BGP Best Current Practices

ISP/IXP Workshops

Configuring BGP

Where do we start?

IOS Good Practices

- ISPs should start off with the following BGP commands as a basic template:

```
router bgp 64511
 bgp deterministic-med
 distance bgp 200 200 200
 no synchronization
 no auto-summary
```

Replace with public ASN

Make ebgp and ibgp distance the same

- If supporting more than just IPv4 unicast neighbours

```
 no bgp default ipv4 unicast
```

  is also very important and required

IOS Good Practices

- BGP in Cisco IOS is **permissive** by default
- Configuring BGP peering without using filters means:
  - All best paths on the local router are passed to the neighbour
  - All routes announced by the neighbour are received by the local router
  - Can have disastrous consequences
- **Good practice is to ensure that each eBGP neighbour has inbound and outbound filter applied:**

```
router bgp 64511
 neighbour 1.2.3.4 remote-as 64510
 neighbour 1.2.3.4 prefix-list as64510-in in
 neighbour 1.2.3.4 prefix-list as64510-out out
```

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

```
router bgp 64511
 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 special circumstances (more later)
- Aggregate should be generated internally

  Not on the network borders!

Announcing Aggregate – Cisco IOS

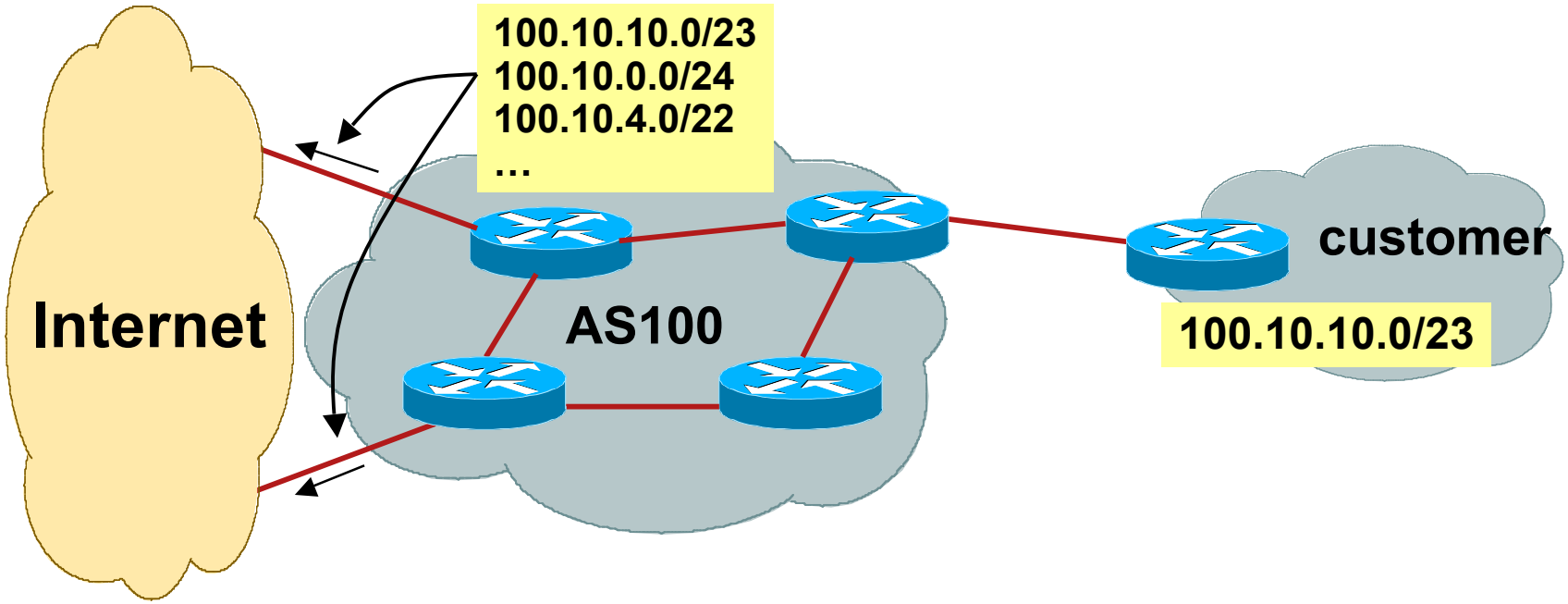
- Configuration Example

```
router bgp 64511
 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 publish their minimum allocation size
  - Anything from a /20 to a /22 depending on RIR
  - Different sizes for different address blocks
- No real reason to see anything longer than a /22 prefix in the Internet
  - BUT there are currently >174000 /24s!
- But: APNIC changed (Oct 2010) its minimum allocation size on all blocks to /24
  - IPv4 run-out is starting to have an impact

Aggregation – Example

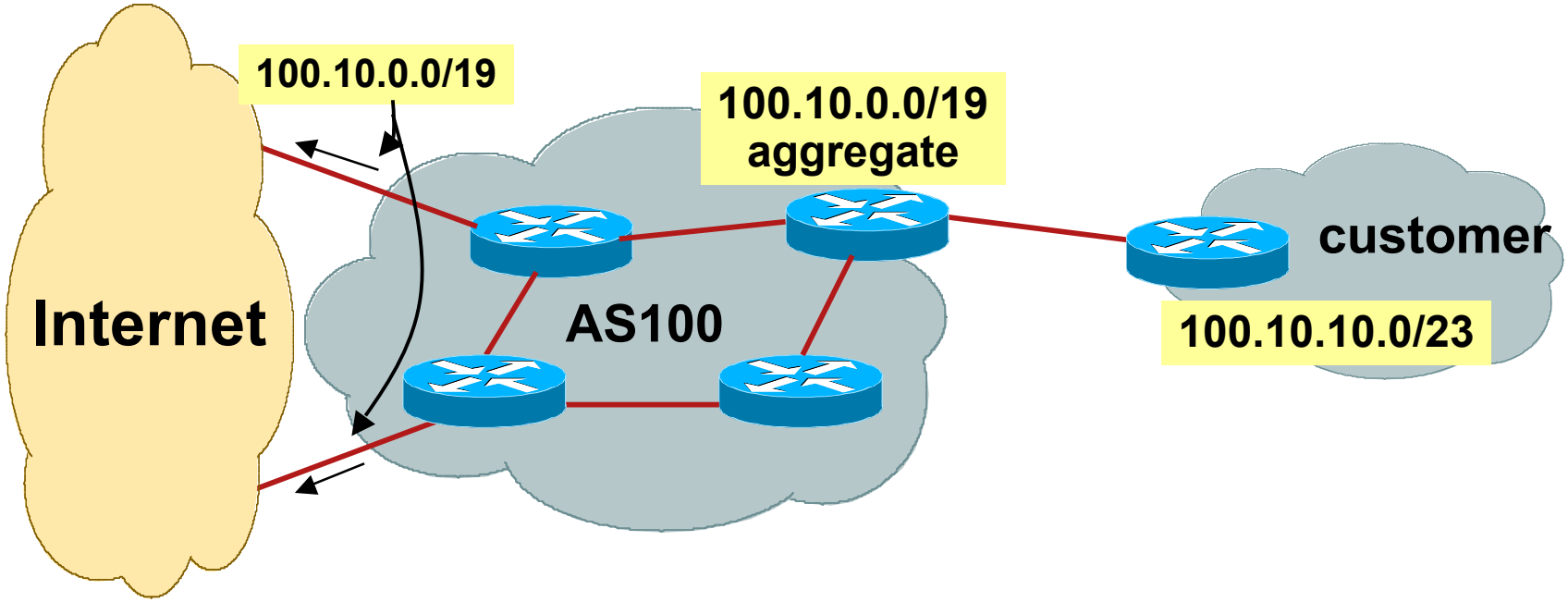


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

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???

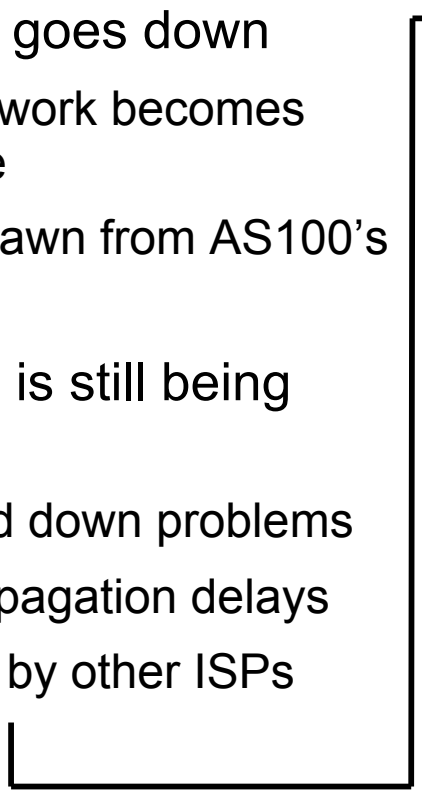
Aggregation – Example



- 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 – 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?
  - Laziness?

Separation of iBGP and eBGP

- Many ISPs do not understand the importance of separating iBGP and eBGP
  - iBGP is where all customer prefixes are carried
  - eBGP is used for announcing aggregate to Internet and for Traffic Engineering
- Do **NOT** do traffic engineering with customer originated iBGP prefixes
  - Leads to instability similar to that mentioned in the earlier bad example
  - Even though aggregate is announced, a flapping subprefix will lead to instability for the customer concerned
- **Generate traffic engineering prefixes on the Border Router**

The Internet Today (October 2010)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	333886
Prefixes after maximum aggregation	153111
Unique prefixes in Internet	163895
Prefixes smaller than registry alloc	137164
/24s announced	173994
ASes in use	34981

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
- RIPE Routing WG aggregation recommendation
  - **RIPE-399 — http://www.ripe.net/ripe/docs/ripe-399.html**

Efforts to Improve Aggregation
The CIDR Report

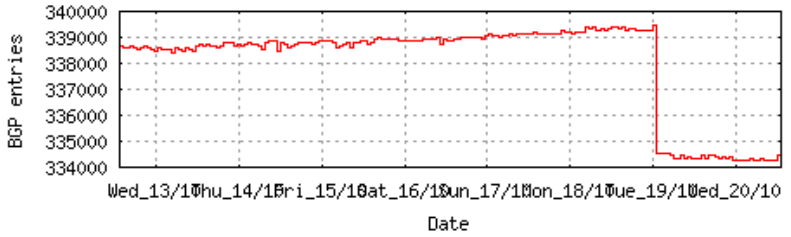
- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
  - Flexible and powerful tool to aid ISPs
  - Intended to show how greater efficiency in terms of BGP table size can be obtained 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

A list of advertisements of address blocks and Autonomous System numbers where there is no matching allocation data.

Status Summary

Table History

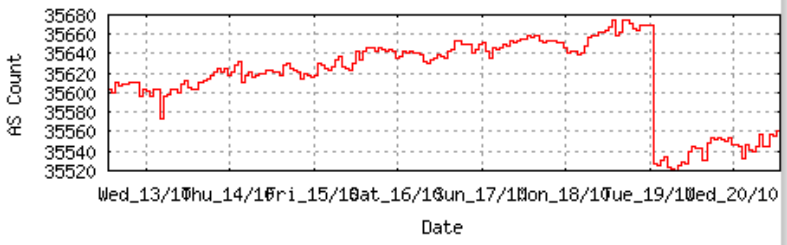
Date	Prefixes	CIDR Aggregated
13-10-10	338465	209416
14-10-10	338701	209568
15-10-10	338841	210332
16-10-10	338847	210836
17-10-10	339053	210881
18-10-10	339193	210990
19-10-10	339234	194438
20-10-10	334281	194615



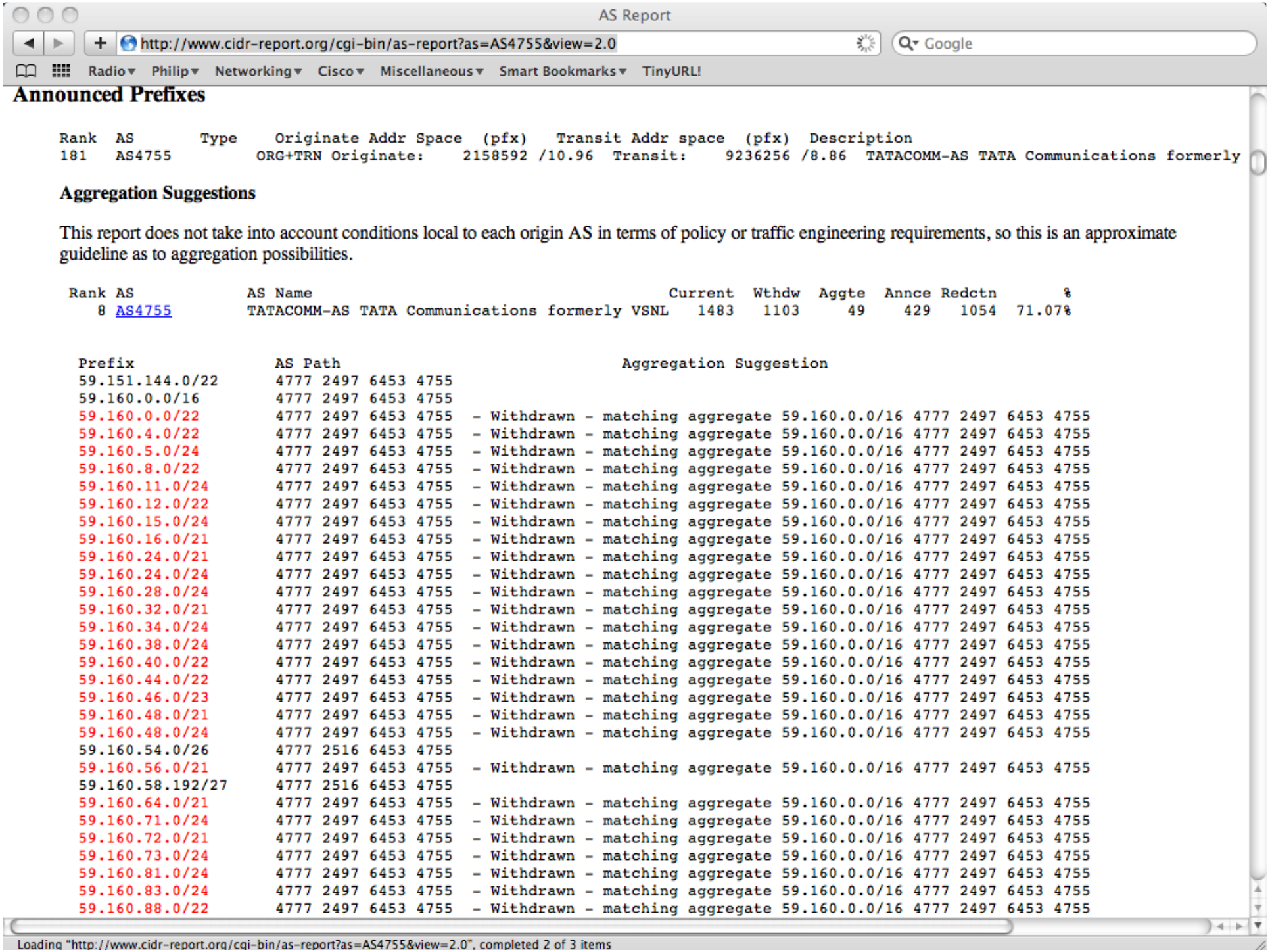
Plot: BGP Table Size

AS Summary

35559	Number of ASes in routing system
15267	Number of ASes announcing only one prefix
4489	Largest number of prefixes announced by an AS AS4323: TWTC - tw telecom holdings, inc.
101030656	Largest address span announced by an AS (/32s) AS4134: CHINANET-BACKBONE No.31,Jin-rong Street



Plot: AS count
Plot: Average announcements per origin AS
Report: ASes ordered by originating address span
Report: ASes ordered by transit address span
Report: Autonomous System number-to-name mapping (from Registry WHOIS data)



Announced Prefixes

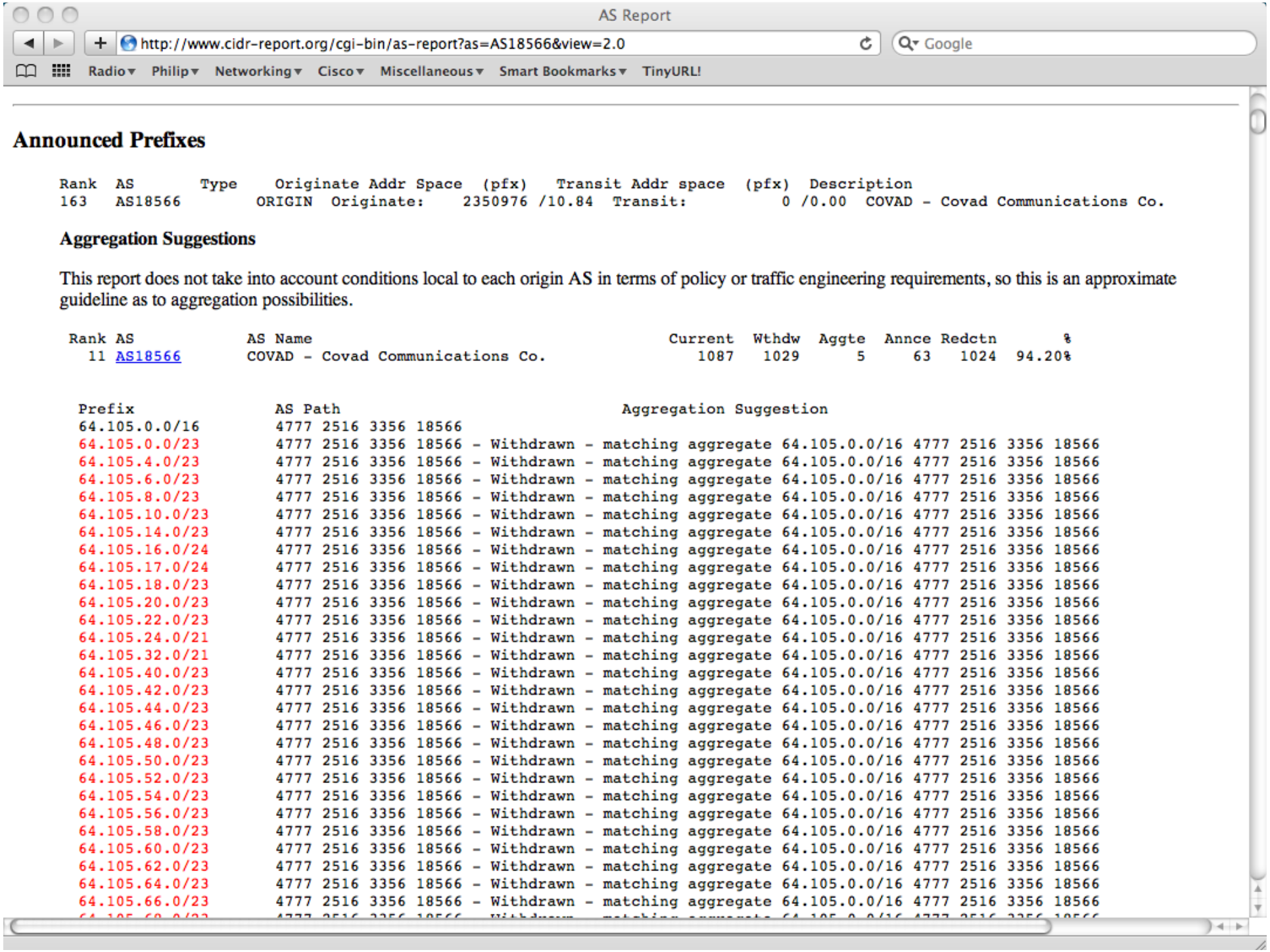
```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
181   AS4755         ORG+TRN Originate:    2158592 /10.96  Transit:    9236256 /8.86  TATACOMM-AS TATA Communications formerly
```

Aggregation Suggestions

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
Rank AS            AS Name                                          Current  Wthdw  Aggte  Annce Redctn       %
   8 AS4755        TATACOMM-AS TATA Communications formerly VSNL     1483   1103     49    429   1054  71.07%
```

Prefix	AS Path	Aggregation Suggestion
59.151.144.0/22	4777 2497 6453 4755	
59.160.0.0/16	4777 2497 6453 4755	
59.160.0.0/22	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.4.0/22	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.5.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.8.0/22	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.11.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.12.0/22	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.15.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.16.0/21	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.24.0/21	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.24.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.28.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.32.0/21	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.34.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.38.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.40.0/22	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.44.0/22	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.46.0/23	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.48.0/21	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.48.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.54.0/26	4777 2516 6453 4755	
59.160.56.0/21	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.58.192/27	4777 2516 6453 4755	
59.160.64.0/21	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.71.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.72.0/21	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.73.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.81.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.83.0/24	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755
59.160.88.0/22	4777 2497 6453 4755	- Withdrawn - matching aggregate 59.160.0.0/16 4777 2497 6453 4755



Announced Prefixes

```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
163   AS18566        ORIGIN  Originate:    2350976 /10.84  Transit:          0 /0.00  COVAD - Covad Communications Co.
```

Aggregation Suggestions

This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
Rank AS           AS Name                                    Current  Wthdw  Aggte  Annce Redctn      %
  11 AS18566      COVAD - Covad Communications Co.              1087   1029      5     63   1024  94.20%
```

Prefix	AS Path	Aggregation Suggestion
64.105.0.0/16	4777 2516 3356 18566	
64.105.0.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.4.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.6.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.8.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.10.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.14.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.16.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.17.0/24	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.18.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.20.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.22.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.24.0/21	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.32.0/21	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.40.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.42.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.44.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.46.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.48.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.50.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.52.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.54.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.56.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.58.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.60.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.62.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.64.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566
64.105.66.0/23	4777 2516 3356 18566	- Withdrawn - matching aggregate 64.105.0.0/16 4777 2516 3356 18566

Importance of Aggregation

- Size of routing table
  - Router Memory is not so much of a problem as it was in the 1990s
  - Routers can be specified to carry 1 million+ prefixes
- Convergence of the Routing System
  - This is a problem
  - Bigger table takes longer for CPU to process
  - BGP updates take longer to deal with
  - BGP Instability Report tracks routing system update activity
  - **http://bgpupdates.potaroo.net/instability/bgpupd.html**

The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 20 October 2010 06:10 (UTC+1000)

50 Most active ASes for the past 7 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	9476	26245	3.05%	30	874.83	INTRAPOWER-AS-AP IntraPower Pty. Ltd.
2	7315	24477	2.85%	70	349.67	COLOMBIA TELECOMUNICACIONES S.A. ESP
3	8151	21070	2.45%	2294	9.18	Uninet S.A. de C.V.
4	3464	19872	2.31%	45	441.60	ASC-NET - Alabama Supercomputer Network
5	4538	16669	1.94%	5378	3.10	ERX-CERNET-BKB China Education and Research Network Center
6	32528	14759	1.72%	8	1844.88	ABBOTT Abbot Labs
7	14420	12917	1.50%	544	23.74	CORPORACION NACIONAL DE TELECOMUNICACIONES - CNT EP
8	5536	12751	1.48%	112	113.85	Internet-Egypt
9	7552	12638	1.47%	653	19.35	VIETEL-AS-AP Vietel Corporation
10	33363	10210	1.19%	1414	7.22	BHN-TAMPA - BRIGHT HOUSE NETWORKS, LLC
11	6128	10200	1.19%	2030	5.02	CABLE-NET-1 - Cablevision Systems Corp.
12	9829	9053	1.05%	826	10.96	BSNL-NIB National Internet Backbone
13	6517	8896	1.03%	574	15.50	RELIANCEGLOBALCOM - Reliance Globalcom Services, Inc
14	5800	8357	0.97%	205	40.77	DNIC-ASBLK-05800-06055 - DoD Network Information Center
15	35931	8147	0.95%	6	1357.83	ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
16	12332	7437	0.87%	61	121.92	PRIMORYE-AS Open Joint Stock Company "Far East Telecommunications Company"
17	3549	6732	0.78%	839	8.02	GBLX Global Crossing Ltd.
18	3816	6298	0.73%	487	12.93	COLOMBIA TELECOMUNICACIONES S.A. ESP
19	21017	5761	0.67%	10	576.10	VSI-AS VSI AS
20	45899	5677	0.66%	263	21.59	VNPT-AS-VN VNPT Corp
21	8452	5612	0.65%	1252	4.48	TE-AS TE-AS
22	24863	4627	0.54%	739	6.26	LINKdotNET-AS
23	45595	4598	0.53%	372	12.36	PKTELECOM-AS-PK Pakistan Telecom Company Limited

The BGP Instability Report

http://bgpupdates.potaroo.net/instability/bgpupd.html

Google

Radio Philip Networking Cisco Miscellaneous Smart Bookmarks TinyURL!

50 Most active Prefixes for the past 7 days

RANK	PREFIX	UPDs	%	Origin AS -- AS NAME
1	203.1.14.0/24	14542	1.53%	9476 -- INTRAPOWER-AS-AP IntraPower Pty. Ltd.
2	203.1.13.0/24	11697	1.23%	9476 -- INTRAPOWER-AS-AP IntraPower Pty. Ltd.
3	129.66.0.0/17	10138	1.06%	3464 -- ASC-NET - Alabama Supercomputer Network
4	129.66.128.0/17	9657	1.01%	3464 -- ASC-NET - Alabama Supercomputer Network
5	76.74.88.0/23	8806	0.92%	6517 -- RELIANCEGLOBALCOM - Reliance Globalcom Services, Inc
6	130.36.34.0/24	7372	0.77%	32528 -- ABBOTT Abbot Labs
7	130.36.35.0/24	7371	0.77%	32528 -- ABBOTT Abbot Labs
8	63.211.68.0/22	5367	0.56%	35931 -- ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
9	190.65.228.0/22	5158	0.54%	3816 -- COLOMBIA TELECOMUNICACIONES S.A. ESP
10	72.31.122.0/24	5062	0.53%	33363 -- BHN-TAMPA - BRIGHT HOUSE NETWORKS, LLC
11	72.31.98.0/24	4994	0.52%	33363 -- BHN-TAMPA - BRIGHT HOUSE NETWORKS, LLC
12	201.134.18.0/24	4980	0.52%	8151 -- Uninet S.A. de C.V.
13	95.32.192.0/18	3010	0.32%	21017 -- VSI-AS VSI AS
14	95.32.128.0/18	2735	0.29%	21017 -- VSI-AS VSI AS
15	198.140.43.0/24	2734	0.29%	35931 -- ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
16	208.54.82.0/24	2708	0.28%	701 -- UUNET - MCI Communications Services, Inc. d/b/a Verizon Business
17	202.92.235.0/24	2448	0.26%	9498 -- BBIL-AP BHARTI Airtel Ltd.
18	66.187.234.0/24	2393	0.25%	22753 -- REDHAT-STUTTGART REDHAT Stuttgart
19	206.184.16.0/24	1827	0.19%	174 -- COGENT Cogent/PSI
20	64.214.66.0/23	1463	0.15%	3549 -- GBLX Global Crossing Ltd.
21	64.214.116.0/23	1463	0.15%	3549 -- GBLX Global Crossing Ltd.
22	113.29.32.0/20	1461	0.15%	3549 -- GBLX Global Crossing Ltd.
23	113.29.0.0/20	1461	0.15%	3549 -- GBLX Global Crossing Ltd.
24	214.15.65.0/24	1390	0.15%	5863 -- DNIC-ASBLK-05800-06055 - DoD Network Information Center
25	189.85.51.0/24	1382	0.14%	28175 --
26	143.74.20.0/24	1069	0.11%	6006 -- DNIC-ASBLK-05800-06055 - DoD Network Information Center
27	202.167.253.0/24	987	0.10%	17819 -- ASN-EQUINIX-AP Equinix Asia Pacific
28	202.177.223.0/24	987	0.10%	17819 -- ASN-EQUINIX-AP Equinix Asia Pacific

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 five 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:

```
$ 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 100.50.0.0/20 block

  should only announce this to upstreams

  upstreams should only accept this from them

- Configuration on upstream

```
router bgp 100
 neighbor 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
```

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.isc.org/sw/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 102.102.10.1 remote-as 101
 neighbor 102.102.10.1 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 101.10.0.0 mask 255.255.224.0
 neighbor 101.5.7.1 remote-as 101
 neighbor 101.5.7.1 prefix-list infilter in
 neighbor 101.5.7.1 prefix-list outfilter out
!
ip prefix-list infilter permit 0.0.0.0/0
!
ip prefix-list outfilter permit 101.10.0.0/19
```

Receiving Prefixes: From Upstream/Transit Provider

- 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-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 private (RFC1918) and certain special use prefixes:
    - **http://www.rfc-editor.org/rfc/rfc5735.txt**
  - don't accept your own prefixes
  - don't accept default (unless you need it)
  - don't accept prefixes longer than /24
- Check Team Cymru's list of "bogons"
  - **www.cymru.com/Documents/bogon-list.html**
  - **www.team-cymru.org/Services/Bogons/routeserver.html**

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
```

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 redistribute static routes into BGP rather than using the network statement

  Only do 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

Why use BGP rather than IGP?

- IGP has Limitations:
  - The more routing information in the network
    - Periodic updates/flooding “overload”
    - Long convergence times
    - Affects the core first
  - Policy definition
    - Not easy to do

Preparing the Network

- We want to deploy BGP now…
- BGP will be used therefore an ASN is required
- If multihoming to different ISPs is intended in the near future, a public ASN should be obtained:

  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
Initial Assumptions

- 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
First Step: IGP

- Decide on an IGP: OSPF or ISIS ☺
- Assign loopback interfaces and /32 address 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
  - e.g. OSPF distance might be 110; static distance is 1
  - **Smallest distance wins**

Preparing the Network IGP (cont)

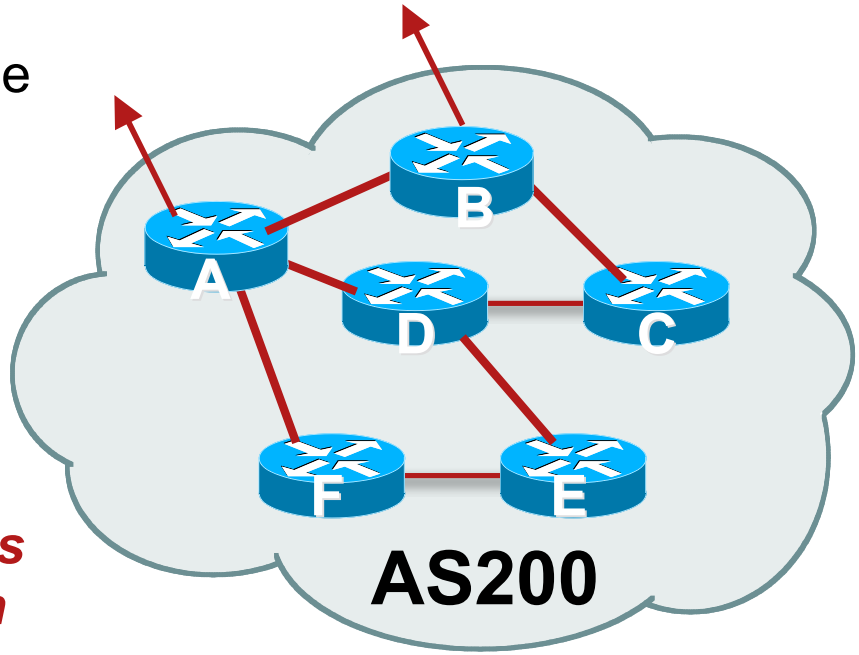
- Be prudent deploying IGP – keep the Link State Database Lean!
  - Router loopbacks go in IGP
  - WAN point to point links go in IGP
  - (In fact, any link where IGP dynamic routing will be run should go into IGP)
  - Summarise on area/level boundaries (if possible) – i.e. think about your IGP address 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

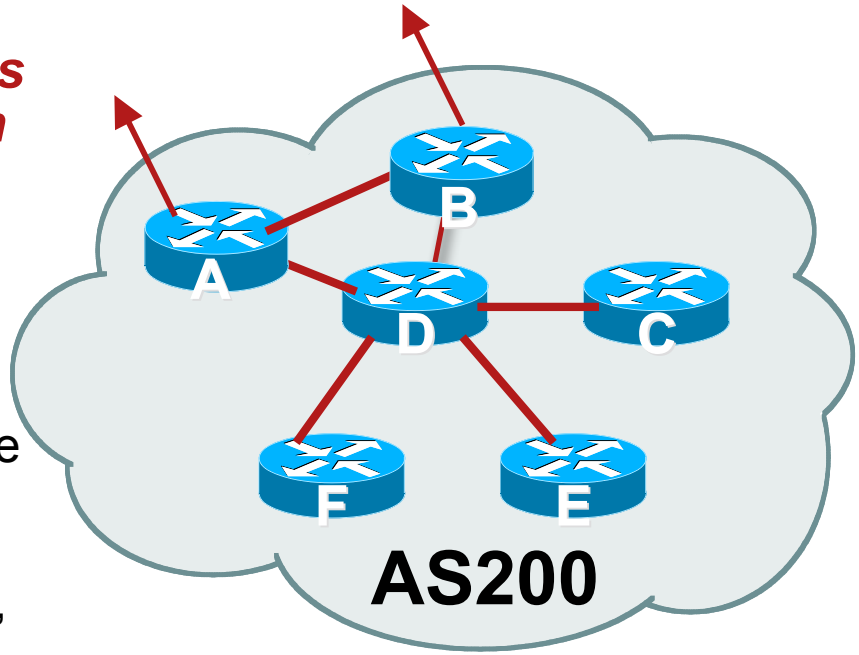
Preparing the Network
Second Step: iBGP

- 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
Second Step: 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 by the 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 Best Current Practices

ISP/IXP Workshops