Multihoming Complex Cases & Caveats

ISP Workshops

Last updated 6 October 2011

Complex Cases & Caveats

- Complex Cases
  - Multiple Transits
  - Multi-exit backbone
  - Disconnected Backbone
  - IDC Multihoming
- Caveats
  - No default route on:
    - Private peer edge router
    - IXP peering router
  - Separating transit and local paths
  - Backup and non-backup
  - Avoiding backbone hijack

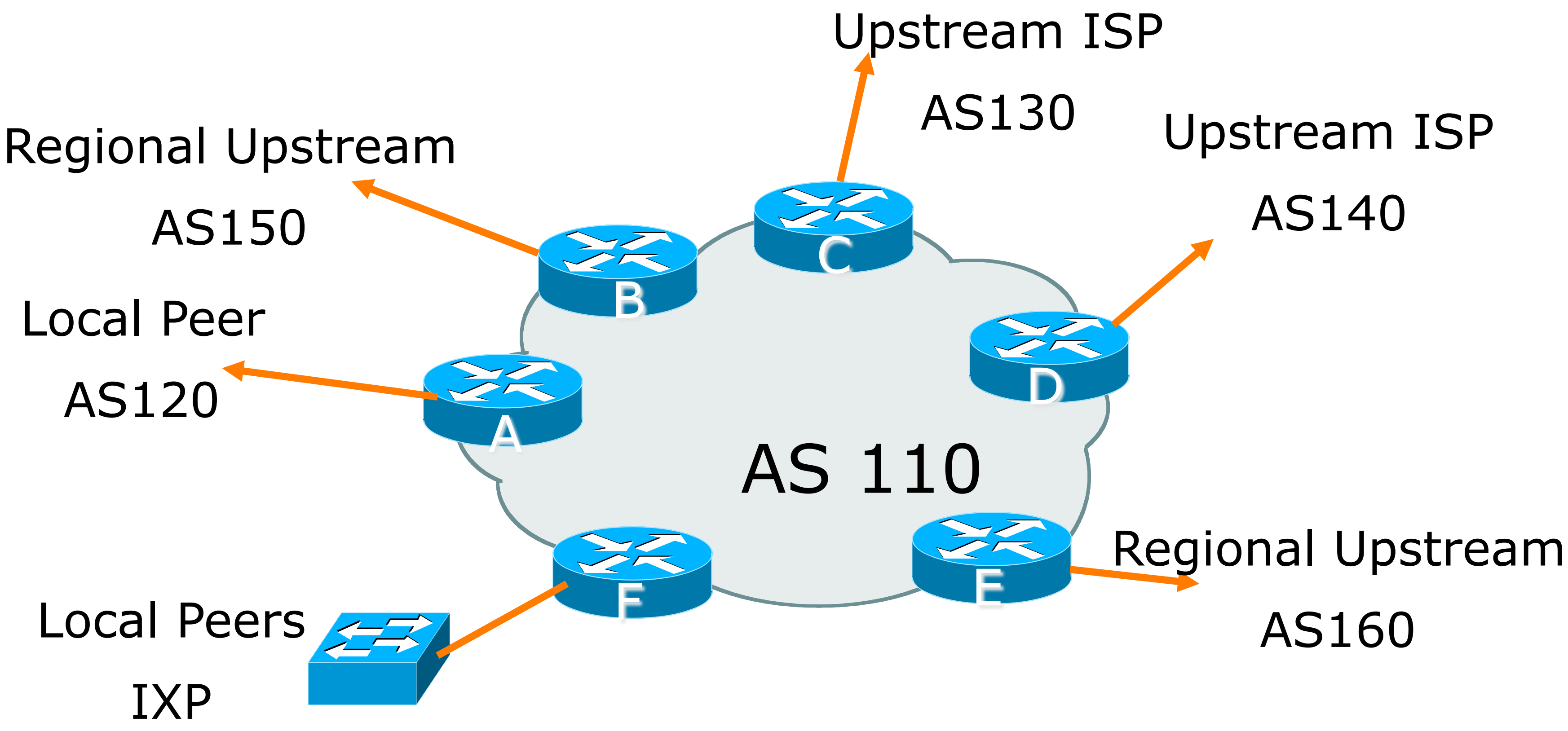
Complex Cases

Two Tier-1 upstreams, two regional upstreams, and local peers

Tier-1 & Regional Upstreams, Local Peers

- This is a complex example, bringing together all the concepts learned so far
- Connect to both upstream transit providers to see the "Internet"
  - Provides external redundancy and diversity – the reason to multihome
- Connect to regional upstreams
  - Hopefully a less expensive and lower latency view of the regional internet than is available through upstream transit provider
- Connect to private peers for local peering purposes
- Connect to the local Internet Exchange Point so that local traffic stays local
  - Saves spending valuable $ on upstream transit costs for local traffic

Tier-1 & Regional Upstreams, Local Peers



Tier-1 & Regional Upstreams, Local Peers

- Announce /19 aggregate on each link
- Accept partial/default routes from upstreams
  - For default, use 0.0.0.0/0 or a network which can be used as default
- Accept all routes from local peer
- Accept all partial routes from regional upstreams
- This is more complex, but a very typical scenario

Tier-1 & Regional Upstreams, Local Peers: Detail

- Router A – local private peer
  - Accept all (local) routes
  - Local traffic stays local
  - Use prefix and/or AS-path filters
  - Use local preference (if needed)
- Router F – local IXP peering
  - Accept all (local) routes
  - Local traffic stays local
  - Use prefix and/or AS-path filters

Tier-1 & Regional Upstreams, Local Peers: Detail

- Router B – regional upstream
  - They provide transit to Internet, but longer AS path than Tier-1s
  - Accept all regional routes from them
    - e.g. ^150_[0-9]+$
  - Ask them to send default, or send a network you can use as default
    - Set local pref on "default" to 60
  - Will provide backup to Internet only when direct Tier-1 links go down

Tier-1 & Regional Upstreams, Local Peers: Detail

- Router E – regional upstream
  - They provide transit to Internet, but longer AS path than Tier-1s
  - Accept all regional routes from them
    - e.g. ^160_[0-9]+$
  - Ask them to send default, or send a network you can use as default
    - Set local pref on “default” to 70
  - Will provide backup to Internet only when direct Tier-1 links go down

Tier-1 & Regional Upstreams, Local Peers: Detail

- Router C – first Tier-1
  - Accept all their customer and AS neighbour routes from them
    - e.g. ^130_[0-9]+$
  - Ask them to send default, or send a network you can use as default
    - Set local pref on "default" to 80
  - Will provide backup to Internet only when link to second Tier-1 goes down

Tier-1 & Regional Upstreams, Local Peers: Detail

- Router D – second Tier-1
  - Ask them to send default, or send a network you can use as default
    - This has local preference 100 by default
  - All traffic without any more specific path will go out this way

Tier-1 & Regional Upstreams, Local Peers: Summary

- Local traffic goes to local peer and IXP
- Regional traffic goes to two regional upstreams
- Everything else is shared between the two Tier-1s
- To modify loadsharing adjust what is heard from the two regionals and the first Tier-1
  - Best way is through modifying the AS-path filter

Tier-1 & Regional Upstreams, Local Peers

- What about outbound announcement strategy?
  - This is to determine incoming traffic flows
  - /19 aggregate must be announced to everyone!
  - /20 or /21 more specifics can be used to improve or modify loadsharing
  - See earlier for hints and ideas

Tier-1 & Regional Upstreams, Local Peers

- What about unequal circuit capacity?
  - AS-path filters are very useful
- What if upstream will only give me full routing table or nothing
  - AS-path and prefix filters are very useful

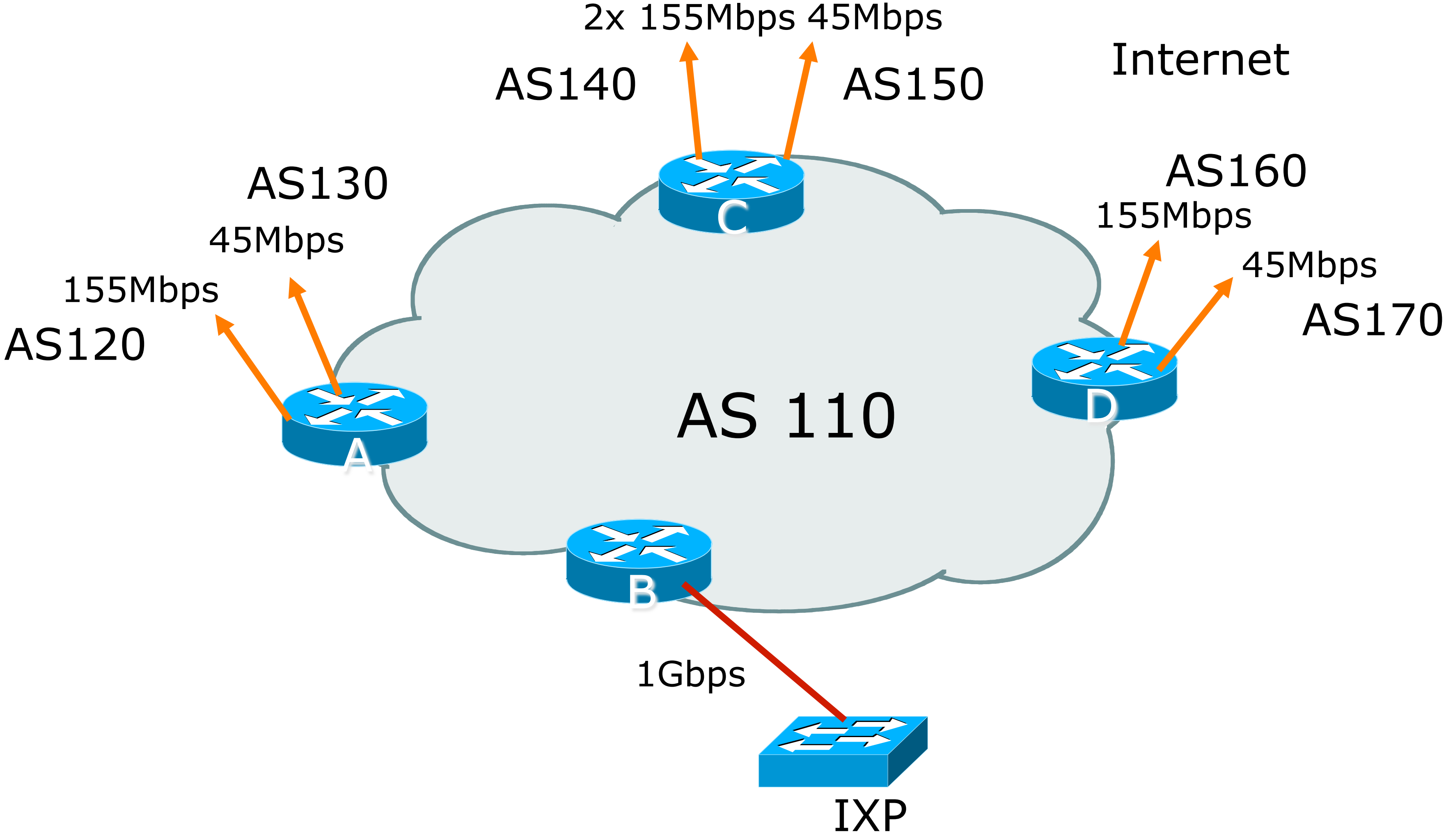
Complex Cases

Multi-exit backbone

Multi-exit backbone

- ISP with many exits to different service providers
  - Could be large transit carrier
  - Could be large regional ISP with a variety of international links to different continental locations
- Load-balancing can be painful to set up
  - Outbound traffic is often easier to balance than inbound

Multi-exit backbone



Multi-exit backbone Step One

- How to approach this?
  - Simple steps
- Step One:
  - The IXP is easy!
  - Will usually be non-transit – so preferred path for all prefixes learned this way
  - Outbound announcement – send our address block
  - Inbound announcement – accept everything originated by IXP peers, high local-pref

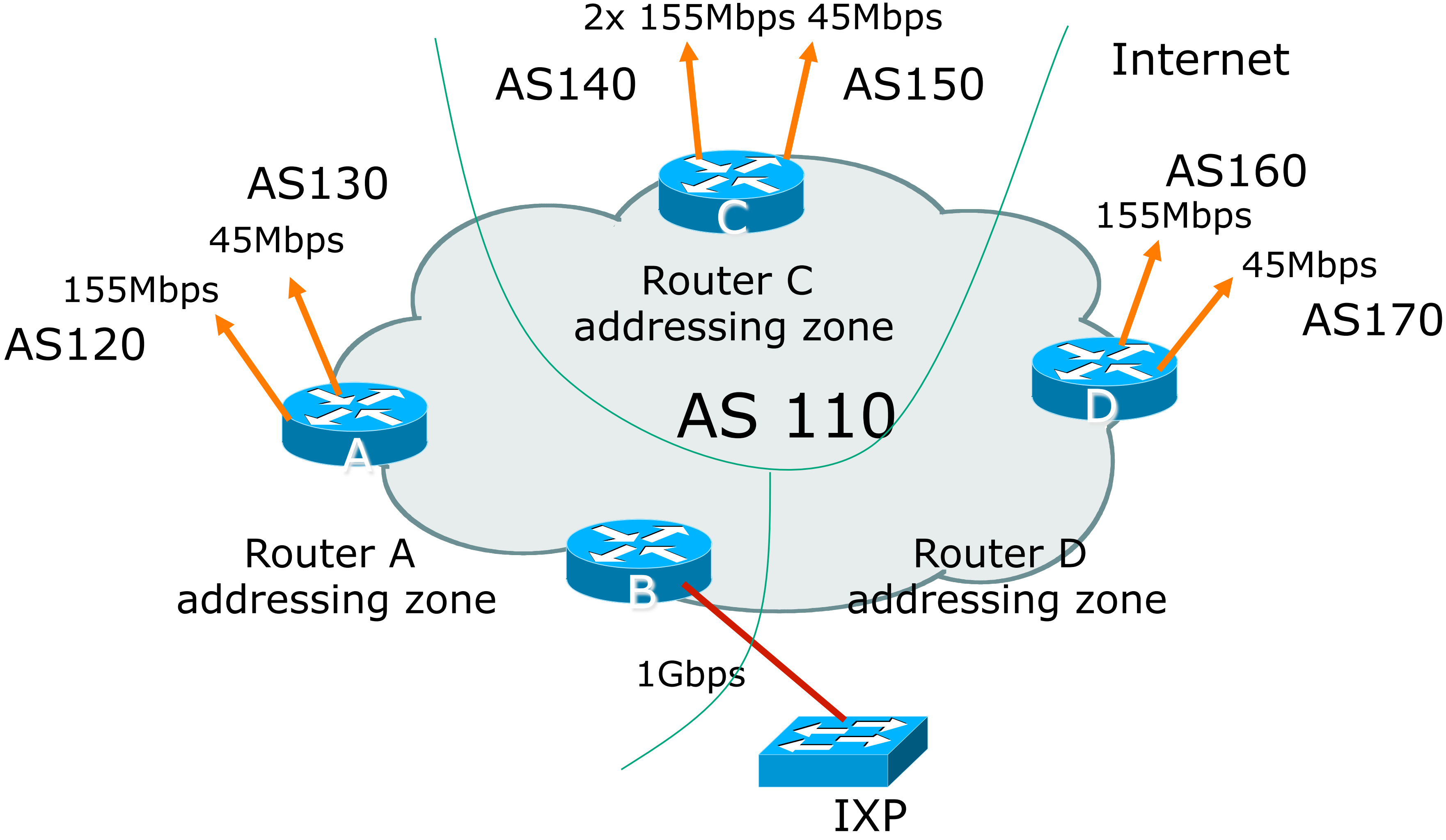
Multi-exit backbone Step Two

- Where does most of the inbound traffic come from?
  - Go to that source location, and check Looking Glass trace and AS-PATHs back to the neighbouring ASNs
  - i.e. which of AS120 through AS170 is the closest to "the source"
- Apply AS-path prepends such that the path through AS140 is one AS-hop closer than the other ASNs
  - AS140 is the ISP's biggest "pipe" to the Internet
  - This makes AS140 the preferred path to get from "the source" to AS110

Multi-exit backbone
Step Three

- Addressing plan now helps
  - Customers in vicinity of each of Router A, C and D addressed from contiguous address block assigned to each Router
  - Announcements from Router A address block sent out to AS120 and AS130
  - Announcements from Router C address block sent out to AS140 and AS150
  - Announcements from Router D address block sent out to AS160 and AS170

Multi-exit backbone
Addressing Plan Assists Multihoming



Multi-exit backbone
Step Four

- Customer type assists zone load balancing
  - Two customer classes: Commercial & Consumer
  - Commercial announced on T3 links
  - Consumer announced on STM-1 links
- Commercial
  - Numbered from one address block in each zone
- Consumer
  - Numbered from the other address block in each zone

Multi-exit backbone
Example Summary (1)

- Address block: 100.10.0.0/16
- Router A zone: 100.10.0.0/18
  - Commercial: 100.10.0.0/19
  - Consumer: 100.10.32.0/19
- Router C zone: 100.10.128.0/17
  - Commercial: 100.10.128.0/18
  - Consumer: 100.10.192.0/18
- Router D zone: 100.10.64.0/18
  - Commercial: 100.10.64.0/19
  - Consumer: 100.10.96.0/19

Multi-exit backbone Example Summary (2)

- Router A announcement:
  - 100.10.0.0/16 with 3x AS-path prepend
  - 100.10.0.0/19 to AS130
  - 100.10.32.0/19 to AS120
- Router B announcement:
  - 100.10.0.0/16
- Router C announcement:
  - 100.10.0.0/16
  - 100.10.128.0/18 to AS150
  - 100.10.192.0/18 to AS140
- Router D announcement:
  - 100.10.0.0/16 with 3x AS-path prepend
  - 100.10.64.0/19 to AS170
  - 100.10.96.0/19 to AS160

Multi-exit backbone
Summary

- This is an example strategy
  - Your mileage will vary
- Example shows:
  - where to start,
  - what the thought processes are, and
  - what the strategies could be

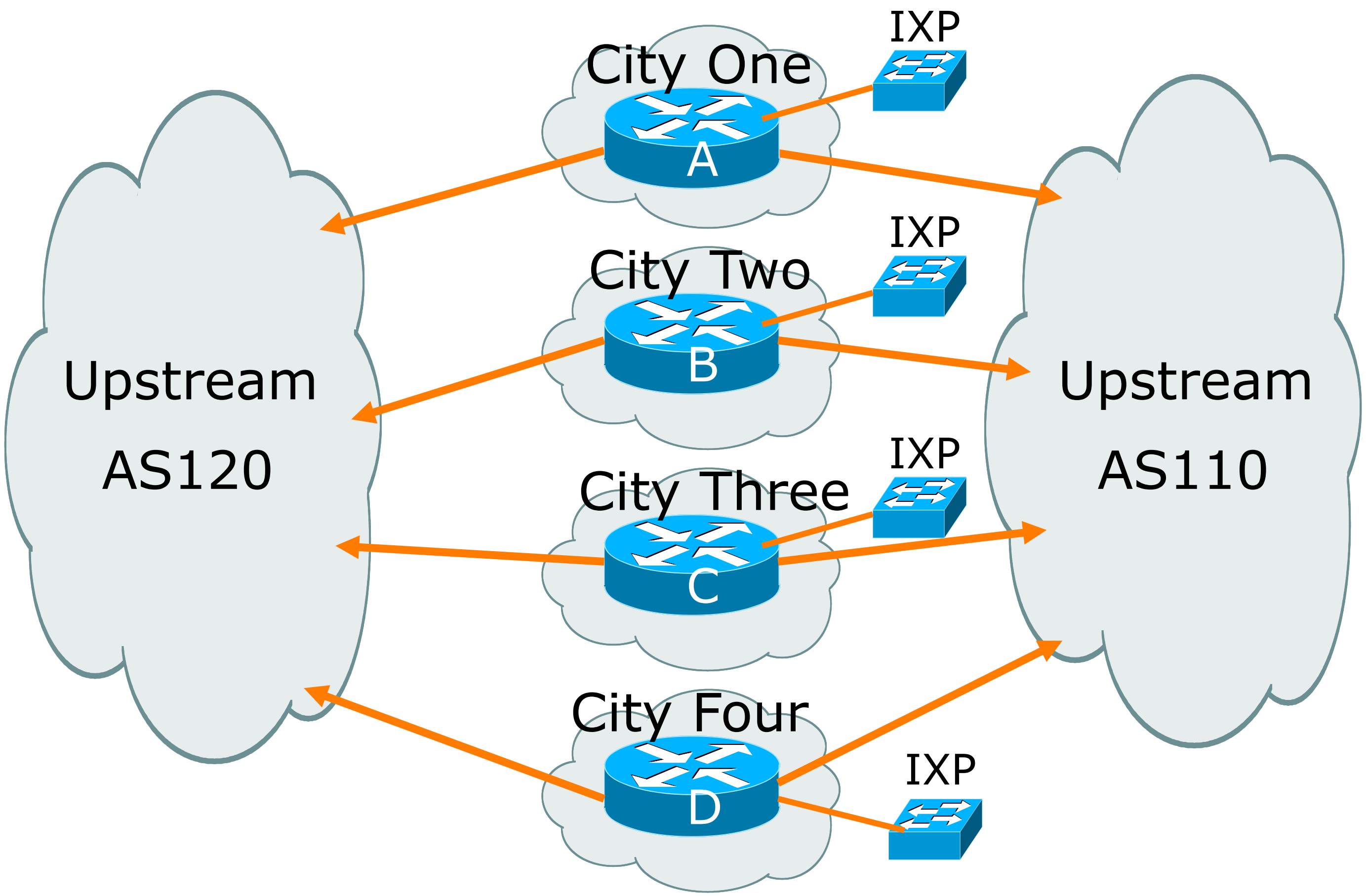
Service Provider Multihoming

Disconnected Backbone

Disconnected Backbone

- ISP runs large network
  - Network has no backbone, only large PoPs in each location
  - Each PoP multihomes to upstreams
  - Common in some countries where backbone circuits are hard to obtain
- This is to show how it could be done
  - Not impossible, nothing “illegal”

Disconnected Backbone



Disconnected Backbone

- Works with one AS number
  - Not four – no BGP loop detection problem
- Each city operates as separate network
  - Uses defaults and selected leaked prefixes for loadsharing
  - Peers at local exchange point

Disconnected Backbone

- Router A Configuration

```
router bgp 100
 network 121.10.0.0 mask 255.255.248.0
 neighbor 122.100.0.1 remote-as 120
 neighbor 122.100.0.1 description AS120 – Serial 0/0
 neighbor 122.100.0.1 prefix-list default in
 neighbor 122.102.0.1 prefix-list my-block out
 neighbor 122.102.10.1 remote-as 110
 neighbor 122.102.10.1 description AS110 – Serial 1/0
 neighbor 122.102.10.1 prefix-list rfc1918-sua in
 neighbor 122.102.10.1 prefix-list my-block out
 neighbor 122.102.10.1 filter-list 10 in
!
```

…continued on next page…

Disconnected Backbone

```
ip prefix-list my-block permit 121.10.0.0/21
ip prefix-list default permit 0.0.0.0/0
!
ip as-path access-list 10 permit ^(110_)+$
ip as-path access-list 10 permit ^(110_)+_[0-9]+$
!…etc to achieve outbound loadsharing
!
ip route 0.0.0.0 0.0.0.0 Serial 1/0 250
ip route 121.10.0.0 255.255.248.0 null0
!
```

Disconnected Backbone

- Peer with AS120
  - Receive just default route
  - Announce /22 address
- Peer with AS110
  - Receive full routing table – filter with AS-path filter
  - Announce /22 address
  - Point backup static default – distance 252 – in case AS120 goes down

Disconnected Backbone

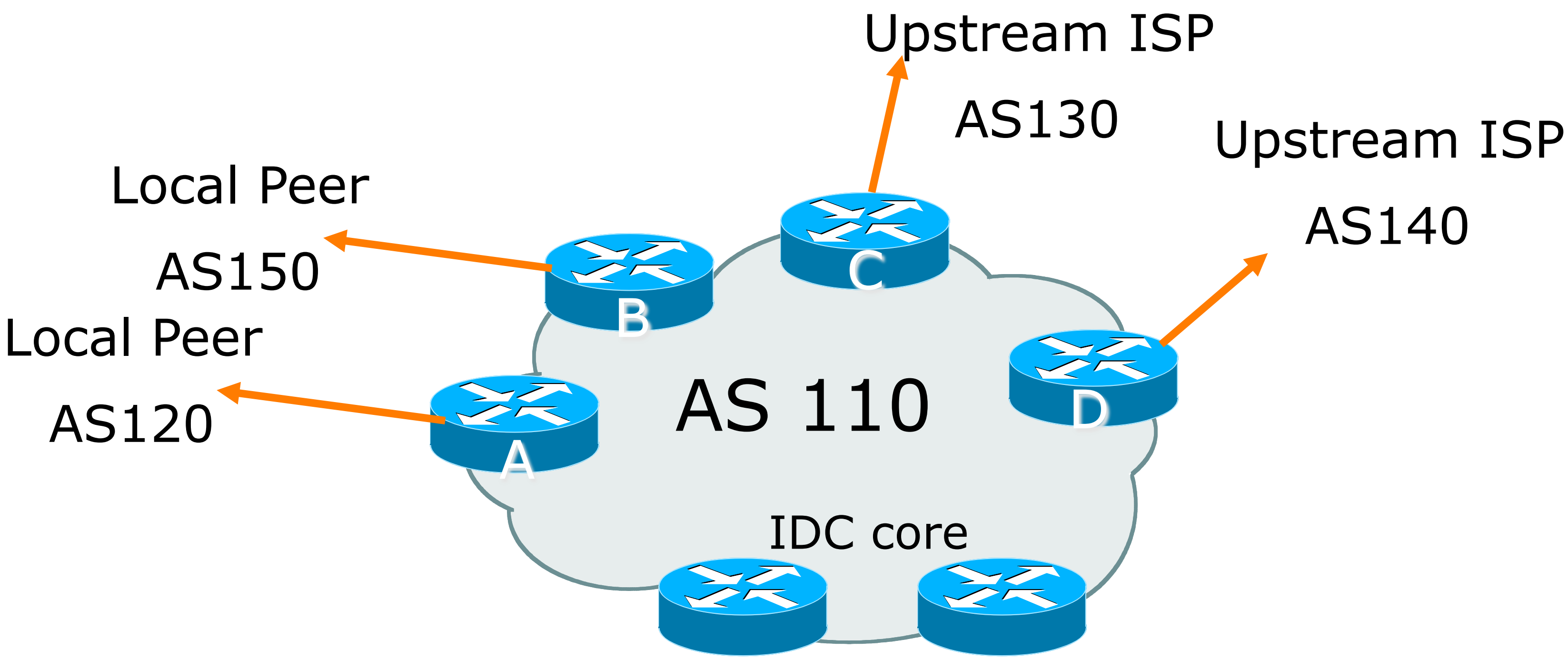
- Default ensures that disconnected parts of AS100 are reachable
  - Static route backs up AS120 default
  - No BGP loop detection – relying on default route
- Do not announce /19 aggregate
  - No advantage in announcing /19 and could lead to problems

IDC Multihoming

IDC Multihoming

- IDCs typically are not registry members so don’t get their own address block
  - Situation also true for small ISPs and “Enterprise Networks”
- Smaller address blocks being announced
  - Address space comes from both upstreams
  - Should be apportioned according to size of circuit to upstream
- Outbound traffic paths matter

Two Upstreams, Two Local Peers IDC



Assigned /24 from AS130 and /23 from AS140.

Circuit to AS130 is 2Mbps, circuit to AS140 is 4Mbps

IDC Multihoming

- Router A and B configuration
  - In: Should accept all routes from AS120 and AS150
  - Out: Should announce all address space to AS120 and AS150
  - Straightforward

IDC Multihoming

- Router C configuration
  - In: Accept partial routes from AS130
    - e.g. ^130_[0-9]+$
  - In: Ask for a route to use as default
    - set local preference on default to 80
  - Out: Send /24, and send /23 with AS-PATH prepend of one AS

IDC Multihoming

- Router D configuration
  - In: Ask for a route to use as default
    - Leave local preference of default at 100
  - Out: Send /23, and send /24 with AS-PATH prepend of one AS

IDC Multihoming Fine Tuning

- For local fine tuning, increase circuit capacity
  - Local circuits usually are cheap
  - Otherwise…
- For longer distance fine tuning
  - In: Modify as-path filter on Router C
  - Out: Modify as-path prepend on Routers C and D
  - Outbound traffic flow is usual critical for an IDC so inbound policies need to be carefully thought out

IDC Multihoming
Other Details

- Redundancy
  - Circuits are terminated on separate routers
- Apply thought to address space use
  - Request from both upstreams
  - Utilise address space evenly across IDC
    - Don't start with /23 then move to /24 – use both blocks at the same time in the same proportion
    - Helps with loadsharing – yes, really!

IDC Multihoming
Other Details

- What about failover?
  - /24 and /23 from upstreams' blocks announced to the Internet routing table all the time
  - No obvious alternative at the moment
    - Conditional advertisement can help in steady state, but subprefixes still need to be announced in failover condition

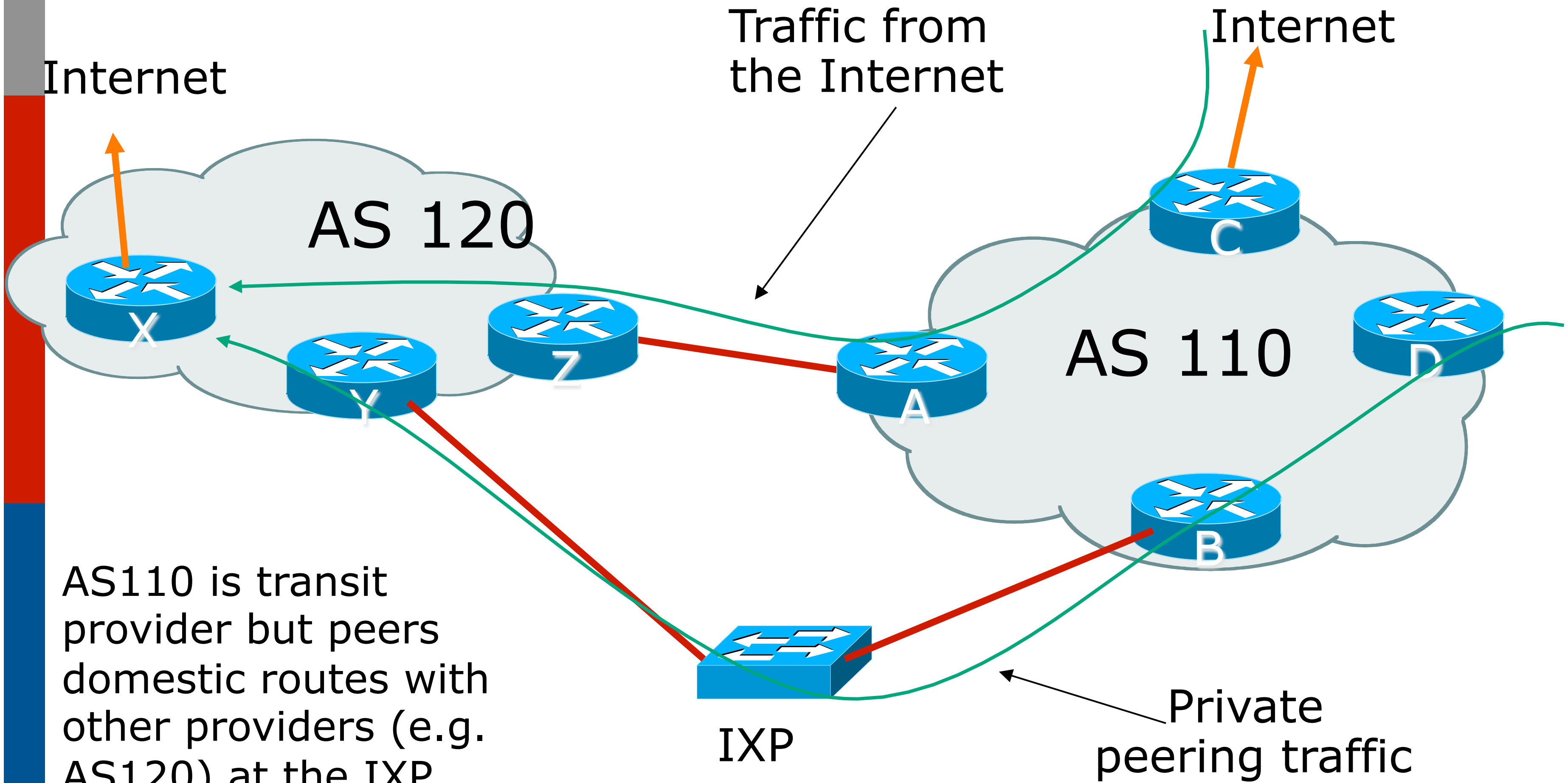
Caveats

Separating Transit and Local Paths

Transit and Local paths

- Common problem is separating transit and local traffic for BGP customers
- Transit provider:
  - Provides internet access for BGP customer over one path
  - Provides domestic access for BGP customer over another path
  - Usually required for commercial reasons
    - Inter-AS traffic is unmetered
    - Transit traffic is metered

Transit and Local paths



AS110 is transit provider but peers domestic routes with other providers (e.g. AS120) at the IXP

Transit and Local paths

- Assume Router X is announcing 192.168/16 prefix
- Router C and D see two entries for 192.168/16 prefix:

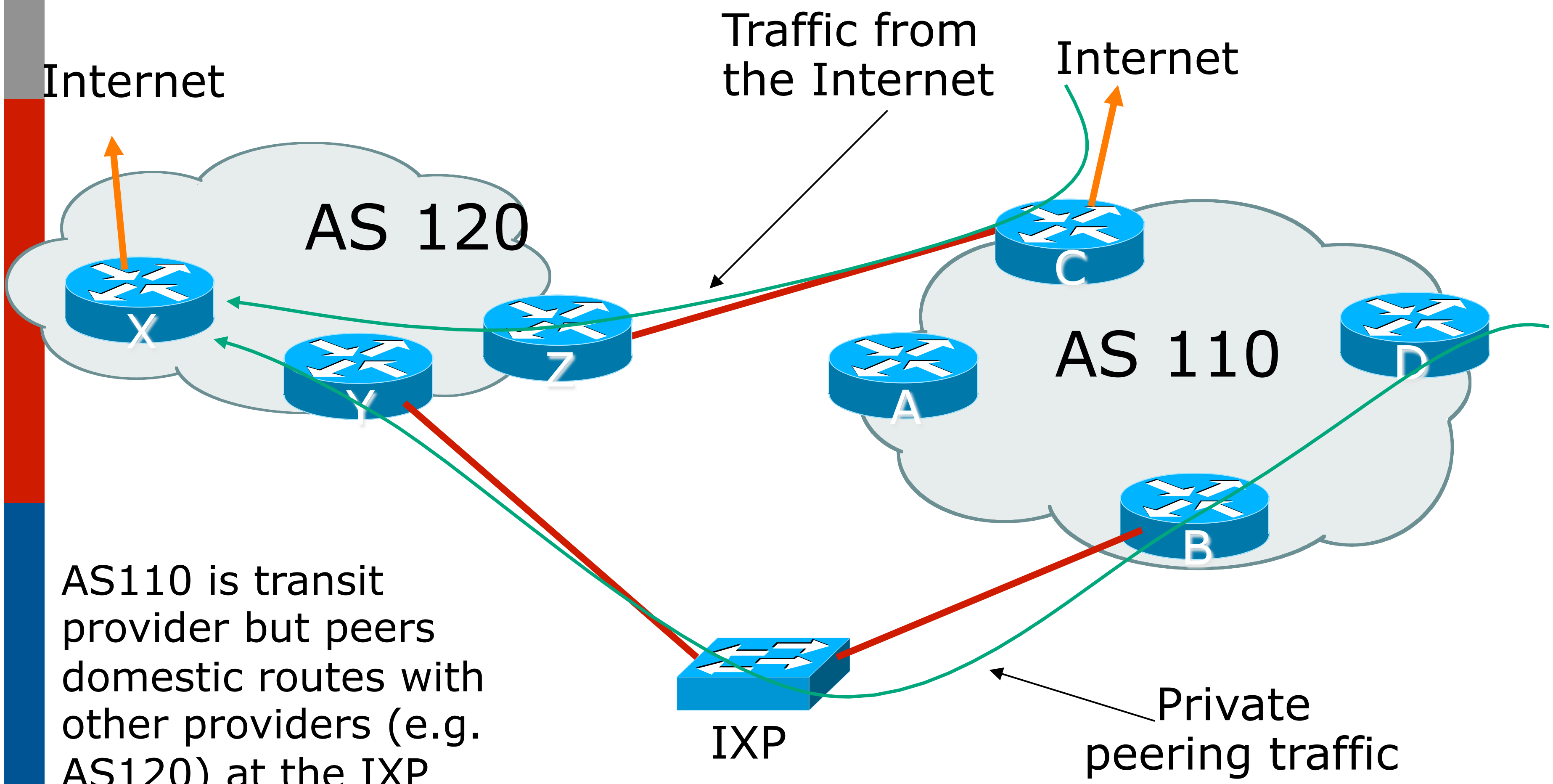
```
RouterC#show ip bgp
   Network          Next Hop            Metric LocPrf Weight Path
* i192.168.0.0/16   10.0.1.1                      100      0 120 i
*>i                 10.0.1.5                      100      0 120 i
```

- BGP path selection rules pick the highest next hop address
  - So this could be Router A or Router B!
  - No exit path selection here…

Transit and Local paths

- There are a few solutions to this problem
  - Policy Routing on Router A according to packet source address
  - GRE tunnels (gulp)
- Preference is to keep it simple
  - Minor redesign and use of BGP weight is a simple solution

Transit and Local paths (Network Revision)



AS110 is transit provider but peers domestic routes with other providers (e.g. AS120) at the IXP

Transit and Local paths

- Router B hears 192.168/16 from Router Y across the IXP
- Router C hears 192.168/16 from Router Z across the private peering link
- Router B sends 192.168/16 by iBGP to Router C:

```
RouterC#show ip bgp
   Network          Next Hop            Metric LocPrf Weight Path
*> 192.168.0.0/16   10.1.5.7                      100      0 120 i
* i                 10.0.1.5                      100      0 120 i
```

- Best path is by eBGP to Router Z
  - So Internet transit traffic to AS120 will go through private peering link

Transit and Local paths

- Router D hears prefix by iBGP from both Router B and Router C
- BGP best path selection might pick either path, depending on IGP metric, or next hop address, etc
- Solution to force local traffic over the IXP link:
  - Apply high local preference on Router B for all routes learned from the IXP peers

```
RouterD#show ip bgp
   Network          Next Hop            Metric LocPrf Weight Path
* i192.168.0.0/16   10.0.1.3                      100      0 120 i
*>i                 10.0.1.5                      120      0 120 i
```

Transit and Local paths

- High local preference on B is visible throughout entire iBGP
  - Including on Router C

```
RouterC#show ip bgp
   Network          Next Hop            Metric LocPrf Weight Path
*  192.168.0.0/16   10.1.5.7                      100      0 120 i
*>i                 10.0.1.5                      120      0 120 i
```

- As a result, Internet traffic now goes through the IX, not the private peering link as intended

Transit and Local paths

- Solution: Use BGP weight on Router C for prefixes heard from AS120:

```
RouterC#show ip bgp
   Network          Next Hop            Metric LocPrf Weight Path
*> 192.168.0.0/16   10.1.5.7                      100  50000 120 i
* i                 10.0.1.5                      120      0 120 i
```

- So Router C prefers private link to AS120 for traffic coming from Internet
- Rest of AS110 prefers Router B exit through the IXP for local traffic

Transit and Local paths
Summary

- Transit customer private peering connects to Border router
  - Transit customer routes get high weight
- Local traffic on IXP peering router gets high local preference
- Internet return traffic goes on private interconnect
- Domestic return traffic crosses IXP

Caveats

Backup and Non-backup

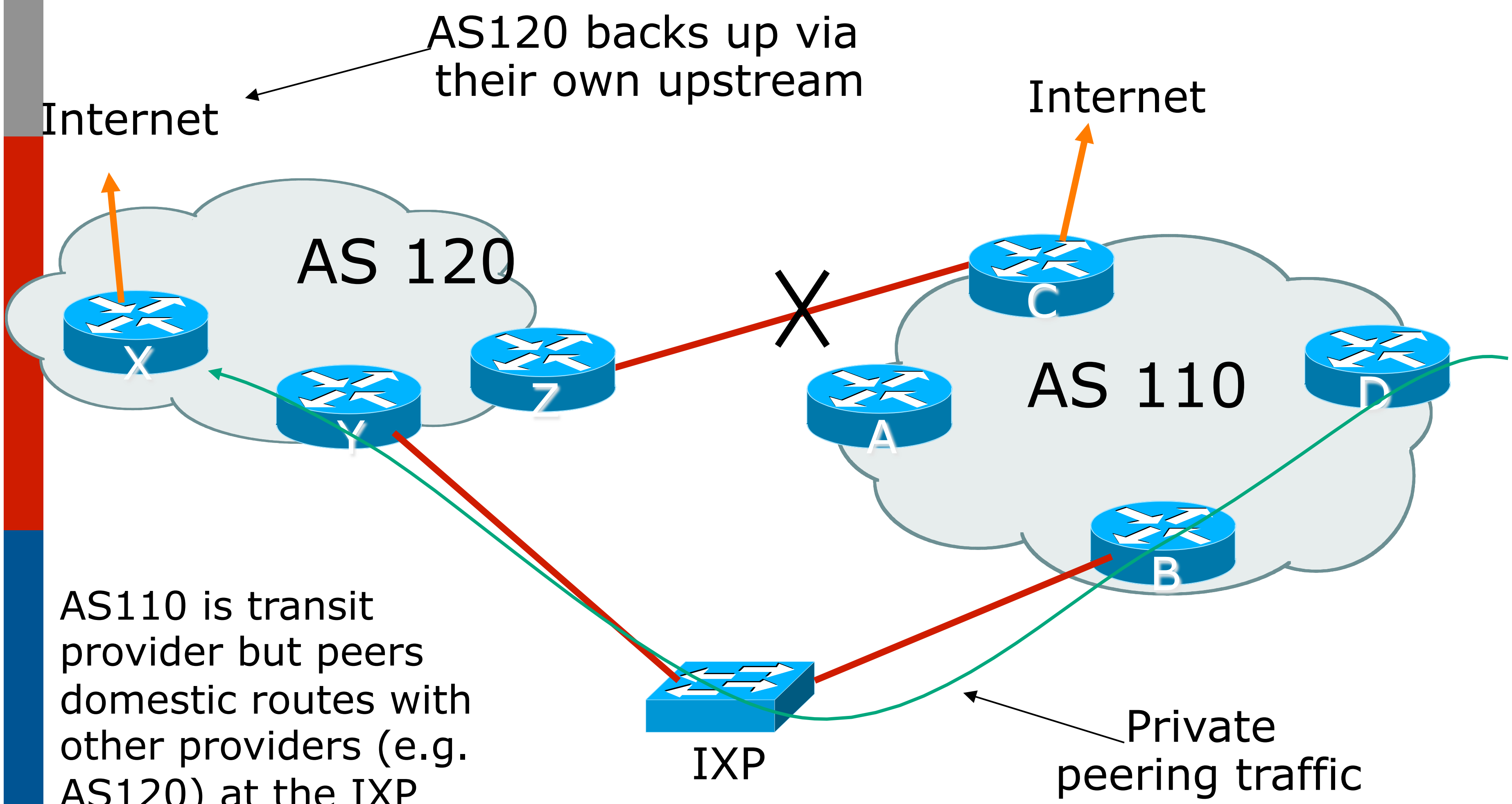
Transit and Local paths
Backups

- For the previous scenario, what happens if private peering link breaks?
  - Traffic backs up across the IXP
- What happens if the IXP breaks?
  - Traffic backs up across the private peering
- Some ISPs find this backup arrangement acceptable
  - It is a backup, after all

Transit and Local paths IXP Non-backup

- IXP actively does not allow transit
- ISP solution:
  - 192.168/16 via IX tagged one community
  - 192.168/16 via PP tagged other community
  - Using community tags, iBGP on IX router (Router B) does not send 192.168/16 to upstream border (Router C)
    - Therefore Router C only hears 192.168/16 via private peering
    - If the link breaks, backup is via AS110 and AS120 upstream ISPs

Transit and Local paths
IXP Non-backup

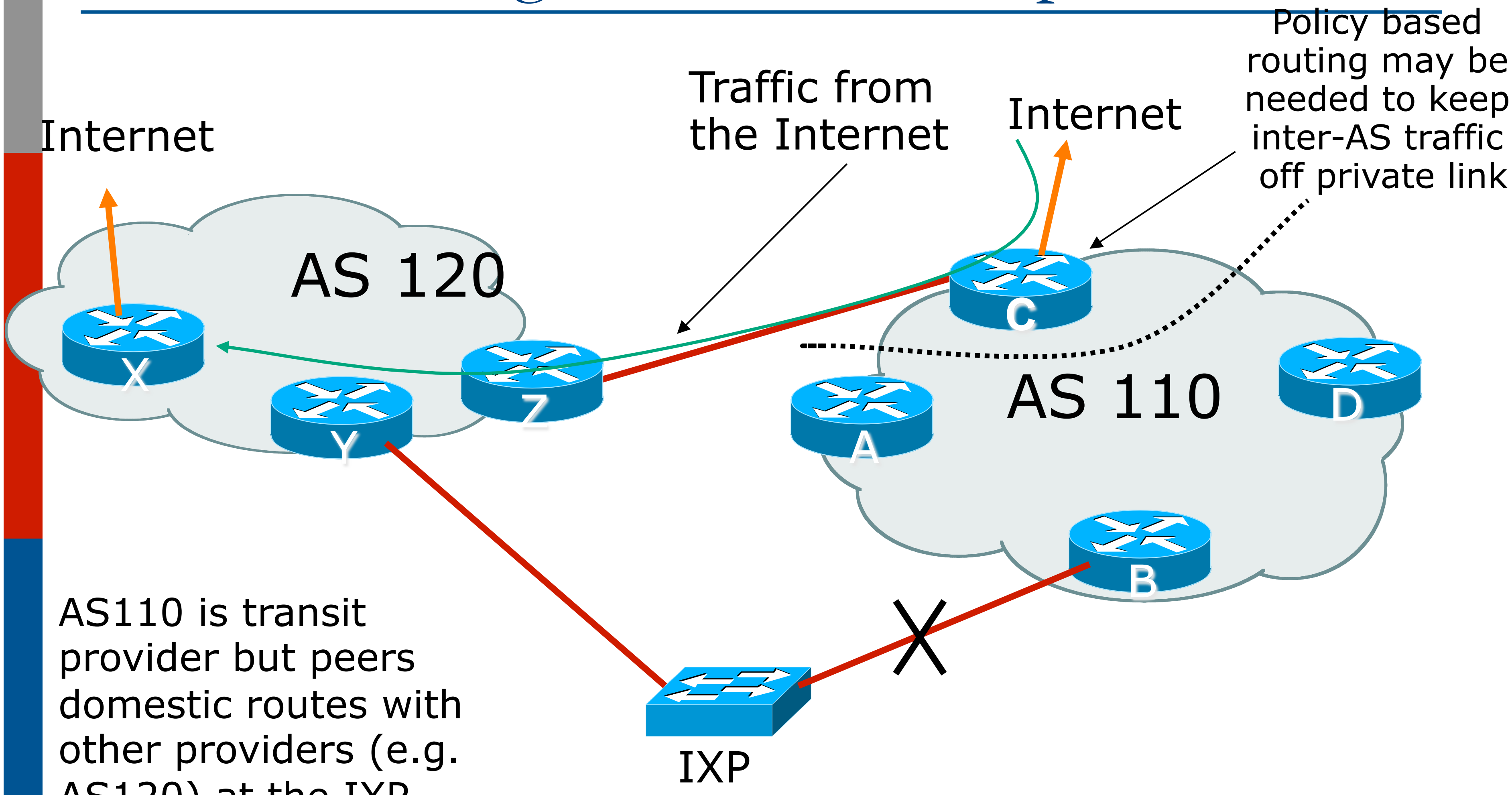


AS110 is transit provider but peers domestic routes with other providers (e.g. AS120) at the IXP

Transit and Local paths
Private Peering link Non-backup

- With this solution, a breakage in the IX means that local peering traffic will still back up over private peering link
  - This link may be metered
- AS110 Solution:
  - Router C does not announce 192.168/16 by iBGP to the other routers in AS110
  - If IX breaks, there is no route to AS120
  - Unless Router C is announcing a default route
    - Whereby traffic will get to Router C anyway, and policy based routing will have to be used to avoid ingress traffic from AS110 going on the private peering link

Transit and Local paths
Private Peering link Non-backup



AS110 is transit provider but peers domestic routes with other providers (e.g. AS120) at the IXP

Transit and Local paths Summary

- Not allowing BGP backup to “do the right thing” can rapidly get messy
- But previous two scenarios are requested quite often
  - Billing of traffic seems to be more important than providing connectivity
  - But thinking through the steps required shows that there is usually a solution without having to resort to extreme measures

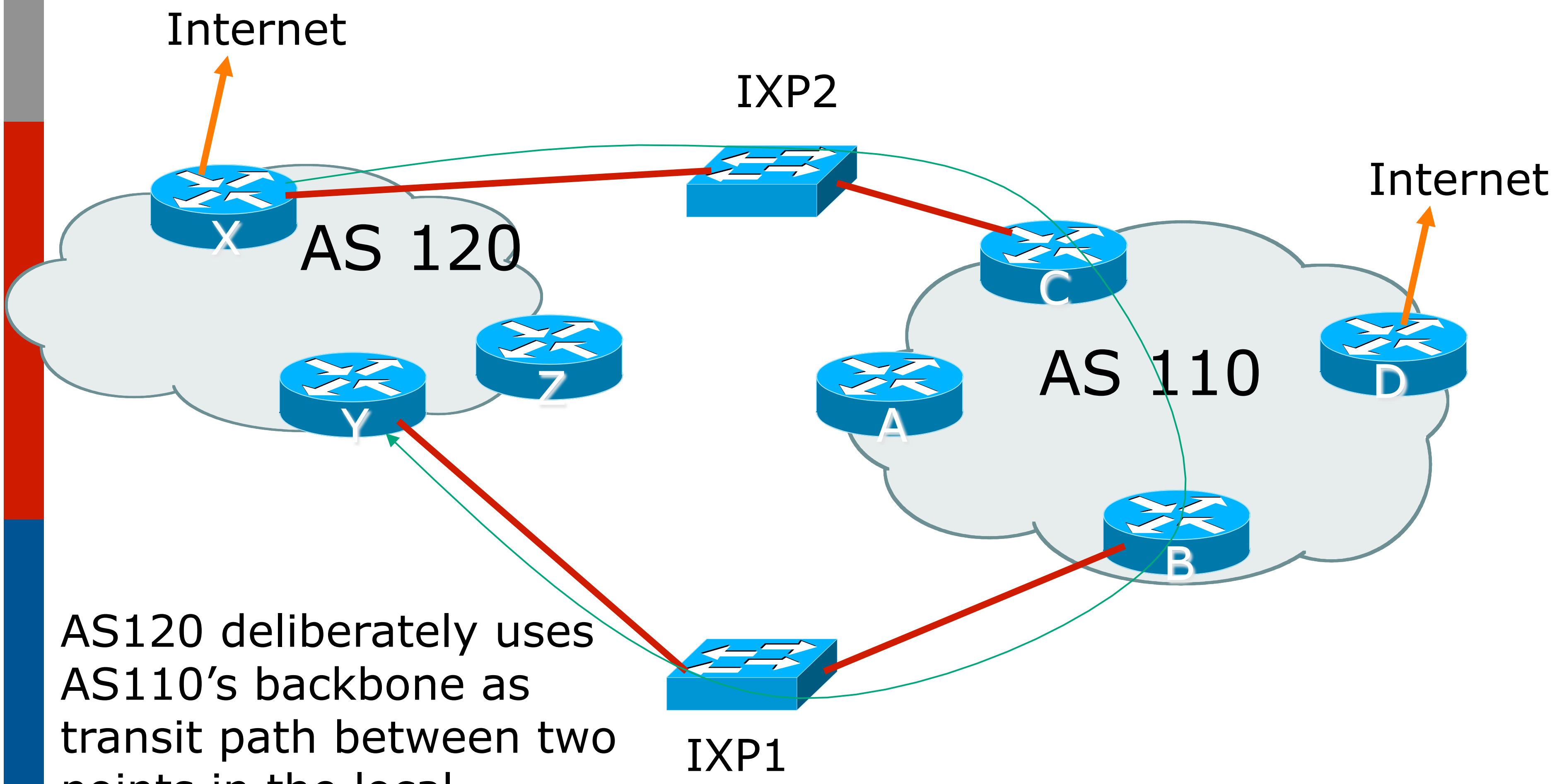
Caveats

Avoiding "Backbone Hijack"

Backbone Hijacks

- Can happen when peering ISPs:
  - are present at two or more IXPs
  - have two or more private peering links
- Usually goes undetected
  - Can be spotted by traffic flow monitoring tools
- Done because:
  - "Their backbone is cheaper than mine"
- Caused by misconfiguration of private peering routers

Avoiding "Backbone Hijack"



AS120 deliberately uses AS110's backbone as transit path between two points in the local network

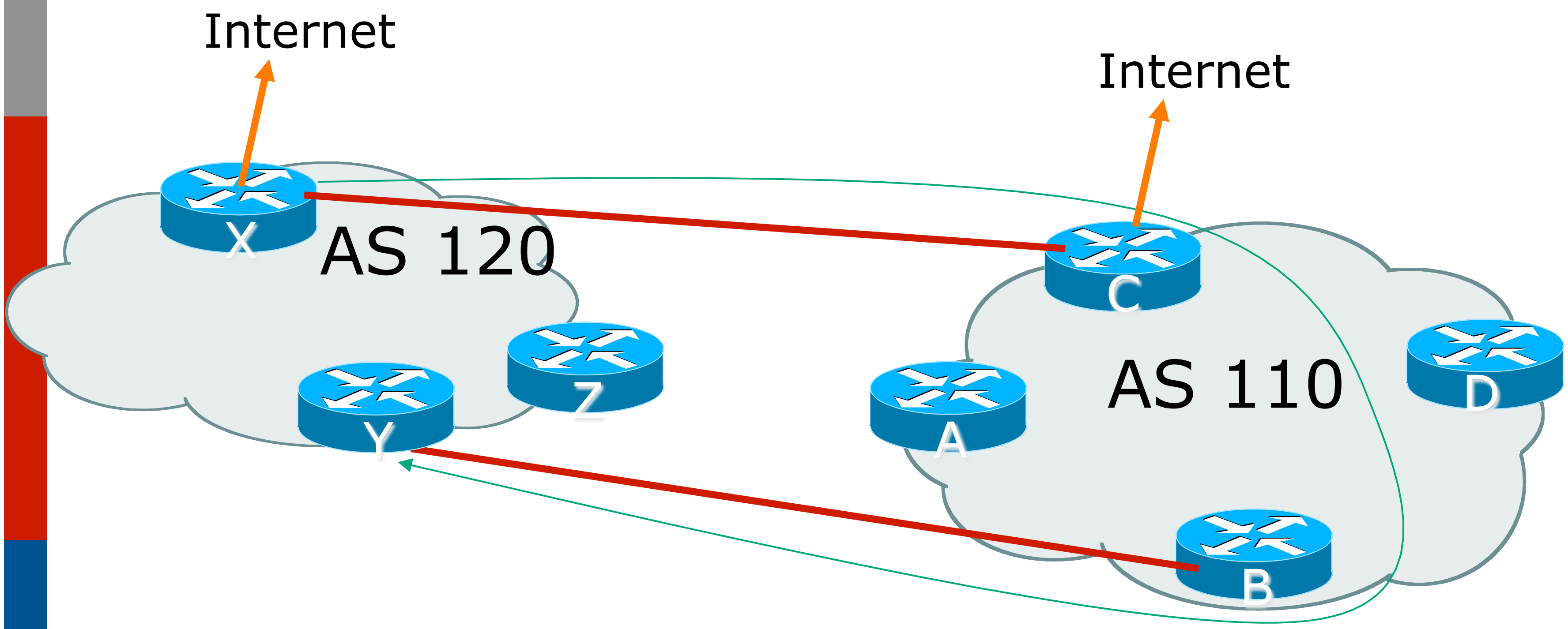
Avoiding "Backbone Hijack"

- AS110 peering routers at the IXPs should only carry AS110 originated routes
  - When AS120 points static route for an AS120 destination to AS110, the peering routers have no destination apart from back towards AS120, so the packets will oscillate until TTL expiry
  - When AS120 points static route for a non-AS110 destination to AS110, the peering routers have no destination at all, so the packet is dropped

Avoiding "Backbone Hijack"

- Same applies for private peering scenarios
  - Private peering routers should only carry the prefixes being exchanged in the peering
  - Otherwise abuses are possible
- What if AS110 is providing the full routing table to AS120?
  - AS110 is the transit provider for AS120

Avoiding "Backbone Hijack"



AS120 deliberately uses AS110's backbone as transit path between two points in the local network

Avoiding "Backbone Hijack"

- Router C carries a full routing table on it
  - So we can't use the earlier trick of only carrying AS110 prefixes
- Reverse path forwarding check?
  - But that only checks the packet source address, not the destination – and the source is fine!
- BGP Weight
  - Recall that BGP weight was used to separate local and transit traffic in the previous example
  - If all prefixes learned from AS120 on Router C had local weight increased, then destination is back out the incoming interface
  - And the same can be done on Router B

Avoiding "Backbone Hijack" Summary

- These are but two examples of many possible scenarios which have become frequently asked questions
- Solution is often a lot simpler than imagined
  - BGP Weight, selective announcement by iBGP, simple network redesigns…

Summary

- Complex Cases
  - Multiple Transits
  - Multi-exit backbone
  - Disconnected Backbone
  - IDC Multihoming
- Caveats
  - No default route on:
    - Private peer edge router
    - IXP peering router
  - Separating transit and local paths
  - Backup and non-backup
  - Avoiding backbone hijack

Multihoming Complex Cases & Caveats

ISP Workshops