



BGP Best Current Practices

Philip Smith

E2 Workshop, AfNOG2006



What is BGP for??

What is an IGP not for?



BGP versus OSPF/ISIS

- Internal Routing Protocols (IGPs)
 - examples are ISIS and OSPF
 - used for carrying **infrastructure** addresses
 - **NOT** used for carrying Internet prefixes or customer prefixes
 - design goal is to **minimise** number of prefixes in IGP to aid scalability and rapid convergence

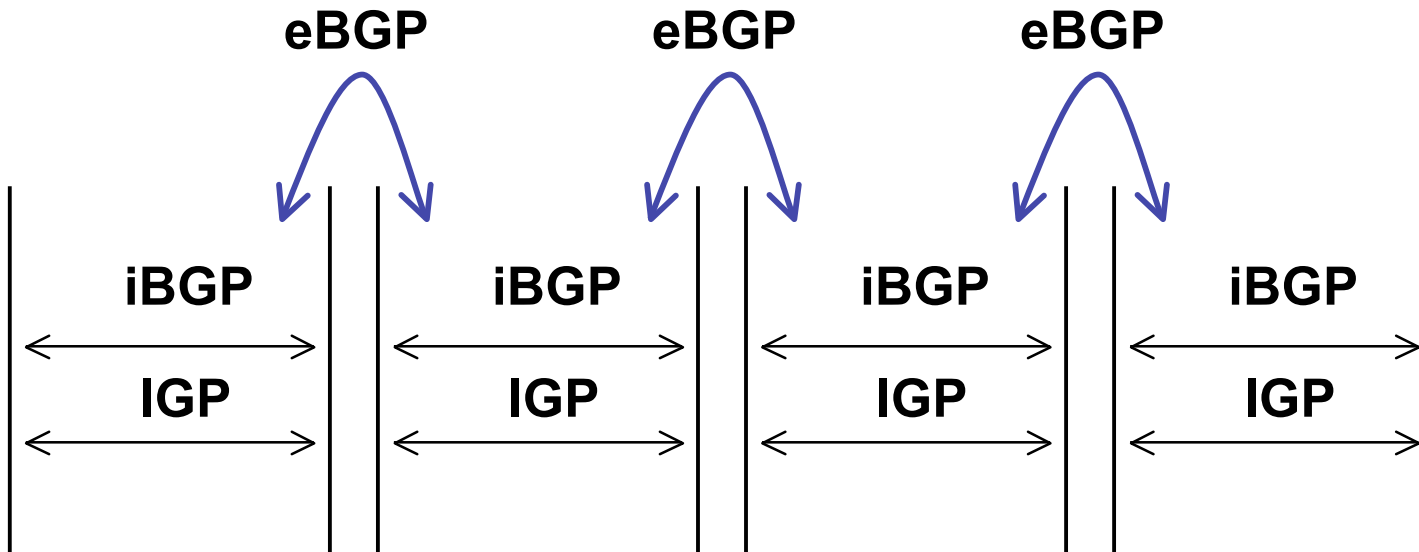


BGP versus OSPF/ISIS

- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
 - some/all Internet prefixes across backbone
 - customer prefixes
- eBGP used to
 - exchange prefixes with other ASes
 - implement routing policy

BGP/IGP model used in ISP networks

- Model representation





BGP versus OSPF/ISIS

- DO NOT:
 - distribute BGP prefixes into an IGP
 - distribute IGP routes into BGP
 - use an IGP to carry customer prefixes
- **YOUR NETWORK WILL NOT SCALE**



Aggregation

Quality, not Quantity!



Aggregation

- ISPs receive address block from Regional Registry or upstream provider
- **Aggregation** means announcing the **address block** only, not subprefixes
- Aggregate should be generated internally



Configuring Aggregation: Cisco IOS

- ISP has 101.10.0.0/19 address block
- To put into BGP as an aggregate:

```
router bgp 100
  network 101.10.0.0 mask 255.255.224.0
  ip route 101.10.0.0 255.255.224.0 null0
```
- The static route is a “pull up” route
 - more specific prefixes within this address block ensure connectivity to ISP’s customers
 - “longest match lookup”



Aggregation

- Address block should be announced to the Internet as an aggregate
- Subprefixes of address block should NOT be announced to Internet unless fine-tuning multihoming
 - And even then care and frugality is required – don't announce more subprefixes than absolutely necessary



Announcing Aggregate: Cisco IOS

- Configuration Example

```
router bgp 100
  network 101.10.0.0 mask 255.255.224.0
  neighbor 102.102.10.1 remote-as 101
  neighbor 102.102.10.1 prefix-list out-filter out
!
ip route 101.10.0.0 255.255.224.0 null0
!
ip prefix-list out-filter permit 101.10.0.0/19
ip prefix-list out-filter deny 0.0.0.0/0 le 32
```



Announcing an Aggregate

- ISPs who don't and won't aggregate are held in poor regard by community
- Registries' minimum allocation size is now at least a /21
 - no real reason to see anything much longer than a /22 prefix in the Internet
 - BUT there are currently >101000 /24s!



The Internet Today (May 2006)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	187255
Prefixes after maximum aggregation	103563
Unique prefixes in Internet	91865
Prefixes smaller than registry alloc	92110
/24s announced	101414
only 5719 /24s are from 192.0.0.0/8	
ASes in use	22089



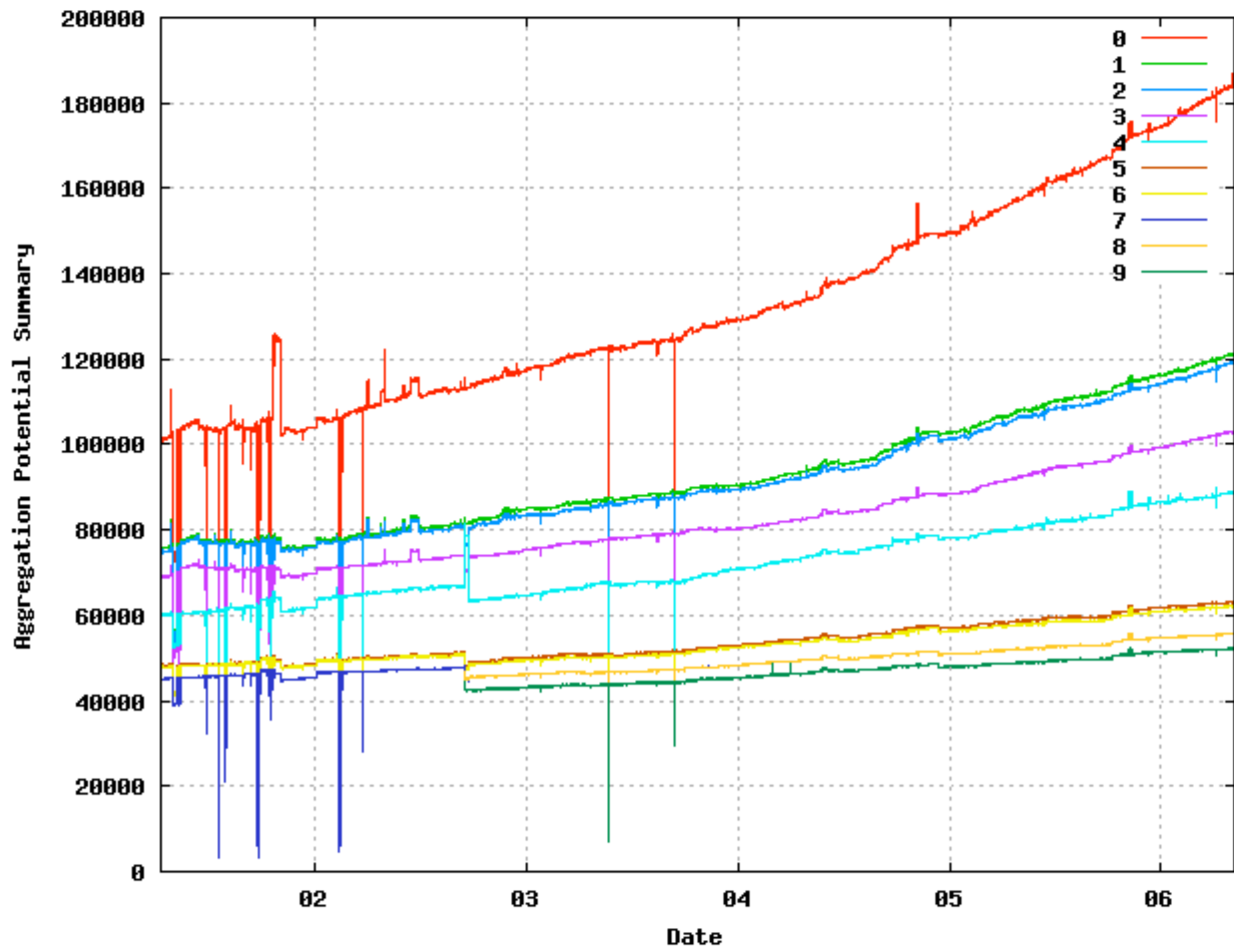
Efforts to Improve Aggregation: The CIDR Report

- Initiated and operated for many years by Tony Bates
- Now combined with Geoff Huston's routing analysis
www.cidr-report.org
- Results e-mailed on a weekly basis to most operations lists around the world
- Lists the top 30 service providers who could do better at aggregating



Efforts to Improve Aggregation: The CIDR Report

- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
 - Flexible and powerful tool to aid ISPs
 - Intended to show how greater efficiency in terms of BGP table size can be obtained without loss of routing and policy information
 - Shows what forms of origin AS aggregation could be performed and the potential benefit of such actions to the total table size
 - Very effectively challenges the traffic engineering excuse





Aggregation: Summary

- Aggregation on the Internet could be **MUCH** better
 - 35% saving on Internet routing table size is quite feasible
 - Tools **are** available
 - Commands on the router are not hard
 - CIDR-Report webpage



Receiving Prefixes



Receiving Prefixes from downstream peers

- ISPs should only accept prefixes which have been assigned or allocated to their downstream peer
- For example
 - downstream has 100.50.0.0/20 block
 - should only announce this to peers
 - peers should only accept this from them



Receiving Prefixes: Cisco IOS

- Configuration Example on upstream

```
router bgp 100
```

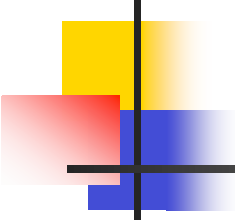
```
neighbor 102.102.10.1 remote-as 101
```

```
neighbor 102.102.10.1 prefix-list customer in
```

```
!
```

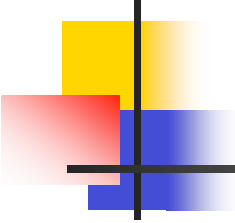
```
ip prefix-list customer permit 100.50.0.0/20
```

```
ip prefix-list customer deny 0.0.0.0/0 le 32
```



Receiving Prefixes from upstream peers

- Not desirable unless really necessary
 - special circumstances
- Ask upstream to either:
 - originate a default-route
 - announce one prefix you can use as default



Receiving Prefixes from upstream peers

- Downstream Router Configuration

```
router bgp 100
  network 101.10.0.0 mask 255.255.224.0
  neighbor 101.5.7.1 remote-as 101
  neighbor 101.5.7.1 prefix-list infilt in
  neighbor 101.5.7.1 prefix-list outfilt out
!
ip prefix-list infilt permit 0.0.0.0/0
ip prefix-list infilt deny 0.0.0.0/0 le 32
!
ip prefix-list outfilt permit 101.10.0.0/19
ip prefix-list outfilt deny 0.0.0.0/0 le 32
```



Receiving Prefixes from upstream peers

- Upstream Router Configuration

```
router bgp 101
```

```
neighbor 101.5.7.2 remote-as 100
```

```
neighbor 101.5.7.2 default-originate
```

```
neighbor 101.5.7.2 prefix-list cust-in in
```

```
neighbor 101.5.7.2 prefix-list cust-out out
```

```
!
```

```
ip prefix-list cust-in permit 101.10.0.0/19
```

```
ip prefix-list cust-in deny 0.0.0.0/0 le 32
```

```
!
```

```
ip prefix-list cust-out permit 0.0.0.0/0
```

```
ip prefix-list cust-out deny 0.0.0.0/0 le 32
```



Receiving Prefixes from upstream peers

- If necessary to receive prefixes from upstream provider, care is required
 - don't accept RFC1918 etc prefixes
 - don't accept your own prefix
 - don't accept default (unless you need it)
 - don't accept prefixes longer than /24



Receiving Prefixes

```
router bgp 100
 network 101.10.0.0 mask 255.255.224.0
 neighbor 101.5.7.1 remote-as 101
 neighbor 101.5.7.1 prefix-list in-filter in
!
ip prefix-list in-filter deny 0.0.0.0/0           ! Block default
ip prefix-list in-filter deny 0.0.0.0/8 le 32
ip prefix-list in-filter deny 10.0.0.0/8 le 32
ip prefix-list in-filter deny 101.10.0.0/19 le 32 ! Block local prefix
ip prefix-list in-filter deny 127.0.0.0/8 le 32
ip prefix-list in-filter deny 169.254.0.0/16 le 32
ip prefix-list in-filter deny 172.16.0.0/12 le 32
ip prefix-list in-filter deny 192.0.2.0/24 le 32
ip prefix-list in-filter deny 192.168.0.0/16 le 32
ip prefix-list in-filter deny 224.0.0.0/3 le 32   ! Block multicast
ip prefix-list in-filter deny 0.0.0.0/0 ge 25     ! Block prefixes >/24
ip prefix-list in-filter permit 0.0.0.0/0 le 32
```



Generic ISP BGP prefix filter

- This prefix-list MUST be applied to all external BGP peerings, in and out!
- RFC3330 lists many special use addresses
- Check Rob Thomas' list of "bogons"
 - <http://www.cymru.com/Documents/bogon-list.html>

```
ip prefix-list rfc1918-sua deny 0.0.0.0/8 le 32
ip prefix-list rfc1918-sua deny 10.0.0.0/8 le 32
ip prefix-list rfc1918-sua deny 127.0.0.0/8 le 32
ip prefix-list rfc1918-sua deny 169.254.0.0/16 le 32
ip prefix-list rfc1918-sua deny 172.16.0.0/12 le 32
ip prefix-list rfc1918-sua deny 192.0.2.0/24 le 32
ip prefix-list rfc1918-sua deny 192.168.0.0/16 le 32
ip prefix-list rfc1918-sua deny 224.0.0.0/3 le 32
ip prefix-list rfc1918-sua deny 0.0.0.0/0 ge 25
ip prefix-list rfc1918-sua permit 0.0.0.0/0 le 32
```



Prefixes into iBGP



Injecting prefixes into iBGP

- Use iBGP to carry customer prefixes
 - don't use IGP
- Point static route to customer interface
- Use BGP network statement
- As long as static route exists (interface active), prefix will be in BGP



Router configuration: network statement

- Example:

```
interface loopback 0
  ip address 215.17.3.1 255.255.255.255
!
interface Serial 5/0
  ip unnumbered loopback 0
  ip verify unicast reverse-path
!
ip route 215.34.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
  network 215.34.10.0 mask 255.255.252.0
```



Injecting prefixes into iBGP

- interface flap will result in prefix withdraw and reannounce
 - use "ip route...permanent"
- many ISPs use redistribute static rather than network statement
 - only use this if you understand why



Router Configuration: redistribute static

- Example:

```
ip route 215.34.10.0 255.255.252.0 Serial 5/0
!
router bgp 100
  redistribute static route-map static-to-bgp
<snip>
!
route-map static-to-bgp permit 10
  match ip address prefix-list ISP-block
  set origin igp
<snip>
!
ip prefix-list ISP-block permit 215.34.10.0/22 le 30
!
```



Injecting prefixes into iBGP

- Route-map ISP-block can be used for many things:
 - setting communities and other attributes
 - setting origin code to IGP, etc
- Be careful with prefix-lists and route-maps
 - absence of either/both means all statically routed prefixes go into iBGP



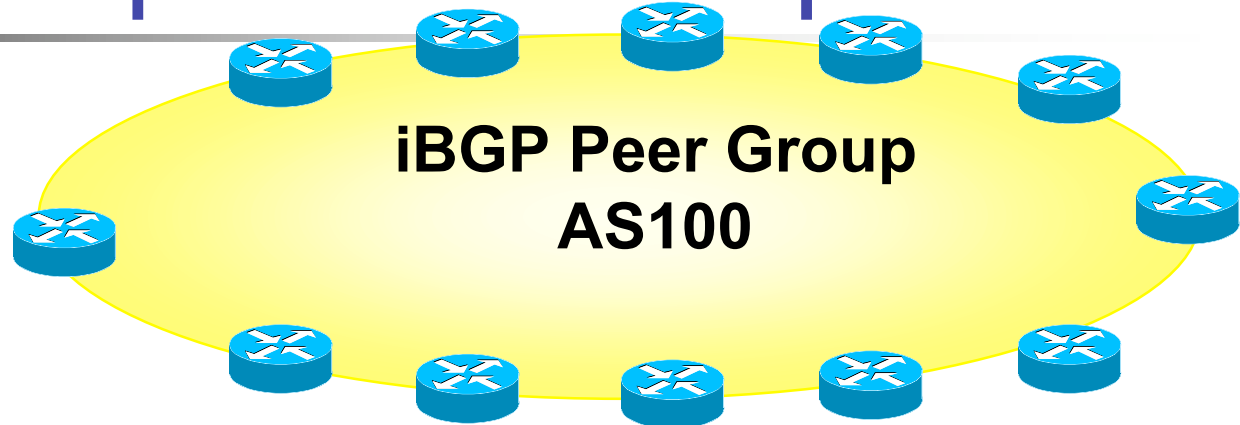
Configuration Tips



Templates

- Good practice to configure templates for everything
 - Vendor defaults tend not to be optimal or even very useful for ISPs
 - ISPs create their own defaults by using configuration templates
 - Sample iBGP and eBGP templates follow for Cisco IOS

BGP Template – iBGP peers



```
router bgp 100
neighbor internal peer-group
neighbor internal description ibgp peers
neighbor internal remote-as 100
neighbor internal update-source Loopback0
neighbor internal next-hop-self
neighbor internal send-community
neighbor internal version 4
neighbor internal password 7 03085A09
neighbor 1.0.0.1 peer-group internal
neighbor 1.0.0.2 peer-group internal
```



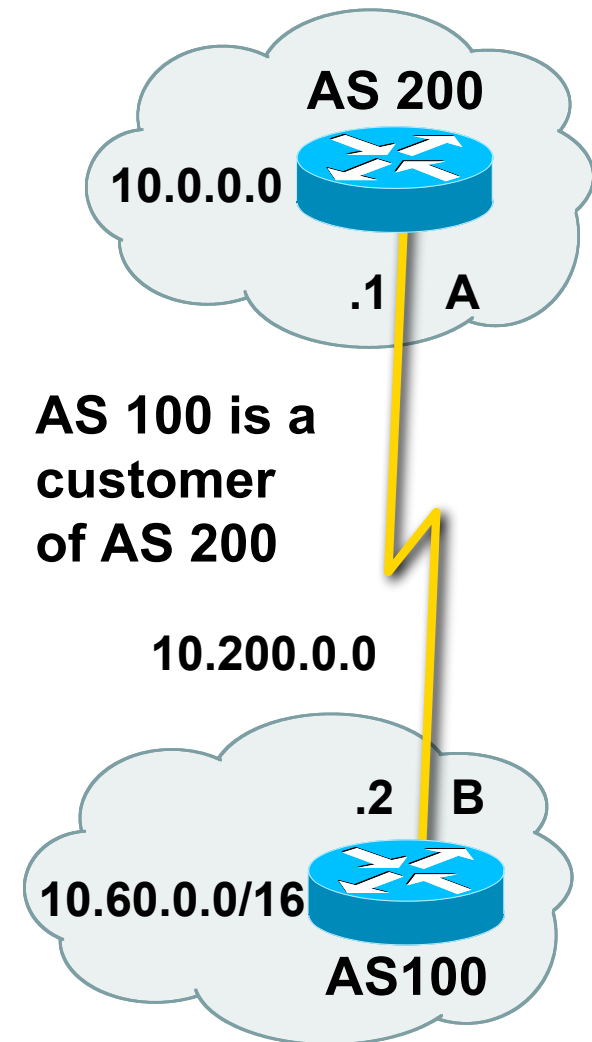
BGP Template – iBGP peers

- Use peer-groups
- iBGP between loopbacks!
- Next-hop-self
 - Keep DMZ and point-to-point out of IGP
- Always send communities in iBGP
 - Otherwise accidents will happen
- Hardwire BGP to version 4
 - Yes, this is being paranoid!
- Use passwords on iBGP session
 - Not being paranoid, **VERY** necessary

BGP Template – eBGP peers

Router B:

```
router bgp 100
network 10.60.0.0 mask 255.255.0.0
neighbor external peer-group
neighbor external remote-as 200
neighbor external description ISP connection
neighbor external remove-private-AS
neighbor external version 4
neighbor external prefix-list ispout out ! "real" filter
neighbor external filter-list 1 out      ! "accident" filter
neighbor external route-map ispout out
neighbor external prefix-list ispin in
neighbor external filter-list 2 in
neighbor external route-map ispin in
neighbor external password 7 020A0559
neighbor external maximum-prefix 220000 [warning-only]
neighbor 10.200.0.1 peer-group external
!
ip route 10.60.0.0 255.255.0.0 null0 254
```





BGP Template – eBGP peers

- Remove private ASes from announcements
 - Common omission today
- Use extensive filters, with “backup”
 - Use as-path filters to backup prefix-lists
 - Use route-maps for policy
- Use password agreed between you and peer on eBGP session
- Use maximum-prefix tracking
 - Router will warn you if there are sudden increases in BGP table size, bringing down eBGP if desired



More BGP “defaults”

- Log neighbour changes
 - Log neighbour changes
 - `bgp log-neighbor-changes`
- Enable deterministic MED
 - `bgp deterministic-med`
 - Otherwise bestpath could be different every time BGP session is reset
- Make BGP admin distance higher than any IGP
 - `distance bgp 200 200 200`



Configuration Tips Summary

- Use configuration templates
- Standardise the configuration
- Anything to make your life easier, network less prone to errors, network more likely to scale
- It's all about scaling – if your network won't scale, then it won't be successful



Summary – BGP BCP

- BGP versus IGP
- Aggregation
- Sending & Receiving Prefixes
- Injecting Prefixes into iBGP
- Configuration Tips