BGP Best Practices

Scalable Infrastructure
Workshop

AfNOG 2010

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

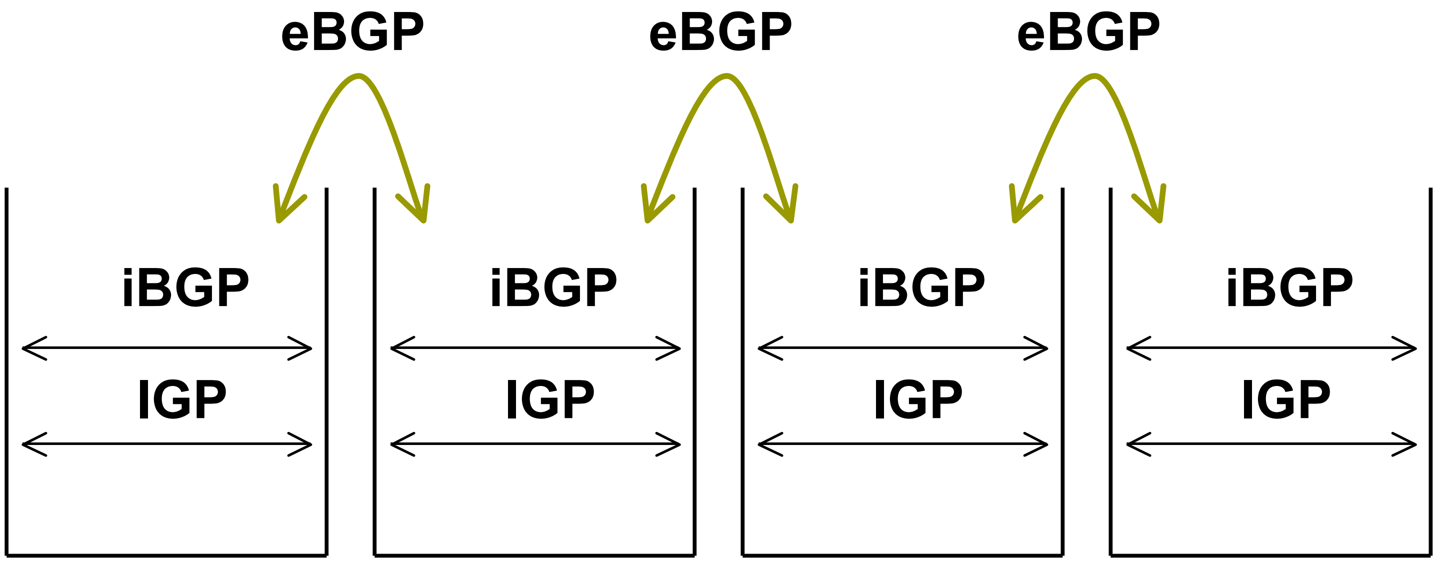
- Internal Routing Protocols (IGPs)
  - examples are ISIS and OSPF
  - used for carrying infrastructure addresses
  - NOT used for carrying Internet prefixes or customer prefixes
  - design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

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/IGP model used in ISP networks

- Model representation



BGP versus OSPF/ISIS

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

Aggregation

Quality, not Quantity!

Aggregation

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

Configuring Aggregation: Cisco IOS

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

```
router bgp 100
 network 101.10.0.0 mask 255.255.224.0
ip route 101.10.0.0 255.255.224.0 null0
```

- The static route is a "pull up" route
  - more specific prefixes within this address block ensure connectivity to ISP's customers
  - "longest match lookup"

Aggregation

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

Announcing Aggregate: Cisco IOS

- Configuration Example

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

Announcing an Aggregate

- ISPs who don’t and won’t aggregate are held in poor regard by community
- Registries’ minimum allocation size is now at least a /21 or /22
  - no real reason to see anything much longer than a /22 prefix in the Internet
  - BUT there are currently ~168000 /24s!

The Internet during AfNOG 2009 (April 2009)

- Internet Routing Table Statistics

BGP Routing Table Entries	288336
Prefixes after maximum aggregation	136251
Unique prefixes in Internet	140888
Prefixes smaller than registry alloc	142536
/24s announced	150651
only 5797 /24s are from 192.0.0.0/8	
ASes in use	31224

The Internet Today (May 2010)

- Current Internet Routing Table Statistics

BGP Routing Table Entries	321324
Prefixes after maximum aggregation	147948
Unique prefixes in Internet	155831
Prefixes smaller than registry alloc	154125
/24s announced	168259
only 5730 /24s are from 192.0.0.0/8	
ASes in use	33989

Efforts to Improve Aggregation: The CIDR Report

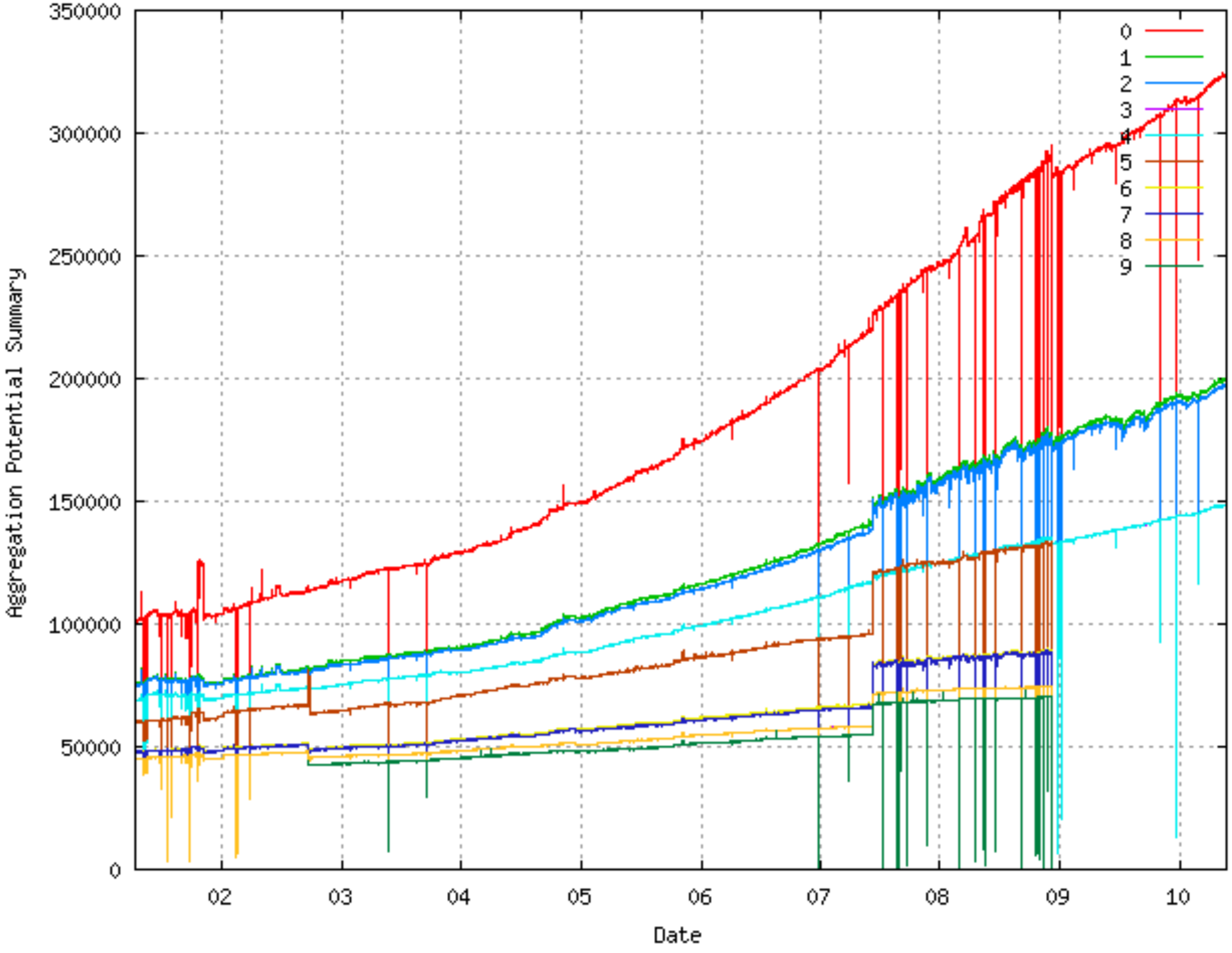
- Initiated and operated for many years by Tony Bates
- Now combined with Geoff Huston’s routing analysis

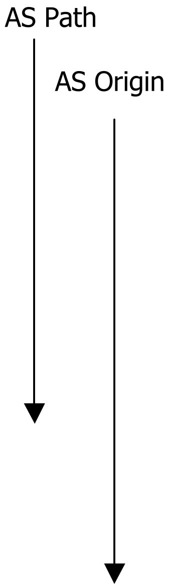
  www.cidr-report.org
- Results e-mailed on a weekly basis to most operations lists around the world
- Lists the top 30 service providers who could do better at aggregating

Efforts to Improve Aggregation: The CIDR Report

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

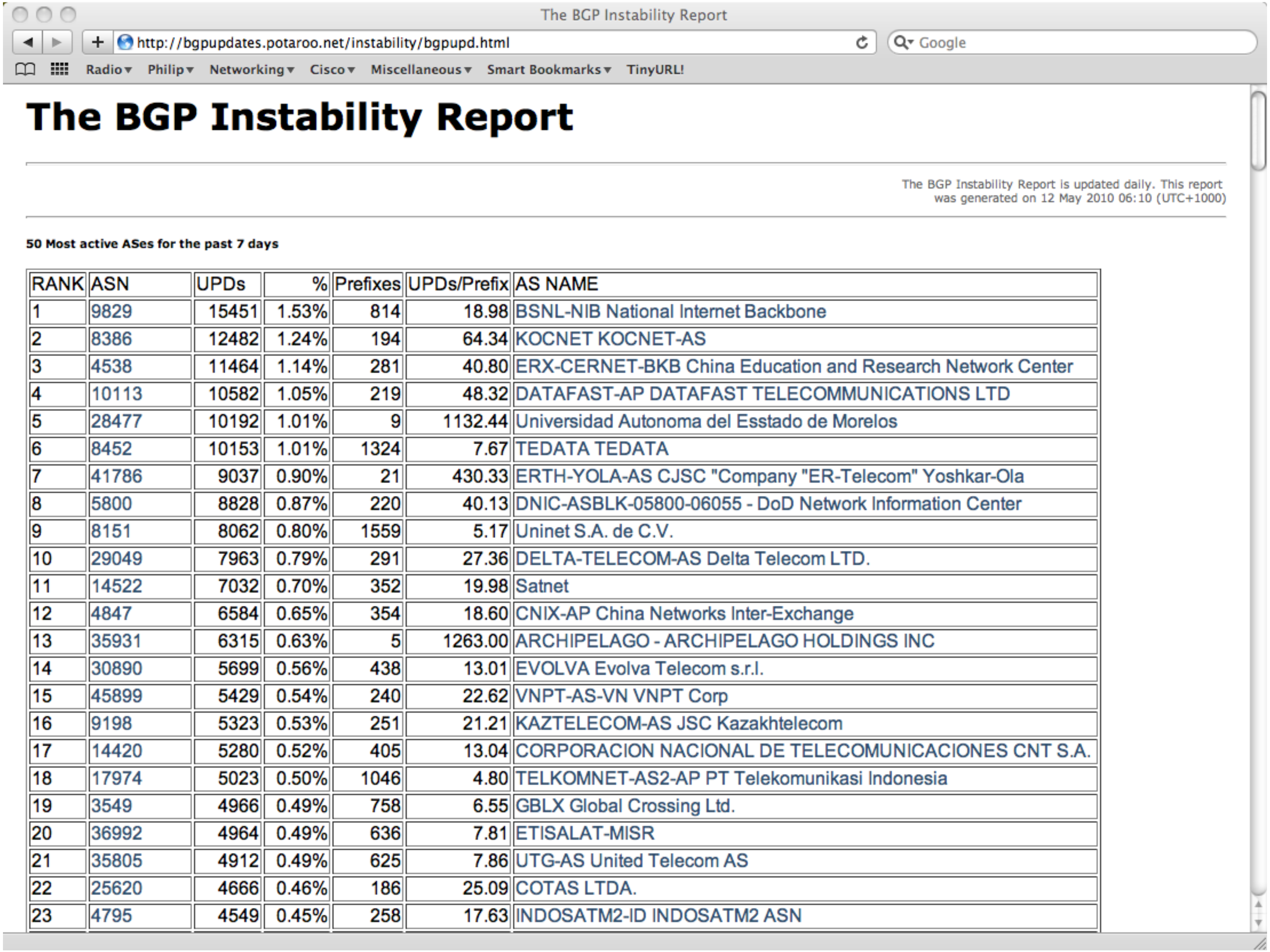
Aggregation Potential





Importance of Aggregation

- Size of routing table
  - Memory is no longer the problem
  - 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

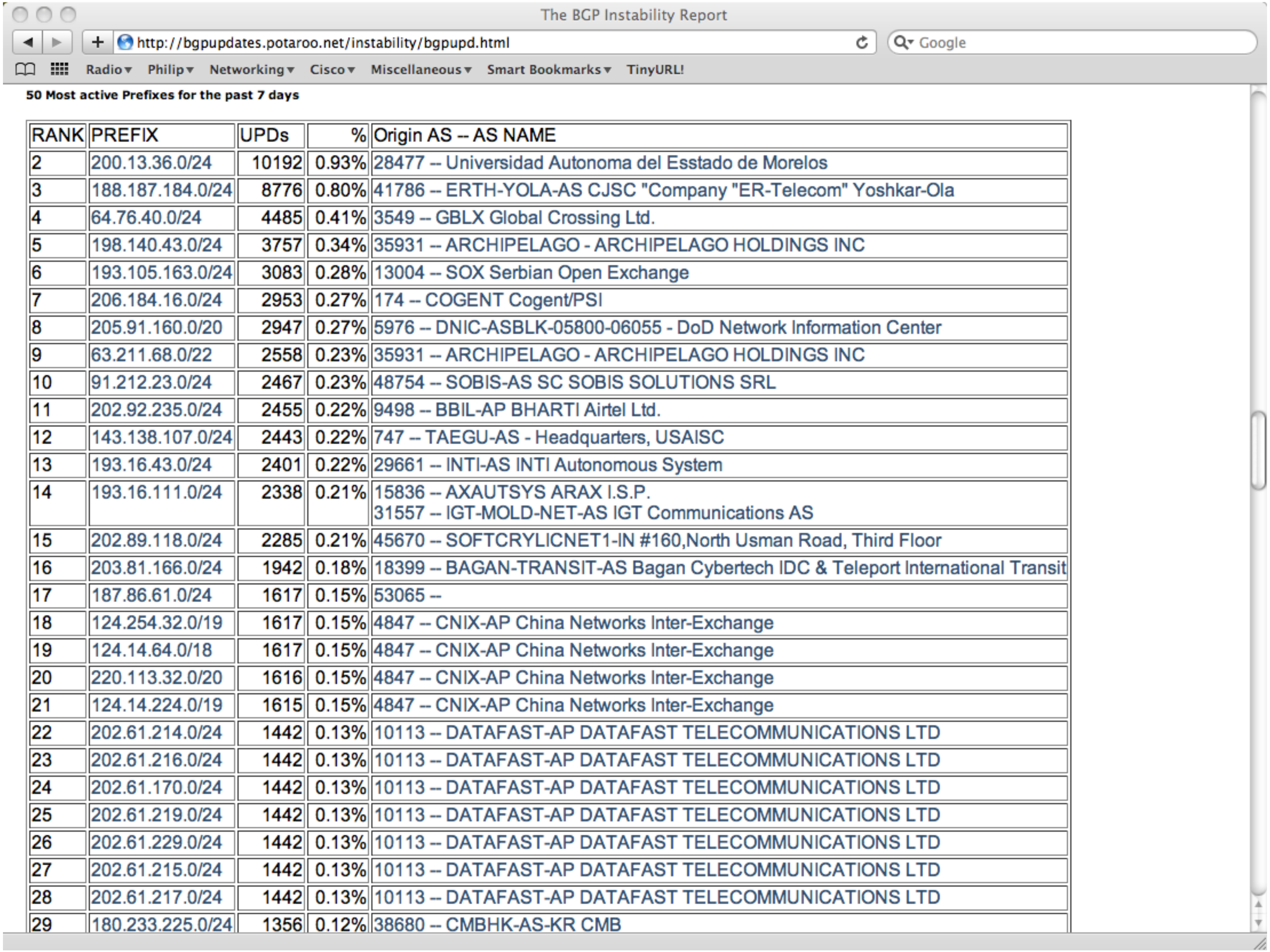

The BGP Instability Report
http://bgpupdates.potaroo.net/instability/bgpupd.html
Google
Radio
Philip
Networking
Cisco
Miscellaneous
Smart Bookmarks
TinyURL!

The BGP Instability Report

The BGP Instability Report is updated daily. This report was generated on 12 May 2010 06:10 (UTC+1000)

50 Most active ASes for the past 7 days

RANK	ASN	UPDs	%	Prefixes	UPDs/Prefix	AS NAME
1	9829	15451	1.53%	814	18.98	BSNL-NIB National Internet Backbone
2	8386	12482	1.24%	194	64.34	KOCNET KOCNET-AS
3	4538	11464	1.14%	281	40.80	ERX-CERNET-BKB China Education and Research Network Center
4	10113	10582	1.05%	219	48.32	DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
5	28477	10192	1.01%	9	1132.44	Universidad Autonoma del Esstado de Morelos
6	8452	10153	1.01%	1324	7.67	TEDATA TEDATA
7	41786	9037	0.90%	21	430.33	ERTH-YOLA-AS CJSC "Company "ER-Telecom" Yoshkar-Ola
8	5800	8828	0.87%	220	40.13	DNIC-ASBLK-05800-06055 - DoD Network Information Center
9	8151	8062	0.80%	1559	5.17	Uninet S.A. de C.V.
10	29049	7963	0.79%	291	27.36	DELTA-TELECOM-AS Delta Telecom LTD.
11	14522	7032	0.70%	352	19.98	Satnet
12	4847	6584	0.65%	354	18.60	CNIX-AP China Networks Inter-Exchange
13	35931	6315	0.63%	5	1263.00	ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
14	30890	5699	0.56%	438	13.01	EVOLVA Evolva Telecom s.r.l.
15	45899	5429	0.54%	240	22.62	VNPT-AS-VN VNPT Corp
16	9198	5323	0.53%	251	21.21	KAZTELECOM-AS JSC Kazakhtelecom
17	14420	5280	0.52%	405	13.04	CORPORACION NACIONAL DE TELECOMUNICACIONES CNT S.A.
18	17974	5023	0.50%	1046	4.80	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia
19	3549	4966	0.49%	758	6.55	GBLX Global Crossing Ltd.
20	36992	4964	0.49%	636	7.81	ETISALAT-MISR
21	35805	4912	0.49%	625	7.86	UTG-AS United Telecom AS
22	25620	4666	0.46%	186	25.09	COTAS LTDA.
23	4795	4549	0.45%	258	17.63	INDOSATM2-ID INDOSATM2 ASN


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
2	200.13.36.0/24	10192	0.93%	28477 -- Universidad Autonoma del Esstado de Morelos
3	188.187.184.0/24	8776	0.80%	41786 -- ERTH-YOLA-AS CJSC "Company "ER-Telecom" Yoshkar-Ola
4	64.76.40.0/24	4485	0.41%	3549 -- GBLX Global Crossing Ltd.
5	198.140.43.0/24	3757	0.34%	35931 -- ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
6	193.105.163.0/24	3083	0.28%	13004 -- SOX Serbian Open Exchange
7	206.184.16.0/24	2953	0.27%	174 -- COGENT Cogent/PSI
8	205.91.160.0/20	2947	0.27%	5976 -- DNIC-ASBLK-05800-06055 - DoD Network Information Center
9	63.211.68.0/22	2558	0.23%	35931 -- ARCHIPELAGO - ARCHIPELAGO HOLDINGS INC
10	91.212.23.0/24	2467	0.23%	48754 -- SOBIS-AS SC SOBIS SOLUTIONS SRL
11	202.92.235.0/24	2455	0.22%	9498 -- BBIL-AP BHARTI Airtel Ltd.
12	143.138.107.0/24	2443	0.22%	747 -- TAEGU-AS - Headquarters, USAISC
13	193.16.43.0/24	2401	0.22%	29661 -- INTI-AS INTI Autonomous System
14	193.16.111.0/24	2338	0.21%	15836 -- AXAUTSYS ARAX I.S.P. 31557 -- IGT-MOLD-NET-AS IGT Communications AS
15	202.89.118.0/24	2285	0.21%	45670 -- SOFTCRYLICNET1-IN #160,North Usman Road, Third Floor
16	203.81.166.0/24	1942	0.18%	18399 -- BAGAN-TRANSIT-AS Bagan Cybertech IDC & Teleport International Transit
17	187.86.61.0/24	1617	0.15%	53065 --
18	124.254.32.0/19	1617	0.15%	4847 -- CNIX-AP China Networks Inter-Exchange
19	124.14.64.0/18	1617	0.15%	4847 -- CNIX-AP China Networks Inter-Exchange
20	220.113.32.0/20	1616	0.15%	4847 -- CNIX-AP China Networks Inter-Exchange
21	124.14.224.0/19	1615	0.15%	4847 -- CNIX-AP China Networks Inter-Exchange
22	202.61.214.0/24	1442	0.13%	10113 -- DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
23	202.61.216.0/24	1442	0.13%	10113 -- DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
24	202.61.170.0/24	1442	0.13%	10113 -- DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
25	202.61.219.0/24	1442	0.13%	10113 -- DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
26	202.61.229.0/24	1442	0.13%	10113 -- DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
27	202.61.215.0/24	1442	0.13%	10113 -- DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
28	202.61.217.0/24	1442	0.13%	10113 -- DATAFAST-AP DATAFAST TELECOMMUNICATIONS LTD
29	180.233.225.0/24	1356	0.12%	38680 -- CMBHK-AS-KR CMB

Aggregation: Summary

- Aggregation on the Internet could be MUCH better
  - 35% saving on Internet routing table size is quite feasible
  - Tools are available
  - Commands on the router are not hard
  - CIDR-Report webpage
- RIPE Routing WG aggregation recommendation
  - RIPE-399 — www.ripe.net/docs/ripe-399.html

Receiving Prefixes

Receiving Prefixes from downstream peers

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

Receiving Prefixes: Cisco IOS

- Configuration Example on upstream

```
router bgp 100
 neighbor 102.102.10.1 remote-as 101
 neighbor 102.102.10.1 prefix-list customer in
!
ip prefix-list customer permit 100.50.0.0/20
ip prefix-list customer deny 0.0.0.0/0 le 32
```

Receiving Prefixes from upstream peers

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

Receiving Prefixes from upstream peers

- Downstream Router Configuration

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

Receiving Prefixes from upstream peers

- Upstream Router Configuration

```
router bgp 101
 neighbor 101.5.7.2 remote-as 100
 neighbor 101.5.7.2 default-originate
 neighbor 101.5.7.2 prefix-list cust-in in
 neighbor 101.5.7.2 prefix-list cust-out out
!
ip prefix-list cust-in permit 101.10.0.0/19
ip prefix-list cust-in deny 0.0.0.0/0 le 32
!
ip prefix-list cust-out permit 0.0.0.0/0
ip prefix-list cust-out deny 0.0.0.0/0 le 32
```

Receiving Prefixes from upstream peers

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

Receiving Prefixes

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

Generic ISP BGP prefix filter

- This prefix-list MUST be applied to all external BGP peerings, in and out!
- RFC5735 lists many special use addresses
- Check Team Cymru's bogon pages
  - http://www.cymru.com/Bogons
  - http://www.cymru.com/BGP/bogon-rs.html – bogon route server

Prefixes into iBGP

Injecting prefixes into iBGP

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

Router configuration: network statement

- Example:

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

Injecting prefixes into iBGP

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

Router Configuration: redistribute static

- Example:

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

Injecting prefixes into iBGP

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

Configuration Tips

Templates

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

BGP Template – iBGP peers

iBGP Peer Group
AS100

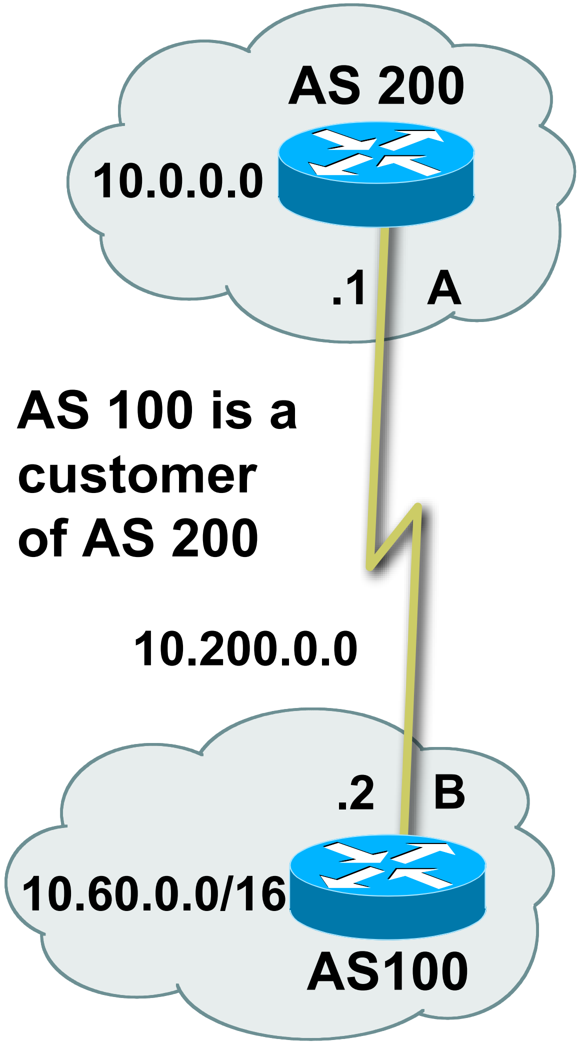
```
router bgp 100
neighbor internal peer-group
neighbor internal description ibgp peers
neighbor internal remote-as 100
neighbor internal update-source Loopback0
neighbor internal next-hop-self
neighbor internal send-community
neighbor internal version 4
neighbor internal password 7 03085A09
neighbor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

BGP Template – iBGP peers

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

BGP Template – eBGP peers

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



AS 100 is a customer of AS 200

BGP Template – eBGP peers

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

More BGP “defaults”

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

Configuration Tips Summary

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

Summary – BGP BCP

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