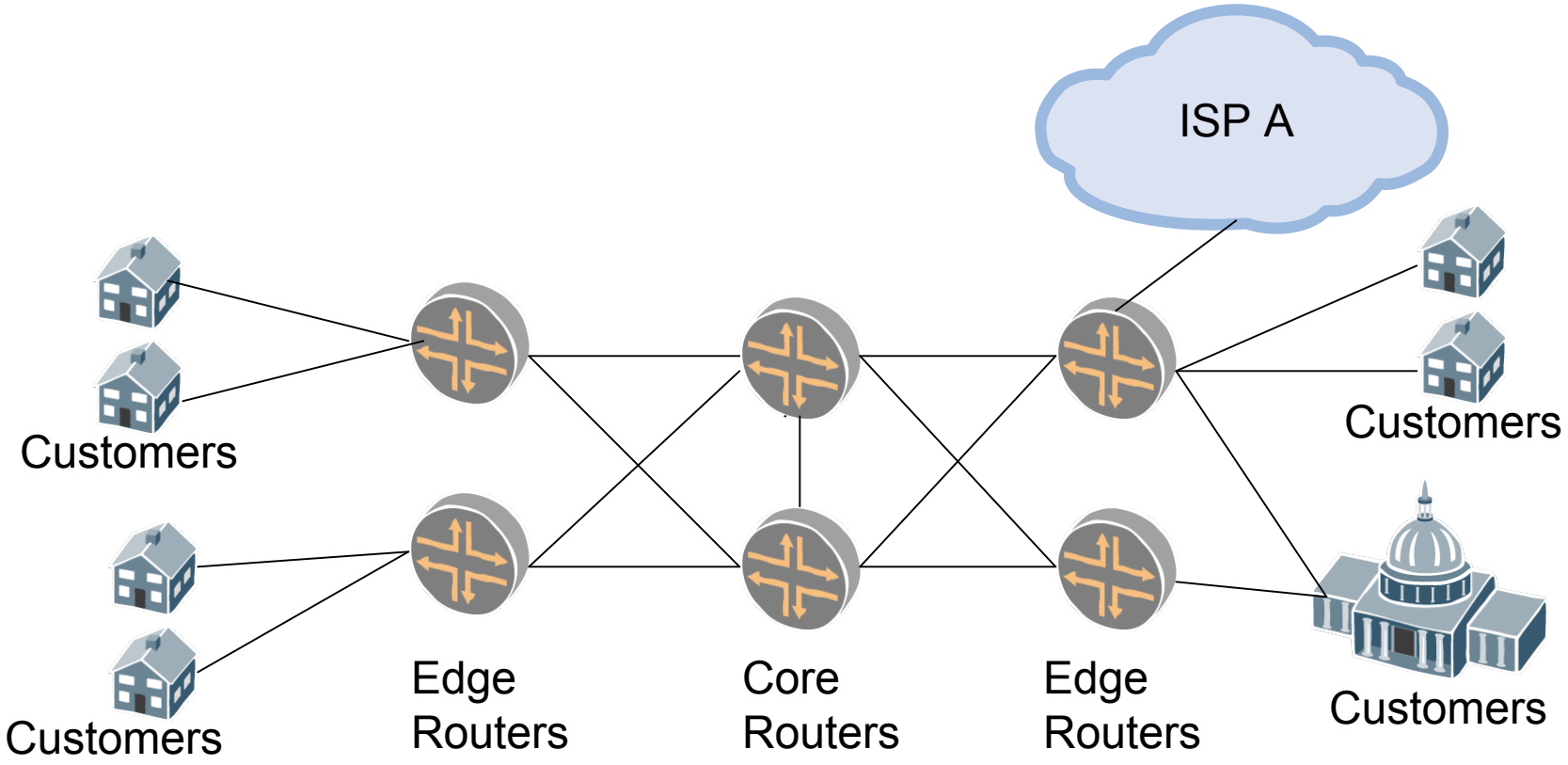


ISP B doesn’t want to provide transit between A and C

Tricks of the Trade

- If you wanted to defeat the policy implemented in the last slide, how would you do it?
  - Some methods require help from a customer of ISP B
  - Some do not

A Closer Look At ISP B



The Routing Problem

- Internal Routes
  - Router interfaces
- External Routes
  - Peers, peer's customers, peer's peers, peer's peer's customers, etc…..
- Customer Routes
  - Some have address space that you provide
  - Some provide their own address space

Peers and Customers

- Two senses of the word
- One implies a financial relationship
- Other has routing implications
  - Relative size of the networks and number of advertisements
- For the purpose of this talk, a peer is any large attached network and a customer is any small attached network, regardless of financial arrangements.

Distributing Internal Routes

- Requirements
  - Granular, fast convergence, loop free
- Intradomain Gateway Protocols
  - Opens Shortest Path First (OSPF)
  - Intermediate System to Intermediate System (ISIS)
- Link State Protocols

Link State Protocols

- Interface administratively assigned to IGP domain
  - Cost / metric
- IGP floods link state and link parameters
- Each node maintains identical copy of Link State Data Base (LSDB)
- Each node uses the LSDB to independently calculate the least cost path between itself and every other interface in the domain

Link State Limitations

- Don’t scale to infinity
- Not designed for interactions outside of administrative domain
  - A single misconfiguration can be deadly
  - Lack sophisticated policy control

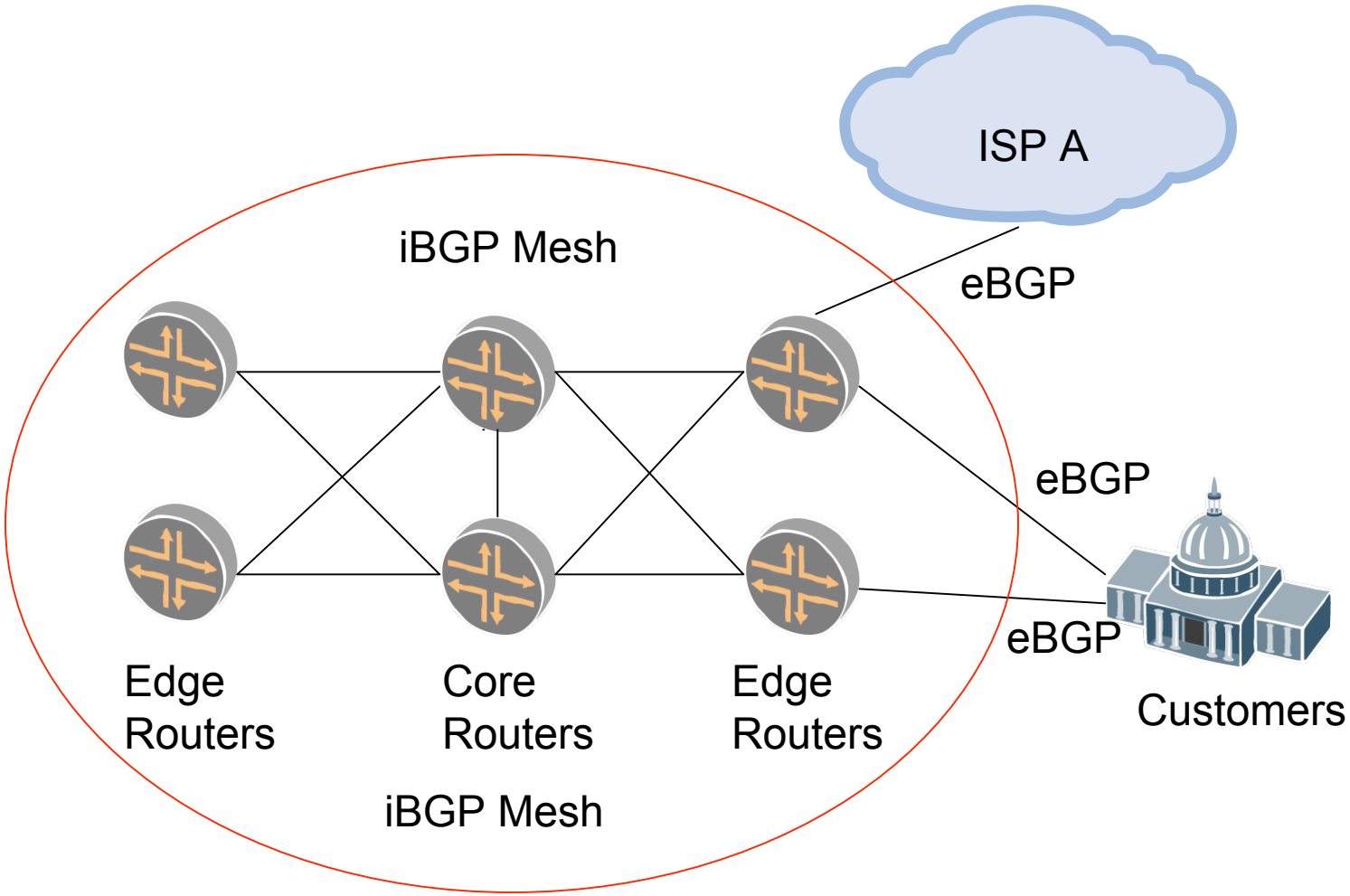
External Routes

- Border Gateway Protocol (BGP-4)
  - Maintained by IDR WG
- Distance Vector Protocol
- Horizons
- Protocol Messages
- Policy
- BGP Best Path

BGP Horizons

- All routers with AS maintain a full mesh of iBGP neighboring sessions
- Autonomous System Border Routers (ASBR) maintain eBGP neighboring sessions with external neightbors
  - In addition to iBGP neighbors
- When a route is received from an eBGP neighbor, it can be repeated to all neighbors
- When a route is received from an iBGP neighbor, it can be repeated to eBGP neighbors only

BGP Horizons



Limitting the iBGP Mesh

- Route Reflectors
  - Several routers configured as reflectors
  - Send iBGP updates to those and they reflect to other clients
- AS Confederations
  - Split AS into regions

BGP-4 Messages

- OPEN
  - Introductions, negotiate capabilities
- UPDATE
  - Here's the beef
- KEEPALIVE
  - Connection termination == route withdrawal
- NOTIFICATION
  - Soon to be followed by connection termination

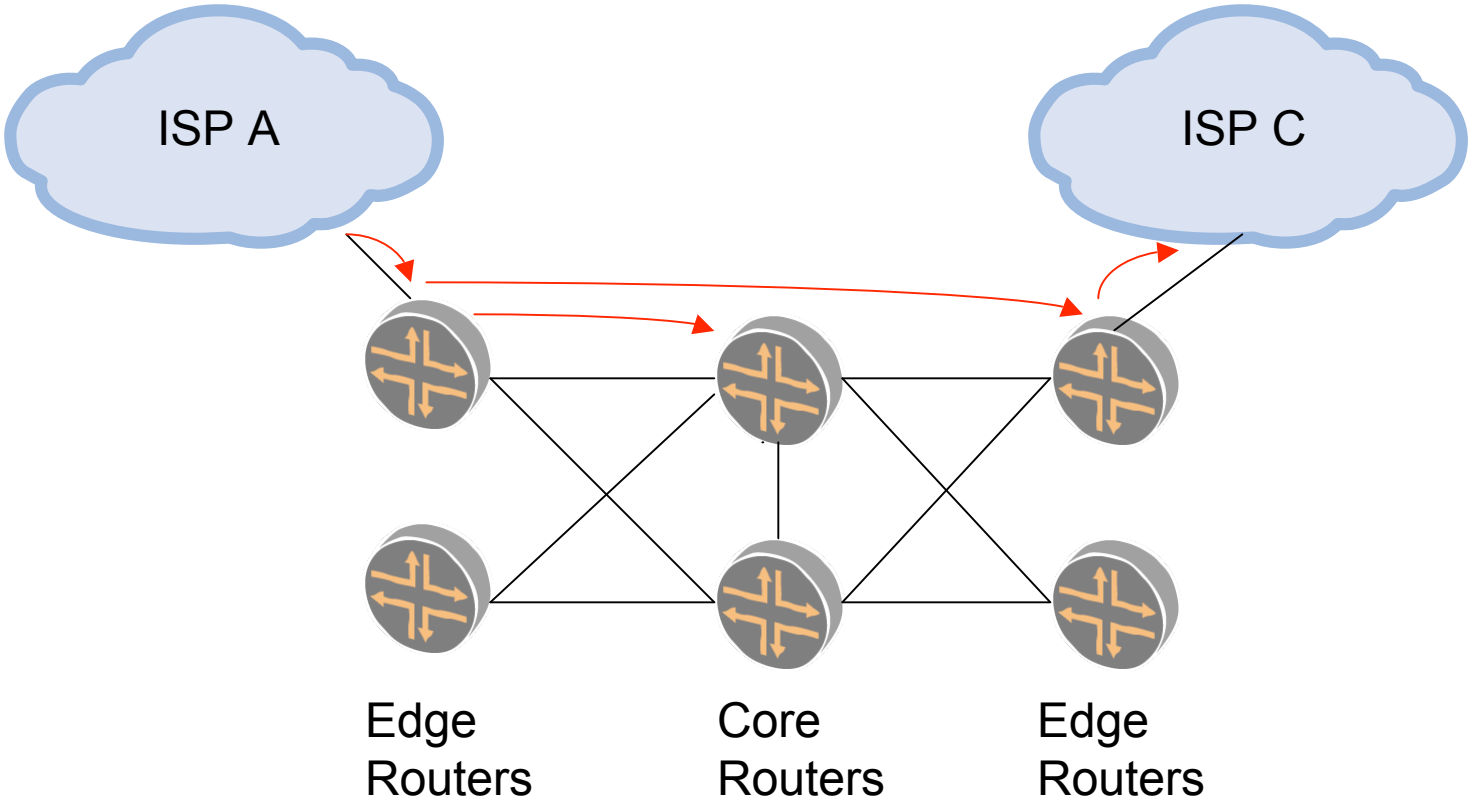
BGP-4 UPDATE Messages

- Prefix List
  - Address families
- Attribute List
  - Mandatory / well-know
  - Discretionary / well-known
  - Optional (transitive and non-transitive)

BGP-4 Attributes

- Next-hop
- AS PATH
- Origin
- Local Preference
- Multi-hop discriminator (MED)
- Atomic aggregator
- Aggregator

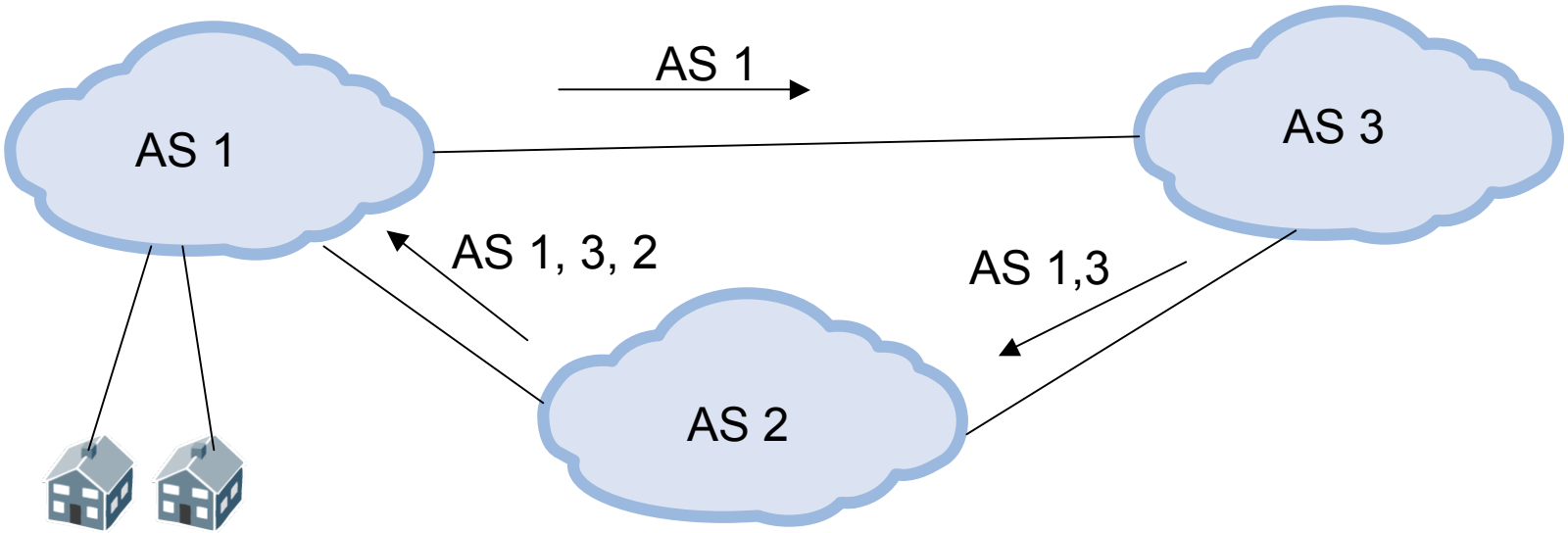
BGP Next-Hop Re-writes



BGP Free Cores

- In the last slide, trace the forwarding path?
- What routing information did the core router require?
- Is there away to avoid that requirement?

AS-PATH and Loop Detection

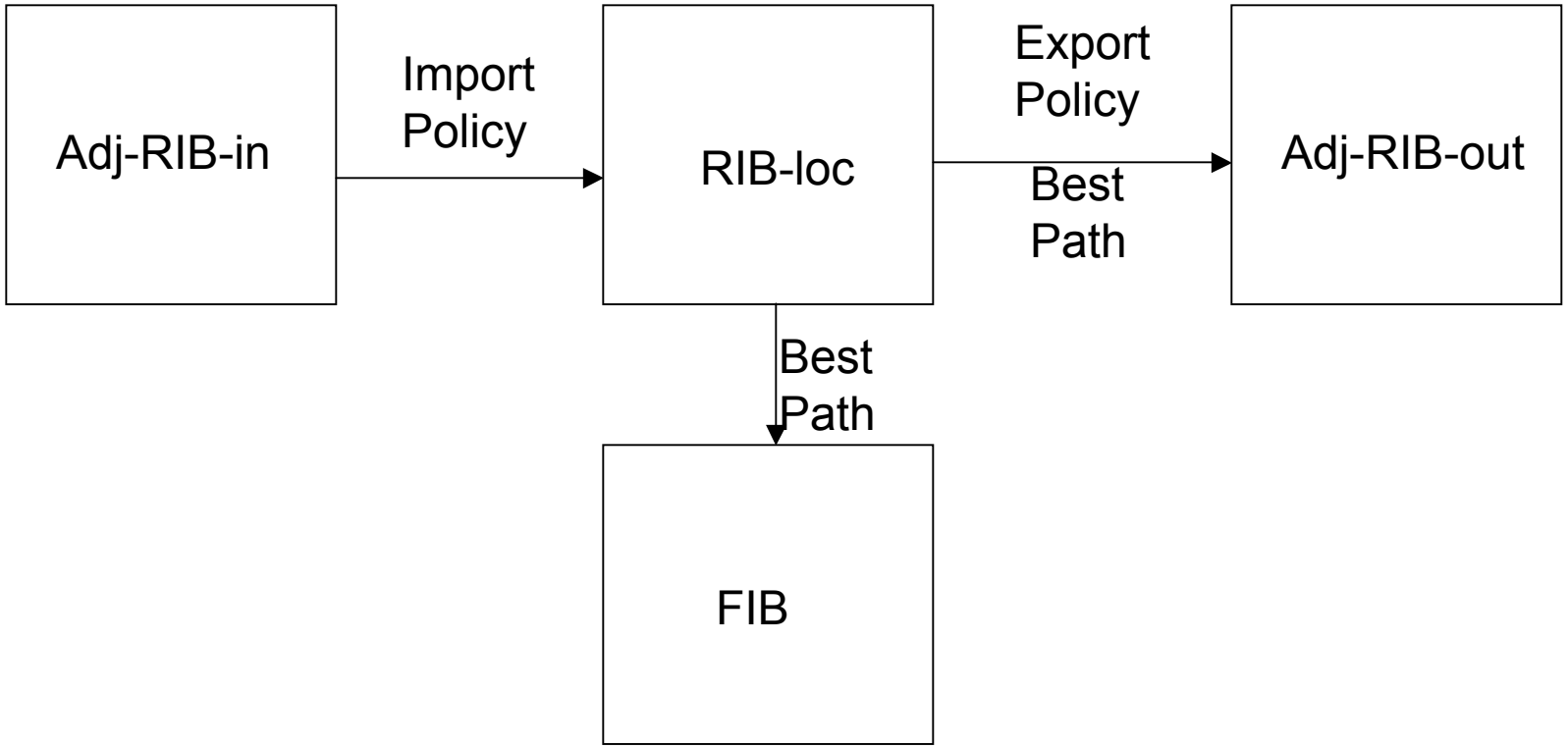


ISP B doesn’t want to provide transit between A and C

Public versus Private AS Numbers

- Public AS numbers assigned by IANA/RIRs
- Range of private AS numbers reserved for special situations (RFC 1930)
  - Example: Customer dual homed to single upstream as running BGP with that AS
  - ISPs typically strip private AS numbers from AS-PATH

BGP Policy


Adj-RIB-in
Import
Policy
RIB-loc
Export
Policy
Best
Path
Adj-RIB-out
Best
Path
FIB

BGP Best Path

- Tie breaker between equally specific prefixes
- Order of evaluation
  - Highest Local PREF
  - Shortest AS-PATH
  - Lowest ORIGIN number
  - Lower MED (from same neighbor)
  - eBGP over iBGP
  - Lowest cost to BGP next-hop
  - Highest BGP peer address

BGP Communities

- Standard versus extended
- Well-known versus locally significant
- Used in policy decisions
  - Return to example in which ISP B doesn't want to provide transit

Customer Routes

- Some are learned from BGP
  - Likely to be dual-homed
  - Likely to have provider independent address space
- Some are directly connected
  - Likely to be single homed
  - Likely to have provider supplied address space

Customer Routes Learned From BGP

- Very specific import policy
  - Permit: 192.0.2.0 exact; reject: all
  - May accept more specifics, but with restrictions
- Possibly remove Private AS number

Directly connected Customer Routes

- Choices regarding distribution to internal nodes
  - Import to IGP
  - Import to iBGP, but do not redistribute to eBGP peers
- Advertise aggregate only to eBGP peers

Import Policy Regarding Peers

- A good ACL is of the form (Permit: good stuff; Reject: all)
- When writing an import policy for a peer, there is too much good stuff
  - You don’t know what that stuff is
  - It changes too often
- This exposes you to risk

The Youtube Incident

- Somebody inappropriately advertises a /24 belonging to Youtube
- That customer's ISP doesn't deploy an input policy to catch the inappropriate advertisement
  - Accepts and re-advertises
- /24 is more specific than real advertisement
  - Oh bad!

The Best You Can Do

- Reject: stuff you recognize to be bad; accept: all
- Reject advertisements for
  - Your own resources
  - RFC 1918 address space
  - Bogons
  - Prefixes too specific (/25 or more specific)
- Limit number of advertisements accepted

Work on The Horizon

- draft-pmohapat-sidr-pfx-validate
  - RIRs maintain accurate registry of prefixes and announcing Autonomous Systems
  - Operators build peering policy from that data base
- Draft-ietf-idr-flow-spec
  - BGP used to distribute fowarding plane filters