# 6.829 BGP Recitation

Rob Beverly <rbeverly@mit>

September 29, 2006

# Addressing and Assignment

# Area-Routing

- Review…
- Why does Internet Scale?
  - Hierarchical Addressing
- How are addresses assigned?
- Classfull Addressing:
  - e.g. class A -> first bit 0, 7 bits network, 24 bits host
  - What's wrong with classes?
- Classless Interdomain Routing (CIDR)

# CIDR

- Stop-gap measure to prevent:
  - Address depletion
  - Route table growth
- Arbitrary network boundaries (not byte)
- Allows for proper sizing (not just 2^{8,16,24})
- Allows for aggregation
- Stroke Format: prefix/mask
- e.g. 18.0.0.0/8

# CIDR

- Example:
  - – 198.61.4.0/24 (class C)
  - – 198.61.5.0/24 (class C)
  - – Aggregate as: 198.61.4.0/23
- What about:
  - – 198.61.3.0/24 (class C)
  - – 198.61.4.0/24 (class C)
  - – Can this be aggregated as: 198.61.3.0/23? No!
  - – 3 = (binary) 00000011
  - – 4 = (binary) 00000100
  - – Differ in first 7 bits, so cannot aggregate

# Routing Nomenclature

- We use lots of acronyms, keep them straight:
  - – IGP: interior gateway protocol,
    - e.g. OSPF, ISIS, RIP
    - Optimized for: Shortest Path, loop-free
  - – EGP: exterior gateway protocol,
    - e.g. BGP
    - Optimized for: scalability, policy
  - – BGP types:
    - iBGP: internal BGP
    - eBGP: external BGP
- iBGP != IGP

## Brief Tangent: AS & IP Assignment

- Useful information for pset and debugging
- Who assigns IP addresses?
  - ARIN to regional registries (RIRs) who subdelegate
  - Lookup: `athena$ whois –h whois.arin.net 18.26.0.25`
- Who assigns AS numbers?
  - ARIN
  - Range? 2^16
  - Lookup: `athena$ whois –h whois.arin.net “AS3”`
- Who maintains IP->AS (or AS->IP) mapping?
  - Not centralized
  - Lookup: routing table

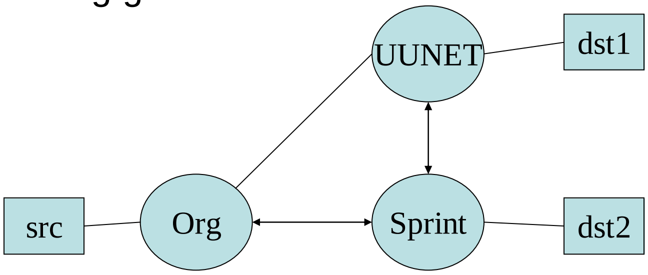
## BGP Policy

# BGP

- Autonomous System Numbers (ASNs)
- Routing preference:
  - Customers: advertise all routes to customers, import their routes
  - Peers: advertise my customers to my peers, import their routes
  - Providers: advertise my customers, import their routes

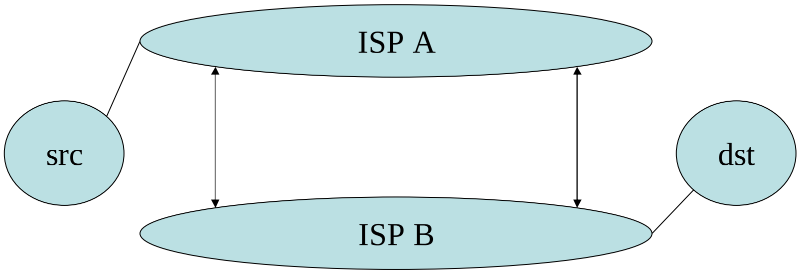
# BGP Policy

- Motivation for peering rules
- Org peers with Sprint
- Sprint peers with UUNET
- Org gets transit from UUNET



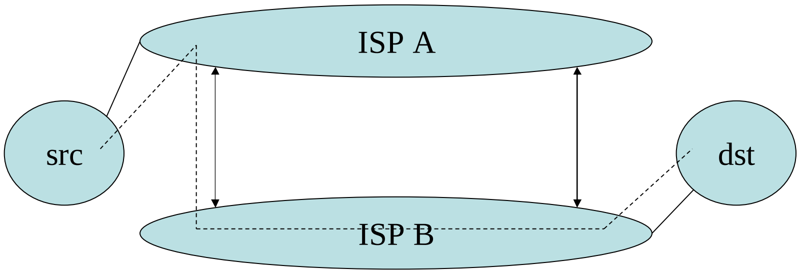
Why would Sprint not advertise UUNET routes to Org?

# BGP by Example: Hot-Potato



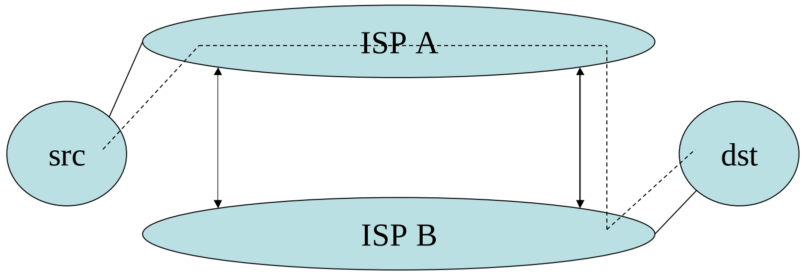
ISP A and B peer on both east and west coasts

# BGP by Example: Hot-Potato



Consider src to dst conversation

## BGP by Example: Hot-Potato

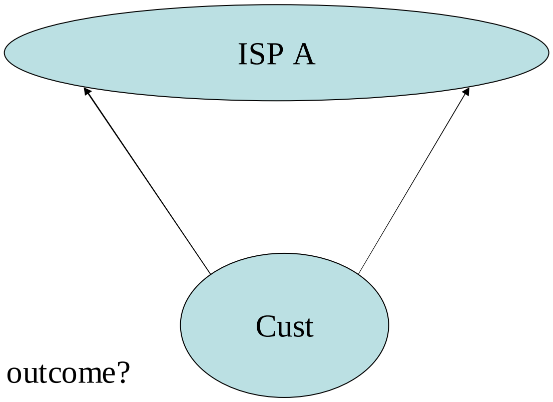


Note Asymmetric path! – Very Common

## BGP by Example: Multihoming

- Most customers don’t run BGP:
  - Simply default route to ISP
  - ISP injects customer route into BGP (or customer’s address space is from ISP)
  - Why BGP for multihoming?
- Scenarios:
  - Customer has own address space
  - Customer has provider address space
  - Customer multihomes with single ISP
  - Customer multihomes with two ISPs

# Multihoming to Single Provider

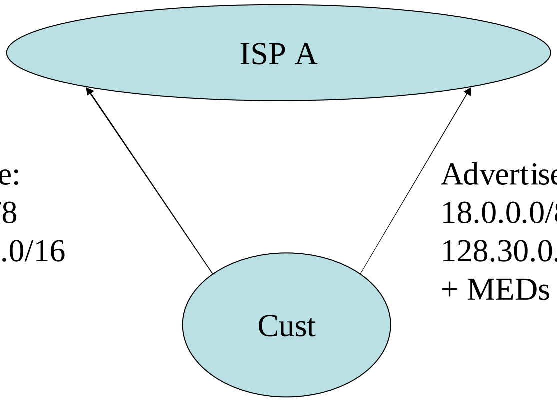


Consider
two default
routes here.

What is the outcome?

ISP and changes within ISP affect customer’s inbound traffic!
Load share outbound traffic: reordering!

# Multihoming to Single Provider



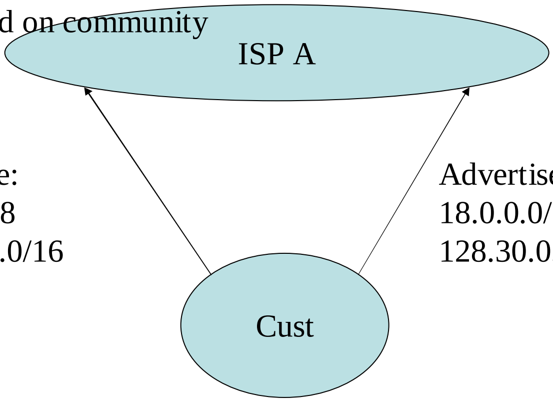
Advertise:
18.0.0.0/8
128.30.0.0/16
+ MEDs

Advertise:
18.0.0.0/8
128.30.0.0/16
+ MEDs

18.0.0.0/8
128.30.0.0/16

# Multihoming to Single Provider

ISP sets localprefs
differently on inbound customer routes
perhaps based on community



ISP A

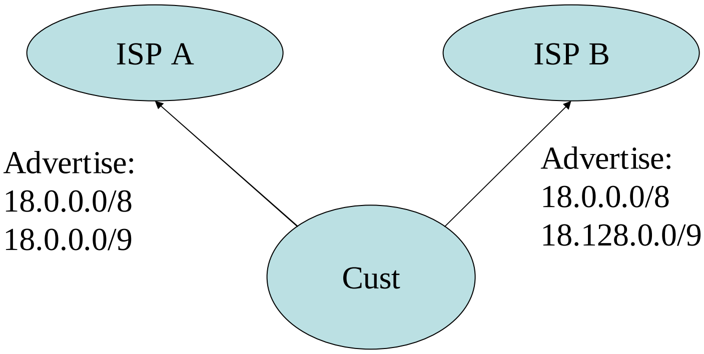
Advertise:
18.0.0.0/8
128.30.0.0/16

Advertise:
18.0.0.0/8
128.30.0.0/16

Cust

18.0.0.0/8
128.30.0.0/16

# Multihoming: Own Address Space



ISP A

ISP B

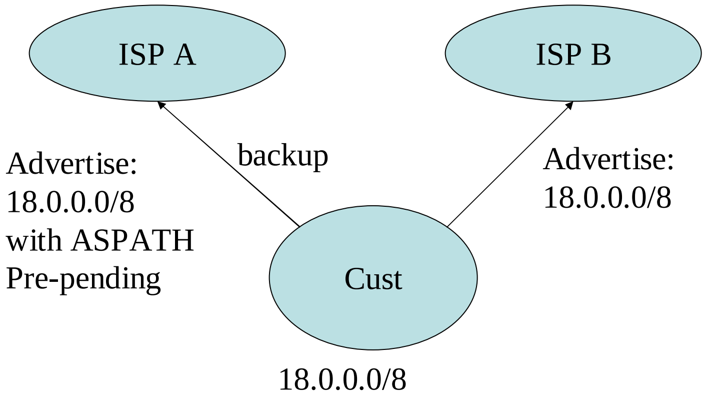
Advertise:
18.0.0.0/8
18.0.0.0/9

Advertise:
18.0.0.0/8
18.128.0.0/9

Cust

18.0.0.0/8

# Multihoming: Own Address Space

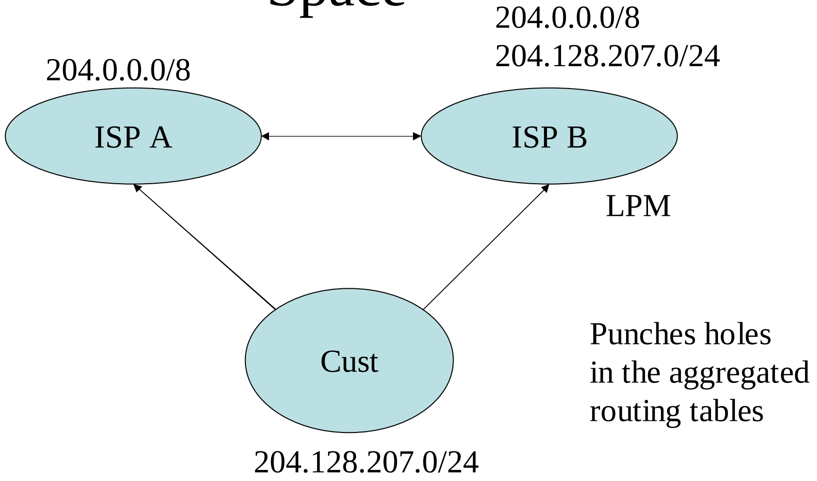


Advertise:
18.0.0.0/8
with ASPATH
Pre-pending

Advertise:
18.0.0.0/8

18.0.0.0/8

# Multihoming: Provider Address Space



204.0.0.0/8

204.0.0.0/8
204.128.207.0/24

LPM

Punches holes
in the aggregated
routing tables

204.128.207.0/24

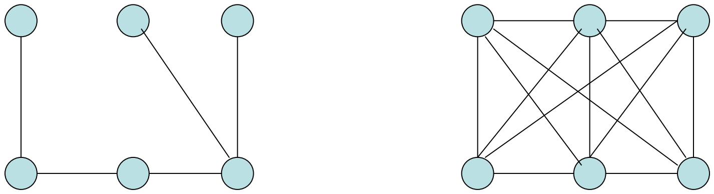
## Route Reflection

## iBGP

- Need to distribute routes within the AS
- Why not inject into IGP?
  - How many BGP routes? Lots, ~100-200k
  - Scalability of link-state database
  - Too much control traffic flooding
- Use iBGP full mesh internally
- Never redistribute a route heard via iBGP to other iBGP neighbors

# iBGP

- Physical and logical topology may be very different
- Must have an IGP running first to establish TCP-based BGP sessions

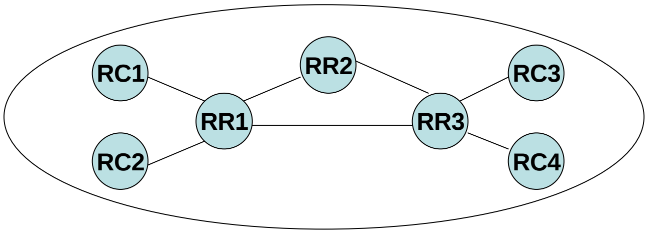


**Physical Topology** **Logical Topology**

# Route Reflection

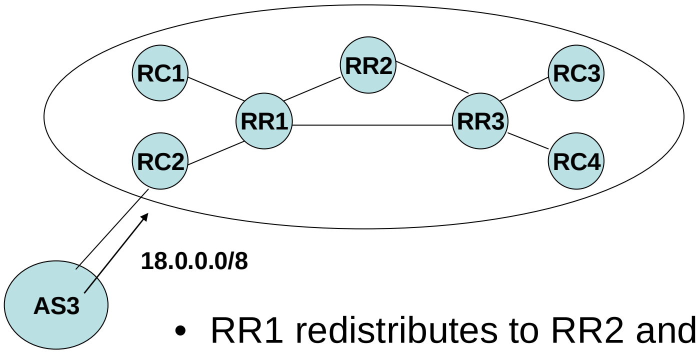
- Full-mesh of iBGP sessions
- Requires (n(n-1)/2) iBGP sessions
- Not scalable: e.g. 50 routers = 1225 sessions
- Solution: hierarchy plus minor tweak to BGP protocol
- New types:
  - Route Reflector (RR)
  - Route Reflector Client (RC) (no change)

# Route Reflection by Example



- RRs redistribute routes from RCs to all iBGP neighbors (other RRs)
- RRs redistribute routes from all iBGP neighbors to their RCs

# Route Reflection by Example



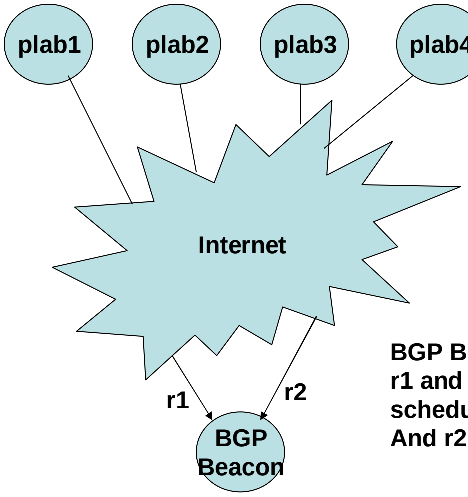
- RR1 redistributes to RR2 and RR3
- RR1 redistributes to her client RC1
- RR3 redistributes to RC3 and RC4

# BGP Badness

# SIGCOMM06: Wang et. al

- “A Measurement Study on the Impact of Routing Events on E2E Internet Path Performance”
- Experimental Methodology:
  – BGP Beacon multihomed to 2 AS
  – Advertises and withdraws on predetermined schedule
  – Planetlab Active Measurement 37-to-1
  – UDP, Ping, Traceroute

## Experimental Methodology

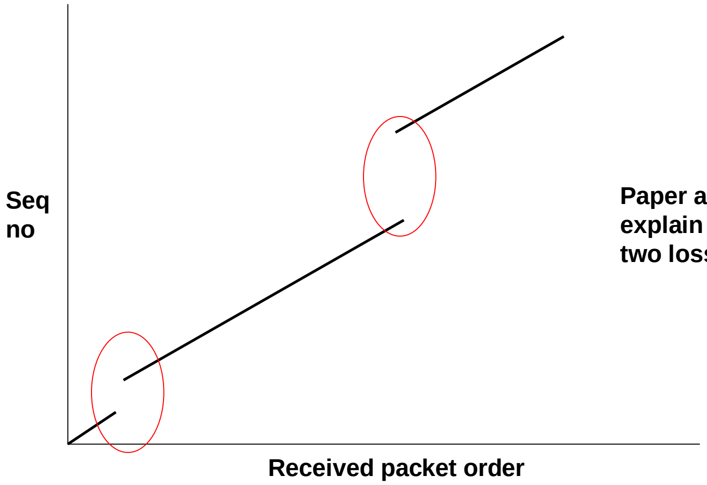


**BGP Beacon withdraws r1 and r2 on predetermined schedule. Restores r1 And r2 on schedule.**

## Routing Events + Path Perf

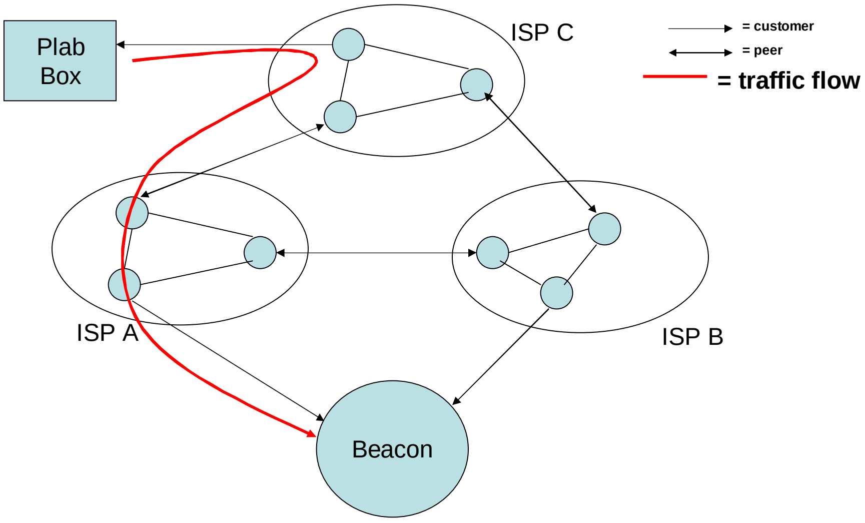
- Loss correlated to *both* withdrawals and restores
- Observe two periods of loss on a withdrawal
- Observe loss even when second path is restored – non-intuitive
- High-level reasons: BGP policy limiting advertisements, MRAI timer

# E2E Traffic Probe

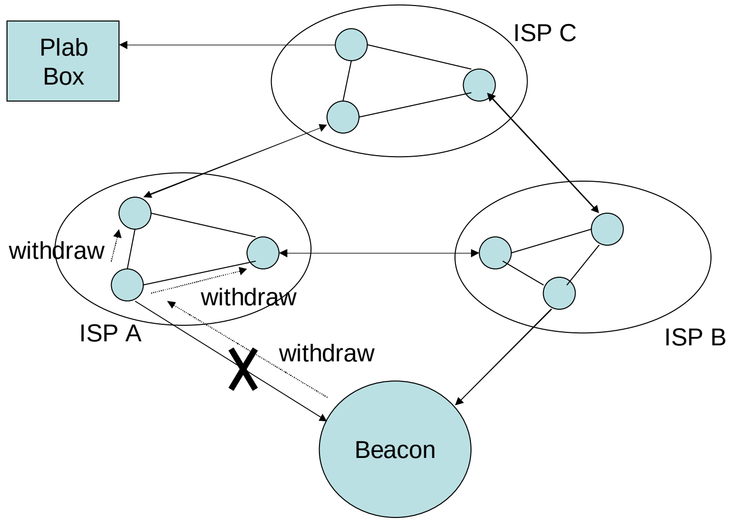


Paper attempts to
explain these
two loss events

# Route Withdrawal == loss



# Route Withdrawal == loss



Plab Box

ISP C

withdraw

withdraw

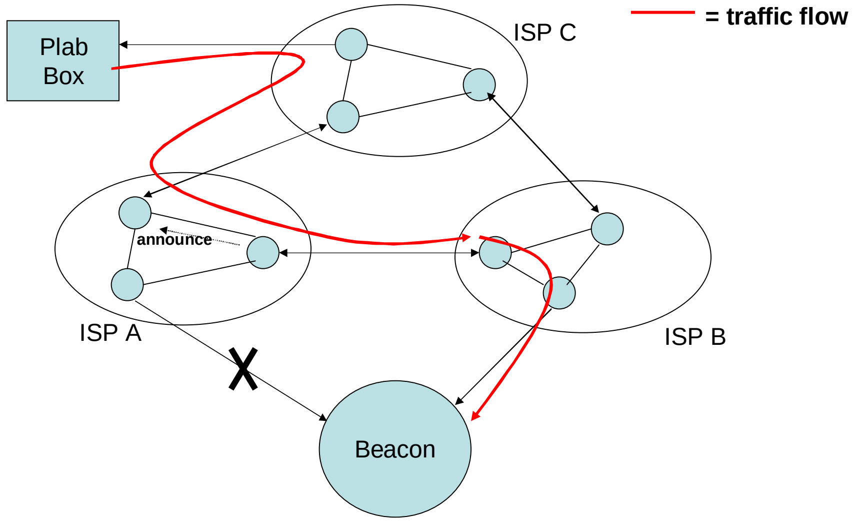
ISP A

withdraw

ISP B

Beacon

# Route Withdrawal == loss



= traffic flow

Plab Box

ISP C

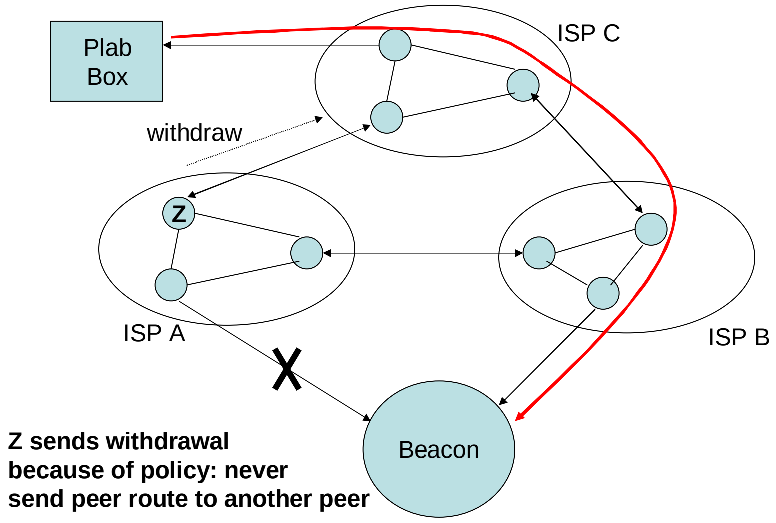
announce

ISP A

ISP B

Beacon

# Route Withdrawal == loss



Plab Box

ISP C

withdraw

Z

ISP A

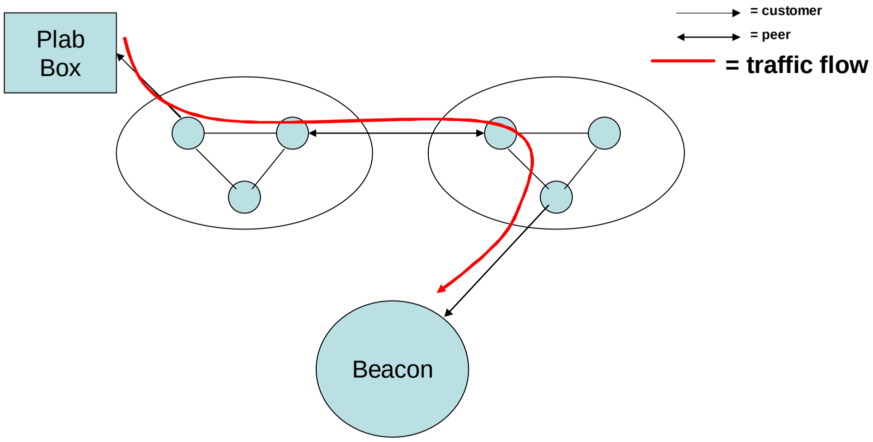
ISP B

X

Beacon

**Z sends withdrawal because of policy: never send peer route to another peer**

# Route Restoration == loss?!?



Plab Box

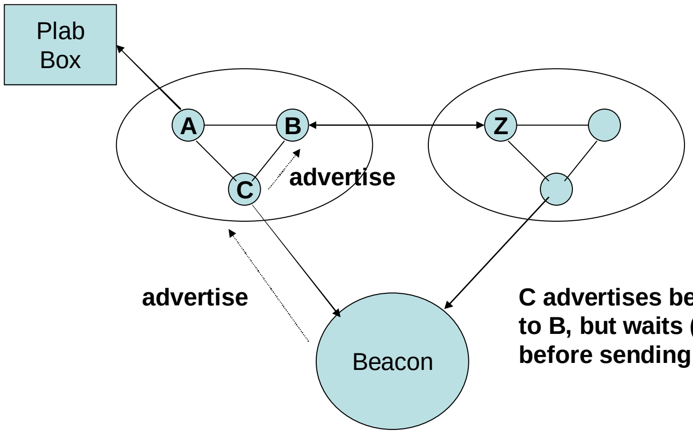
= customer

= peer

**= traffic flow**

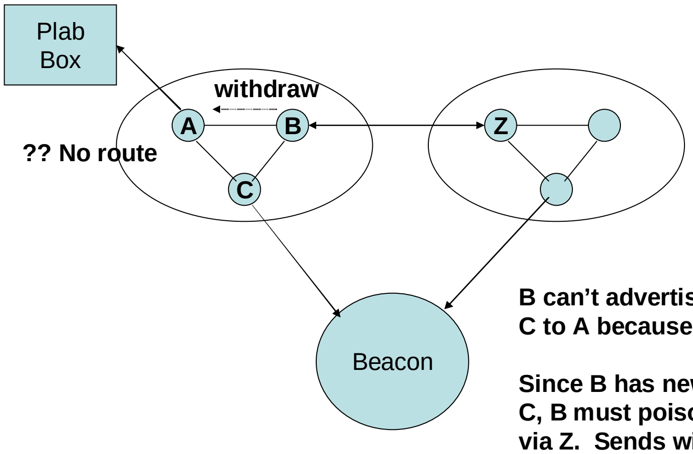
Beacon

# Route Restoration == loss?!?



**C advertises beacon route to B, but waits (due to MRAI) before sending to A**

# Route Restoration == loss?!?

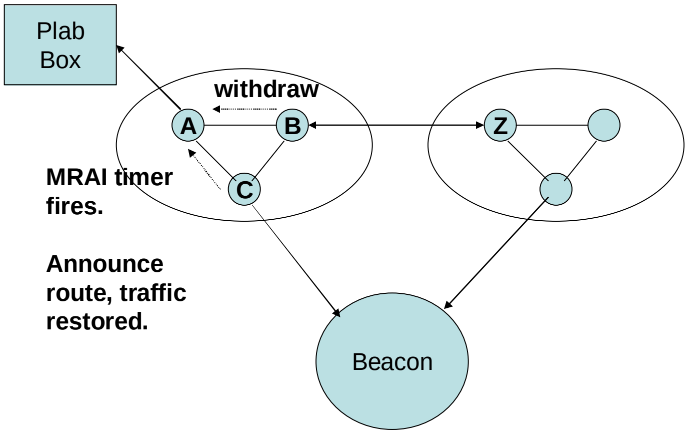


**B can't advertise route from C to A because of iBGP rules.**

**Since B has new best route via C, B must poison previous route via Z. Sends withdrawal to A.**

**A has no route, traffic dropped.**

# Route Restoration == loss?!?



**MRAI timer fires.**

**Announce route, traffic restored.**