BGP Multihoming Techniques

Philip Smith <pfs@cisco.com>

NANOG35

23-25 October 2005

Los Angeles

Presentation Slides

- **Available on**

  ftp://ftp-eng.cisco.com

  /pfs/seminars/NANOG35-Multihoming.pdf

  And on the NANOG 35 meeting pages at
  http://www.nanog.org/mtg-0510/pdf/smith.pdf

Preliminaries

- Presentation has many configuration examples
  - Uses Cisco IOS CLI
- Aimed at Service Providers
  - Techniques can be used by many enterprises too
- Feel free to ask questions

BGP Multihoming Techniques

- Why Multihome?
- Definition & Options
- Preparing the Network
- Basic Multihoming
- Service Provider Multihoming
- Complex Cases & Caveats

Why Multihome?

It’s all about redundancy, diversity & reliability

Why Multihome?

- Redundancy

  One connection to internet means the network is dependent on:

  Local router (configuration, software, hardware)

  WAN media (physical failure, carrier failure)

  Upstream Service Provider (configuration, software, hardware)

Why Multihome?

- Reliability

  Business critical applications demand continuous availability

  Lack of redundancy implies lack of reliability implies loss of revenue

Why Multihome?

- Supplier Diversity
  - Many businesses demand supplier diversity as a matter of course
  - Internet connection from two or more suppliers
    - With two or more diverse WAN paths
    - With two or more exit points
    - With two or more international connections
    - Two of everything

Why Multihome?

- Not really a reason, but oft quoted…
- Leverage:

    Playing one ISP off against the other for:

        Service Quality

        Service Offerings

        Availability

Why Multihome?

- **Summary:**

  Multihoming is easy to demand as requirement for any service provider or end-site network

  But what does it really mean:

    In real life?

    For the network?

    For the Internet?

  And how do we do it?

BGP Multihoming Techniques

- Why Multihome?
- Definition & Options
- Preparing the Network
- Basic Multihoming
- Service Provider Multihoming
- Complex Cases & Caveats

Multihoming: Definitions & Options

What does it mean, what do we need, and how do we do it?

Multihoming Definition

- More than one link external to the local network
  - two or more links to the same ISP
  - two or more links to different ISPs
- Usually two external facing routers
  - one router gives link and provider redundancy only

AS Numbers

- An Autonomous System Number is required by BGP
- Obtained from upstream ISP or Regional Registry (RIR)
  - AfriNIC, APNIC, ARIN, LACNIC, RIPE NCC
- Necessary when you have links to more than one ISP or to an exchange point
- 16 bit integer, ranging from 1 to 65534
  - Zero and 65535 are reserved
  - 64512 through 65534 are called Private ASNs

Private-AS – Application

- **Applications**

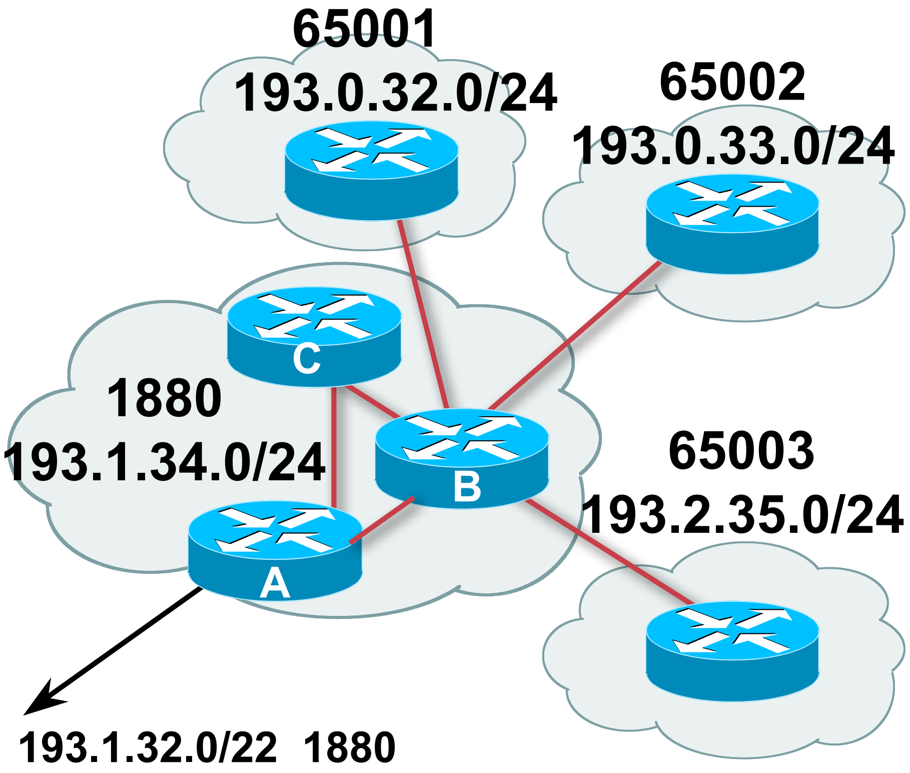
  An ISP with customers multihomed on their backbone (RFC2270)

  -or-

  A corporate network with several regions but connections to the Internet only in the core

  -or-

  Within a BGP Confederation



Private-AS – Removal

- Private ASNs MUST be removed from all prefixes announced to the public Internet

  Include configuration to remove private ASNs in the eBGP template

- As with RFC1918 address space, private ASNs are intended for internal use

  They should not be leaked to the public Internet

- Cisco IOS

  neighbor x.x.x.x remove-private-AS

Configuring Policy

- **Three BASIC Principles for IOS configuration examples throughout presentation:**

  prefix-lists to filter prefixes

  filter-lists to filter ASNs

  route-maps to apply policy

- **Route-maps can be used for filtering, but this is more “advanced” configuration**

Policy Tools

- **Local preference**
  - outbound traffic flows
- **Metric (MED)**
  - inbound traffic flows (local scope)
- **AS-PATH prepend**
  - inbound traffic flows (Internet scope)
- **Communities**
  - specific inter-provider peering

Originating Prefixes: Assumptions

- MUST announce assigned address block to Internet
- MAY also announce subprefixes – reachability is not guaranteed
- Current RIR minimum allocation is /21

  Several ISPs filter RIR blocks on this boundary

  Several ISPs filter the rest of address space according to the IANA assignments

  This activity is called “Net Police” by some

Originating Prefixes

- **Some ISPs publish their minimum allocation sizes per /8 address block**

AfriNIC:	www.afrinic.net/docs/policies/afpol-v4200407-000.htm
APNIC:	www.apnic.net/db/min-alloc.html
ARIN:	www.arin.net/reference/ip_blocks.html
LACNIC:	lacnic.net/en/registro/index.html
RIPE NCC:	www.ripe.net/ripe/docs/smallest-alloc-sizes.html

  Note that AfriNIC only publishes its current minimum allocation size, not the allocation size for its address blocks

- **IANA publishes the address space it has assigned to end-sites and allocated to the RIRs:**

  www.iana.org/assignments/ipv4-address-space

- **Several ISPs use this published information to filter prefixes on:**

  What should be routed (from IANA)

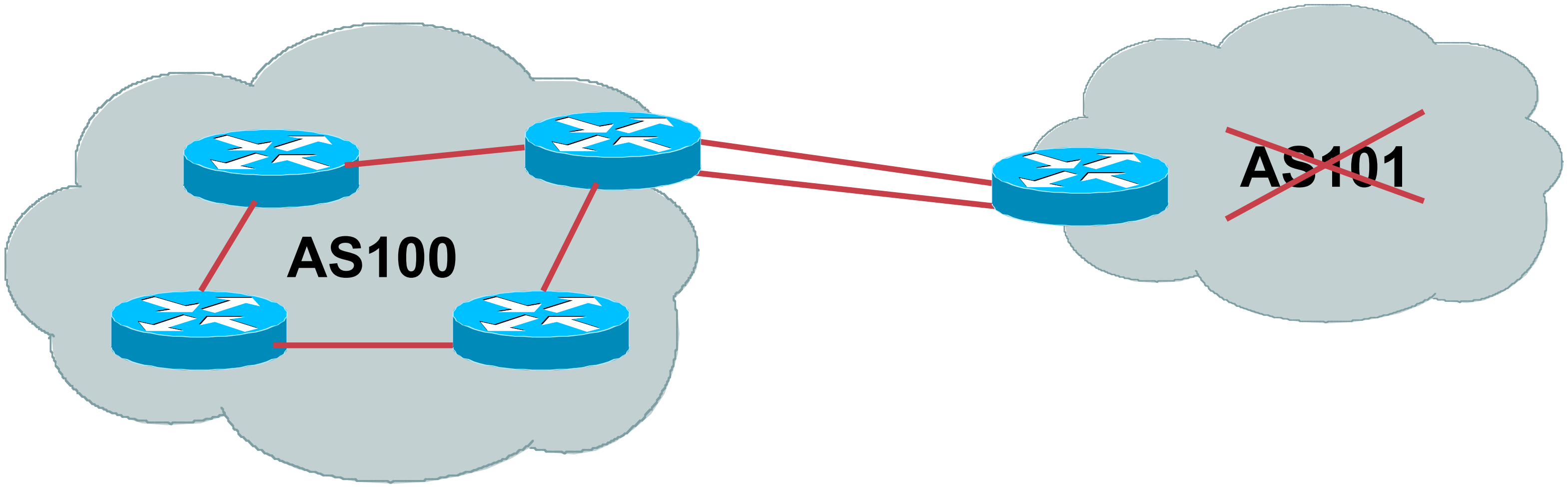
  The minimum allocation size from the RIRs

“Net Police” prefix list issues

- meant to “punish” ISPs who pollute the routing table with specifics rather than announcing aggregates
- impacts legitimate multihoming especially at the Internet’s edge
- impacts regions where domestic backbone is unavailable or costs $$$ compared with international bandwidth
- hard to maintain – requires updating when RIRs start allocating from new address blocks
- don’t do it unless consequences understood and you are prepared to keep the list current

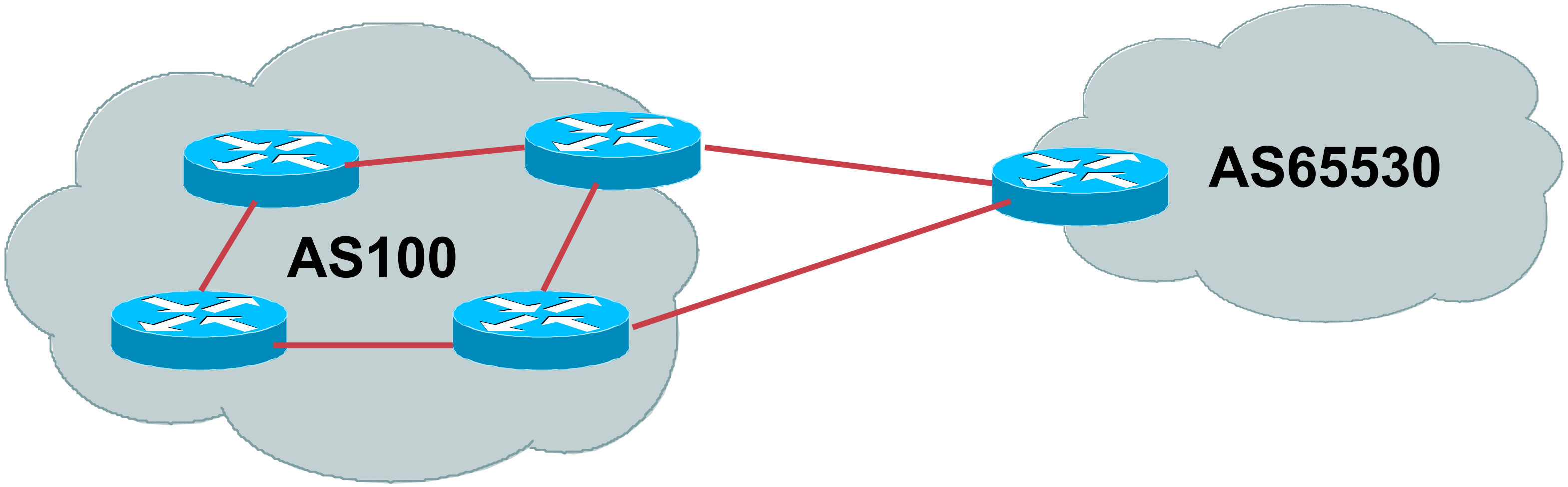
Consider using the Project Cymru bogon BGP feed instead
http://www.cymru.com/BGP/bogon-rs.html

Multihoming Scenarios: Stub Network



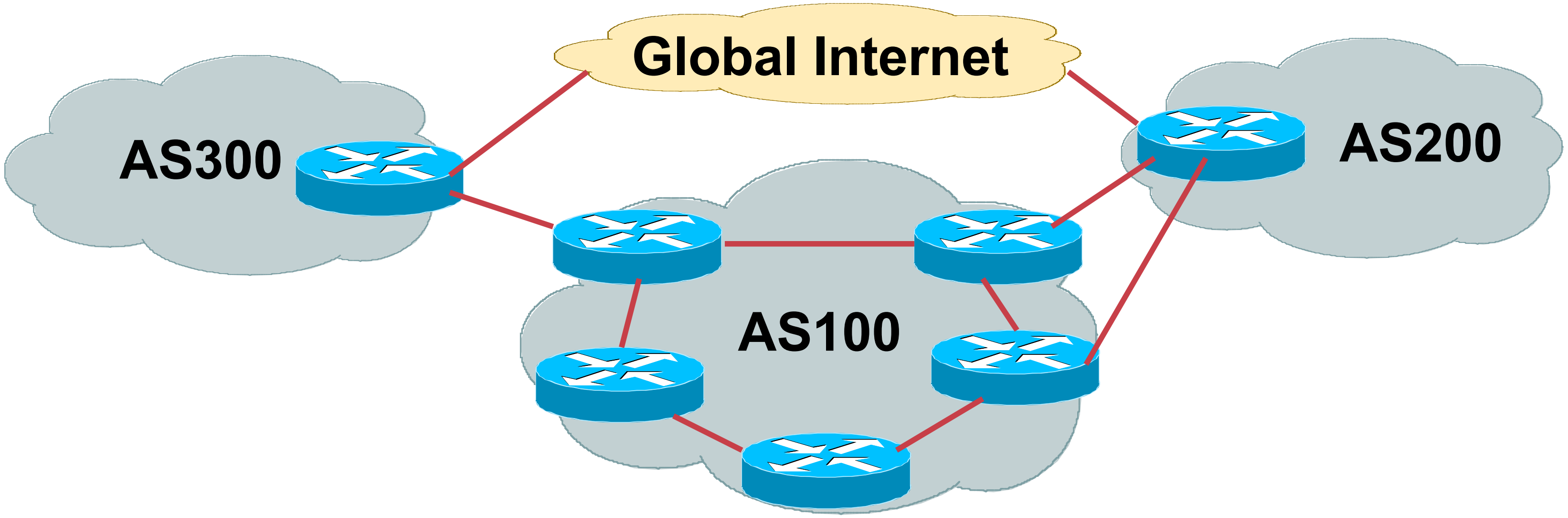
- No need for BGP
- Point static default to upstream ISP
- Router will load share on the two parallel circuits
- Upstream ISP advertises stub network
- Policy confined within upstream ISP’s policy

Multihoming Scenarios:
Multi-homed Stub Network



- Use BGP (not IGP or static) to loadshare
- Use private AS (ASN > 64511)
- Upstream ISP advertises stub network
- Policy confined within upstream ISP’s policy

Multihoming Scenarios: Multi-Homed Network



- **Many situations possible**

  multiple sessions to same ISP

  secondary for backup only

  load-share between primary and secondary

  selectively use different ISPs

Multihoming Scenarios: Multiple Sessions to an ISP

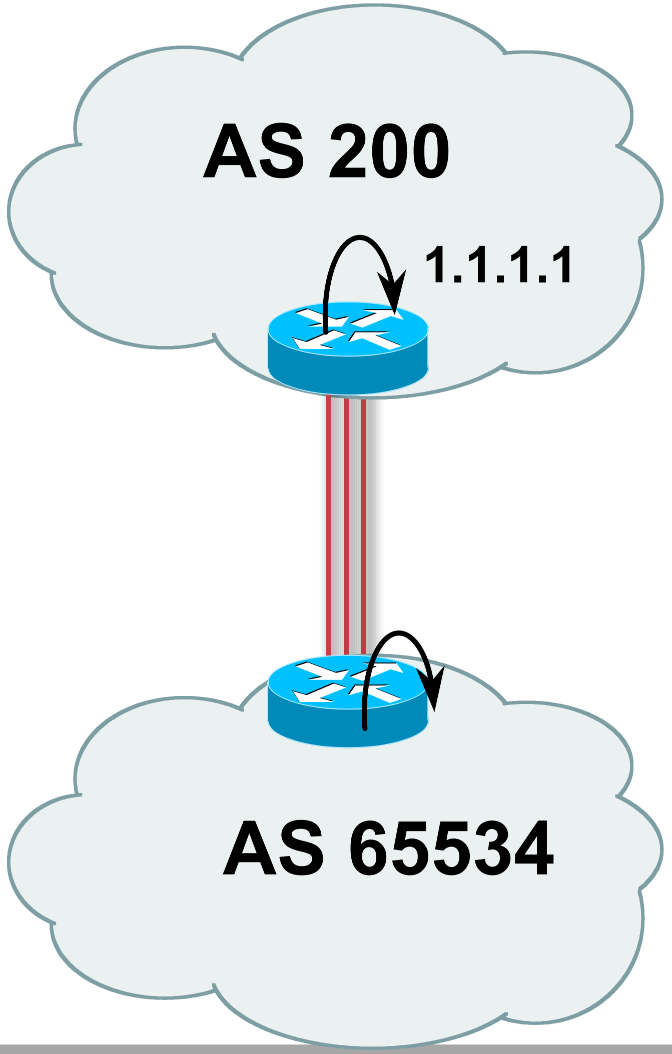
- Use eBGP multihop

  eBGP to loopback addresses

  eBGP prefixes learned with loopback address as next hop

- Cisco IOS

```
router bgp 65534
 neighbor 1.1.1.1 remote-as 200
 neighbor 1.1.1.1 ebgp-multihop 2
!
ip route 1.1.1.1 255.255.255.255 serial 1/0
ip route 1.1.1.1 255.255.255.255 serial 1/1
ip route 1.1.1.1 255.255.255.255 serial 1/2
```



Multiple Sessions to an ISP
eBGP multihop

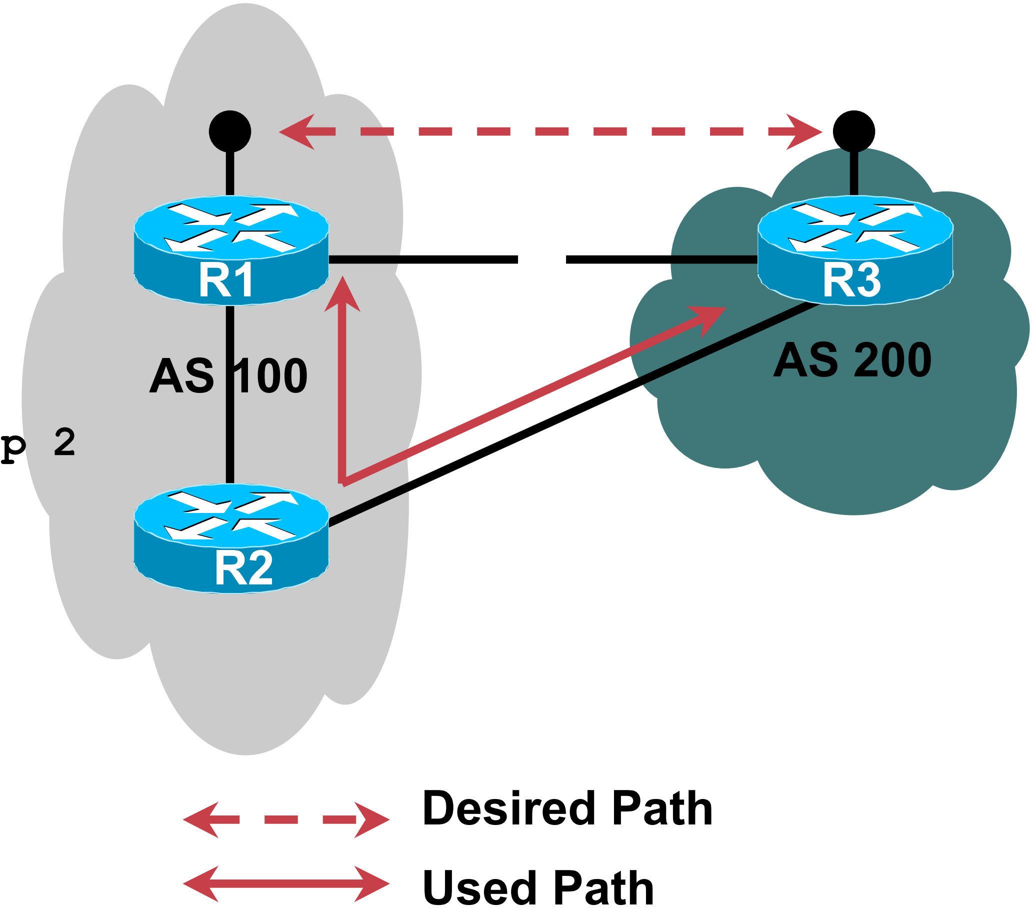
- One eBGP-multihop gotcha:

  R1 and R3 are eBGP peers that are loopback peering

  Configured with:

```
neighbor x.x.x.x ebgp-multihop 2
```

  If the R1 to R3 link goes down the session could establish via R2



Multiple Sessions to an ISP
eBGP multihop

- **Try and avoid use of ebgp-multihop unless:**

  It’s absolutely necessary –or–

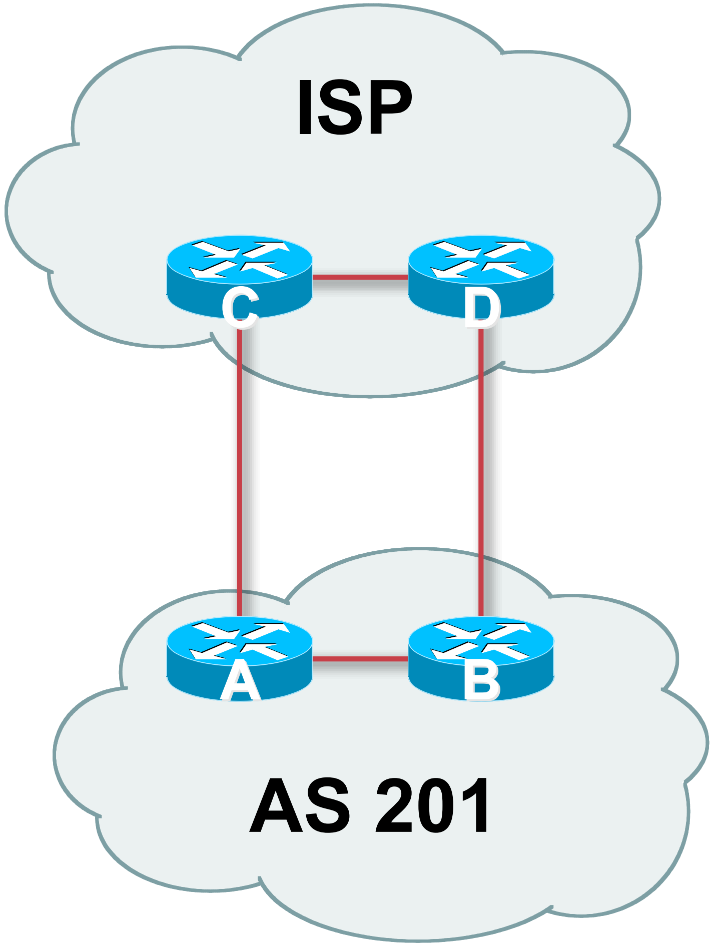
  Loadsharing across multiple links

- **Many ISPs discourage its use, for example:**

> **We will run eBGP multihop, but do not support it as a standard offering because customers generally have a hard time managing it due to:**
>
> **• routing loops**
>
> **• failure to realise that BGP session stability problems are usually due connectivity problems between their CPE and their BGP speaker**

Multihoming Scenarios: Multiple Sessions to an ISP

- Simplest scheme is to use defaults
- Learn/advertise prefixes for better control
- Planning and some work required to achieve loadsharing
  - Point default towards one ISP
  - Learn selected prefixes from second ISP
  - Modify the number of prefixes learnt to achieve acceptable load sharing
- No magic solution



BGP Multihoming Techniques

- Why Multihome?
- Definition & Options
- Preparing the Network
- Basic Multihoming
- Service Provider Multihoming
- Complex Cases & Caveats

Preparing the Network

Putting our own house in order first…

Preparing the Network

- We will deploy BGP across the network before we try and multihome
- BGP will be used therefore an ASN is required
- If multihoming to different ISPs, public ASN needed:

    Either go to upstream ISP who is a registry member, or

    Apply to the RIR yourself for a one off assignment, or

    Ask an ISP who is a registry member, or

    Join the RIR and get your own IP address allocation too (this option strongly recommended)!

Preparing the Network

- The network is not running any BGP at the moment

  single statically routed connection to upstream ISP

- The network is not running any IGP at all

  Static default and routes through the network to do “routing”

Preparing the Network
IGP

- Decide on IGP: OSPF or ISIS ☺
- Assign loopback interfaces and /32 addresses to each router which will run the IGP
    - Loopback is used for OSPF and BGP router id anchor
    - Used for iBGP and route origination
- Deploy IGP (e.g. OSPF)
    - IGP can be deployed with NO IMPACT on the existing static routing
        - OSPF distance is 110, static distance is 1
        - Smallest distance wins

Preparing the Network
IGP (cont)

- **Be prudent deploying IGP – keep the Link State Database Lean!**

  Router loopbacks go in IGP

  WAN point to point links go in IGP

  (In fact, any link where IGP dynamic routing will be run should go into IGP)

  Summarise on area/level boundaries (if possible) – i.e. think about your IGP address plan

Preparing the Network
IGP (cont)

- Routes which don’t go into the IGP include:

  Dynamic assignment pools (DSL/Cable/Dial)

  Customer point to point link addressing

    (using next-hop-self in iBGP ensures that these do NOT need to be in IGP)

  Static/Hosting LANs

  Customer assigned address space

  Anything else not listed in the previous slide

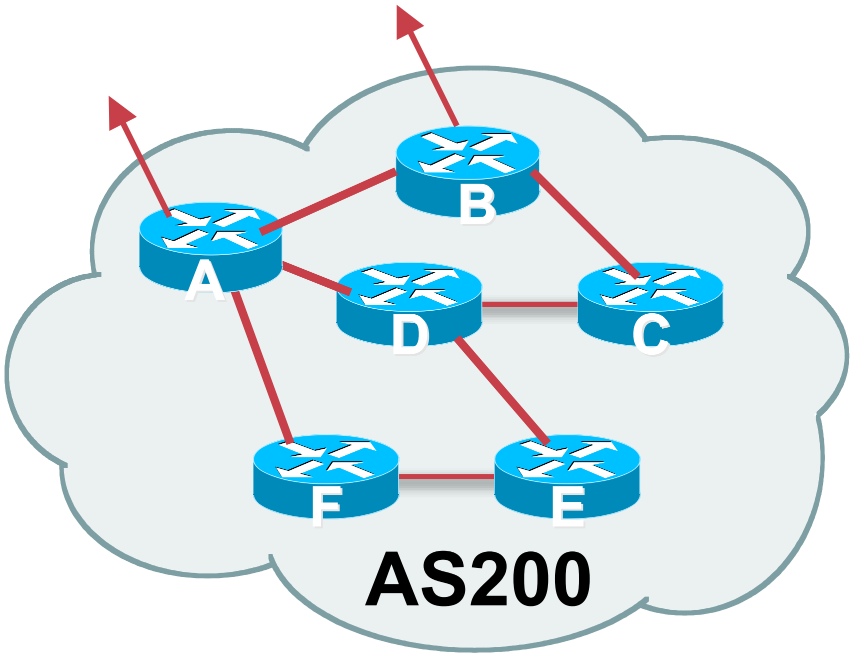
Preparing the Network
iBGP

- **Second step is to configure the local network to use iBGP**
- **iBGP can run on**

  all routers, or

  a subset of routers, or

  just on the upstream edge
- ***iBGP must run on all routers which are in the transit path between external connections***



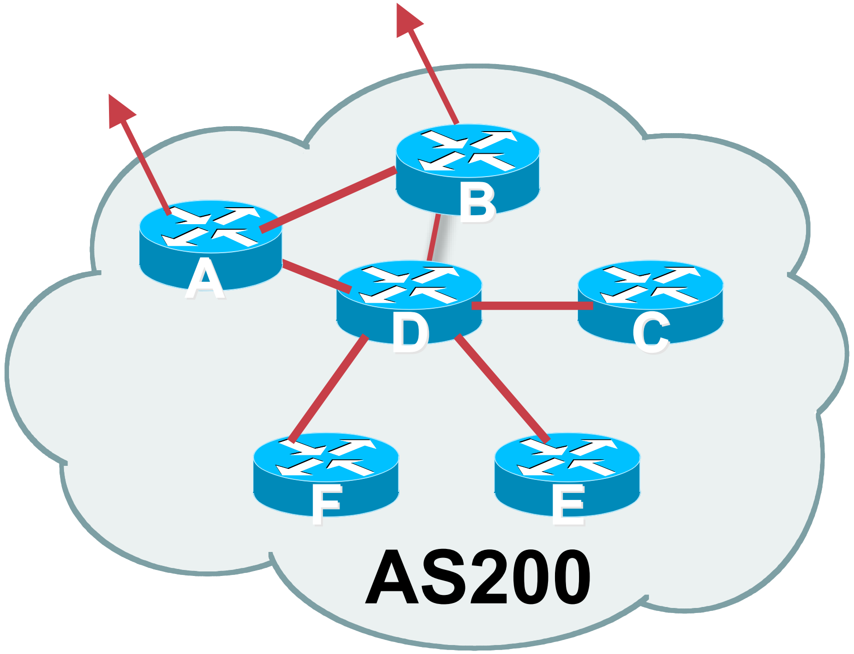
Preparing the Network
iBGP (Transit Path)

- ***iBGP must run on all routers which are in the transit path between external connections***
- **Routers C, E and F are not in the transit path**

  Static routes or IGP will suffice

- **Router D is in the transit path**

  Will need to be in iBGP mesh, otherwise routing loops will result



Preparing the Network
Layers

- Typical SP networks have three layers:

  Core – the backbone, usually the transit path

  Distribution – the middle, PoP aggregation layer

  Aggregation – the edge, the devices connecting customers

Preparing the Network Aggregation Layer

- iBGP is optional
    - Many ISPs run iBGP here, either partial routing (more common) or full routing (less common)
    - Full routing is not needed unless customers want full table
    - Partial routing is cheaper/easier, might usually consist of internal prefixes and, optionally, external prefixes to aid external load balancing
        - Communities and peer-groups make this administratively easy
- Many aggregation devices can’t run iBGP
    - Static routes from distribution devices for address pools
    - IGP for best exit

Preparing the Network Distribution Layer

- **Usually runs iBGP**
  - **Partial or full routing (as with aggregation layer)**
- **But does not have to run iBGP**
  - **IGP is then used to carry customer prefixes (does not scale)**
  - **IGP is used to determine nearest exit**
- **Networks which plan to grow large should deploy iBGP from day one**
  - **Migration at a later date is extra work**
  - **No extra overhead in deploying iBGP, indeed IGP benefits**

Preparing the Network
Core Layer

- Core of network is usually the transit path
- iBGP necessary between core devices
    - Full routes or partial routes:
        - Transit ISPs carry full routes in core
        - Edge ISPs carry partial routes only
- Core layer includes AS border routers

Preparing the Network
iBGP Implementation

Decide on:

- Best iBGP policy

  Will it be full routes everywhere, or partial, or some mix?

- iBGP scaling technique

  Community policy?

  Route-reflectors?

  Techniques such as peer groups and peer templates?

Preparing the Network
iBGP Implementation

- Then deploy iBGP:

  Step 1: Introduce iBGP mesh on chosen routers

  make sure that iBGP distance is greater than IGP distance (it usually is)

  Step 2: Install "customer" prefixes into iBGP

  Check! Does the network still work?

  Step 3: Carefully remove the static routing for the prefixes now in IGP and iBGP

  Check! Does the network still work?

  Step 4: Deployment of eBGP follows

Preparing the Network
iBGP Implementation

Install "customer" prefixes into iBGP?

- Customer assigned address space

  Network statement/static route combination

  Use unique community to identify customer assignments

- Customer facing point-to-point links

  Redistribute connected through filters which only permit point-to-point link addresses to enter iBGP

  Use a unique community to identify point-to-point link addresses (these are only required for your monitoring system)

- Dynamic assignment pools & local LANs

  Simple network statement will do this

  Use unique community to identify these networks

Preparing the Network
iBGP Implementation

Carefully remove static routes?

- Work on one router at a time:

  Check that static route for a particular destination is also learned either by IGP or by iBGP

  If so, remove it

  If not, establish why and fix the problem

  (Remember to look in the RIB, not the FIB!)

- Then the next router, until the whole PoP is done
- Then the next PoP, and so on until the network is now dependent on the IGP and iBGP you have deployed

Preparing the Network Completion

- Previous steps are NOT flag day steps

  Each can be carried out during different maintenance periods, for example:

  Step One on Week One

  Step Two on Week Two

  Step Three on Week Three

  And so on

  And with proper planning will have NO customer visible impact at all

Preparing the Network Configuration Summary

- IGP essential networks are in IGP
- Customer networks are now in iBGP
  - iBGP deployed over the backbone
  - Full or Partial or Upstream Edge only
- BGP distance is greater than any IGP
- Now ready to deploy eBGP

BGP Multihoming Techniques

- Why Multihome?
- Definition & Options
- Preparing the Network
- Basic Multihoming
- “BGP Traffic Engineering”
- Complex Cases & Caveats

Basic Multihoming

Learning to walk before we try running

Basic Multihoming

- No frills multihoming
- Will look at two cases:
  - Multihoming with the same ISP
  - Multihoming to different ISPs
- Will keep the examples easy
  - Understanding easy concepts will make the more complex scenarios easier to comprehend
  - All examples assume that the site multihoming has a /19 address block

Basic Multihoming

- This type is most commonplace at the edge of the Internet
  - Networks here are usually concerned with inbound traffic flows
  - Outbound traffic flows being “nearest exit” is usually sufficient
- Can apply to the leaf ISP as well as Enterprise networks

Basic Multihoming

Multihoming to the Same ISP

Basic Multihoming: Multihoming to the same ISP

- Use BGP for this type of multihoming

  use a private AS (ASN > 64511)

  There is no need or justification for a public ASN

  Making the nets of the end-site visible gives no useful information to the Internet

- Upstream ISP proxy aggregates

  in other words, announces only your address block to the Internet from their AS (as would be done if you had one statically routed connection)

Two links to the same ISP

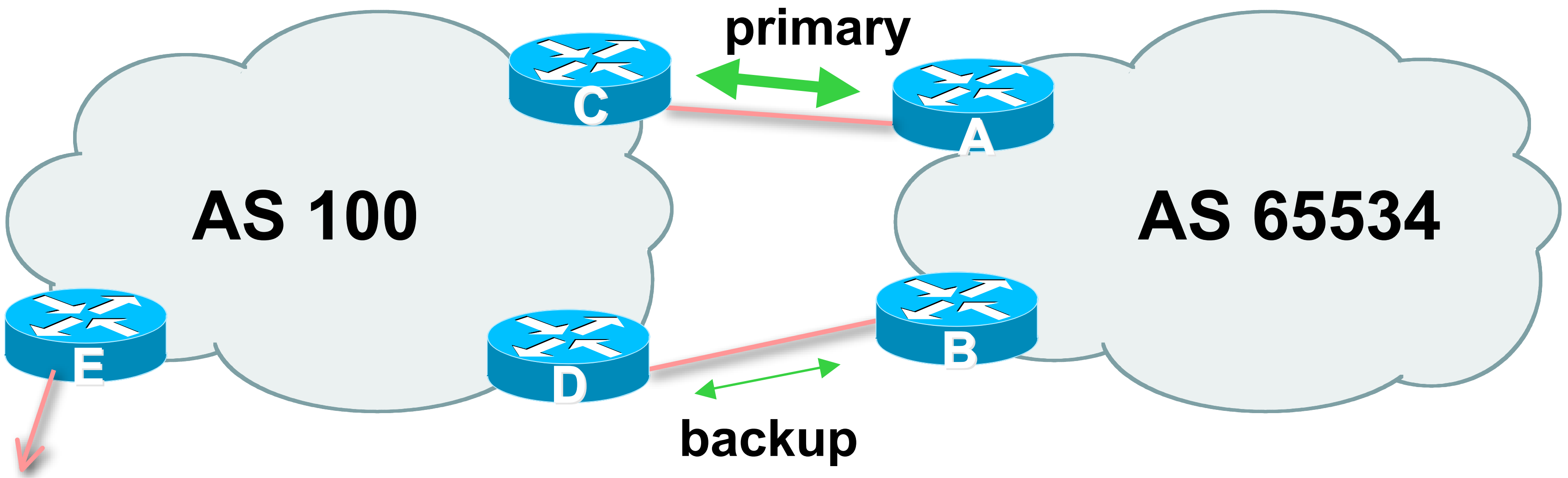
One link primary, the other link backup only

Two links to the same ISP (one as backup only)

- **Applies when end-site has bought a large primary WAN link to their upstream a small secondary WAN link as the backup**

  For example, primary path might be an E1, backup might be 64kbps

Two links to the same ISP (one as backup only)



- **Border router E in AS100 removes private AS and any customer subprefixes from Internet announcement**

Two links to the same ISP (one as backup only)

- **Announce /19 aggregate on each link**

  primary link:

  Outbound – announce /19 unaltered

  Inbound – receive default route

  backup link:

  Outbound – announce /19 with increased metric

  Inbound – received default, and reduce local preference

- **When one link fails, the announcement of the /19 aggregate via the other link ensures continued connectivity**

Two links to the same ISP (one as backup only)

- **Router A Configuration**

```
router bgp 65534
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.2 remote-as 100
 neighbor 122.102.10.2 description RouterC
 neighbor 122.102.10.2 prefix-list aggregate out
 neighbor 122.102.10.2 prefix-list default in
!
ip prefix-list aggregate permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
```

Two links to the same ISP
(one as backup only)

- **Router B Configuration**

```
router bgp 65534
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.6 remote-as 100
 neighbor 122.102.10.6 description RouterD
 neighbor 122.102.10.6 prefix-list aggregate out
 neighbor 122.102.10.6 route-map routerD-out out
 neighbor 122.102.10.6 prefix-list default in
 neighbor 122.102.10.6 route-map routerD-in in
!
..next slide
```

Two links to the same ISP (one as backup only)

```
ip prefix-list aggregate permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
route-map routerD-out permit 10
 match ip address prefix-list aggregate
 set metric 10
route-map routerD-out permit 20
!
route-map routerD-in permit 10
 set local-preference 90
!
```

Two links to the same ISP
(one as backup only)

- **Router C Configuration (main link)**

```
router bgp 100
 neighbor 122.102.10.1 remote-as 65534
 neighbor 122.102.10.1 default-originate
 neighbor 122.102.10.1 prefix-list Customer in
 neighbor 122.102.10.1 prefix-list default out
!
ip prefix-list Customer permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
```

Two links to the same ISP (one as backup only)

- **Router D Configuration (backup link)**

```
router bgp 100
 neighbor 122.102.10.5 remote-as 65534
 neighbor 122.102.10.5 default-originate
 neighbor 122.102.10.5 prefix-list Customer in
 neighbor 122.102.10.5 prefix-list default out
!
ip prefix-list Customer permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
```

Two links to the same ISP (one as backup only)

- Router E Configuration

```
router bgp 100
 neighbor 122.102.10.17 remote-as 110
 neighbor 122.102.10.17 remove-private-AS
 neighbor 122.102.10.17 prefix-list Customer out
!
ip prefix-list Customer permit 121.10.0.0/19
```

- Router E removes the private AS and customer’s subprefixes from external announcements
- Private AS still visible inside AS100

Two links to the same ISP

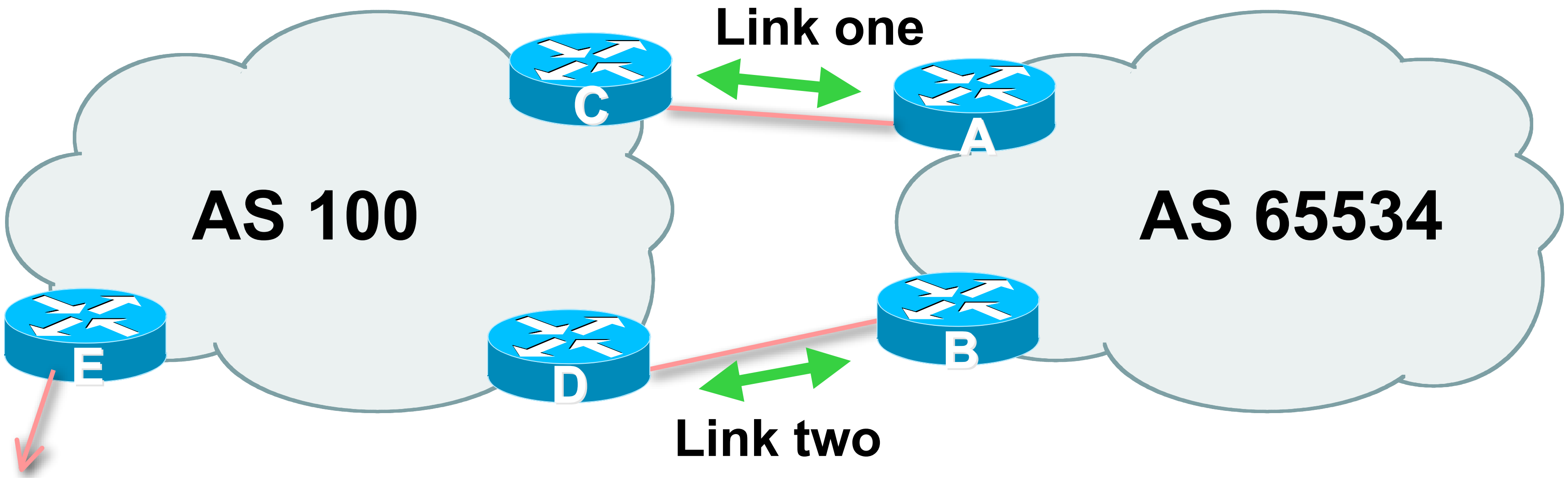
With Loadsharing

Loadsharing to the same ISP

- More common case
- End sites tend not to buy circuits and leave them idle, only used for backup as in previous example
- This example assumes equal capacity circuits

    Unequal capacity circuits requires more refinement – see later

Loadsharing to the same ISP



- Border router E in AS100 removes private AS and any customer subprefixes from Internet announcement

Loadsharing to the same ISP

- Announce /19 aggregate on each link
- Split /19 and announce as two /20s, one on each link

  basic inbound loadsharing

  assumes equal circuit capacity and even spread of traffic across address block
- Vary the split until "perfect" loadsharing achieved
- Accept the default from upstream

  basic outbound loadsharing by nearest exit

  okay in first approx as most ISP and end-site traffic is inbound

Loadsharing to the same ISP

- **Router A Configuration**

```
router bgp 65534
 network 121.10.0.0 mask 255.255.224.0
 network 121.10.0.0 mask 255.255.240.0
 neighbor 122.102.10.2 remote-as 100
 neighbor 122.102.10.2 prefix-list routerC out
 neighbor 122.102.10.2 prefix-list default in
!
ip prefix-list default permit 0.0.0.0/0
ip prefix-list routerC permit 121.10.0.0/20
ip prefix-list routerC permit 121.10.0.0/19
!
ip route 121.10.0.0 255.255.240.0 null0
ip route 121.10.0.0 255.255.224.0 null0
```

Router B configuration is similar but with the other /20

Loadsharing to the same ISP

- **Router C Configuration**

```
router bgp 100
 neighbor 122.102.10.1 remote-as 65534
 neighbor 122.102.10.1 default-originate
 neighbor 122.102.10.1 prefix-list Customer in
 neighbor 122.102.10.1 prefix-list default out
!
ip prefix-list Customer permit 121.10.0.0/19 le 20
ip prefix-list default permit 0.0.0.0/0
```

- **Router C only allows in /19 and /20 prefixes from customer block**
- **Router D configuration is identical**

Loadsharing to the same ISP

- Loadsharing configuration is only on customer router
- Upstream ISP has to

  remove customer subprefixes from external announcements

  remove private AS from external announcements
- Could also use BGP communities

Two links to the same ISP

Multiple Dualhomed Customers (RFC2270)

Multiple Dualhomed Customers (RFC2270)

- **Unusual for an ISP just to have one dualhomed customer**

  Valid/valuable service offering for an ISP with multiple PoPs

  Better for ISP than having customer multihome with another provider!

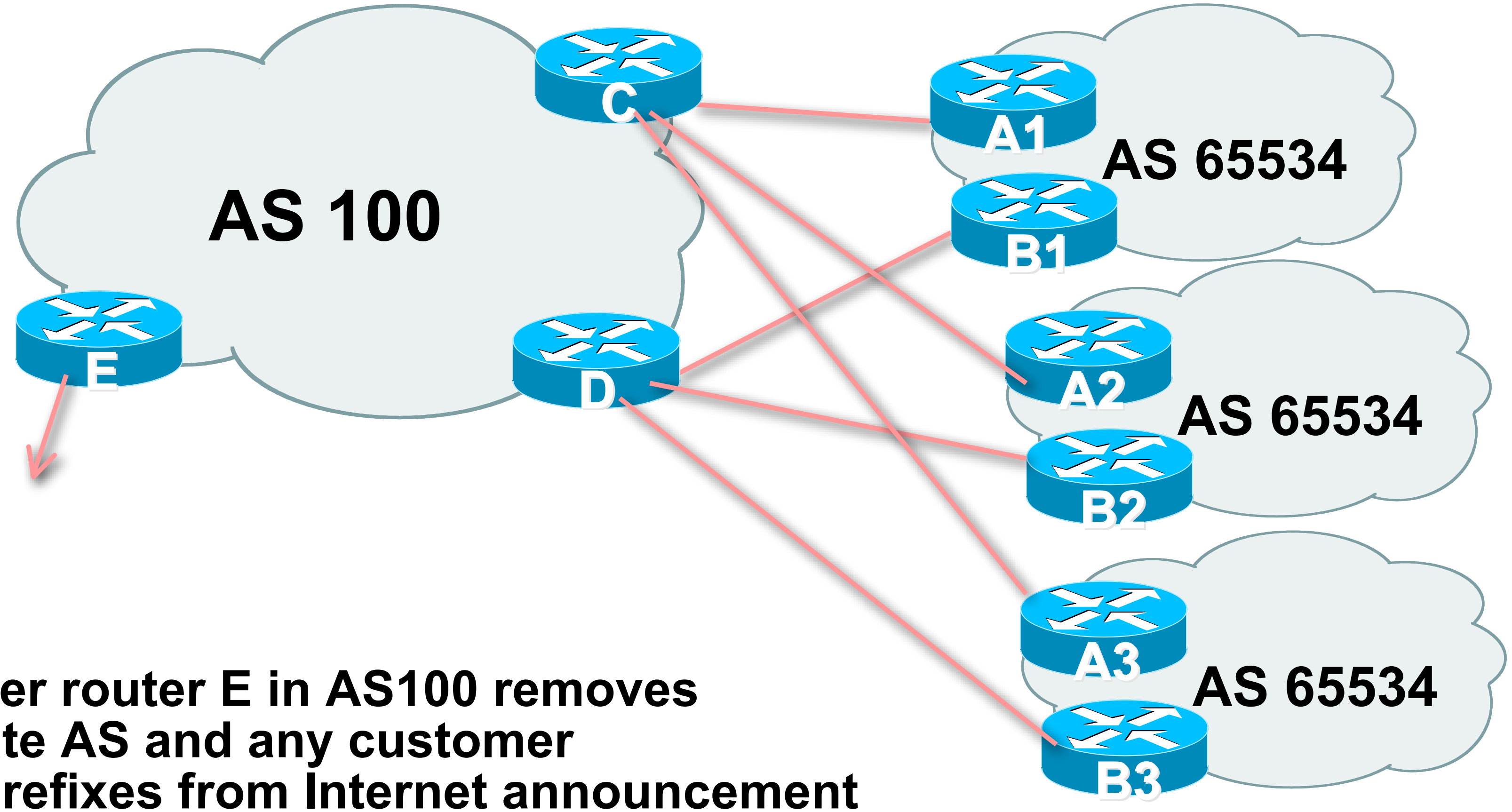
- **Look at scaling the configuration**

  ⇒ Simplifying the configuration

  Using templates, peer-groups, etc

  Every customer has the same configuration (basically)

Multiple Dualhomed Customers (RFC2270)



- **Border router E in AS100 removes private AS and any customer subprefixes from Internet announcement**

Multiple Dualhomed Customers

- Customer announcements as per previous example
- Use the *same* private AS for each customer

  documented in RFC2270

  address space is not overlapping

  each customer hears default only
- Router A*n* and B*n* configuration same as Router A and B previously

Multiple Dualhomed Customers

- **Router A1 Configuration**

```
router bgp 65534
 network 121.10.0.0 mask 255.255.224.0
 network 121.10.0.0 mask 255.255.240.0
 neighbor 122.102.10.2 remote-as 100
 neighbor 122.102.10.2 prefix-list routerC out
 neighbor 122.102.10.2 prefix-list default in
!
ip prefix-list default permit 0.0.0.0/0
ip prefix-list routerC permit 121.10.0.0/20
ip prefix-list routerC permit 121.10.0.0/19
!
ip route 121.10.0.0 255.255.240.0 null0
ip route 121.10.0.0 255.255.224.0 null0
```

Router B1 configuration is similar but for the other /20

Multiple Dualhomed Customers

- **Router C Configuration**

```
router bgp 100
 neighbor bgp-customers peer-group
 neighbor bgp-customers remote-as 65534
 neighbor bgp-customers default-originate
 neighbor bgp-customers prefix-list default out
 neighbor 122.102.10.1 peer-group bgp-customers
 neighbor 122.102.10.1 description Customer One
 neighbor 122.102.10.1 prefix-list Customer1 in
 neighbor 122.102.10.9 peer-group bgp-customers
 neighbor 122.102.10.9 description Customer Two
 neighbor 122.102.10.9 prefix-list Customer2 in
```

Multiple Dualhomed Customers

```
 neighbor 122.102.10.17 peer-group bgp-customers
 neighbor 122.102.10.17 description Customer Three
 neighbor 122.102.10.17 prefix-list Customer3 in
!
ip prefix-list Customer1 permit 121.10.0.0/19 le 20
ip prefix-list Customer2 permit 121.16.64.0/19 le 20
ip prefix-list Customer3 permit 121.14.192.0/19 le 20
ip prefix-list default permit 0.0.0.0/0
```

- Router C only allows in /19 and /20 prefixes from customer block
- Router D configuration is almost identical

Multiple Dualhomed Customers

- **Router E Configuration**

  assumes customer address space is not part of upstream's address block

```
router bgp 100
 neighbor 122.102.10.17 remote-as 110
 neighbor 122.102.10.17 remove-private-AS
 neighbor 122.102.10.17 prefix-list Customers out
!
ip prefix-list Customers permit 121.10.0.0/19
ip prefix-list Customers permit 121.16.64.0/19
ip prefix-list Customers permit 121.14.192.0/19
```

- **Private AS still visible inside AS100**

Multiple Dualhomed Customers

- If customers' prefixes come from ISP's address block

    do NOT announce them to the Internet

    announce ISP aggregate only

- Router E configuration:

```
router bgp 100
 neighbor 122.102.10.17 remote-as 110
 neighbor 122.102.10.17 prefix-list my-aggregate out
!
ip prefix-list my-aggregate permit 121.8.0.0/13
```

Basic Multihoming

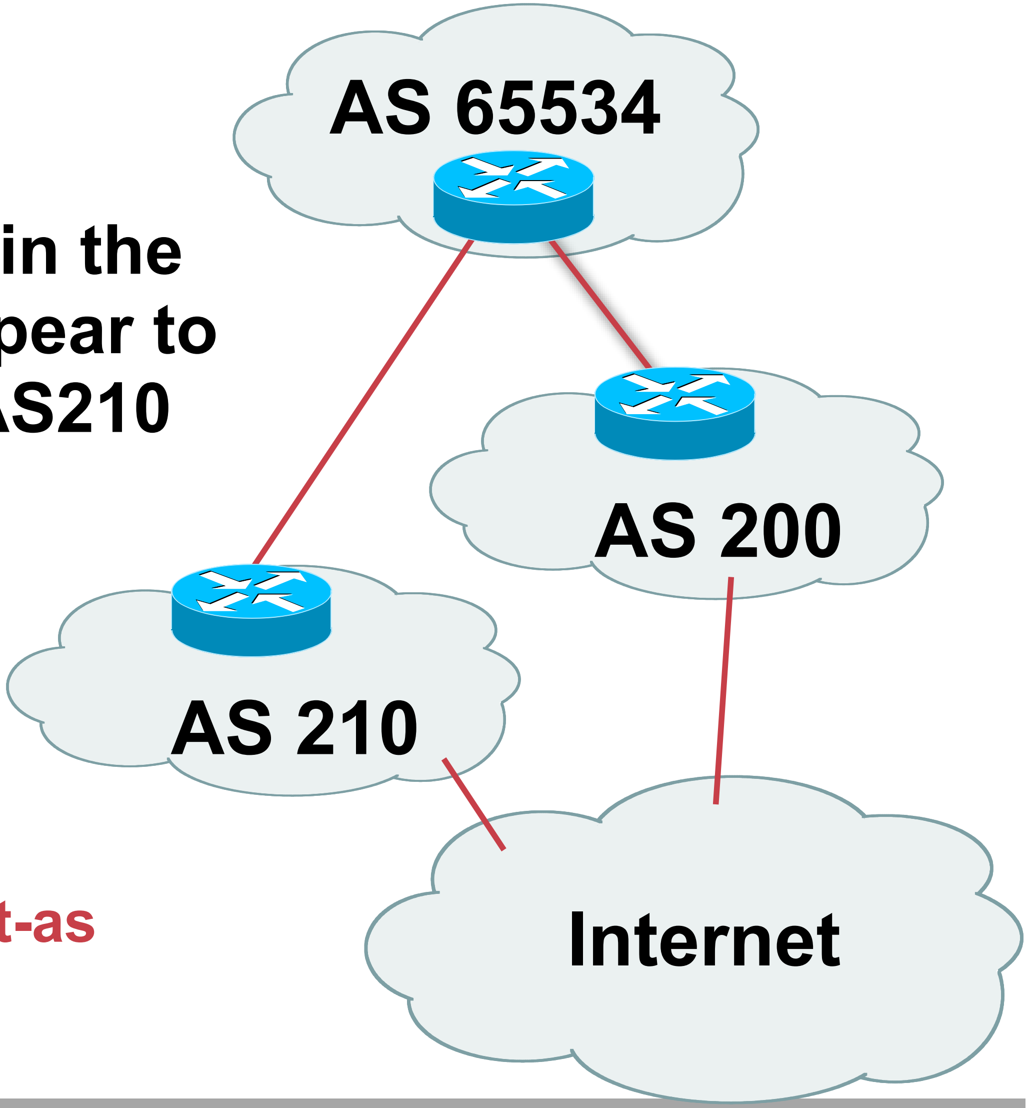
Multihoming to different ISPs

Two links to different ISPs

- Use a Public AS
  - Or use private AS if agreed with the other ISP
  - But some people don't like the "inconsistent-AS" which results from use of a private-AS
- Address space comes from
  - both upstreams or
  - Regional Internet Registry
- Configuration concepts very similar

Inconsistent-AS?

- Viewing the prefixes originated by AS65534 in the Internet shows they appear to be originated by both AS210 and AS200
  - This is NOT bad
  - Nor is it illegal
- Cisco IOS command is
  - show ip bgp inconsistent-as



Two links to different ISPs

One link primary, the other link backup only

Two links to different ISPs (one as backup only)

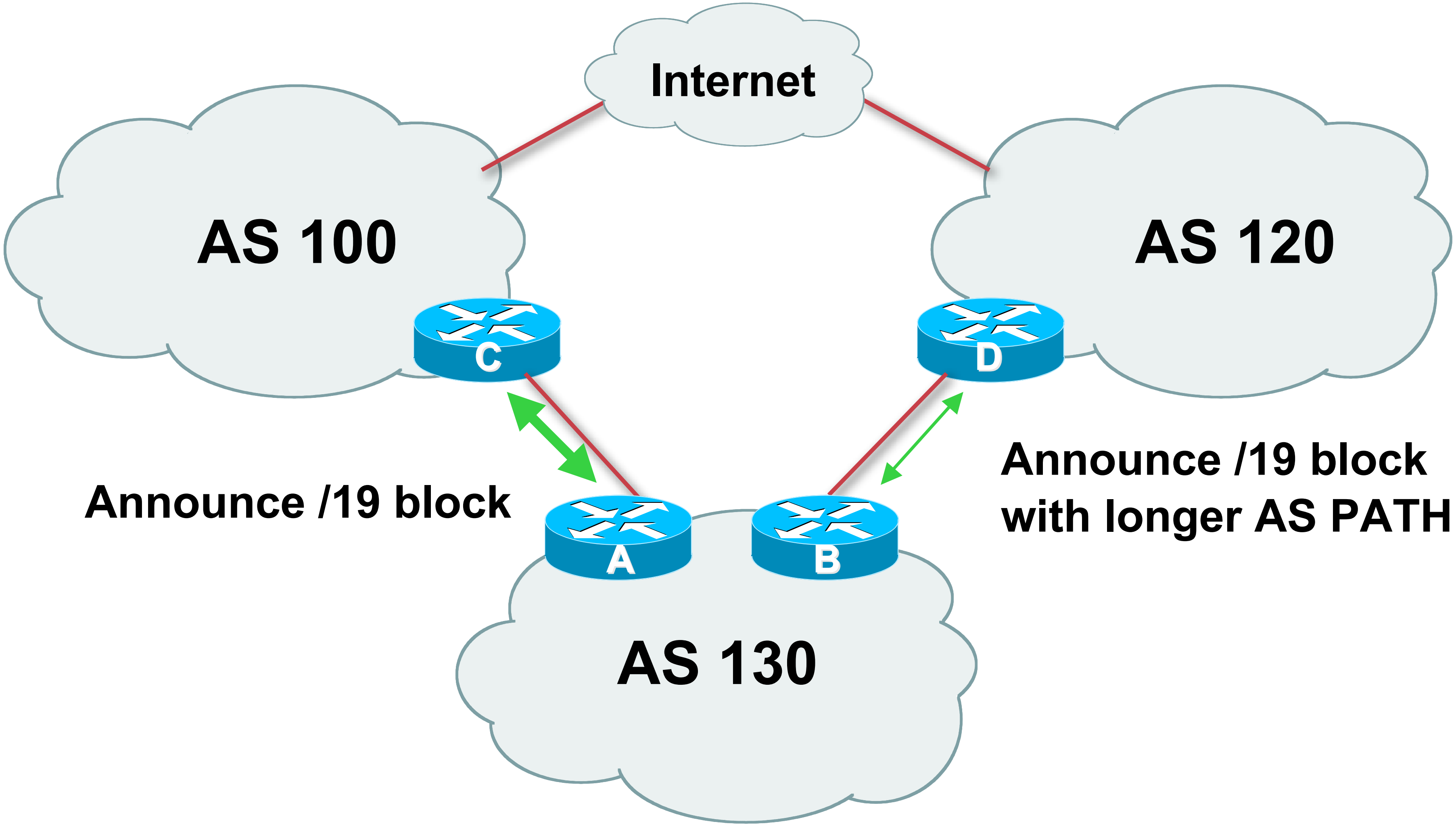
- Announce /19 aggregate on each link

  primary link makes standard announcement

  backup link lengthens the AS PATH by using AS PATH prepend

- When one link fails, the announcement of the /19 aggregate via the other link ensures continued connectivity

Two links to different ISPs (one as backup only)



Announce /19 block

Announce /19 block with longer AS PATH

Two links to different ISPs
(one as backup only)

- **Router A Configuration**

```
router bgp 130
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.1 remote-as 100
 neighbor 122.102.10.1 prefix-list aggregate out
 neighbor 122.102.10.1 prefix-list default in
!
ip prefix-list aggregate permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
```

Two links to different ISPs (one as backup only)

- **Router B Configuration**

```
router bgp 130
 network 121.10.0.0 mask 255.255.224.0
 neighbor 120.1.5.1 remote-as 120
 neighbor 120.1.5.1 prefix-list aggregate out
 neighbor 120.1.5.1 route-map routerD-out out
 neighbor 120.1.5.1 prefix-list default in
 neighbor 120.1.5.1 route-map routerD-in in
!
ip prefix-list aggregate permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
route-map routerD-out permit 10
 set as-path prepend 130 130 130
!
route-map routerD-in permit 10
 set local-preference 80
```

Two links to different ISPs (one as backup only)

- Not a common situation as most sites tend to prefer using whatever capacity they have
- But it shows the basic concepts of using local-prefs and AS-path prepends for engineering traffic in the chosen direction

Two links to different ISPs

With Loadsharing

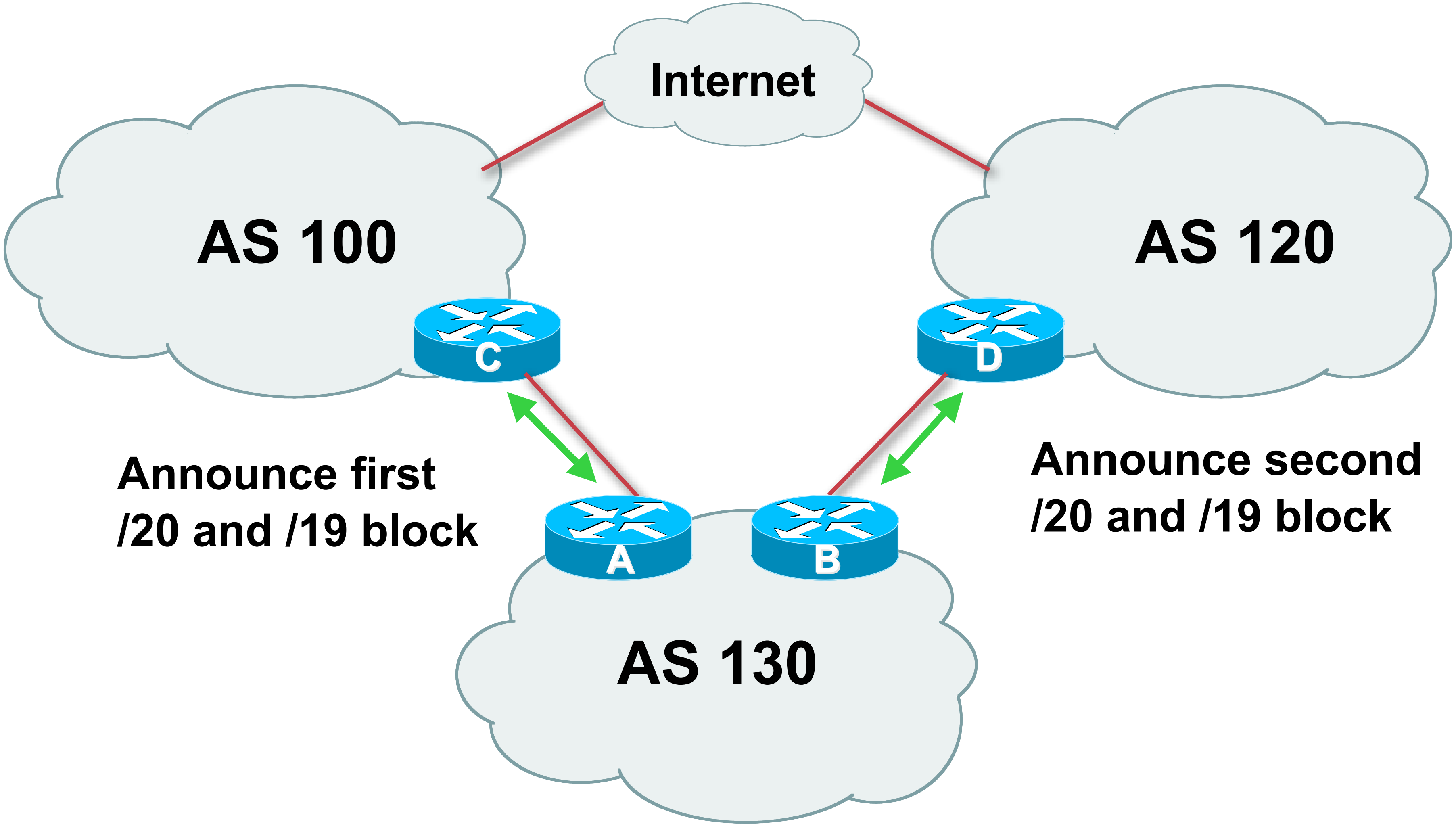
Two links to different ISPs (with loadsharing)

- **Announce /19 aggregate on each link**
- **Split /19 and announce as two /20s, one on each link**

  basic inbound loadsharing

- **When one link fails, the announcement of the /19 aggregate via the other ISP ensures continued connectivity**

Two links to different ISPs (with loadsharing)



Announce first /20 and /19 block

Announce second /20 and /19 block

Two links to different ISPs (with loadsharing)

- **Router A Configuration**

```
router bgp 130
 network 121.10.0.0 mask 255.255.224.0
 network 121.10.0.0 mask 255.255.240.0
 neighbor 122.102.10.1 remote-as 100
 neighbor 122.102.10.1 prefix-list firstblock out
 neighbor 122.102.10.1 prefix-list default in
!
ip prefix-list default permit 0.0.0.0/0
!
ip prefix-list firstblock permit 121.10.0.0/20
ip prefix-list firstblock permit 121.10.0.0/19
```

Two links to different ISPs (with loadsharing)

- **Router B Configuration**

```
router bgp 130
 network 121.10.0.0 mask 255.255.224.0
 network 121.10.16.0 mask 255.255.240.0
 neighbor 120.1.5.1 remote-as 120
 neighbor 120.1.5.1 prefix-list secondblock out
 neighbor 120.1.5.1 prefix-list default in
!
ip prefix-list default permit 0.0.0.0/0
!
ip prefix-list secondblock permit 121.10.16.0/20
ip prefix-list secondblock permit 121.10.0.0/19
```

Two links to different ISPs (with loadsharing)

- Loadsharing in this case is very basic
- But shows the first steps in designing a load sharing solution

  Start with a simple concept

  And build on it…!

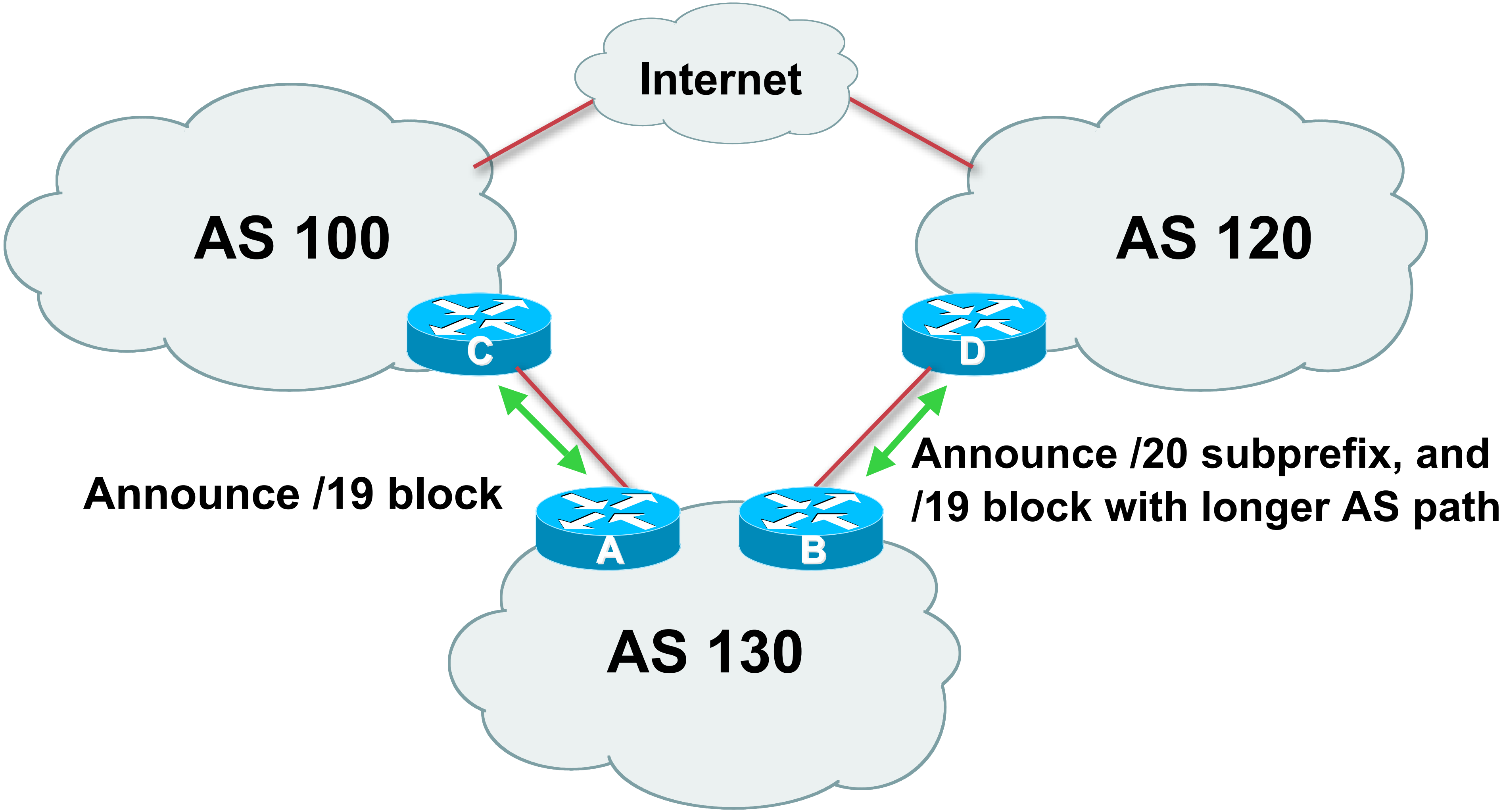
Two links to different ISPs

More Controlled Loadsharing

Loadsharing with different ISPs

- Announce /19 aggregate on each link
  - On first link, announce /19 as normal
  - On second link, announce /19 with longer AS PATH, and announce one /20 subprefix
    - controls loadsharing between upstreams and the Internet
- Vary the subprefix size and AS PATH length until “perfect” loadsharing achieved
- Still require redundancy!

Loadsharing with different ISPs



Announce /19 block

Announce /20 subprefix, and /19 block with longer AS path

Loadsharing with different ISPs

- **Router A Configuration**

```
router bgp 130
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.1 remote-as 100
 neighbor 122.102.10.1 prefix-list default in
 neighbor 122.102.10.1 prefix-list aggregate out
!
ip prefix-list aggregate permit 121.10.0.0/19
```

Loadsharing with different ISPs

- **Router B Configuration**

```
router bgp 130
 network 121.10.0.0 mask 255.255.224.0
 network 121.10.16.0 mask 255.255.240.0
 neighbor 120.1.5.1 remote-as 120
 neighbor 120.1.5.1 prefix-list default in
 neighbor 120.1.5.1 prefix-list subblocks out
 neighbor 120.1.5.1 route-map routerD out
!
route-map routerD permit 10
 match ip address prefix-list aggregate
 set as-path prepend 130 130
route-map routerD permit 20
!
ip prefix-list subblocks permit 121.10.0.0/19 le 20
ip prefix-list aggregate permit 121.10.0.0/19
```

Loadsharing with different ISPs

- This example is more commonplace
- Shows how ISPs and end-sites subdivide address space frugally, as well as use the AS-PATH prepend concept to optimise the load sharing between different ISPs
- Notice that the /19 aggregate block is ALWAYS announced

BGP Multihoming Techniques

- Why Multihome?
- Definition & Options
- Preparing the Network
- Basic Multihoming
- “BGP Traffic Engineering”
- Complex Cases & Caveats

Service Provider Multihoming

BGP Traffic Engineering

Service Provider Multihoming

- **Previous examples dealt with loadsharing inbound traffic**
  - Of primary concern at Internet edge
  - What about outbound traffic?
- **Transit ISPs strive to balance traffic flows in both directions**
  - Balance link utilisation
  - Try and keep most traffic flows symmetric
  - Some edge ISPs try and do this too
- **The original "Traffic Engineering"**

Service Provider Multihoming

- Balancing outbound traffic requires inbound routing information
  - Common solution is "full routing table"
  - Rarely necessary
    - Why use the "routing mallet" to try solve loadsharing problems?
  - "Keep It Simple" is often easier (and $$$ cheaper) than carrying N-copies of the full routing table

Service Provider Multihoming MYTHS!!

- Common MYTHS
- 1: You need the full routing table to multihome
  - People who sell router memory would like you to believe this
  - Only true if you are a transit provider
  - Full routing table can be a significant hindrance to multihoming
- 2: You need a BIG router to multihome
  - Router size is related to data rates, not running BGP
  - In reality, to multihome, your router needs to:
    - Have two interfaces,
    - Be able to talk BGP to at least two peers,
    - Be able to handle BGP attributes,
    - Handle at least one prefix
- 3: BGP is complex
  - In the wrong hands, yes it can be! Keep it Simple!

Service Provider Multihoming: Some Strategies

- Take the prefixes you need to aid traffic engineering

  Look at NetFlow data for popular sites

- Prefixes originated by your immediate neighbours and their neighbours will do more to aid load balancing than prefixes from ASNs many hops away

  Concentrate on local destinations

- Use default routing as much as possible

  Or use the full routing table with care

Service Provider Multihoming

- Examples
  - One upstream, one local peer
  - One upstream, local exchange point
  - Two upstreams, one local peer
- Require BGP and a public ASN
- Examples assume that the local network has their own /19 address block

Service Provider Multihoming

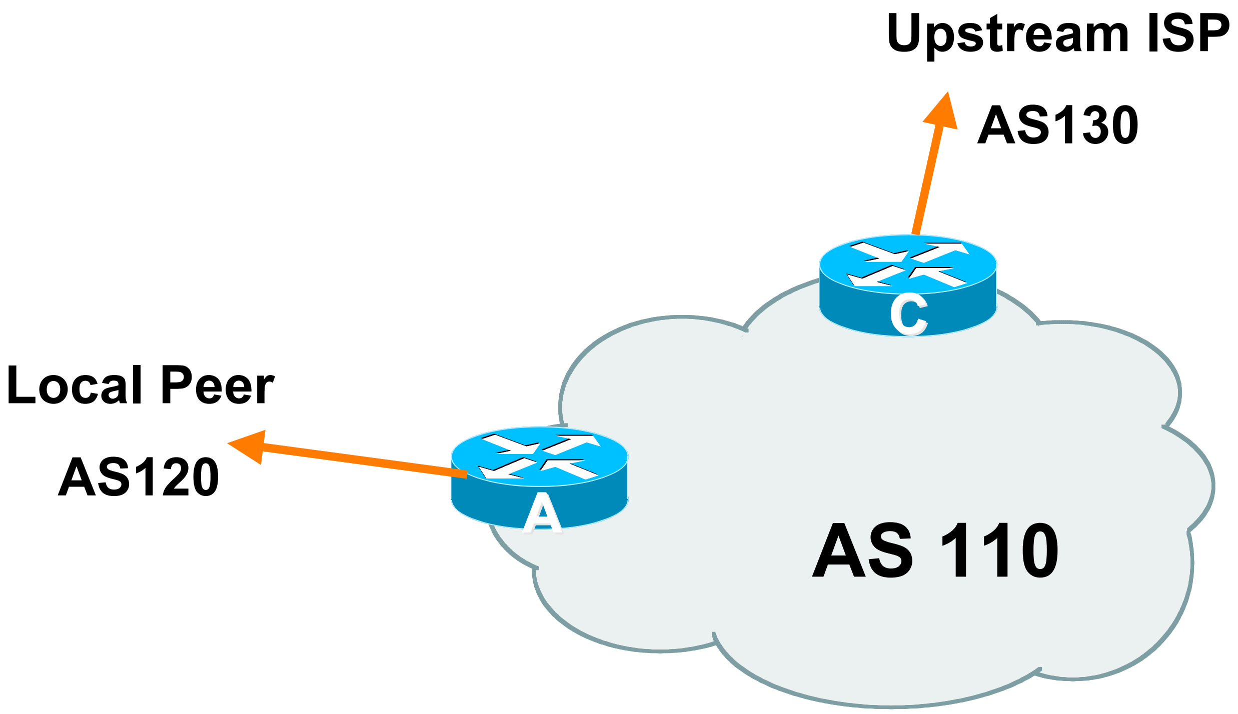
One upstream, one local peer

One Upstream, One Local Peer

- Very common situation in many regions of the Internet
- Connect to upstream transit provider to see the “Internet”
- Connect to the local competition so that local traffic stays local

  Saves spending valuable $ on upstream transit costs for local traffic

One Upstream, One Local Peer



One Upstream, One Local Peer

- **Announce /19 aggregate on each link**
- **Accept default route only from upstream**

  Either 0.0.0.0/0 or a network which can be used as default
- **Accept all routes from local peer**

One Upstream, One Local Peer

- **Router A Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.2 remote-as 120
 neighbor 122.102.10.2 prefix-list my-block out
 neighbor 122.102.10.2 prefix-list AS120-peer in
!
ip prefix-list AS120-peer permit 122.5.16.0/19
ip prefix-list AS120-peer permit 121.240.0.0/20
ip prefix-list my-block permit 121.10.0.0/19
!
ip route 121.10.0.0 255.255.224.0 null0
```

Prefix filters inbound

One Upstream, One Local Peer

- **Router A – Alternative Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.2 remote-as 120
 neighbor 122.102.10.2 prefix-list my-block out
 neighbor 122.102.10.2 filter-list 10 in
!
ip as-path access-list 10 permit ^(120_)+$
!
ip prefix-list my-block permit 121.10.0.0/19
!
ip route 121.10.0.0 255.255.224.0 null0
```

AS Path filters – more “trusting”

One Upstream, One Local Peer

- **Router C Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.1 remote-as 130
 neighbor 122.102.10.1 prefix-list default in
 neighbor 122.102.10.1 prefix-list my-block out
!
ip prefix-list my-block permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 121.10.0.0 255.255.224.0 null0
```

One Upstream, One Local Peer

- Two configurations possible for Router A
  - Filter-lists assume peer knows what they are doing
  - Prefix-list higher maintenance, but safer
  - Some ISPs use both
- Local traffic goes to and from local peer, everything else goes to upstream

Aside: Configuration Recommendation

- **Private Peers**
  - The peering ISPs exchange prefixes they originate
  - Sometimes they exchange prefixes from neighbouring ASNs too
- **Be aware that the private peer eBGP router should carry only the prefixes you want the private peer to receive**
  - Otherwise they could point a default route to you and unintentionally transit your backbone

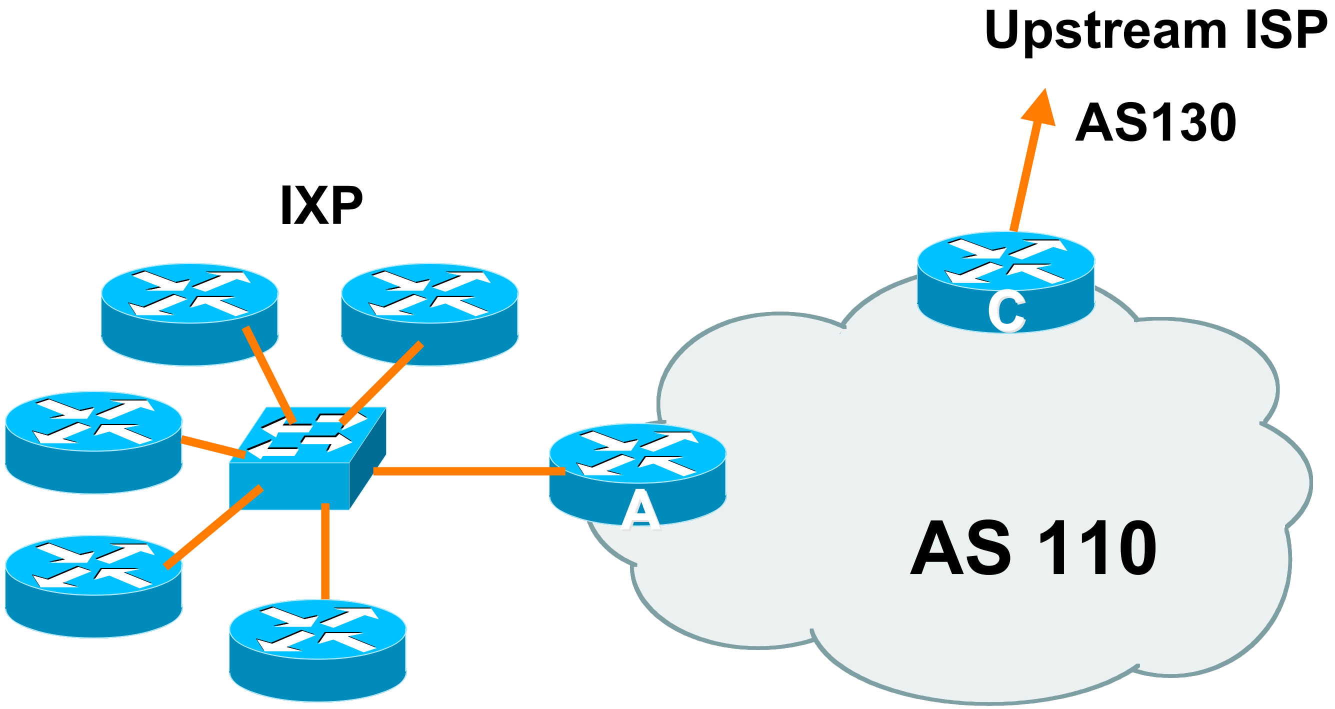
Service Provider Multihoming

One Upstream, Local Exchange Point

One Upstream, Local Exchange Point

- **Very common situation in many regions of the Internet**
- **Connect to upstream transit provider to see the "Internet"**
- **Connect to the local Internet Exchange Point so that local traffic stays local**
  - **Saves spending valuable $ on upstream transit costs for local traffic**

One Upstream, Local Exchange Point



One Upstream, Local Exchange Point

- Announce /19 aggregate to every neighbouring AS
- Accept default route only from upstream
  - Either 0.0.0.0/0 or a network which can be used as default
- Accept all routes originated by IXP peers

One Upstream, Local Exchange Point

- **Router A Configuration**

```
interface fastethernet 0/0
 description Exchange Point LAN
 ip address 120.5.10.1 mask 255.255.255.224
 ip verify unicast reverse-path
!
router bgp 110
 neighbor ixp-peers peer-group
 neighbor ixp-peers prefix-list my-block out
 neighbor ixp-peers remove-private-AS
 neighbor ixp-peers route-map set-local-pref in
..next slide
```

One Upstream, Local Exchange Point

```
neighbor 120.5.10.2 remote-as 100
neighbor 120.5.10.2 peer-group ixp-peers
neighbor 120.5.10.2 prefix-list peer100 in
neighbor 120.5.10.3 remote-as 101
neighbor 120.5.10.3 peer-group ixp-peers
neighbor 120.5.10.3 prefix-list peer101 in
neighbor 120.5.10.4 remote-as 102
neighbor 120.5.10.4 peer-group ixp-peers
neighbor 120.5.10.4 prefix-list peer102 in
neighbor 120.5.10.5 remote-as 103
neighbor 120.5.10.5 peer-group ixp-peers
neighbor 120.5.10.5 prefix-list peer103 in
```

..next slide

One Upstream, Local Exchange Point

```
!
ip prefix-list my-block permit 121.10.0.0/19
ip prefix-list peer100 permit 122.0.0.0/19
ip prefix-list peer101 permit 122.30.0.0/19
ip prefix-list peer102 permit 122.12.0.0/19
ip prefix-list peer103 permit 122.18.128.0/19
!
route-map set-local-pref permit 10
 set local-preference 150
!
```

One Upstream, Local Exchange

- **Note that Router A does not generate the aggregate for AS110**
  - **If Router A becomes disconnected from backbone, then the aggregate is no longer announced to the IX**
  - **BGP failover works as expected**
- **Note the inbound route-map which sets the local preference higher than the default**
  - **This ensures that local traffic crosses the IXP**
  - **(And avoids potential problems with uRPF check)**

One Upstream, Local Exchange Point

- **Router C Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.1 remote-as 130
 neighbor 122.102.10.1 prefix-list default in
 neighbor 122.102.10.1 prefix-list my-block out
!
ip prefix-list my-block permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 121.10.0.0 255.255.224.0 null0
```

One Upstream, Local Exchange Point

- **Note Router A configuration**
  - **Prefix-list higher maintenance, but safer**
  - **uRPF on the IX facing interface**
  - **No generation of AS110 aggregate**
- **IXP traffic goes to and from local IXP, everything else goes to upstream**

Aside: IXP Configuration Recommendation

- IXP peers

  The peering ISPs at the IXP exchange prefixes they originate

  Sometimes they exchange prefixes from neighbouring ASNs too

- Be aware that the IXP border router should carry only the prefixes you want the IXP peers to receive and the destinations you want them to be able to reach

  Otherwise they could point a default route to you and unintentionally transit your backbone

- If IXP router is at IX, and distant from your backbone

  Don’t originate your address block at your IXP router

Service Provider Multihoming

Two Upstreams, One local peer

Two Upstreams, One Local Peer

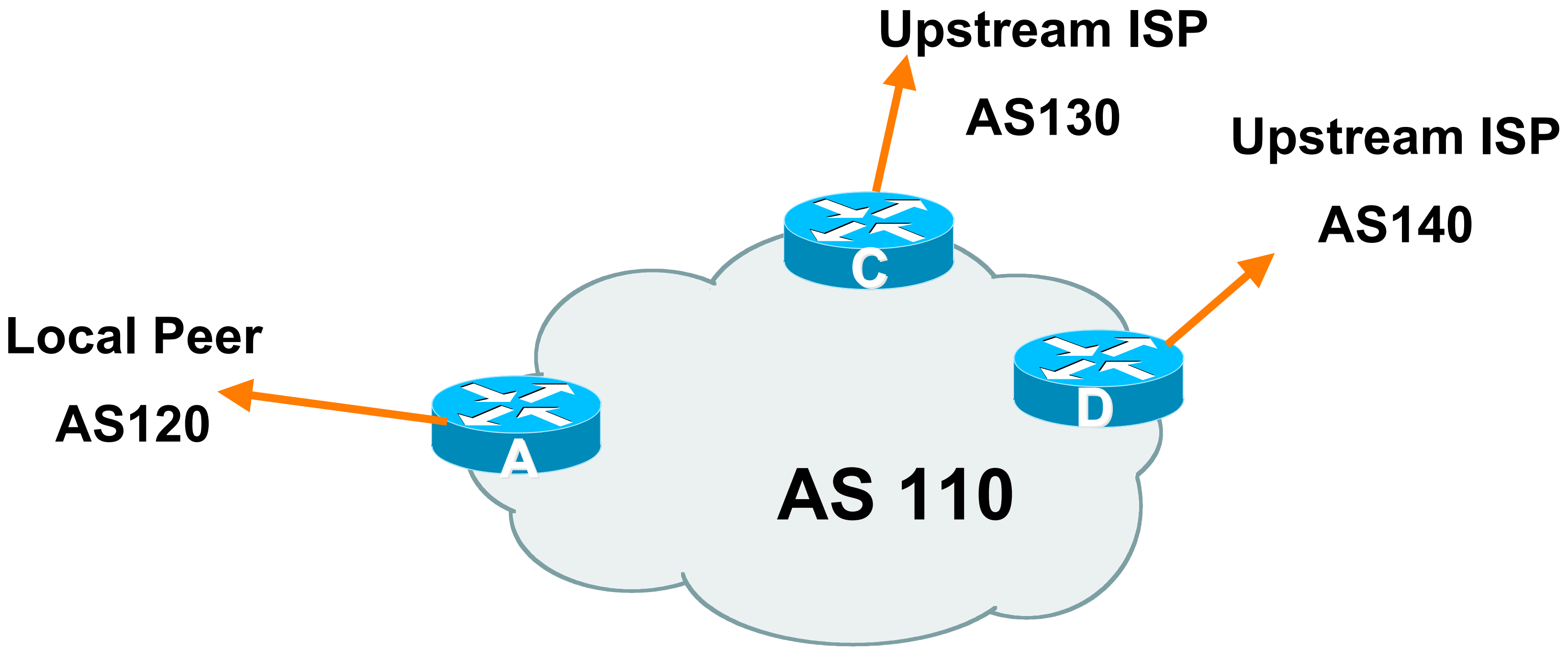
- Connect to both upstream transit providers to see the “Internet”

  Provides external redundancy and diversity – the reason to multihome

- Connect to the local peer so that local traffic stays local

  Saves spending valuable $ on upstream transit costs for local traffic

Two Upstreams, One Local Peer



Upstream ISP

AS130

Upstream ISP

AS140

Local Peer

AS120

AS 110

Two Upstreams, One Local Peer

- Announce /19 aggregate on each link
- Accept default route only from upstreams

  Either 0.0.0.0/0 or a network which can be used as default

- Accept all routes from local peer

Two Upstreams, One Local Peer

- Router A

  Same routing configuration as in example with one upstream and one local peer

  Same hardware configuration

Two Upstreams, One Local Peer

- **Router C Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.1 remote-as 130
 neighbor 122.102.10.1 prefix-list default in
 neighbor 122.102.10.1 prefix-list my-block out
!
ip prefix-list my-block permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 121.10.0.0 255.255.224.0 null0
```

Two Upstreams, One Local Peer

- **Router D Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.5 remote-as 140
 neighbor 122.102.10.5 prefix-list default in
 neighbor 122.102.10.5 prefix-list my-block out
!
ip prefix-list my-block permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 121.10.0.0 255.255.224.0 null0
```

Two Upstreams, One Local Peer

- This is the simple configuration for Router C and D
- Traffic out to the two upstreams will take nearest exit

  Inexpensive routers required

  This is not useful in practice especially for international links

  Loadsharing needs to be better

Two Upstreams, One Local Peer

- **Better configuration options:**
  - **Accept full routing from both upstreams**
    - **Expensive & unnecessary!**
  - **Accept default from one upstream and some routes from the other upstream**
    - **The way to go!**

Two Upstreams, One Local Peer
Full Routes

- **Router C Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.1 remote-as 130
 neighbor 122.102.10.1 prefix-list rfc1918-deny in
 neighbor 122.102.10.1 prefix-list my-block out
 neighbor 122.102.10.1 route-map AS130-loadshare in
!
ip prefix-list my-block permit 121.10.0.0/19
! See www.cymru.com/Documents/bogon-list.html
! ...for “RFC1918 and friends” list
..next slide
```

Allow all prefixes in apart from RFC1918 and friends

Two Upstreams, One Local Peer
Full Routes

```
ip route 121.10.0.0 255.255.224.0 null0
!
ip as-path access-list 10 permit ^(130_)+$
ip as-path access-list 10 permit ^(130_)+_[0-9]+$
!
route-map AS130-loadshare permit 10
 match ip as-path 10
 set local-preference 120
route-map AS130-loadshare permit 20
 set local-preference 80
!
```

Two Upstreams, One Local Peer
Full Routes

- **Router D Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.5 remote-as 140
 neighbor 122.102.10.5 prefix-list rfc1918-deny in
 neighbor 122.102.10.5 prefix-list my-block out
!
ip prefix-list my-block permit 121.10.0.0/19
! See www.cymru.com/Documents/bogon-list.html
! ...for “RFC1918 and friends” list
```

Allow all prefixes in apart from RFC1918 and friends

Two Upstreams, One Local Peer
Full Routes

- **Router C configuration:**
  - **Accept full routes from AS130**
  - **Tag prefixes originated by AS130 and AS130's neighbouring ASes with local preference 120**
    - **Traffic to those ASes will go over AS130 link**
  - **Remaining prefixes tagged with local preference of 80**
    - **Traffic to other all other ASes will go over the link to AS140**
- **Router D configuration same as Router C without the route-map**

Two Upstreams, One Local Peer
Full Routes

- Full routes from upstreams

    Expensive – needs lots of memory and CPU

    Need to play preference games

    Previous example is only an example – real life will need improved fine-tuning!

    Previous example doesn't consider inbound traffic – see earlier in presentation for examples

Two Upstreams, One Local Peer Partial Routes

- **Strategy:**
  - **Ask one upstream for a default route**
    - **Easy to originate default towards a BGP neighbour**
  - **Ask other upstream for a full routing table**
    - **Then filter this routing table based on neighbouring ASN**
    - **E.g. want traffic to their neighbours to go over the link to that ASN**
    - **Most of what upstream sends is thrown away**
    - **Easier than asking the upstream to set up custom BGP filters for you**

Two Upstreams, One Local Peer
Partial Routes

- **Router C Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.1 remote-as 130
 neighbor 122.102.10.1 prefix-list rfc1918-nodef-deny in
 neighbor 122.102.10.1 prefix-list my-block out
 neighbor 122.102.10.1 filter-list 10 in
 neighbor 122.102.10.1 route-map tag-default-low in
!
..next slide
```

Allow all prefixes and default in; deny RFC1918 and friends

AS filter list filters prefixes based on origin ASN

Two Upstreams, One Local Peer
Partial Routes

```
ip prefix-list my-block permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 121.10.0.0 255.255.224.0 null0
!
ip as-path access-list 10 permit ^(130_)+$
ip as-path access-list 10 permit ^(130_)+_[0-9]+$
!
route-map tag-default-low permit 10
 match ip address prefix-list default
 set local-preference 80
route-map tag-default-low permit 20
!
```

Two Upstreams, One Local Peer
Partial Routes

- **Router D Configuration**

```
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.5 remote-as 140
 neighbor 122.102.10.5 prefix-list default in
 neighbor 122.102.10.5 prefix-list my-block out
!
ip prefix-list my-block permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 121.10.0.0 255.255.224.0 null0
```

Two Upstreams, One Local Peer
Partial Routes

- **Router C configuration:**
  - **Accept full routes from AS130**
    - **(or get them to send less)**
  - **Filter ASNs so only AS130 and AS130's neighbouring ASes are accepted**
  - **Allow default, and set it to local preference 80**
  - **Traffic to those ASes will go over AS130 link**
  - **Traffic to other all other ASes will go over the link to AS140**
  - **If AS140 link fails, backup via AS130 – and vice-versa**

Two Upstreams, One Local Peer
Partial Routes

- Partial routes from upstreams

  Not expensive – only carry the routes necessary for loadsharing

  Need to filter on AS paths

  Previous example is only an example – real life will need improved fine-tuning!

  Previous example doesn’t consider inbound traffic – see earlier in presentation for examples

Two Upstreams, One Local Peer

- When upstreams cannot or will not announce default route
  - Because of operational policy against using “default-originate” on BGP peering
  - Solution is to use IGP to propagate default from the edge/peering routers

Two Upstreams, One Local Peer
Partial Routes

- **Router C Configuration**

```
router ospf 110
 default-information originate metric 30
 passive-interface Serial 0/0
!
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.1 remote-as 130
 neighbor 122.102.10.1 prefix-list rfc1918-deny in
 neighbor 122.102.10.1 prefix-list my-block out
 neighbor 122.102.10.1 filter-list 10 in
!
..next slide
```

Two Upstreams, One Local Peer
Partial Routes

```
ip prefix-list my-block permit 121.10.0.0/19
! See www.cymru.com/Documents/bogon-list.html
! ...for “RFC1918 and friends” list
!
ip route 121.10.0.0 255.255.224.0 null0
ip route 0.0.0.0 0.0.0.0 serial 0/0 254
!
ip as-path access-list 10 permit ^(130_)+$
ip as-path access-list 10 permit ^(130_)+_[0-9]+$
!
```

Two Upstreams, One Local Peer
Partial Routes

- **Router D Configuration**

```
router ospf 110
 default-information originate metric 10
 passive-interface Serial 0/0
!
router bgp 110
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.5 remote-as 140
 neighbor 122.102.10.5 prefix-list deny-all in
 neighbor 122.102.10.5 prefix-list my-block out
!
..next slide
```

Two Upstreams, One Local Peer
Partial Routes

```
ip prefix-list deny-all deny 0.0.0.0/0 le 32
ip prefix-list my-block permit 121.10.0.0/19
!
ip route 121.10.0.0 255.255.224.0 null0
ip route 0.0.0.0 0.0.0.0 serial 0/0 254
!
```

Two Upstreams, One Local Peer
Partial Routes

- Partial routes from upstreams

    Use OSPF to determine outbound path

    Router D default has metric 10 – primary outbound path

    Router C default has metric 30 – backup outbound path

    Serial interface goes down, static default is removed from routing table, OSPF default withdrawn

Aside: Configuration Recommendation

- When distributing internal default by iBGP or OSPF

  Make sure that routers connecting to private peers or to IXPs do NOT carry the default route

  Otherwise they could point a default route to you and unintentionally transit your backbone

  Simple fix for Private Peer/IXP routers:

  ```
   ip route 0.0.0.0 0.0.0.0 null0
  ```

BGP Multihoming Techniques

- Why Multihome?
- Definition & Options
- Preparing the Network
- Basic Multihoming
- “BGP Traffic Engineering”
- Complex Cases & Caveats

Complex Cases & Caveats

How not to get stuck; how not to compromise routing system security

Complex Cases & Caveats

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

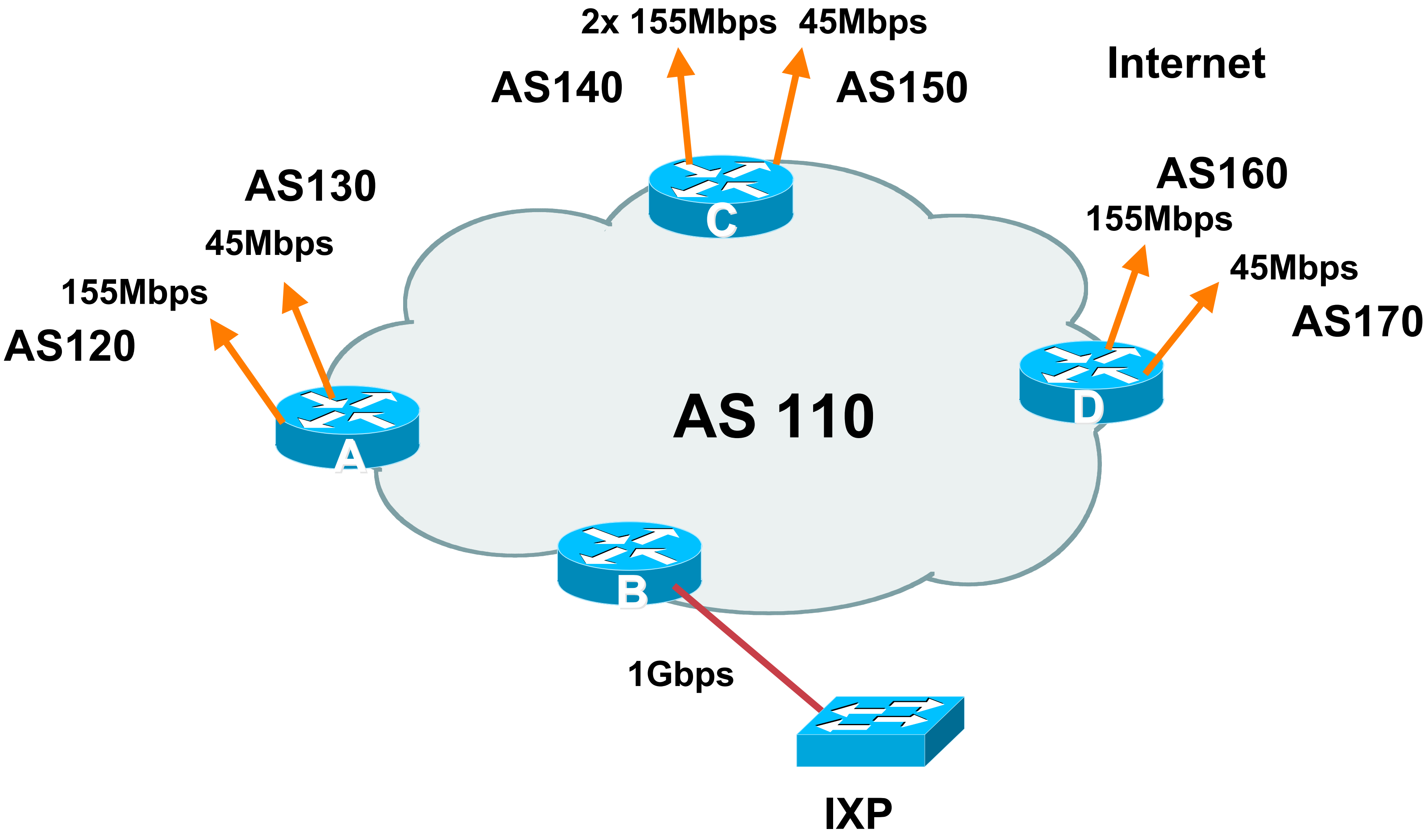
Complex Cases

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

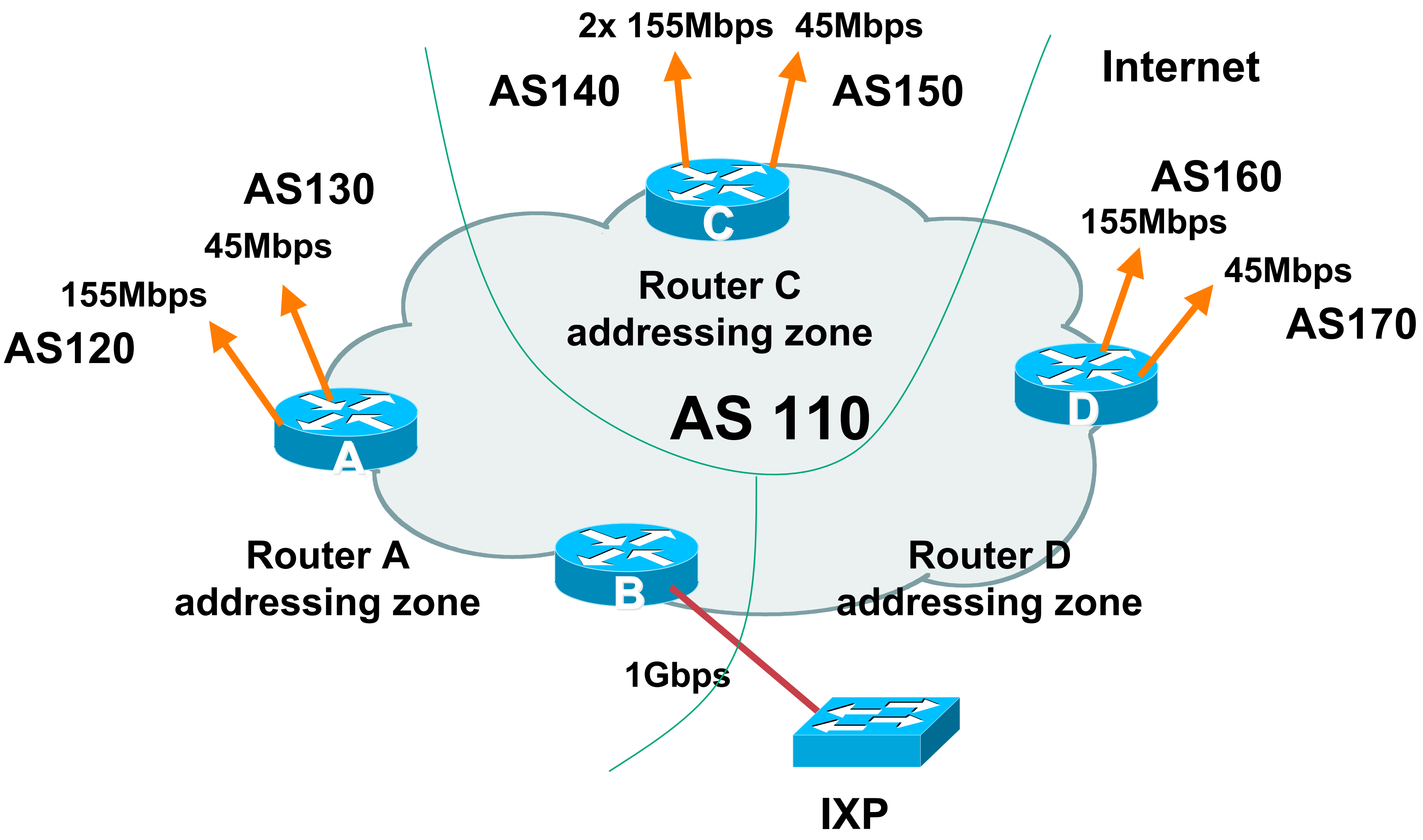
  Customers in vicinity of each of Router A, C and D addressed from contiguous address block assigned to each Router

  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 may vary

- Example shows:

  where to start,

  what the thought processes are, and

  what the strategies could be

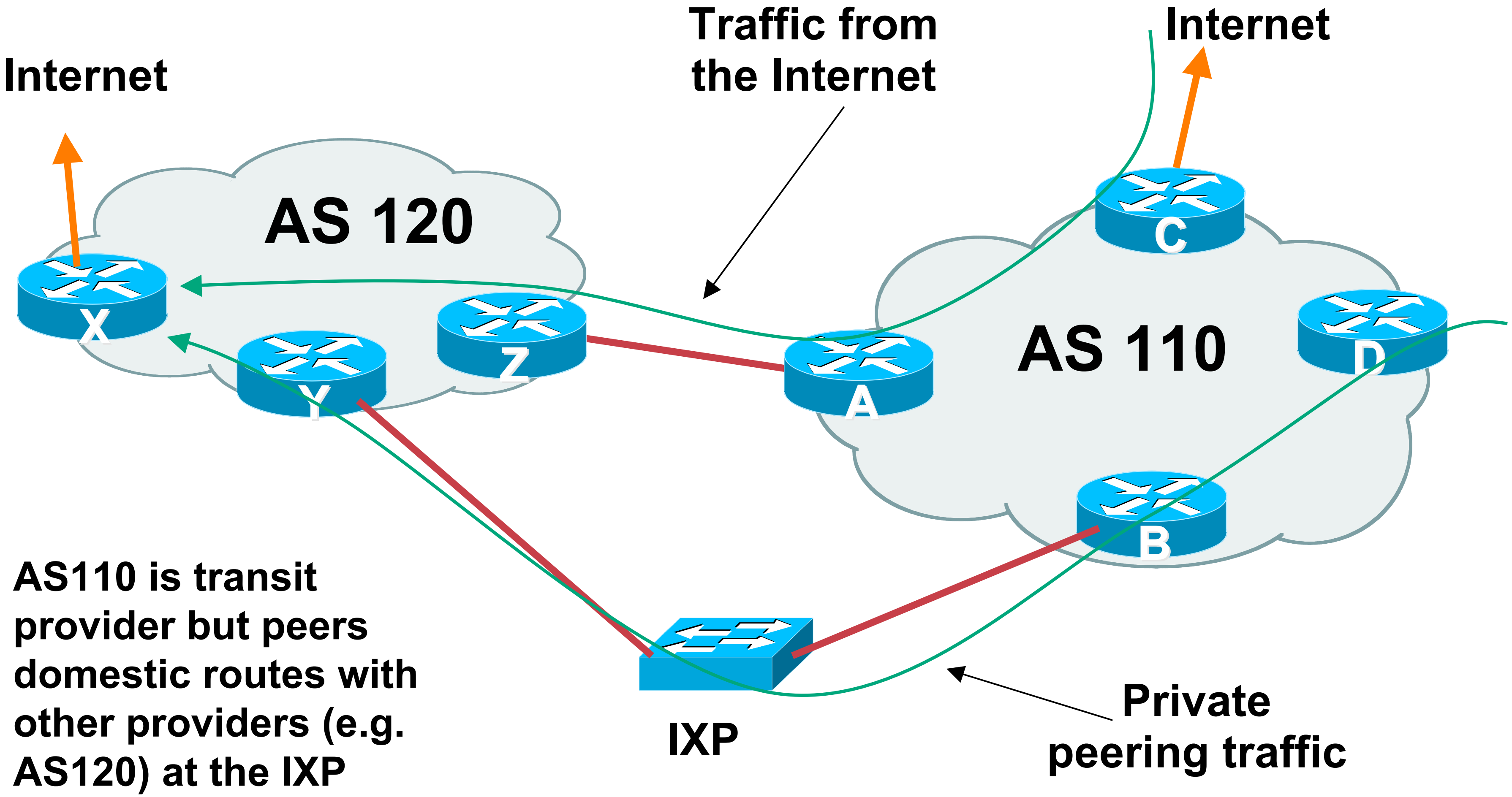
Caveats

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

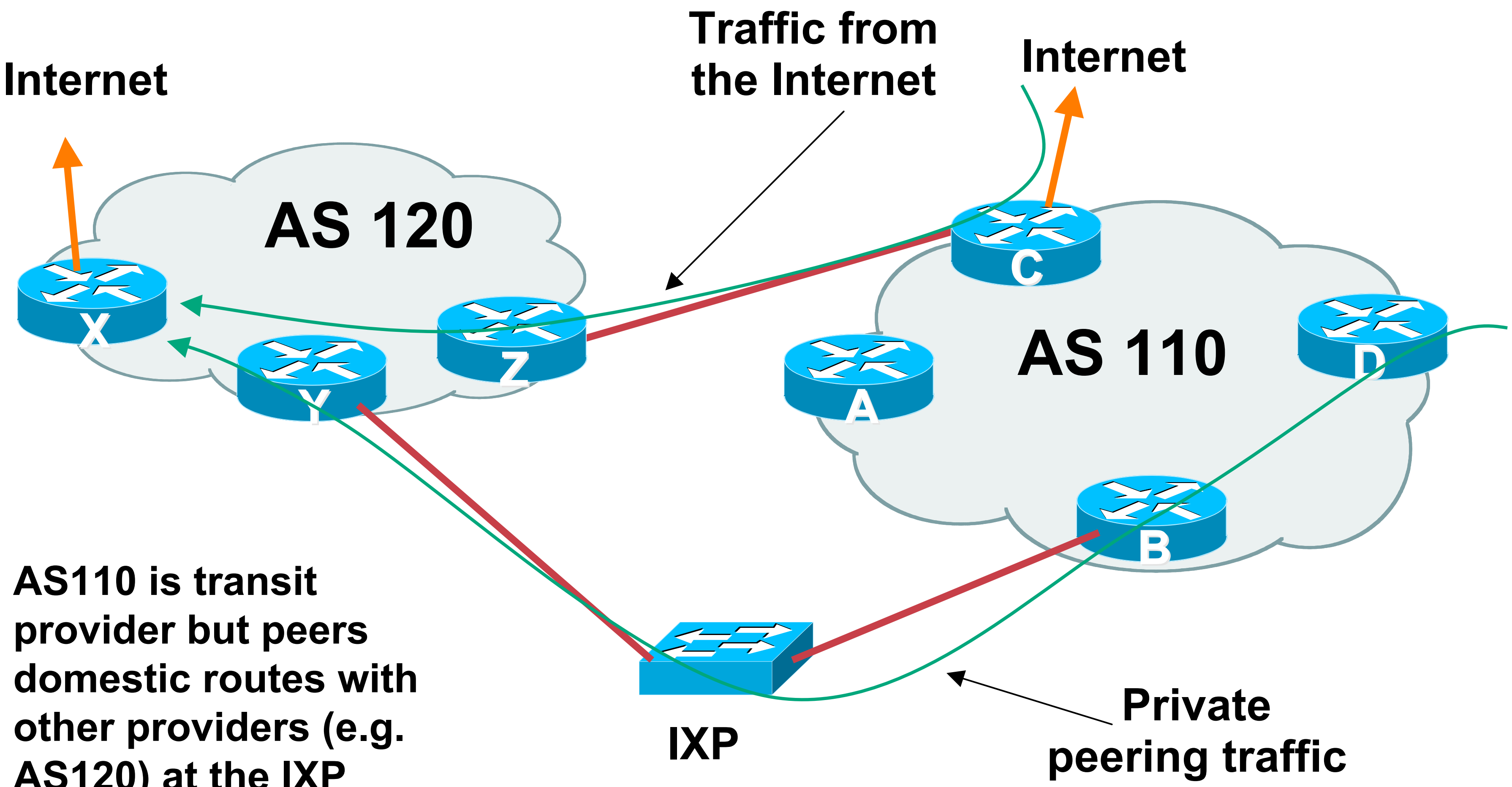
  Policy Routing on Router A according to packet source address

  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

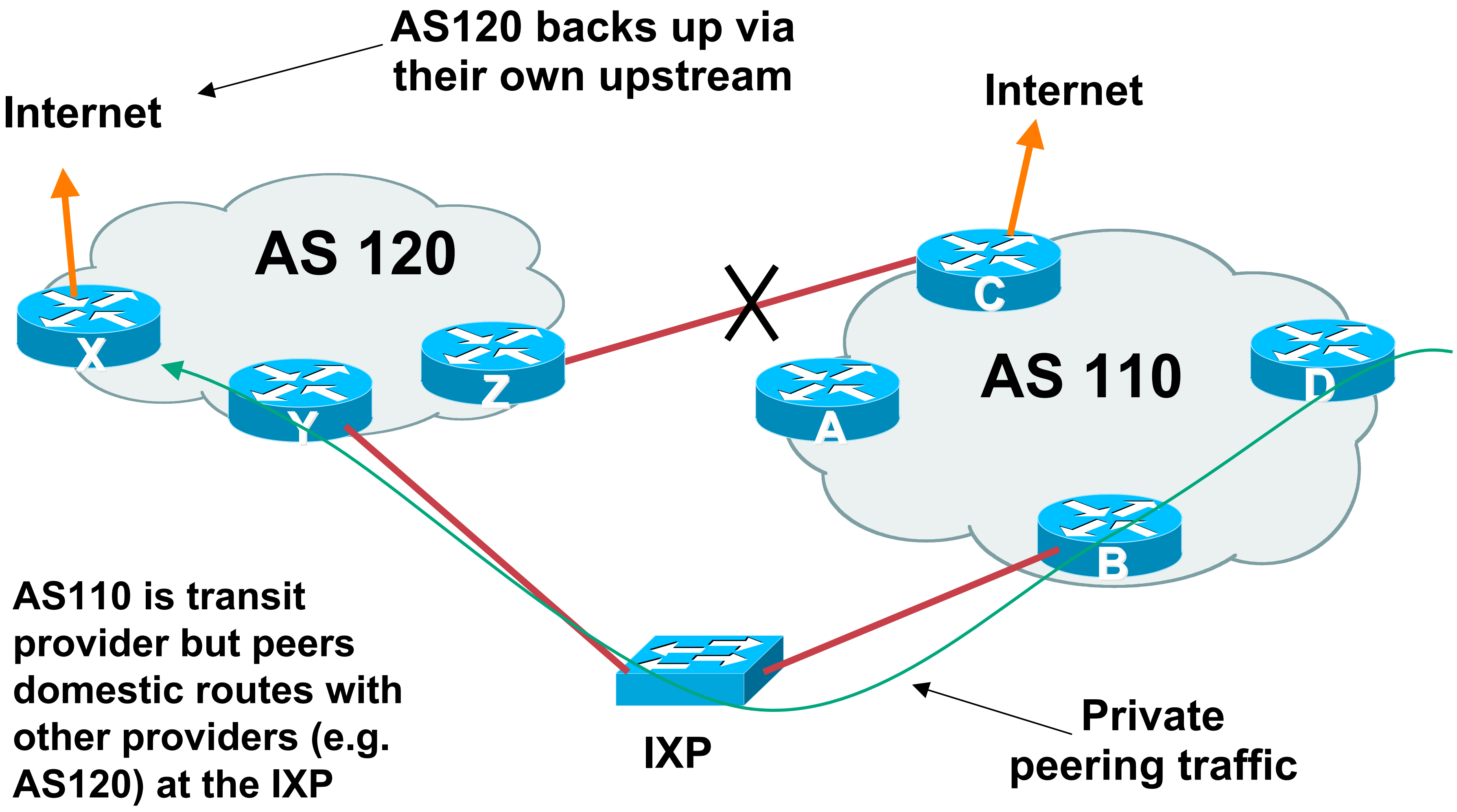
- **Some ISPs find this backup arrangement acceptable**

  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



AS120 backs up via their own upstream

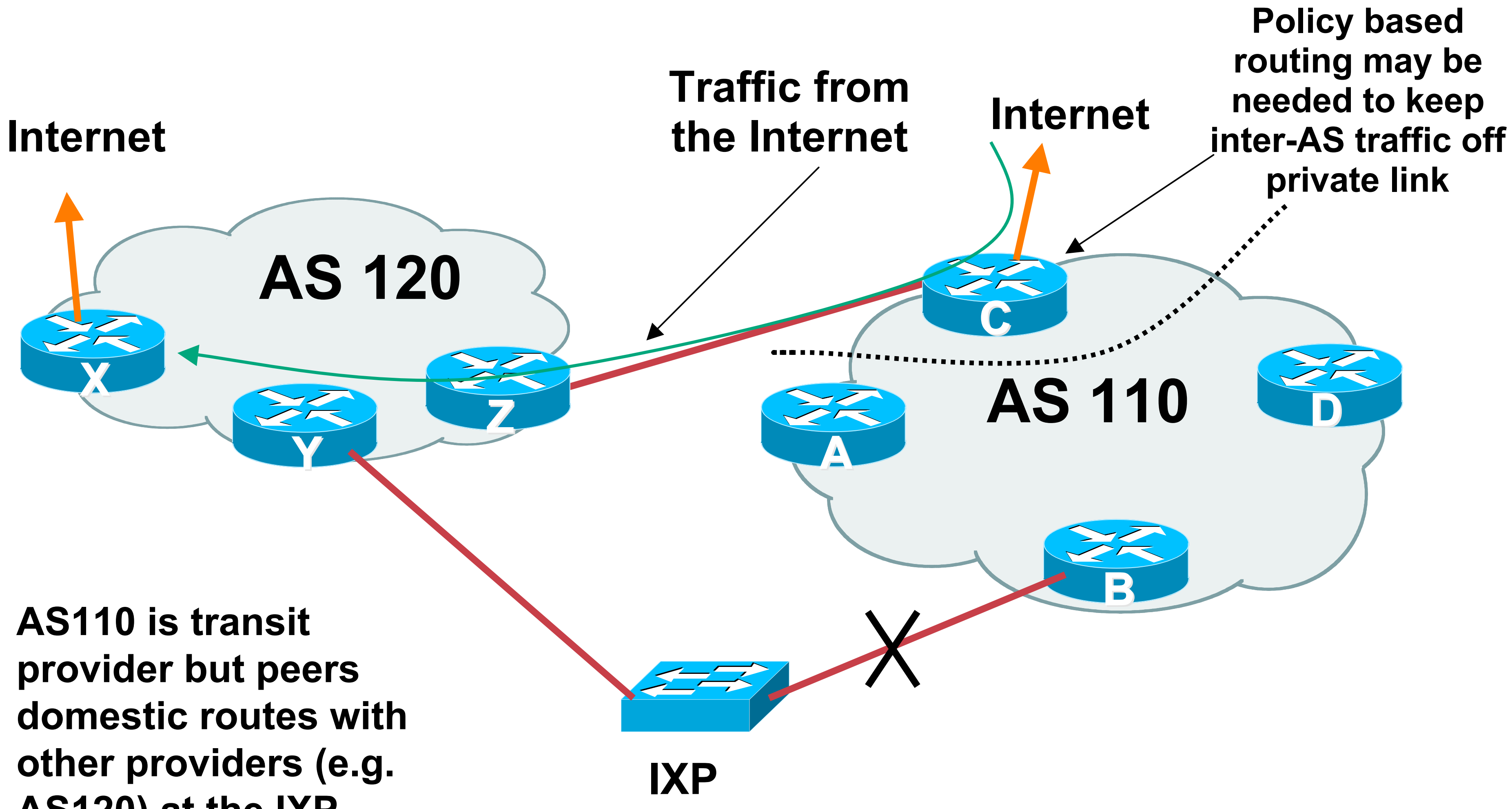
AS110 is transit provider but peers domestic routes with other providers (e.g. AS120) at the IXP

Private peering traffic

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



Policy based routing may be needed to keep inter-AS traffic off private link

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

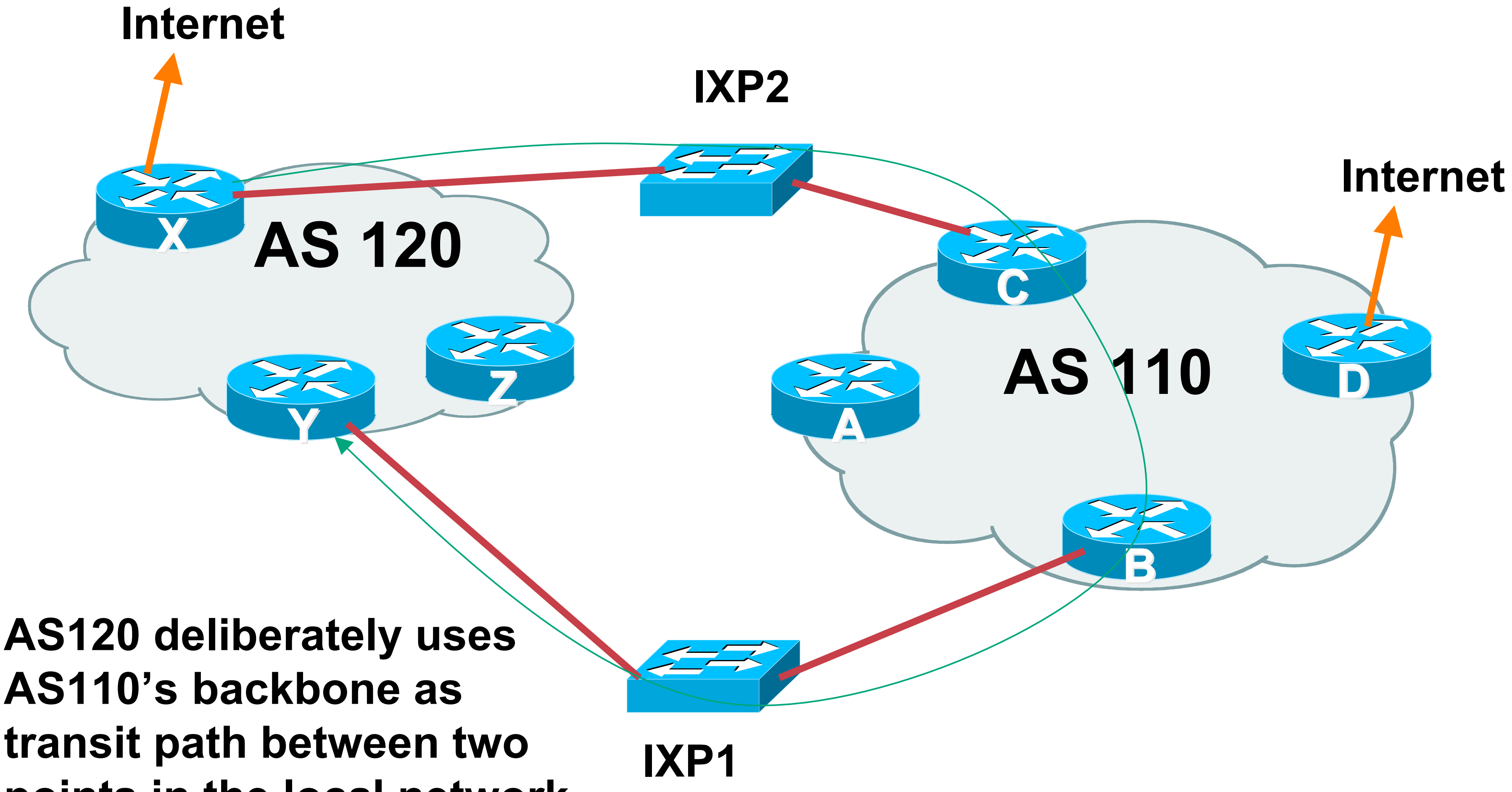
Caveats

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**

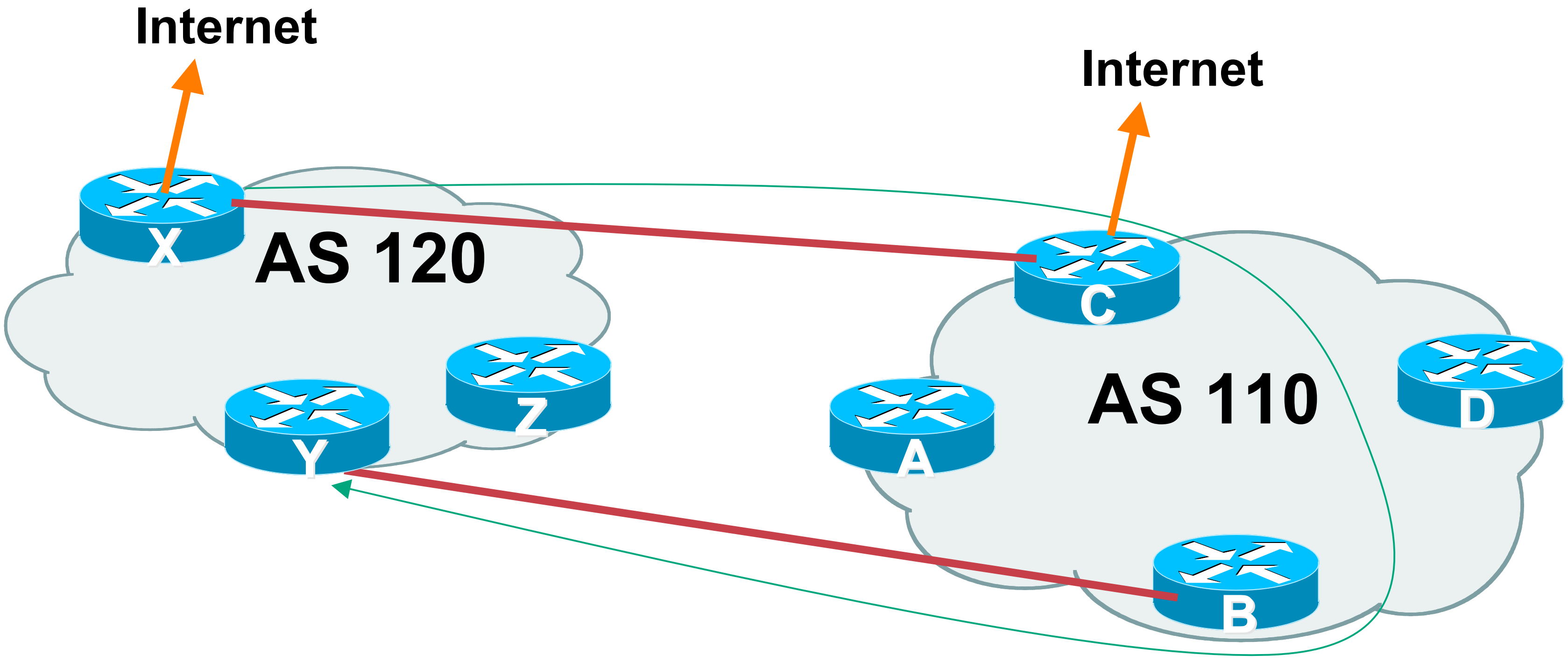
  When AS120 points static route for an AS120 destination to AS110, the peering routers have no destination apart from back towards AS120, so the packets will oscillate until TTL expiry

  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

Summary

- **Multihoming is not hard, really…**

  Keep It Simple & Stupid!

- **Full routing table is rarely required**

  A default is often just as good

  If customers want 170k prefixes, charge them money for it

Presentation Slides

- **Available on**

  ftp://ftp-eng.cisco.com

  /pfs/seminars/NANOG35-Multihoming.pdf

  And on the NANOG 35 meeting pages at
  http://www.nanog.org/mtg-0510/pdf/smith.pdf

BGP Multihoming Techniques

Philip Smith <pfs@cisco.com>

NANOG35

23-25 October 2005

Los Angeles