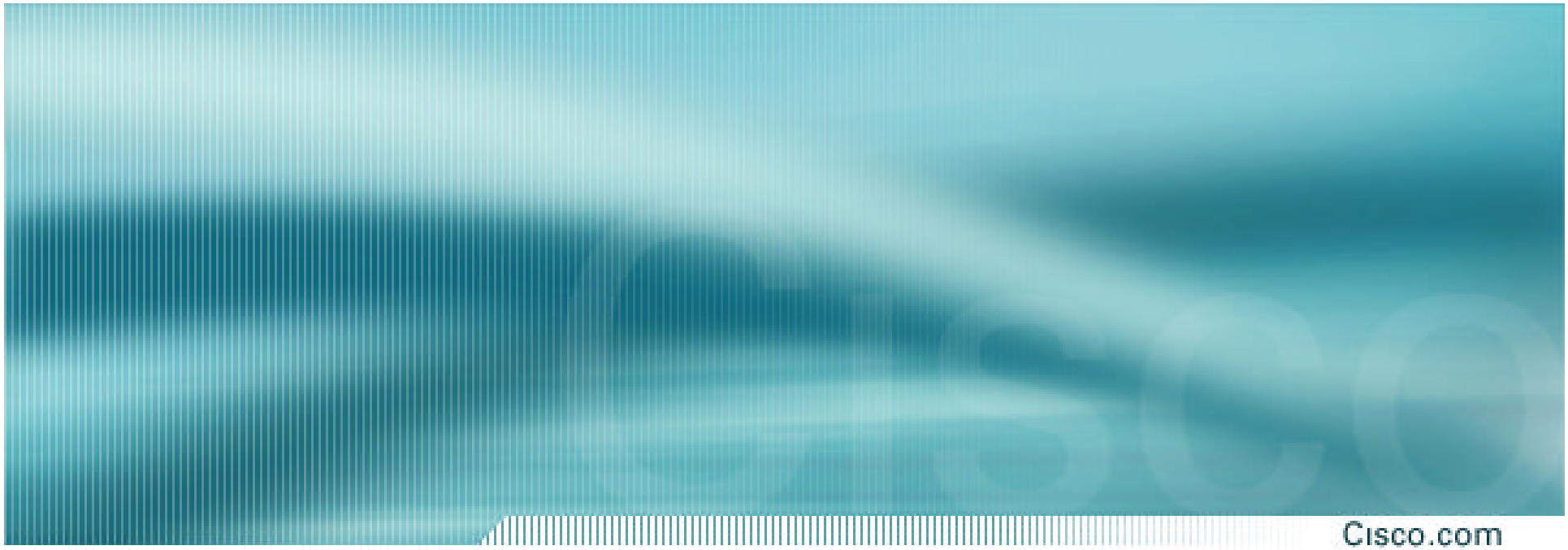
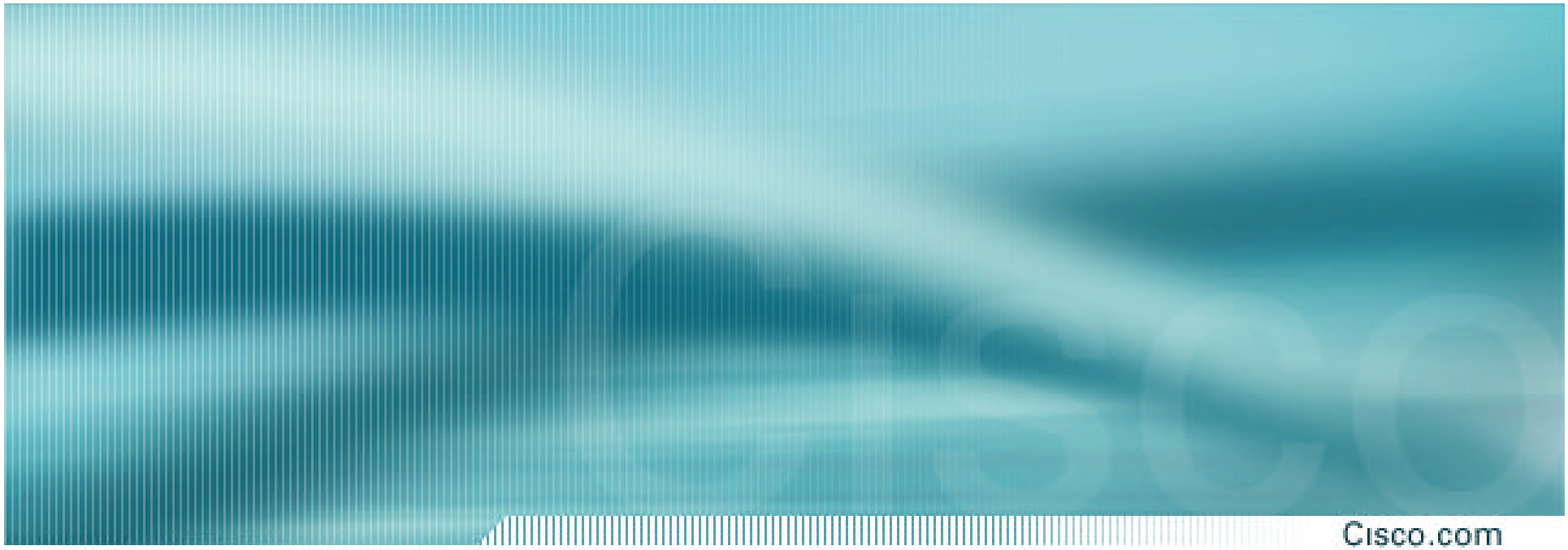


BGP in the Internet

Best Current Practices



Recommended IOS Releases

Which IOS??

Which IOS?

- IOS is a feature rich and highly complex router control system
- ISPs should choose the IOS variant which is most appropriate for the intended application
- There is an exclusive service provider train in IOS
  - This is 12.0S, supporting 7200, 7500, 10000 and 12000
  - Images also available for 2500, 2600, 3600 and 4500, but are completely unsupported
- There is a service provider image in most IOS releases
  - This is the image with –p– in its name, for example:
    - c7200-p-mz.122-8.T1 and c2600-p-mz.121-14
  - The –p– image is IP-only plus ISIS/CLNS

Which IOS?

Cisco.com

- **12.*n* – for example 12.2**
  - **This means the IOS is a mainline image**
  - **NO new features**
  - **ONLY bug fixes**
  - **The aim is stability!**
- **12.*n*T – for example 12.2T**
  - **This means the IOS is the technology release**
  - **NEW features**
  - **Bug fixes**
  - **Avoid unless you need the feature!**

12.2 IOS release images

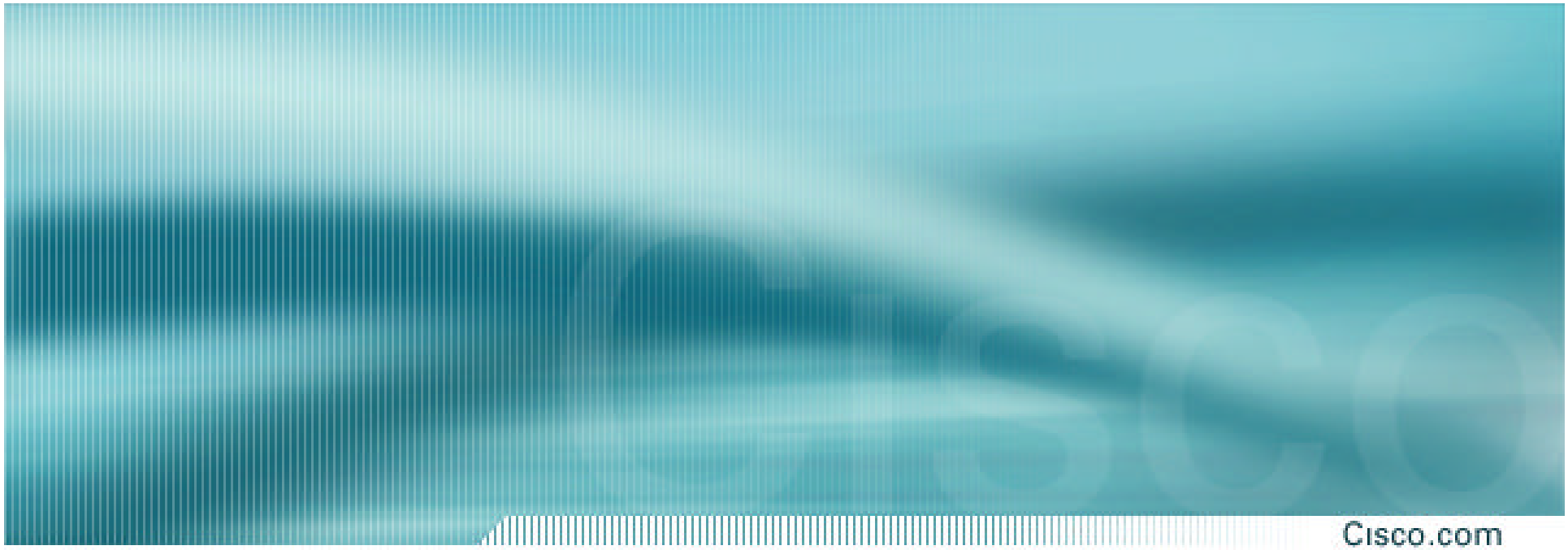
- 12.2 is the old “mainline” train
  - Originated from 12.1T, currently at 12.2(21)
  - Bug fix release only – aiming for stability
  - Supports more platforms and has more features than 12.1
- 12.2T was the old “technology train”
  - new features introduced in IOS 12.2
  - Included IPv6 for the first time
- Available on CCO, supported by TAC

12.3 IOS release images

- **12.3 is the current “mainline” train**
  - **Originated from 12.2T, currently at 12.3(5a)**
  - **Bug fix release only – aiming for stability**
  - **Supports more platforms and has more features than 12.2**
- **12.3T is the current “technology train”**
  - **new features introduced in IOS 12.3**
  - **Currently at 12.3(4)T2**
- **Available on CCO, supported by TAC**

IOS images for ISPs

- **12.0S is the release for all ISPs**
  - **For 7200, 7500, 10000 and GSR/12000**
  - **Replaces 11.1CC and 11.2GS**
  - **Currently at 12.0(26)S1**
- **12.2S is a new ISP release**
  - **For 7x00 series (x = 2 ® 6)**
  - **Combines 12.0S and 12.1E enhancements**
  - **Currently at 12.2(18)S1**
- **Available on CCO, supported by TAC**



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**

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 versus OSPF/ISIS

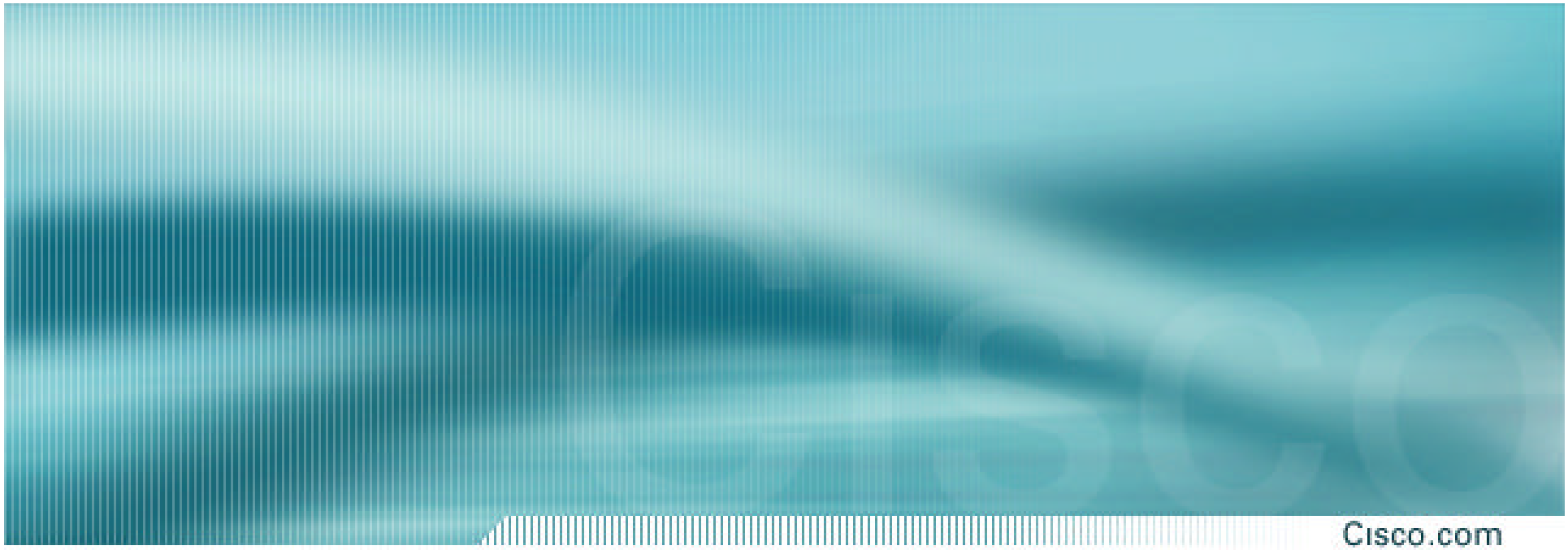
- **DO NOT:**

  distribute BGP prefixes into an IGP

  distribute IGP routes into BGP

  use an IGP to carry customer prefixes

- **YOUR NETWORK WILL NOT SCALE**



Aggregation

Aggregation

- **Aggregation means announcing the address block received from the RIR to the other ASes connected to your network**
- **Subprefixes of this aggregate *may* be:**
  - **Used internally in the ISP network**
  - **Announced to other ASes to aid with multihoming**
- **Unfortunately too many people are still thinking about class Cs, resulting in a proliferation of /24s in the Internet routing table**

Configuring Aggregation – Cisco IOS

- ISP has 221.10.0.0/19 address block
- To put into BGP as an aggregate:

```
router bgp 100
 network 221.10.0.0 mask 255.255.224.0
ip route 221.10.0.0 255.255.224.0 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 special circumstances (more later)
- Aggregate should be generated internally

  Not on the network borders!

Announcing Aggregate – Cisco IOS

- **Configuration Example**

```
router bgp 100
 network 221.10.0.0 mask 255.255.224.0
 neighbor 222.222.10.1 remote-as 101
 neighbor 222.222.10.1 prefix-list out-filter out
!
ip route 221.10.0.0 255.255.224.0 null0
!
ip prefix-list out-filter permit 221.10.0.0/19
ip prefix-list out-filter deny 0.0.0.0/0 le 32
```

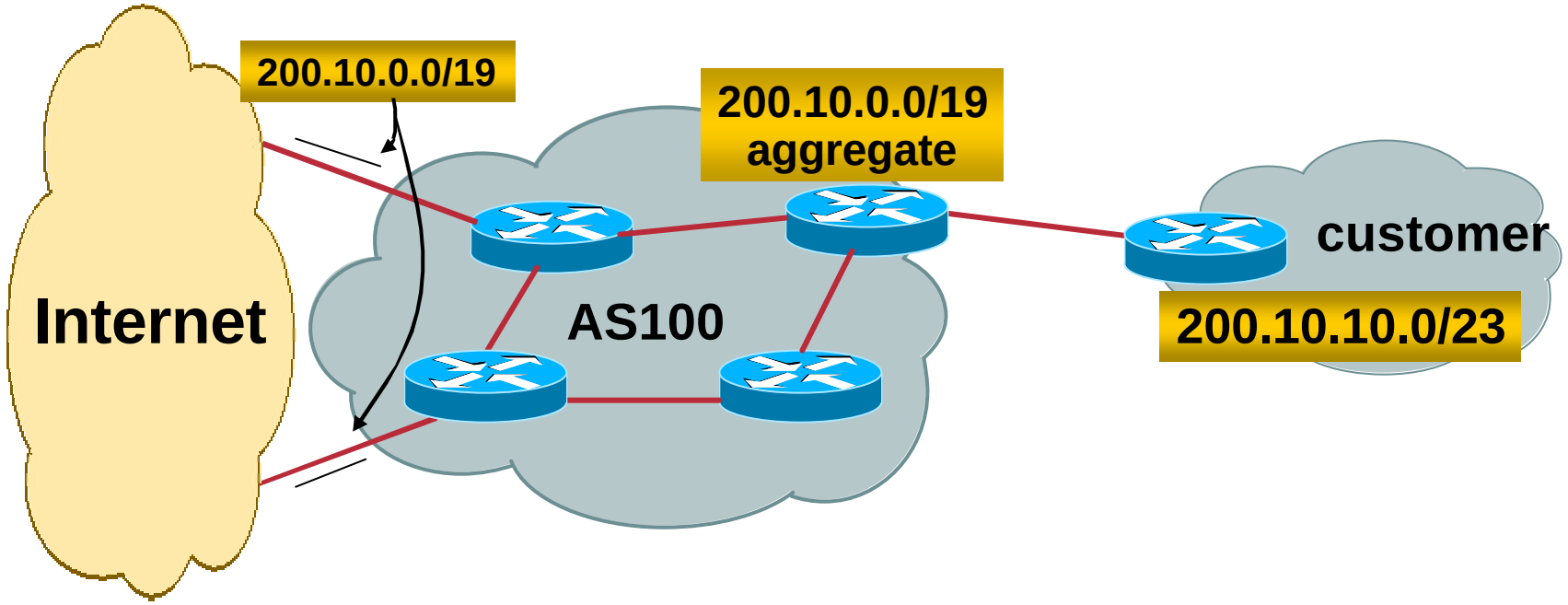
Announcing an Aggregate

- **ISPs who don't and won't aggregate are held in poor regard by community**
- **Registries' minimum allocation size is a /20**

  no real reason to see anything longer than a /21 prefix in the Internet

  BUT there are currently >71000 /24s!

Aggregation – Example



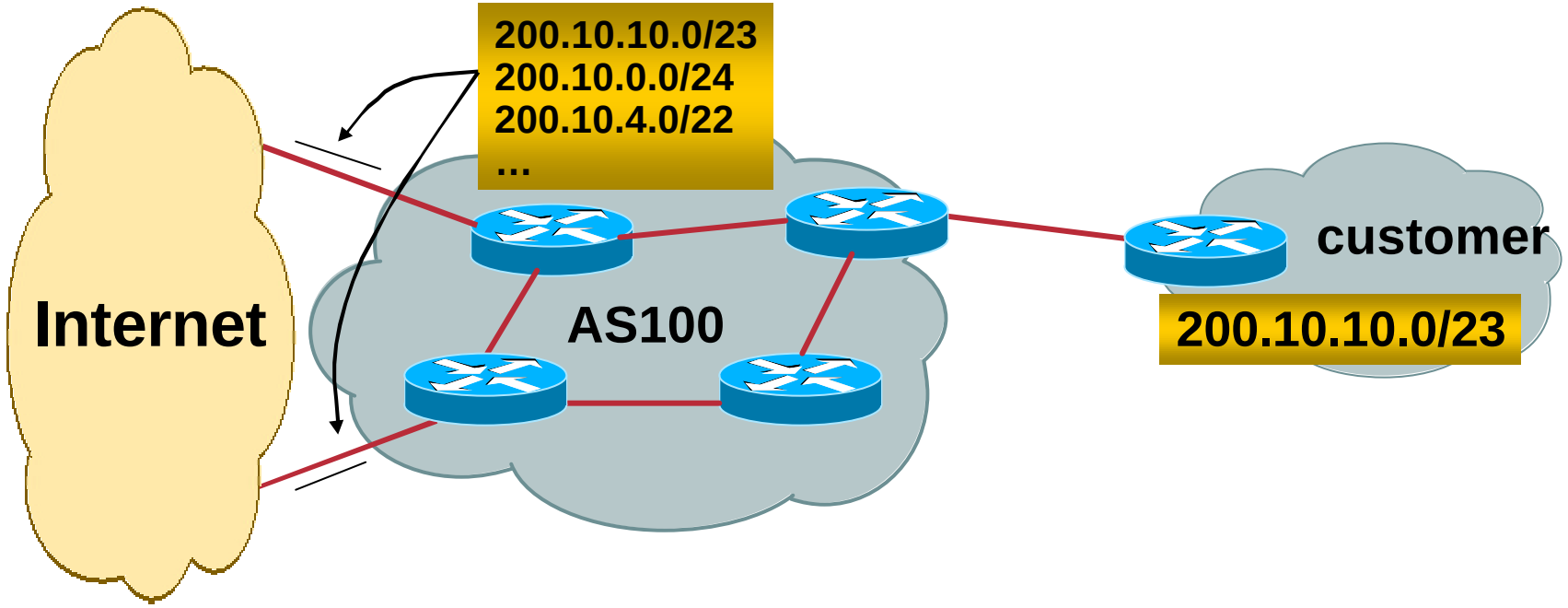
- **Customer has /23 network assigned from AS100's /19 address block**
- **AS100 announced /19 aggregate to the Internet**

Aggregation – Good Example

- Customer link goes down
  - their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- /19 aggregate is still being announced
  - no BGP hold down problems
  - no BGP propagation delays
  - no damping by other ISPs

- Customer link returns
- Their /23 network is visible again
  - The /23 is re-injected into AS100's iBGP
- The whole Internet becomes visible immediately
- Customer has Quality of Service perception

Aggregation – Example



- **Customer has /23 network assigned from AS100's /19 address block**
- **AS100 announces customers' individual networks to the Internet**

Aggregation – Bad Example

- Customer link goes down
  - Their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- Their ISP doesn't aggregate its /19 network block
  - /23 network withdrawal announced to peers
  - starts rippling through the Internet
  - added load on all Internet backbone routers as network is removed from routing table
- Customer link returns
  - Their /23 network is now visible to their ISP
  - Their /23 network is re-advertised to peers
  - Starts rippling through Internet
  - Load on Internet backbone routers as network is reinserted into routing table
  - Some ISP's suppress the flaps
  - Internet may take 10-20 min or longer to be visible
  - Where is the Quality of Service???

Aggregation – Summary

- **Good example is what everyone should do!**
  - **Adds to Internet stability**
  - **Reduces size of routing table**
  - **Reduces routing churn**
  - **Improves Internet QoS for everyone**
- **Bad example is what too many still do!**
  - **Why? Lack of knowledge?**

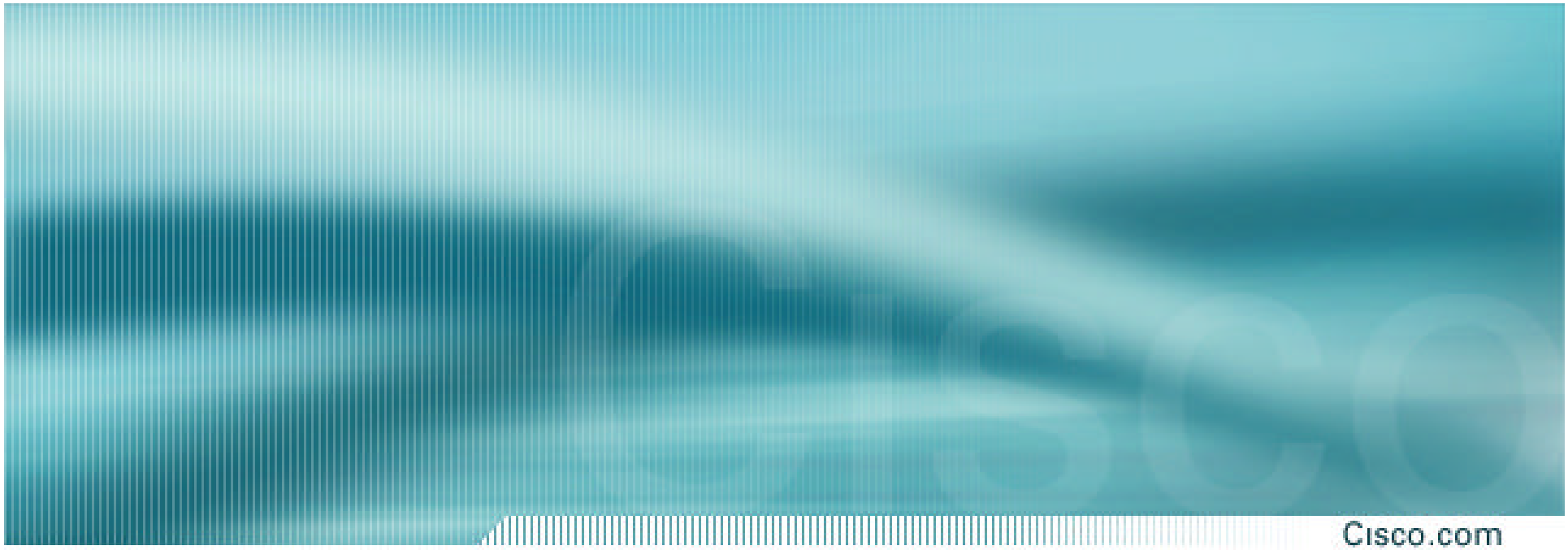
The Internet Today (January 2004)

- **Current Internet Routing Table Statistics**

BGP Routing Table Entries	131486
Prefixes after maximum aggregation	80923
Unique prefixes in Internet	63391
Prefixes smaller than registry alloc	57949
/24s announced	71643
only 5521 /24s are from 192.0.0.0/8	
ASes in use	16426

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**



Receiving Prefixes

Receiving Prefixes

- There are three scenarios for receiving prefixes from other ASNs
  - Customer talking BGP
  - Peer talking BGP
  - Upstream/Transit talking BGP
- Each has different filtering requirements and need to be considered separately

Receiving Prefixes: From Customers

- ISPs should only accept prefixes which have been assigned or allocated to their downstream customer
- If ISP has assigned address space to its customer, then the customer IS entitled to announce it back to his ISP
- If the ISP has NOT assigned address space to its customer, then:

  Check in the four RIR databases to see if this address space really has been assigned to the customer

  The tool: whois –h whois.apnic.net x.x.x.0/24

Receiving Prefixes: From Customers

- **Example use of whois to check if customer is entitled to announce address space:**

```
pfs-pc$ whois -h whois.apnic.net 202.12.29.0
inetnum:      202.12.29.0 - 202.12.29.255
netname:      APNIC-AP-AU-BNE
descr:        APNIC Pty Ltd - Brisbane Offices + Servers
descr:        Level 1, 33 Park Rd
descr:        PO Box 2131, Milton
descr:        Brisbane, QLD.
country:      AU
admin-c:      HM20-AP
tech-c:       NO4-AP
mnt-by:       APNIC-HM
changed:      hm-changed@apnic.net 20030108
status:       ASSIGNED PORTABLE
source:       APNIC
```

Portable – means its an assignment to the customer, the customer can announce it to you

Receiving Prefixes: From Customers

- **Example use of whois to check if customer is entitled to announce address space:**

```
$ whois -h whois.ripe.net 193.128.2.0
inetnum:      193.128.2.0 - 193.128.2.15
descr:        Wood Mackenzie
country:      GB
admin-c:      DB635-RIPE
tech-c:       DB635-RIPE
status:       ASSIGNED PA
mnt-by:       AS1849-MNT
changed:      davids@uk.uu.net 20020211
source:       RIPE

route:        193.128.0.0/14
descr:        PIPEX-BLOCK1
origin:       AS1849
notify:       routing@uk.uu.net
mnt-by:       AS1849-MNT
changed:      beny@uk.uu.net 20020321
source:       RIPE
```

ASSIGNED PA – means that it is Provider Aggregatable address space and can only be used for connecting to the ISP who assigned it

Receiving Prefixes from customer: Cisco IOS

- **For Example:**
  - **downstream has 220.50.0.0/20 block**
  - **should only announce this to upstreams**
  - **upstreams should only accept this from them**
- **Configuration on upstream**

```
router bgp 100
 neighbor 222.222.10.1 remote-as 101
 neighbor 222.222.10.1 prefix-list customer in
!
ip prefix-list customer permit 220.50.0.0/20
```

Receiving Prefixes: From Peers

- A peer is an ISP with whom you agree to exchange prefixes you originate into the Internet routing table

  Prefixes you accept from a peer are only those they have indicated they will announce

  Prefixes you announce to your peer are only those you have indicated you will announce

Receiving Prefixes: From Peers

- **Agreeing what each will announce to the other:**

  Exchange of e-mail documentation as part of the peering agreement, and then ongoing updates

  OR

  Use of the Internet Routing Registry and configuration tools such as the IRRToolSet

  www.ripe.net/ripencc/pub-services/db/irrtoolset/

Receiving Prefixes from peer: Cisco IOS

- **For Example:**

  peer has 220.50.0.0/16, 61.237.64.0/18 and 81.250.128.0/17 address blocks

- **Configuration on local router**

```
router bgp 100
 neighbor 222.222.10.1 remote-as 101
 neighbor 222.222.10.1 prefix-list my-peer in
!
ip prefix-list my-peer permit 220.50.0.0/16
ip prefix-list my-peer permit 61.237.64.0/18
ip prefix-list my-peer permit 81.250.128.0/17
ip prefix-list my-peer deny 0.0.0.0/0 le 32
```

Receiving Prefixes: From Upstream/Transit Provider

- **Upstream/Transit Provider is an ISP who you pay to give you transit to the WHOLE Internet**
- **Receiving prefixes from them is not desirable unless really necessary**

  special circumstances – see later
- **Ask upstream/transit provider to either:**

  originate a default-route

  OR

  announce one prefix you can use as default

Receiving Prefixes: From Upstream/Transit Provider

- **Downstream Router Configuration**

```
router bgp 100
 network 221.10.0.0 mask 255.255.224.0
 neighbor 221.5.7.1 remote-as 101
 neighbor 221.5.7.1 prefix-list infilter in
 neighbor 221.5.7.1 prefix-list outfilter out
!
ip prefix-list infilter permit 0.0.0.0/0
!
ip prefix-list outfilter permit 221.10.0.0/19
```

Receiving Prefixes: From Upstream/Transit Provider

- **Upstream Router Configuration**

```
router bgp 101
 neighbor 221.5.7.2 remote-as 100
 neighbor 221.5.7.2 default-originate
 neighbor 221.5.7.2 prefix-list cust-in in
 neighbor 221.5.7.2 prefix-list cust-out out
!
ip prefix-list cust-in permit 221.10.0.0/19
!
ip prefix-list cust-out permit 0.0.0.0/0
```

Receiving Prefixes: From Upstream/Transit Provider

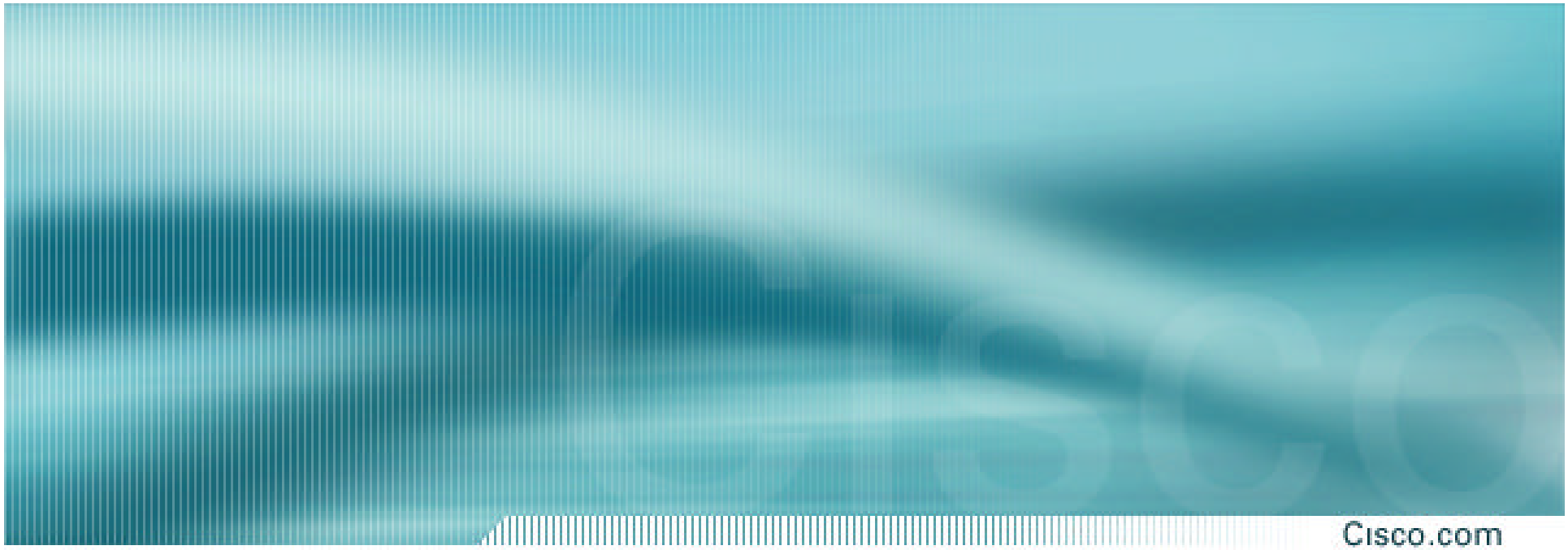
- **If necessary to receive prefixes from any provider, care is required**
  - **don't accept RFC1918 *etc* prefixes**
    - **ftp://ftp.rfc-editor.org/in-notes/rfc3330.txt**
  - **don't accept your own prefixes**
  - **don't accept default (unless you need it)**
  - **don't accept prefixes longer than /24**
- **Check Rob Thomas' list of "bogons"**
  - **http://www.cymru.org/Documents/bogon-list.html**

Receiving Prefixes


```
router bgp 100
 network 221.10.0.0 mask 255.255.224.0
 neighbor 221.5.7.1 remote-as 101
 neighbor 221.5.7.1 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 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 221.10.0.0/19 le 32 ! Block local prefix
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
```

Receiving Prefixes

- **Paying attention to prefixes received from customers, peers and transit providers assists with:**
  - **The integrity of the local network**
  - **The integrity of the Internet**
- **Responsibility of all ISPs to be good Internet citizens**



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

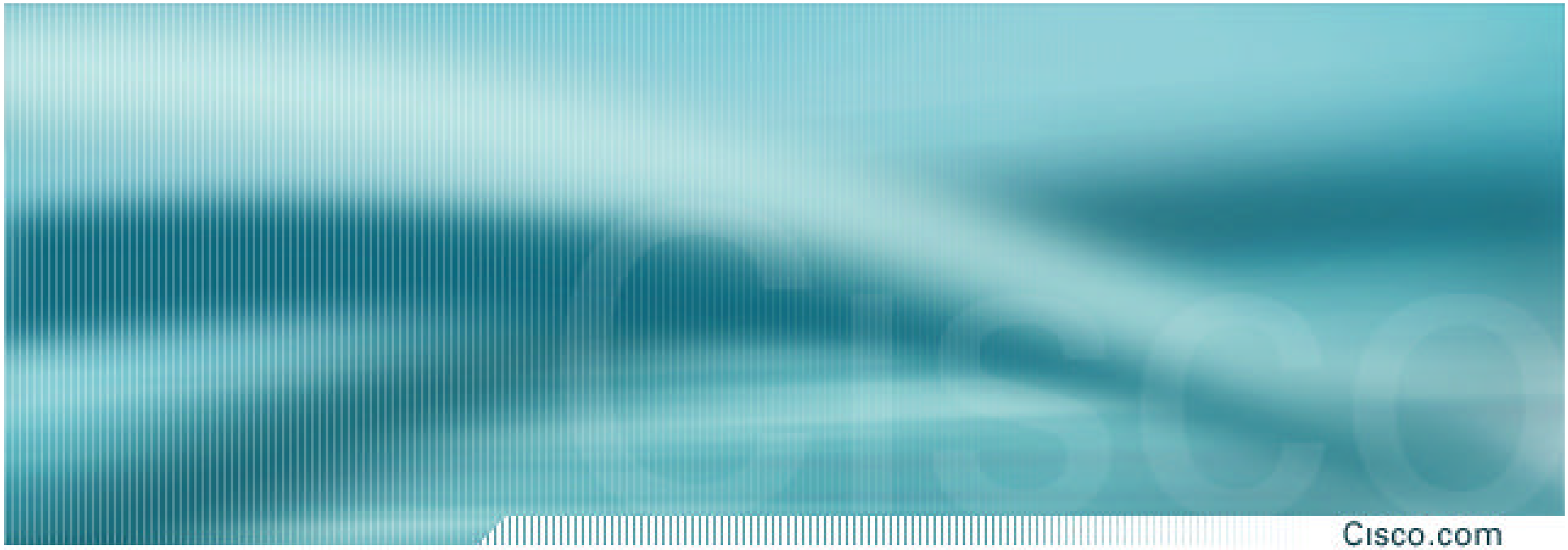
Router Configuration: redistribute static

- **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**



Scaling the network

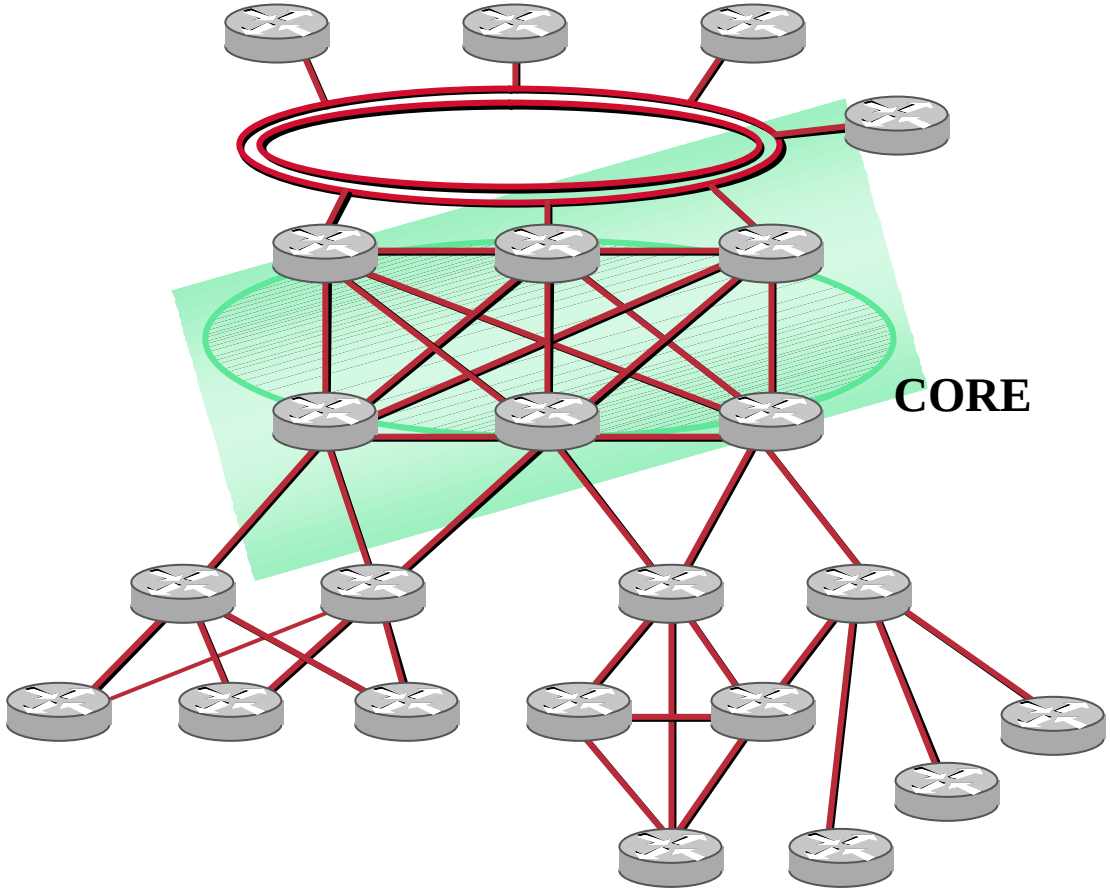
How to get out of carrying all prefixes in IGP

IGP Limitations

- **Amount of routing information in the network**
  - **Periodic updates/flooding**
  - **Long convergence times**
  - **Affects the core first**
- **Policy definition**
  - **Not easy to do**

BGP Cores
Sample Network

- Geographically distributed
- Hierarchical
- Redundant
- Media independent
- A clearly identifiable core



iBGP Core: Migration Plan

- Configure BGP on all the core routers
  - Transit path
  - Turn synchronisation off
  - Turn auto-summarisation off
- Check network borders
  - Ensure eBGP peerings only announce aggregates and won't leak specifics
- Route generation
  - Use static routes to generate summaries if required
  - Redistribution from the IGP is NOT recommended as it will cause instability

iBGP Core
Migration Plan (Cont.)

- **Route Generation – Example:**

```
!
router bgp 109
 network 200.200.200.0
 network 201.201.0.0 mask 255.255.0.0
!
ip route 200.200.200.0 255.255.255.0 null0
ip route 201.201.0.0 255.255.0.0 null0
!
```

iBGP Core Migration Plan (Cont.)

- **Verify consistency of routing information**

  Compare the IGP routing table against the BGP table – they must match!

- **Change the distance parameters so that the BGP routes are preferred**

  distance bgp 20 20 20

  All IGPs have a higher administrative distance

iBGP Core Migration Plan (Cont.)

- **Filter “non-core” IGP routes**
  - **Method will depend on the IGP used**
  - **May require the use of a different IGP process in the core if using a link state protocol**
  - **The routes to reach all the core links plus the BGP peering addresses must be carried by the IGP**

iBGP Core Migration Plan (Cont.)

- **Once iBGP carrying prefixes...**

  apply route-map to IGP redistribute commands so that only infrastructure addresses are in IGP

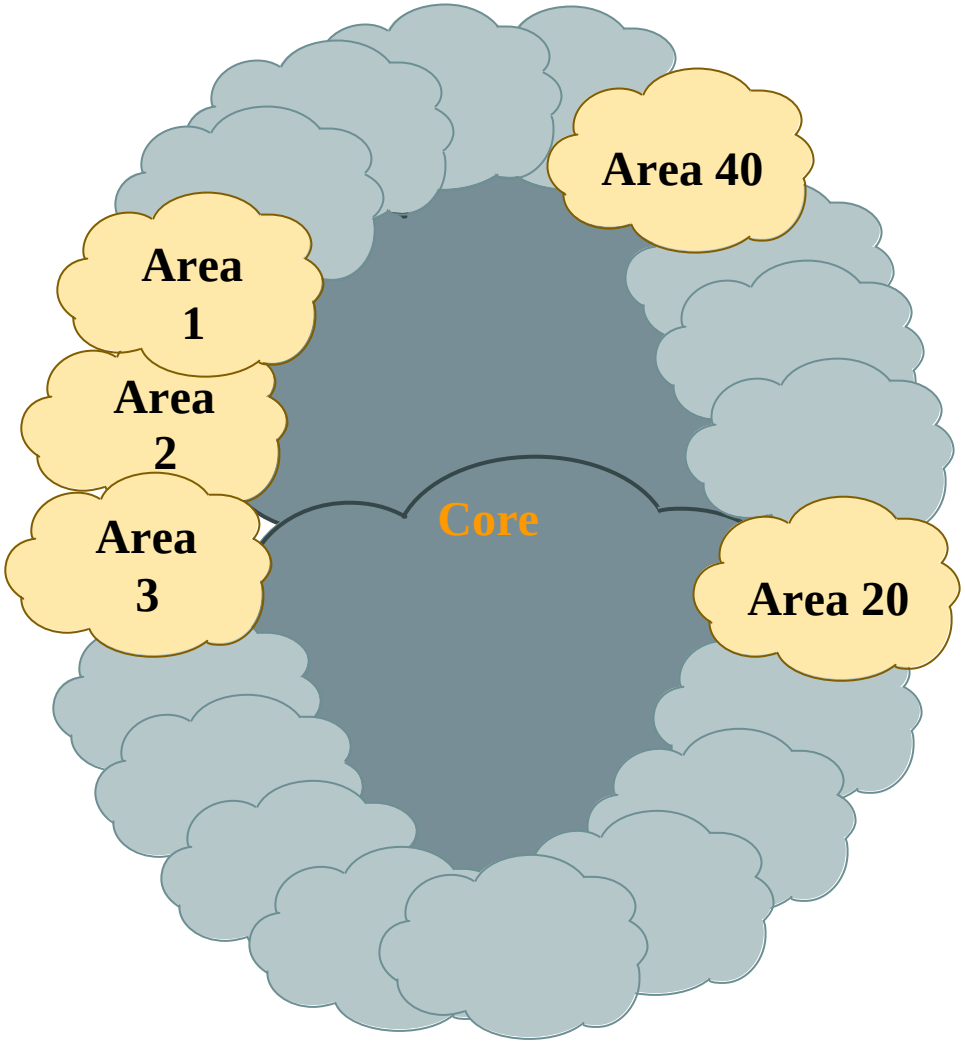
  check that customer routes in IGP have disappeared

  change BGP distance back to default

  no distance bgp 20 20 20

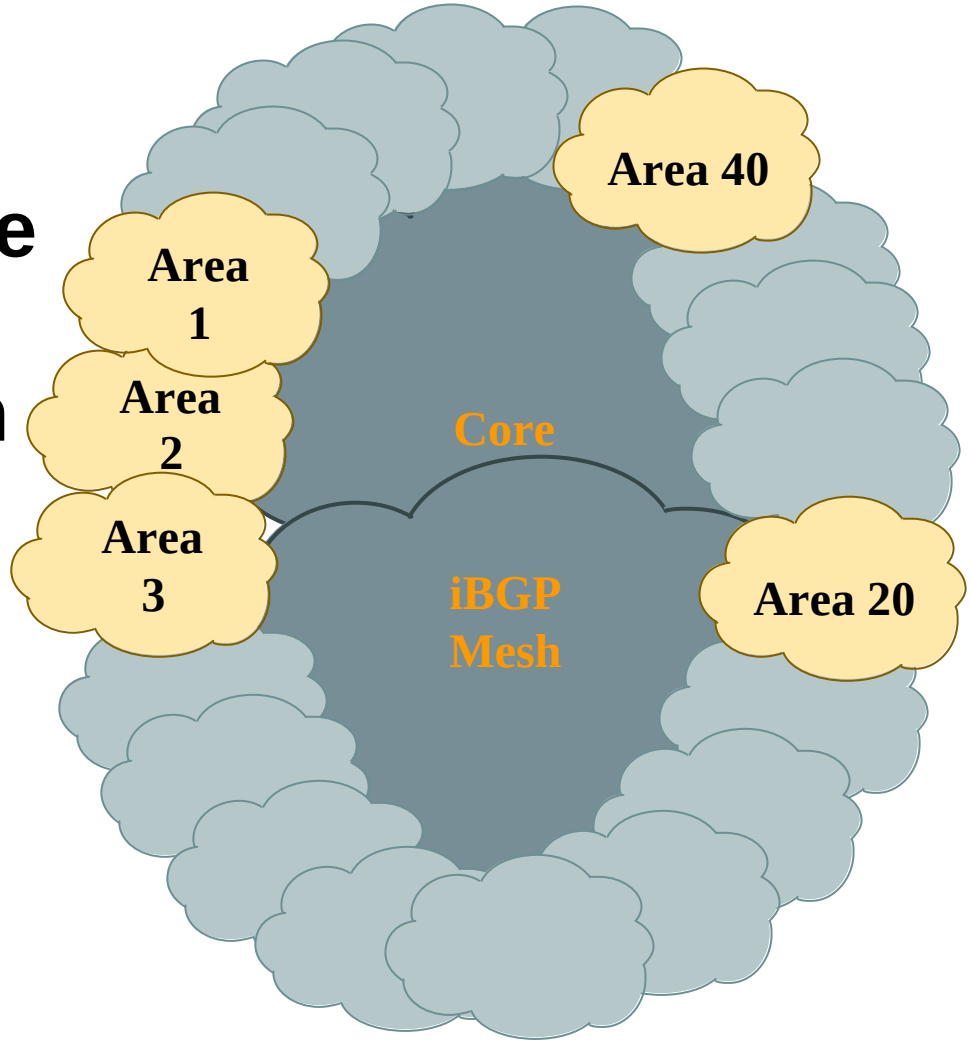
iBGP Core Before...

- IGP carries all the routes
- The core routers may be stressed due to the large number of routes



iBGP Core After...

- IGP carries only core links plus peering address information
- BGP carries all the routes
- Increased Stability!



iBGP Core Results

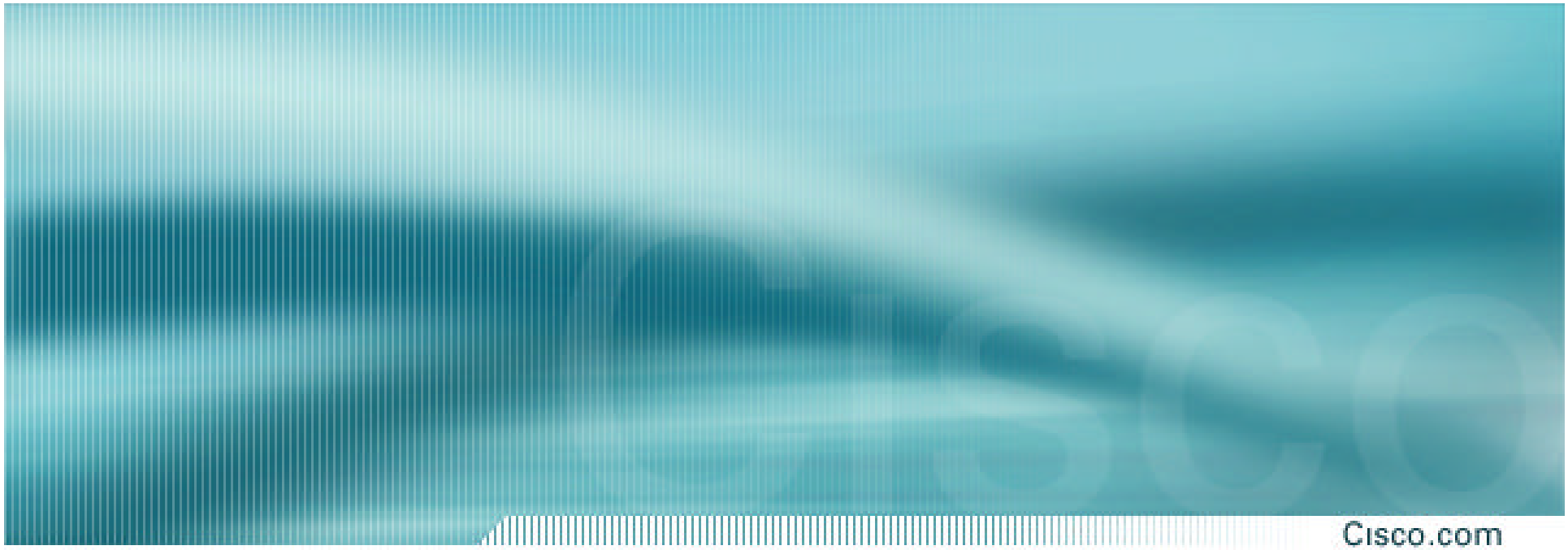
- **The routes from the core cannot be redistributed back into the IGP**
  - **Non-core areas need a default route**
  - **Amount of routing information in non-core areas has been reduced!**
- **Full logical iBGP mesh**
- **External connections must be located in the core**

Scaling Issues

- **Full mesh core**
  - **High number of neighbors**
  - **Update generation**
- **Complex topologies**
  - **Not a “simple” hierarchical network**
  - **Multiple external and/or inter-region connections**
  - **Policy definition and enforcement**

Scaling Issues: Solutions

- **Reduce the number of updates**
  - **Peer groups**
- **Reduce the number of neighbors**
  - **Confederations**
  - **Route reflectors**
- **Use additional information to effectively apply policies**
  - **eBGP provides extra granularity**
  - **Confederations**



BGP in the Internet

Best Current Practices