# BGP Best Current Practices

AfNOG 2011 AR-E Workshop

# 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

# Cisco 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 for traffic engineering
  - See BGP Multihoming presentations
- Aggregate should be generated internally
  - Not on the network borders!

# Announcing Aggregate – Cisco IOS

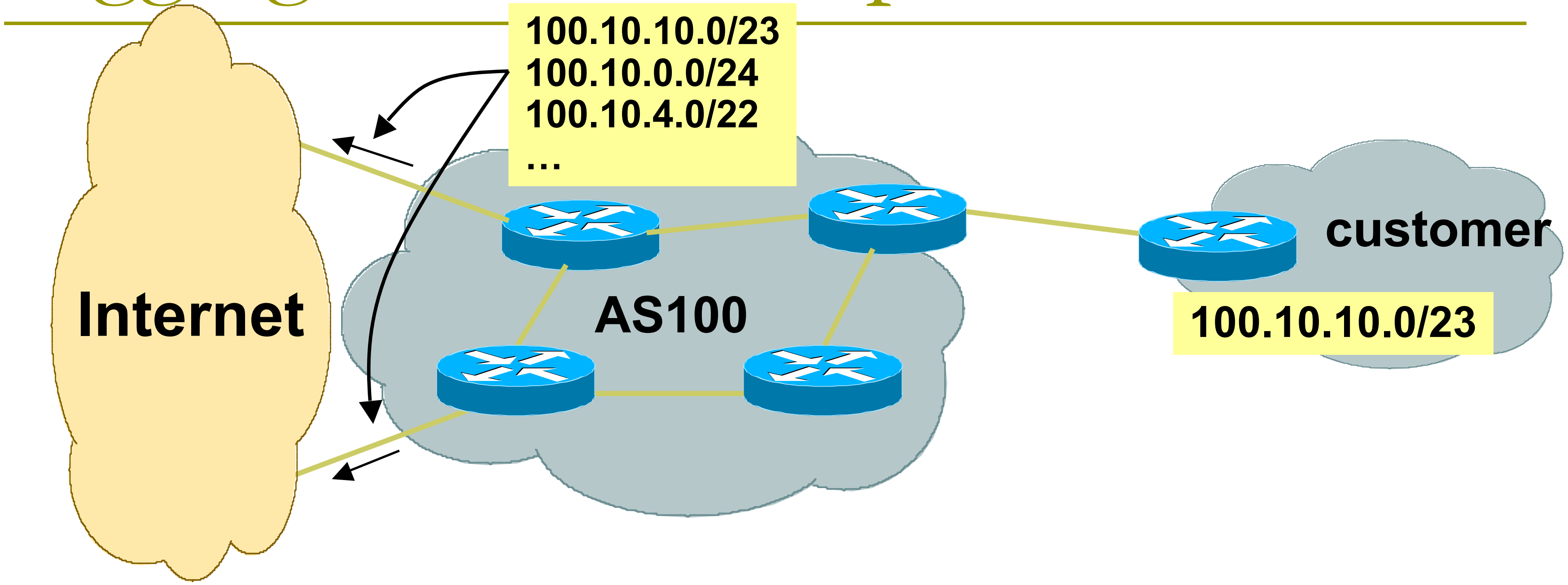
- Configuration Example

```
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 (May 2011) >185000 /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



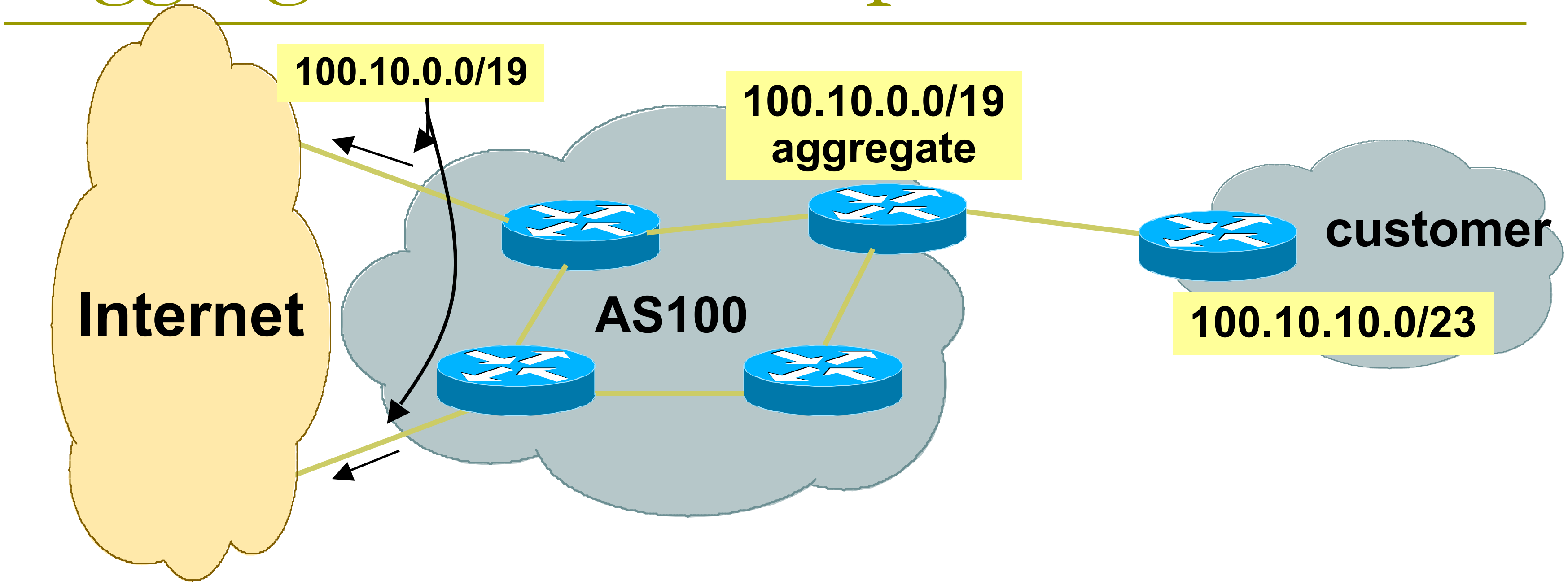
- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

# 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 (May 2011)

- Current Internet Routing Table Statistics

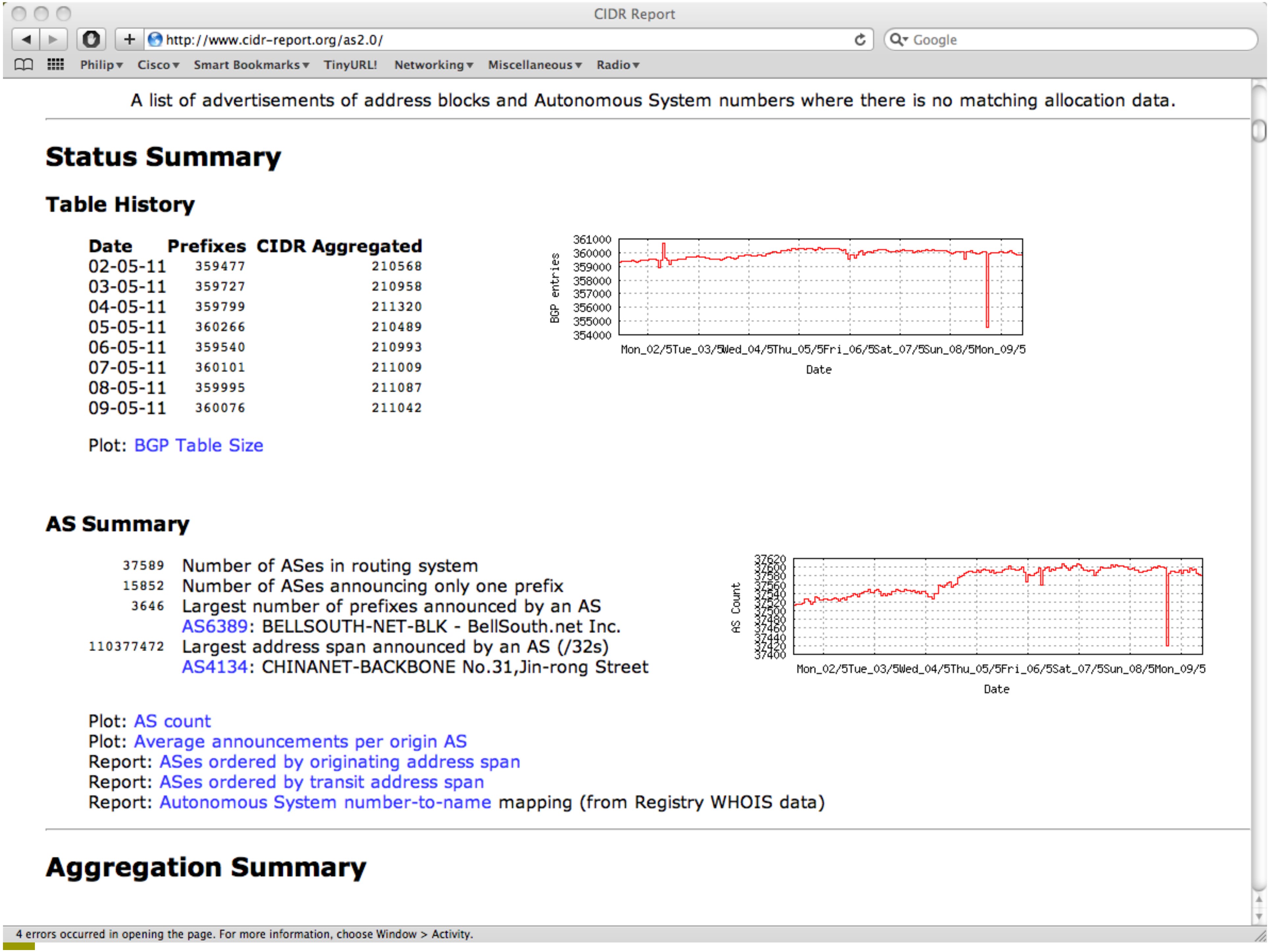
| | |
|---|---|
| BGP Routing Table Entries | 356524 |
| Prefixes after maximum aggregation | 161247 |
| Unique prefixes in Internet | 176062 |
| Prefixes smaller than registry alloc | 148138 |
| /24s announced | 185829 |
| ASes in use | 37487 |

# 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
    - (covers both IPv4 and IPv6 BGP tables)
  - 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 — 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 |
|---|---|---|
| 02-05-11 | 359477 | 210568 |
| 03-05-11 | 359727 | 210958 |
| 04-05-11 | 359799 | 211320 |
| 05-05-11 | 360266 | 210489 |
| 06-05-11 | 359540 | 210993 |
| 07-05-11 | 360101 | 211009 |
| 08-05-11 | 359995 | 211087 |
| 09-05-11 | 360076 | 211042 |

BGP entries
361000
360000
359000
358000
357000
356000
355000
354000
Mon_02/5Tue_03/5Wed_04/5Thu_05/5Fri_06/5Sat_07/5Sun_08/5Mon_09/5
Date

Plot: BGP Table Size

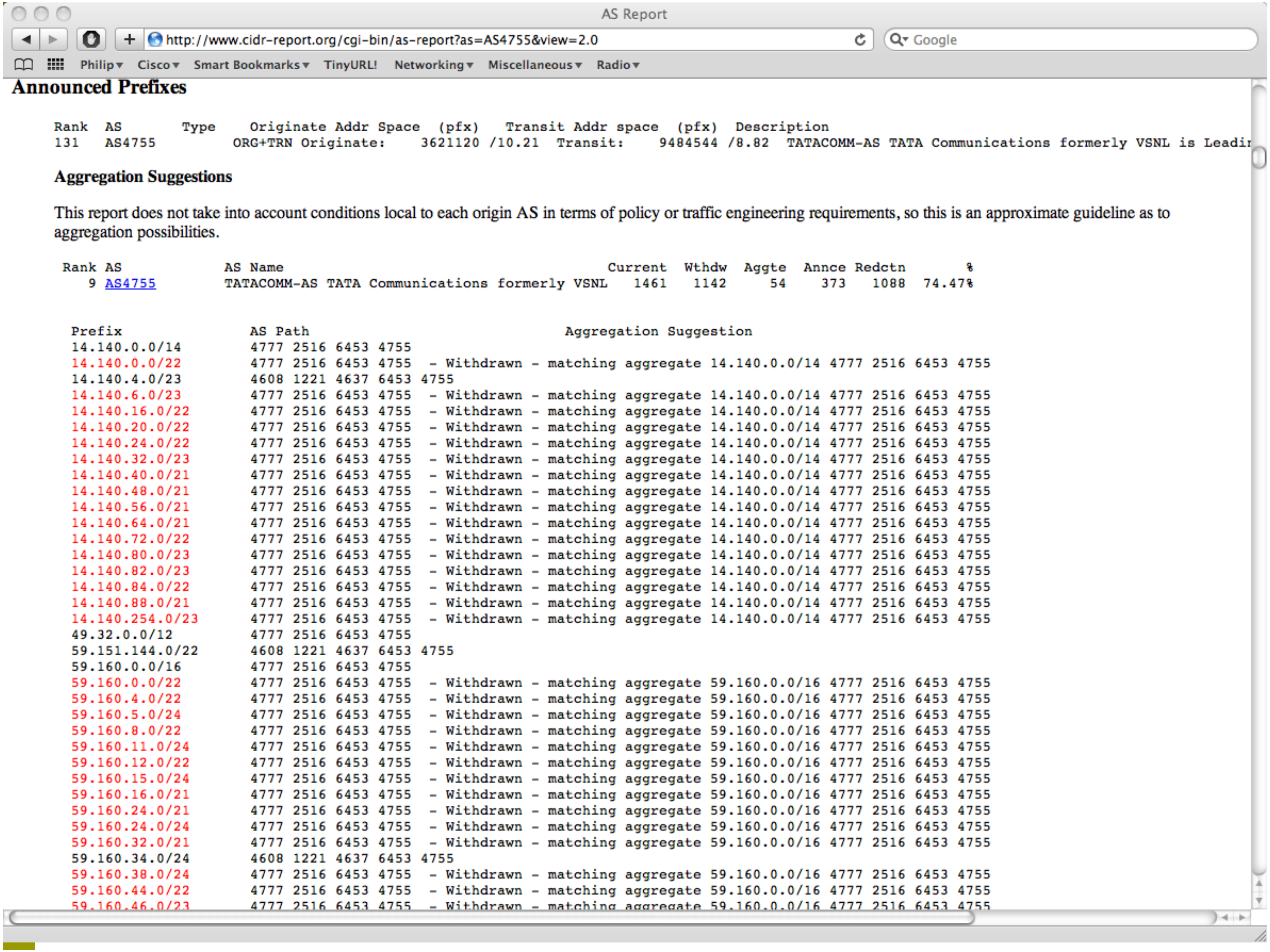
### AS Summary

37589 Number of ASes in routing system
15852 Number of ASes announcing only one prefix
3646 Largest number of prefixes announced by an AS
AS6389: BELLSOUTH-NET-BLK - BellSouth.net Inc.
110377472 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)

## Aggregation Summary

4 errors occurred in opening the page. For more information, choose Window > Activity.



## Announced Prefixes

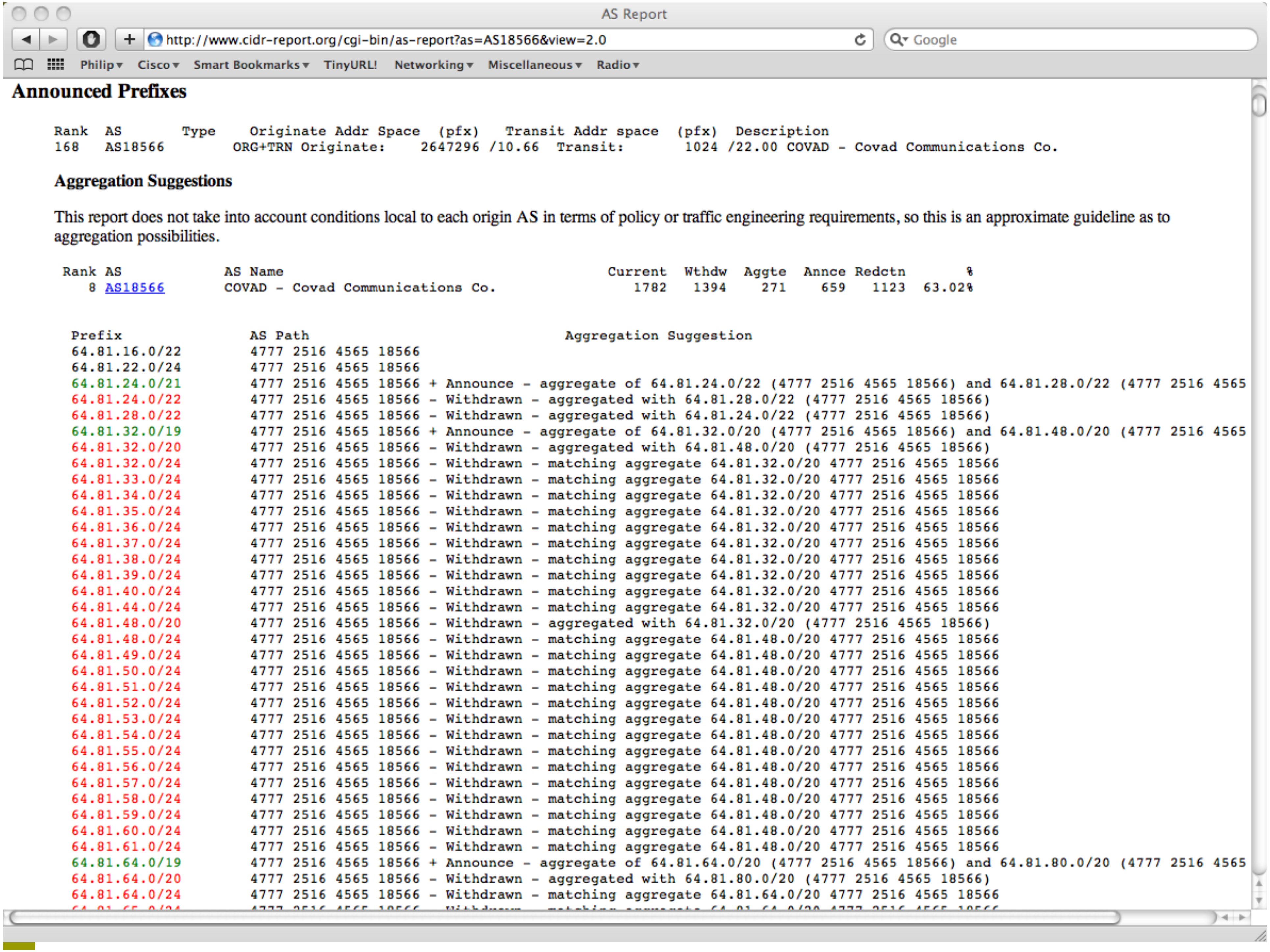
```
Rank  AS       Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
131   AS4755          ORG+TRN Originate:    3621120 /10.21  Transit:    9484544 /8.82  TATACOMM-AS TATA Communications formerly VSNL is Leadir
```

### 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       %
    9 AS4755       TATACOMM-AS TATA Communications formerly VSNL   1461   1142     54    373   1088  74.47%


  Prefix               AS Path                           Aggregation Suggestion
  14.140.0.0/14        4777 2516 6453 4755
  14.140.0.0/22        4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.4.0/23        4608 1221 4637 6453 4755
  14.140.6.0/23        4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.16.0/22       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.20.0/22       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.24.0/22       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.32.0/23       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.40.0/21       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.48.0/21       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.56.0/21       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.64.0/21       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.72.0/22       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.80.0/23       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.82.0/23       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.84.0/22       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.88.0/21       4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  14.140.254.0/23      4777 2516 6453 4755  - Withdrawn - matching aggregate 14.140.0.0/14 4777 2516 6453 4755
  49.32.0.0/12         4777 2516 6453 4755
  59.151.144.0/22      4608 1221 4637 6453 4755
  59.160.0.0/16        4777 2516 6453 4755
  59.160.0.0/22        4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.4.0/22        4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.5.0/24        4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.8.0/22        4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.11.0/24       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.12.0/22       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.15.0/24       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.16.0/21       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.24.0/21       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.24.0/24       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.32.0/21       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.34.0/24       4608 1221 4637 6453 4755
  59.160.38.0/24       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.44.0/22       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
  59.160.46.0/23       4777 2516 6453 4755  - Withdrawn - matching aggregate 59.160.0.0/16 4777 2516 6453 4755
```



## Announced Prefixes

```
Rank  AS        Type    Originate Addr Space  (pfx)   Transit Addr space  (pfx)  Description
168   AS18566       ORG+TRN Originate:     2647296 /10.66   Transit:        1024 /22.00 COVAD - Covad Communications Co.
```

### Aggregation Suggestions

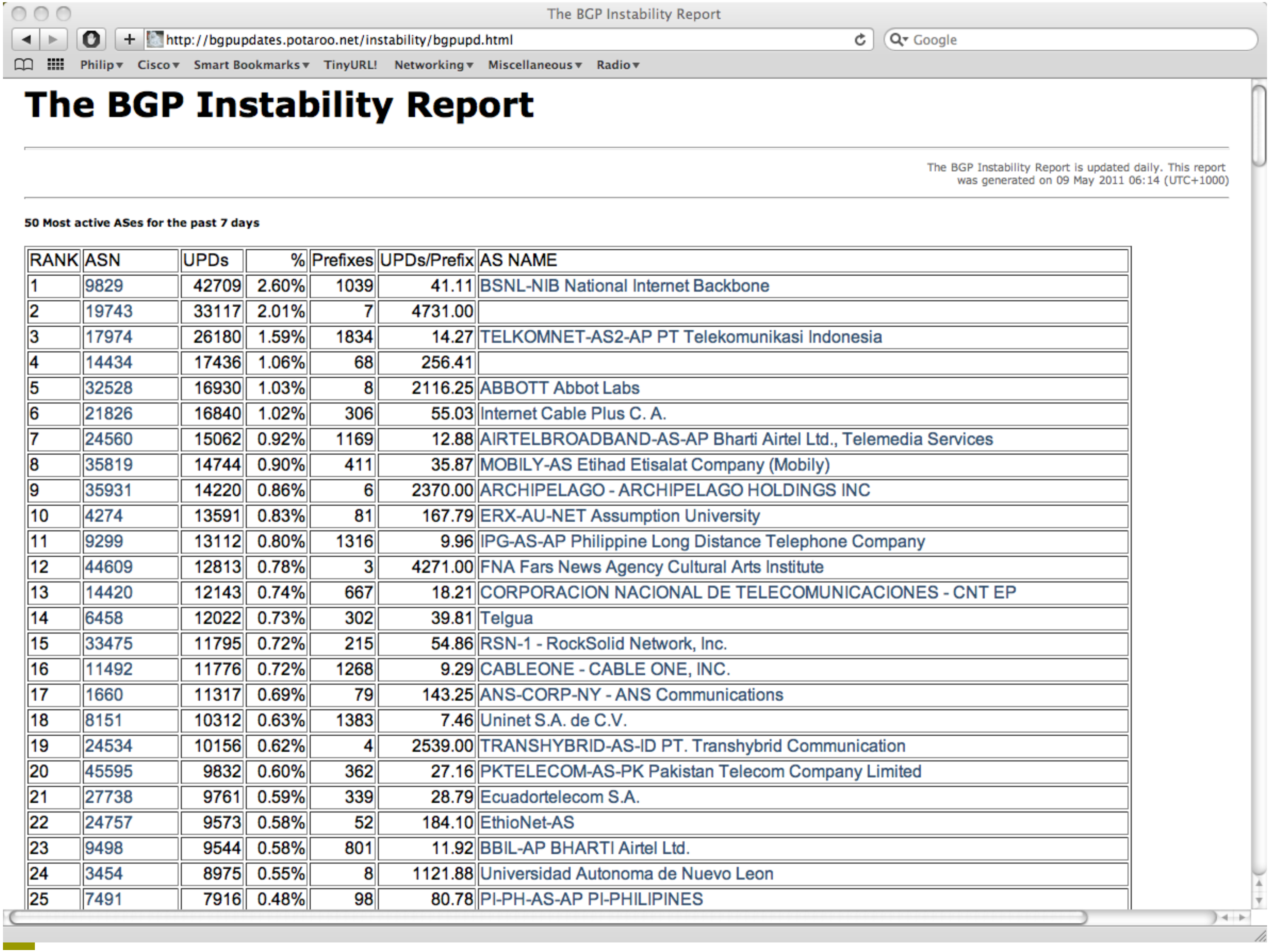
This report does not take into account conditions local to each origin AS in terms of policy or traffic engineering requirements, so this is an approximate guideline as to aggregation possibilities.

```
 Rank AS          AS Name                                     Current  Wthdw  Aggte  Annce Redctn     %
    8 AS18566     COVAD - Covad Communications Co.               1782   1394    271    659   1123  63.02%


  Prefix               AS Path                            Aggregation Suggestion
  64.81.16.0/22        4777 2516 4565 18566
  64.81.22.0/24        4777 2516 4565 18566
  64.81.24.0/21        4777 2516 4565 18566 + Announce - aggregate of 64.81.24.0/22 (4777 2516 4565 18566) and 64.81.28.0/22 (4777 2516 4565
  64.81.24.0/22        4777 2516 4565 18566 - Withdrawn - aggregated with 64.81.28.0/22 (4777 2516 4565 18566)
  64.81.28.0/22        4777 2516 4565 18566 - Withdrawn - aggregated with 64.81.24.0/22 (4777 2516 4565 18566)
  64.81.32.0/19        4777 2516 4565 18566 + Announce - aggregate of 64.81.32.0/20 (4777 2516 4565 18566) and 64.81.48.0/20 (4777 2516 4565
  64.81.32.0/20        4777 2516 4565 18566 - Withdrawn - aggregated with 64.81.48.0/20 (4777 2516 4565 18566)
  64.81.32.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.33.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.34.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.35.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.36.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.37.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.38.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.39.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.40.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.44.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.32.0/20 4777 2516 4565 18566
  64.81.48.0/20        4777 2516 4565 18566 - Withdrawn - aggregated with 64.81.32.0/20 (4777 2516 4565 18566)
  64.81.48.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.49.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.50.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.51.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.52.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.53.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.54.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.55.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.56.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.57.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.58.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.59.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.60.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.61.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.48.0/20 4777 2516 4565 18566
  64.81.64.0/19        4777 2516 4565 18566 + Announce - aggregate of 64.81.64.0/20 (4777 2516 4565 18566) and 64.81.80.0/20 (4777 2516 4565
  64.81.64.0/20        4777 2516 4565 18566 - Withdrawn - aggregated with 64.81.80.0/20 (4777 2516 4565 18566)
  64.81.64.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.64.0/20 4777 2516 4565 18566
  64.81.65.0/24        4777 2516 4565 18566 - Withdrawn - matching aggregate 64.81.64.0/20 4777 2516 4565 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
  - bgpupdates.potaroo.net/instability/bgpupd.html

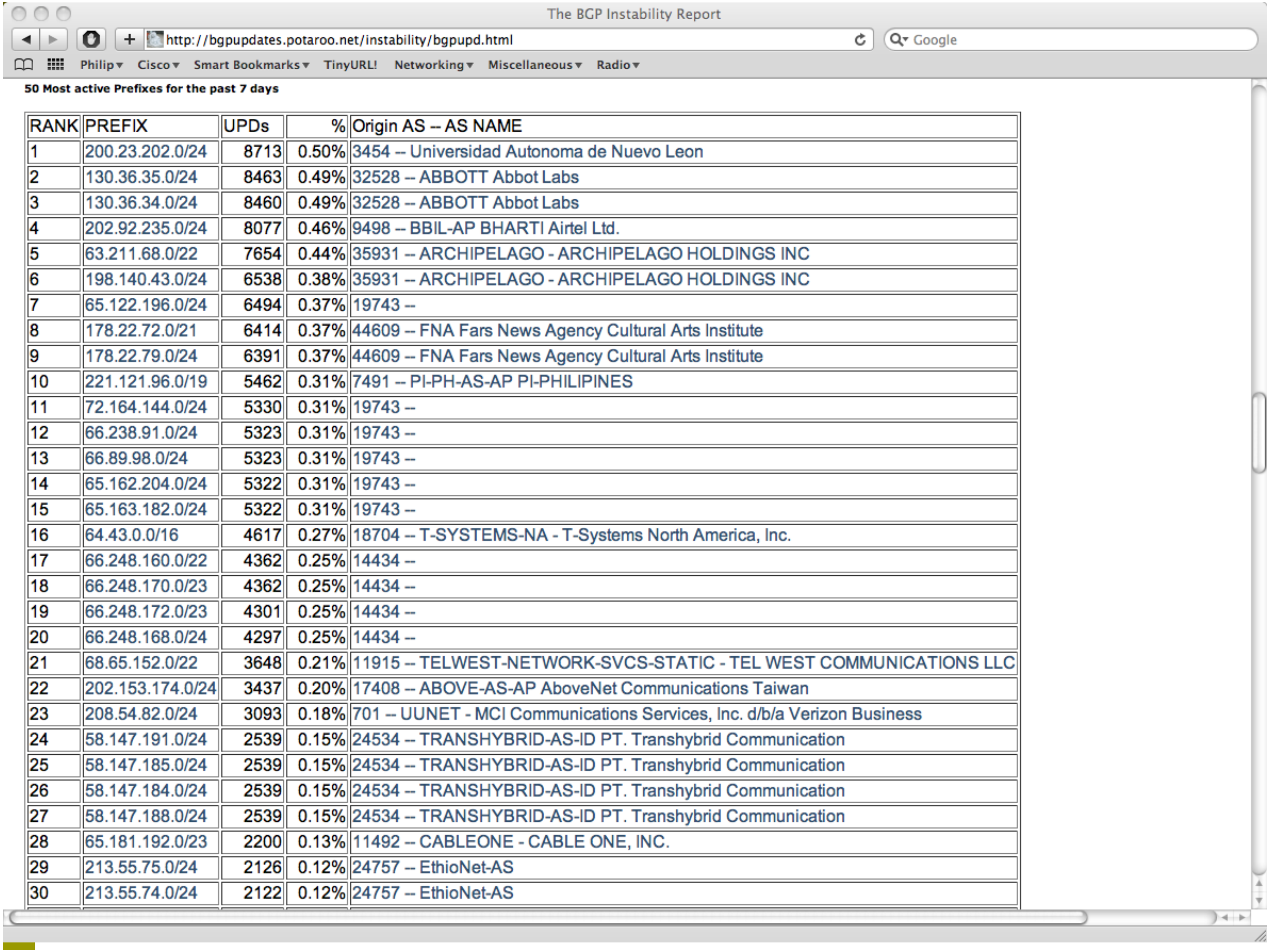

The BGP Instability Report
http://bgpupdates.potaroo.net/instability/bgpupd.html
Google
Philip
Cisco
Smart Bookmarks
TinyURL!
Networking
Miscellaneous
Radio

# The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 09 May 2011 06:14 (UTC+1000)

**50 Most active ASes for the past 7 days**

| RANK | ASN | UPDs | % | Prefixes | UPDs/Prefix | AS NAME |
|---|---|---|---|---|---|---|
| 1 | 9829 | 42709 | 2.60% | 1039 | 41.11 | BSNL-NIB National Internet Backbone |
| 2 | 19743 | 33117 | 2.01% | 7 | 4731.00 | |
| 3 | 17974 | 26180 | 1.59% | 1834 | 14.27 | TELKOMNET-AS2-AP PT Telekomunikasi Indonesia |
| 4 | 14434 | 17436 | 1.06% | 68 | 256.41 | |
| 5 | 32528 | 16930 | 1.03% | 8 | 2116.25 | ABBOTT Abbot Labs |
| 6 | 21826 | 16840 | 1.02% | 306 | 55.03 | Internet Cable Plus C. A. |
| 7 | 24560 | 15062 | 0.92% | 1169 | 12.88 | AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services |
| 8 | 35819 | 14744 | 0.90% | 411 | 35.87 | MOBILY-AS Etihad Etisalat Company (Mobily) |
| 9 | 35931 | 14220 | 0.86% | 6 | 2370.00 | ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC |
| 10 | 4274 | 13591 | 0.83% | 81 | 167.79 | ERX-AU-NET Assumption University |
| 11 | 9299 | 13112 | 0.80% | 1316 | 9.96 | IPG-AS-AP Philippine Long Distance Telephone Company |
| 12 | 44609 | 12813 | 0.78% | 3 | 4271.00 | FNA Fars News Agency Cultural Arts Institute |
| 13 | 14420 | 12143 | 0.74% | 667 | 18.21 | CORPORACION NACIONAL DE TELECOMUNICACIONES - CNT EP |
| 14 | 6458 | 12022 | 0.73% | 302 | 39.81 | Telgua |
| 15 | 33475 | 11795 | 0.72% | 215 | 54.86 | RSN-1 - RockSolid Network, Inc. |
| 16 | 11492 | 11776 | 0.72% | 1268 | 9.29 | CABLEONE - CABLE ONE, INC. |
| 17 | 1660 | 11317 | 0.69% | 79 | 143.25 | ANS-CORP-NY - ANS Communications |
| 18 | 8151 | 10312 | 0.63% | 1383 | 7.46 | Uninet S.A. de C.V. |
| 19 | 24534 | 10156 | 0.62% | 4 | 2539.00 | TRANSHYBRID-AS-ID PT. Transhybrid Communication |
| 20 | 45595 | 9832 | 0.60% | 362 | 27.16 | PKTELECOM-AS-PK Pakistan Telecom Company Limited |
| 21 | 27738 | 9761 | 0.59% | 339 | 28.79 | Ecuadortelecom S.A. |
| 22 | 24757 | 9573 | 0.58% | 52 | 184.10 | EthioNet-AS |
| 23 | 9498 | 9544 | 0.58% | 801 | 11.92 | BBIL-AP BHARTI Airtel Ltd. |
| 24 | 3454 | 8975 | 0.55% | 8 | 1121.88 | Universidad Autonoma de Nuevo Leon |
| 25 | 7491 | 7916 | 0.48% | 98 | 80.78 | PI-PH-AS-AP PI-PHILIPINES |


The BGP Instability Report
http://bgpupdates.potaroo.net/instability/bgpupd.html
Google
Philip
Cisco
Smart Bookmarks
TinyURL!
Networking
Miscellaneous
Radio

**50 Most active Prefixes for the past 7 days**

| RANK | PREFIX | UPDs | % | Origin AS -- AS NAME |
|---|---|---|---|---|
| 1 | 200.23.202.0/24 | 8713 | 0.50% | 3454 -- Universidad Autonoma de Nuevo Leon |
| 2 | 130.36.35.0/24 | 8463 | 0.49% | 32528 -- ABBOTT Abbot Labs |
| 3 | 130.36.34.0/24 | 8460 | 0.49% | 32528 -- ABBOTT Abbot Labs |
| 4 | 202.92.235.0/24 | 8077 | 0.46% | 9498 -- BBIL-AP BHARTI Airtel Ltd. |
| 5 | 63.211.68.0/22 | 7654 | 0.44% | 35931 -- ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC |
| 6 | 198.140.43.0/24 | 6538 | 0.38% | 35931 -- ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC |
| 7 | 65.122.196.0/24 | 6494 | 0.37% | 19743 -- |
| 8 | 178.22.72.0/21 | 6414 | 0.37% | 44609 -- FNA Fars News Agency Cultural Arts Institute |
| 9 | 178.22.79.0/24 | 6391 | 0.37% | 44609 -- FNA Fars News Agency Cultural Arts Institute |
| 10 | 221.121.96.0/19 | 5462 | 0.31% | 7491 -- PI-PH-AS-AP PI-PHILIPINES |
| 11 | 72.164.144.0/24 | 5330 | 0.31% | 19743 -- |
| 12 | 66.238.91.0/24 | 5323 | 0.31% | 19743 -- |
| 13 | 66.89.98.0/24 | 5323 | 0.31% | 19743 -- |
| 14 | 65.162.204.0/24 | 5322 | 0.31% | 19743 -- |
| 15 | 65.163.182.0/24 | 5322 | 0.31% | 19743 -- |
| 16 | 64.43.0.0/16 | 4617 | 0.27% | 18704 -- T-SYSTEMS-NA - T-Systems North America, Inc. |
| 17 | 66.248.160.0/22 | 4362 | 0.25% | 14434 -- |
| 18 | 66.248.170.0/23 | 4362 | 0.25% | 14434 -- |
| 19 | 66.248.172.0/23 | 4301 | 0.25% | 14434 -- |
| 20 | 66.248.168.0/24 | 4297 | 0.25% | 14434 -- |
| 21 | 68.65.152.0/22 | 3648 | 0.21% | 11915 -- TELWEST-NETWORK-SVCS-STATIC - TEL WEST COMMUNICATIONS LLC |
| 22 | 202.153.174.0/24 | 3437 | 0.20% | 17408 -- ABOVE-AS-AP AboveNet Communications Taiwan |
| 23 | 208.54.82.0/24 | 3093 | 0.18% | 701 -- UUNET - MCI Communications Services, Inc. d/b/a Verizon Business |
| 24 | 58.147.191.0/24 | 2539 | 0.15% | 24534 -- TRANSHYBRID-AS-ID PT. Transhybrid Communication |
| 25 | 58.147.185.0/24 | 2539 | 0.15% | 24534 -- TRANSHYBRID-AS-ID PT. Transhybrid Communication |
| 26 | 58.147.184.0/24 | 2539 | 0.15% | 24534 -- TRANSHYBRID-AS-ID PT. Transhybrid Communication |
| 27 | 58.147.188.0/24 | 2539 | 0.15% | 24534 -- TRANSHYBRID-AS-ID PT. Transhybrid Communication |
| 28 | 65.181.192.0/23 | 2200 | 0.13% | 11492 -- CABLEONE - CABLE ONE, INC. |
| 29 | 213.55.75.0/24 | 2126 | 0.12% | 24757 -- EthioNet-AS |
| 30 | 213.55.74.0/24 | 2122 | 0.12% | 24757 -- EthioNet-AS |

# 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:        202.12.28.0 - 202.12.29.255
netname:        APNIC-AP
descr:          Asia Pacific Network Information Centre
descr:          Regional Internet Registry for the Asia-Pacific
descr:          6 Cordelia Street
descr:          South Brisbane, QLD 4101
descr:          Australia
country:        AU
admin-c:        AIC1-AP
tech-c:         NO4-AP
mnt-by:         APNIC-HM
mnt-irt:        IRT-APNIC-AP
changed:        hm-changed@apnic.net
status:         ASSIGNED PORTABLE
changed:        hm-changed@apnic.net 20110309
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 193.128.0.0
inetnum:        193.128.0.0 - 193.133.255.255
netname:        UK-PIPEX-193-128-133
descr:          Verizon UK Limited
country:        GB
org:            ORG-UA24-RIPE
admin-c:        WERT1-RIPE
tech-c:         UPHM1-RIPE
status:         ALLOCATED UNSPECIFIED
remarks:        Please send abuse notification to abuse@uk.uu.net
mnt-by:         RIPE-NCC-HM-MNT
mnt-lower:      AS1849-MNT
mnt-routes:     AS1849-MNT
mnt-routes:     WCOM-EMEA-RICE-MNT
mnt-irt:        IRT-MCI-GB
source:         RIPE # Filtered
```

**ALLOCATED – means that this is Provider Aggregatable address space and can only be announced by the ISP holding the allocation (in this case Verizon UK)**

# 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
  - Traffic Engineering – see BGP Multihoming presentations
- 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 default (unless you need it)
  - Don’t accept your own prefixes
- For IPv4:
  - Don’t accept private (RFC1918) and certain special use prefixes:

    http://www.rfc-editor.org/rfc/rfc5735.txt
  - Don’t accept prefixes longer than /24 (?)
- For IPv6:
  - Don’t accept certain special use prefixes:

    http://www.rfc-editor.org/rfc/rfc5156.txt
  - Don’t accept prefixes longer than /48 (?)

# Receiving Prefixes: From Upstream/Transit Provider

- Check Team Cymru's list of "bogons"
  www.team-cymru.org/Services/Bogons/http.html
- For IPv4 also consult:
  datatracker.ietf.org/doc/draft-vegoda-no-more-unallocated-slash8s
- For IPv6 also consult:
  www.space.net/~gert/RIPE/ipv6-filters.html
- Bogon Route Server:
  www.team-cymru.org/Services/Bogons/routeserver.html
  - Supplies a BGP feed (IPv4 and/or IPv6) of address blocks which should not appear in the BGP table

# Receiving IPv4 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              ! 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   ! Auto-config
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     ! TEST1
ip prefix-list in-filter deny 192.168.0.0/16 le 32
ip prefix-list in-filter deny 198.18.0.0/15 le 32    ! Benchmarking
ip prefix-list in-filter deny 198.51.100.0/24 le 32  ! TEST2
ip prefix-list in-filter deny 203.0.113.0/24 le 32   ! TEST3
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 IPv6 Prefixes

```
router bgp 100
 network 2020:3030::/32
 neighbor 2020:3030::1 remote-as 101
 neighbor 2020:3030::1 prefix-list v6in-filter in
!
ipv6 prefix-list v6in-filter deny ::/0                       ! Default
ipv6 prefix-list v6in-filter deny ::/8 le 128
ipv6 prefix-list v6in-filter permit 2001::/32                         ! Teredo
ipv6 prefix-list v6in-filter deny 2001::/32 le 128
ipv6 prefix-list v6in-filter deny 2001:db8::/32 le 128                ! Documentation
ipv6 prefix-list v6in-filter permit 2002::/16                         ! 6to4
ipv6 prefix-list v6in-filter deny 2002::/16 le 128
ipv6 prefix-list v6in-filter deny 2020:3030::/32 le 128               ! Local Prefix
ipv6 prefix-list v6in-filter deny 3ffe::/16 le 128           ! Old 6bone
ipv6 prefix-list v6in-filter deny fc00::/7 le 128            ! Unique Local
ipv6 prefix-list v6in-filter deny fe80::/10 le 128           ! Link Local
ipv6 prefix-list v6in-filter deny ff00::/8 le 128            ! Multicast
ipv6 prefix-list v6in-filter permit 2000::/3 le 48           ! Global Unicast Block
ipv6 prefix-list v6in-filter deny ::/0 le 128
```

# 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)

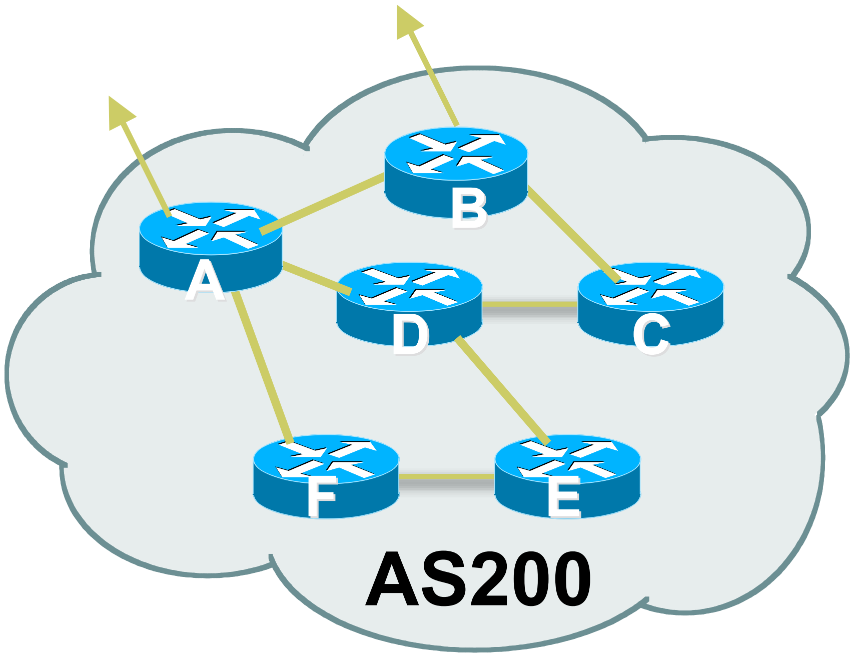
- Be prudent deploying IGP – keep the Link State Database Lean!
  - Router loopbacks go in IGP
  - WAN point to point links go in IGP
  - (In fact, any link where IGP dynamic routing will be run should go into IGP)
  - Summarise on area/level boundaries (if possible) – i.e. think about your IGP address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

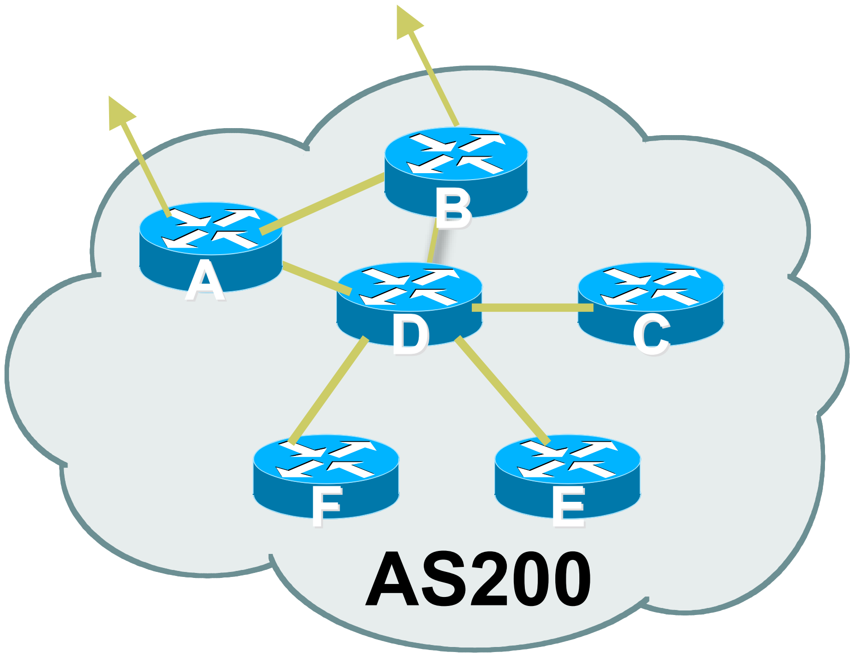
# Preparing the Network
# Second Step: iBGP

- 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

AfNOG 2011 AR-E Workshop