# BGP

BGP protocol
iBGP configuration
BGP Misconfiguration
BGP convergence



## References


1. BGP tutorial - BPG4 case studies by Sam Halabi
2. BGP routing policies in ISP networks by Mathew Caesar and Jennifer Rexford

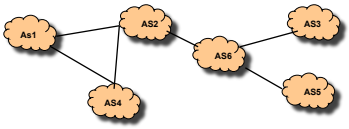



## Autonomous System

- Definition: internet is network of networks glued by IP
- Within a network (intra-domain) any routing policy can be chosen
- A common routing policy is needed when routing between networks or domains
- A Domain is a network that has unified administrative routing policy
- Autonomous System (domain) or AS: Has a number assigned to it and provides routing information to other ASes



## Internet structure



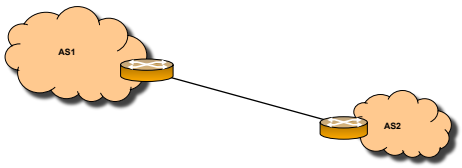

- AS provide reachability information to other ASes
- Within AS, local routing protocols used (optimize path metric)
- Inter-AS concerned with reachability and policy implementation
  - Usually $$ involved with relationships

## Autonomous system

- The actual entity that participates in interdomain routing
- Has a unique 16 bit number assigned
- Examples:
  - RUTGERS: 46, STANFORD;32,MIT: 3, CMU: 9
  - AT&T: 6431, …
  - Quest: 209, …
  - Sprint: 1239, …
- How do ASes interconnect to provide global connectivity?
- How does routing information get exchanged?
- How is policy  specified and implemented?

5

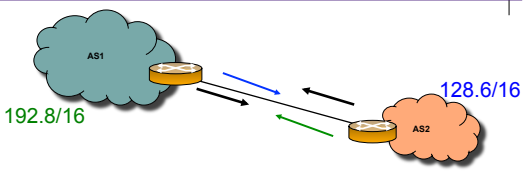
## Internet routing




- Intra domain
  - OSPF, RIP
  - Route on IP addresses
  - Path metrics
- Inter domain
  - BGP
  - Route on AS numbers
  - Policy and business relations based

6

## BGP: basic idea




- AS1 needs to inform AS2 that it can route to 192.8/16 and AS2 needs to inform AS1 that it can route to 128.16/16
- After this, what else
  - Route updates/changes
  - Policy: what is AS1 does not want to route to anyone else but its own domain?
  - What paths should be preferred?
- This is essentially BGP

7

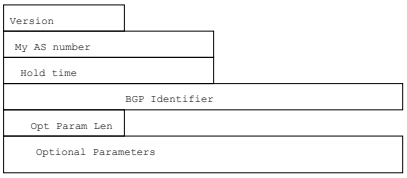
## BGP protocol

- BGP uses TCP as its transport protocol, on port 179. On connection start, BGP peers exchange complete copies of their routing tables, which can be quite large. However, only changes (deltas) are then exchanged, which makes long running BGP sessions more efficient than shorter ones.
- Four Basic messages:
  - *Open*:
    Establishes BGP session (uses TCP port #179)
  - *Notification*:
    Report unusual conditions
  - *Update*:
    Inform neighbor of new routes that become active
    Inform neighbor of old routes that become inactive
  - *Keepalive*:
    Inform neighbor that connection is still viable

8

## OPEN Message

- During session establishment, two BGP speakers exchange their
  - AS numbers
  - BGP identifiers (usually one of the router's IP addresses)
  - Select hold timer : max time before declaring peer is down
- A BGP speaker has option to refuse a session
- authentication information (optional)




9

## NOTIFICATION and KEEPALIVE Messages

- NOTIFICATION
  - Indicates an error
  - terminates the TCP session
  - gives receiver an indication of why BGP session terminated
  - Examples: header errors, hold timer expiry, bad peer AS, bad BGP identifier, malformed attribute list, missing required attribute, AS routing loop, etc.
- KEEPALIVE
  - protocol requires some data to be sent periodically. If no UPDATE to send within the specified time period, then send KEEPALIVE message to assure partner that connection is still alive

10

## UPDATE Message

- used to either advertise and/or withdraw previously announced prefixes
- path attributes: list of attributes that pertain to ALL the prefixes in the Reachability Info field

FORMAT:

| |
|---|
| Withdrawn routes length (2 octets) |
| Withdrawn routes (variable length) |
| Total path attributes length (2 octets) |
| Path Attributes (variable length) |
| (NLRI) Reachability Information (variable length) |

11

## BGP update message

- Withdrawn Routes: Length field 2 Bytes
- Withdrawn route list
- Path attributes: Length field 2 bytes
- Path attributes list
- NLRI list : a list of entries
  - Length field (1 byte), Prefix (variable length)
- Path attributes apply to all the prefixes in the NLRI list

12

## Advertising a prefix

- When a router advertises a prefix to one of its BGP neighbors:
  - information is valid until first router explicitly advertises that the information is no longer valid
  - BGP does not require routing information to be refreshed
  - if node A advertises a path for a prefix to node B, then node B can be sure node A is using that path itself to reach the destination.

13

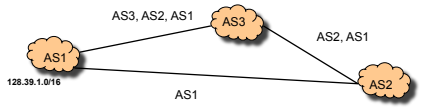
## BGP attributes

- BGP protocol announcements carries with it several attributes
- Attribute describes characteristics of a prefix
- BGP chooses a single path for a given prefix based on attributes (can choose to ignore!)
- BGP always announces the best path to neighbors
- Attributes
  - 1 ORIGIN
  - 2 AS_PATH
  - 3 NEXT_HOP
  - 4 MED
  - 5 LOCAL_PREF
  - 6 WEIGHT
  - 7 COMMUNITY
  - 8 AGGREGATOR

14

## PATH ATTRIBUTES

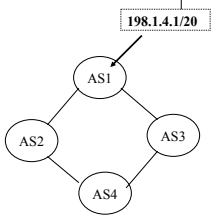
- ORIGIN(TYPE CODE=1):
  - Who originated the announcement? Where was a prefix *injected* into BGP?
  - Manually configured, directly connected, by other intra-routing protocols
  - IGP, EGP, default – incomplete (learnt from some other means)
- AS-PATH (TYPE CODE =2)
  - a list of AS' s through which the announcement for a prefix has passed
  - each AS prepends its AS # to the AS-PATH attribute when forwarding an announcement
  - useful to detect and prevent loops
  - AS length can be used to select among routes unless a LOCAL PREF attribute overrides




15

## Attribute: Local Preference (type code = 5)

- Used to indicate preference among multiple paths for the same prefix *anywhere* in the internet.
- The higher the value the more it is preferred
- Default value is 100
- Local to the AS (non-transitive)
- Often used to select a specific exit point for outbound traffic
- Override influence of AS path length

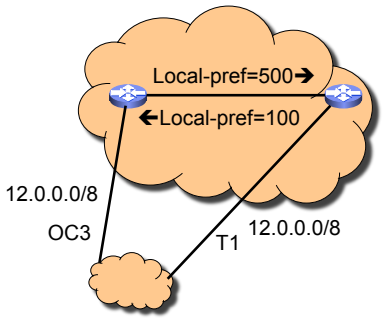



*BGP table at AS4:*

| Destination | AS Path | Local Pref |
|---|---|---|
| 198.1.4/20 | **AS3 AS1** | **300** |
| 198.1.4/20 | **AS2 AS1** | **100** |

16

## Use of local pref




17

## Attribute: NEXT HOP (code=3)

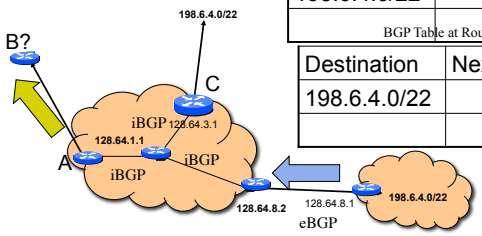
When AS boundaries are Crossed, the next hop field is Changed and replaced with the Address of the border router
**eBGP address of external neighbor**
**iBGP nexthop from eBGP**

BGP Table at Router A:

| Destination | Nexthop |
|---|---|
| 198.6.4.0/22 | 128.64.8.1 |
| | |

BGP Table at Router B:

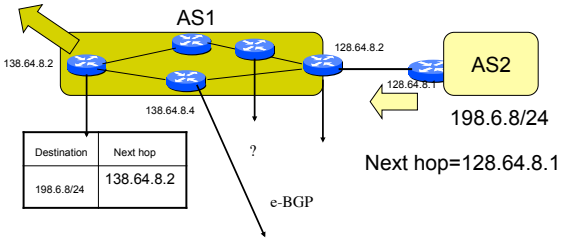
| Destination | Nexthop |
|---|---|
| 198.6.4.0/22 | |
| | |




18

## Use of next hop

Next hop=138.64.8.2 for 198.6.8/24

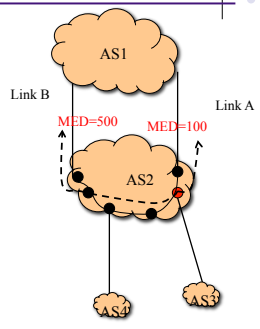



| Destination | Next hop |
|---|---|
| 198.6.8/24 | 138.64.8.2 |

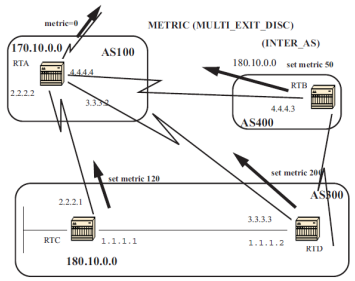
19

## Attribute: Multi-Exit Discriminator (MED) (code=4)

- when AS's interconnected via 2 or more links
- AS path length are same
- AS announcing prefix, sets MED value
- enables AS2 to indicate its preference (lower MED is better)
- AS receiving prefix uses MED to select link
- a way to specify how close a prefix is to the link it is announced on




20

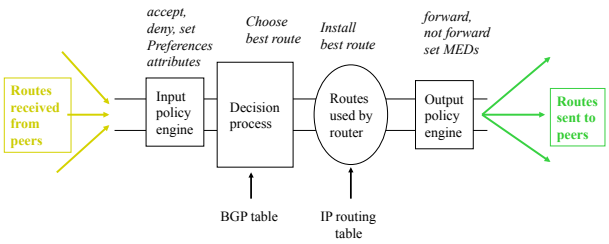



- Med values from the same AS are compared
- A lower MED value is preferred
- MED values exchanged between ASs- non-transitive

Ref: BGP Tutorial – Sam Halabi- Cisco



## Routing Process Overview






## BGP Decision Process: Path Selection on a Router

- Routing Information Base
  - Store all BGP routes for each destination prefix
  - Withdrawal message: remove the route entry
  - Announcement message: update the route entry
- Selecting the best route
  - Consider all BGP routes for the prefix
  - Apply rules for comparing the routes
  - Select the one best route
    - Use this route in the forwarding table
    - Send this route to neighbors



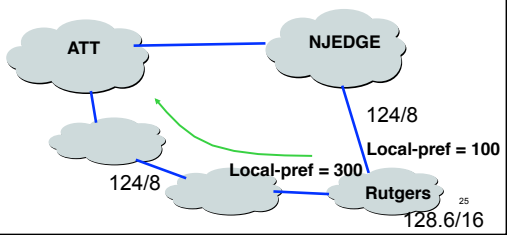
## BGP Decision Process

1. Choose route with highest LOCAL-PREF
2. If have more than 1 route, select route with shortest AS-PATH
3. If have more than 1 route, select according to lowest ORIGIN type where *I-BGP < E-BGP < default*
4. If have more than 1 route, select route with lowest MED value
5. Select e-BGP learned over i-BGP learned path
6. Select min cost path to NEXT HOP using IGP metrics (lowest IGP cost to BGP egress)
7. If have multiple internal paths, use lowest BGP Router ID to break the tie.

- See: BGP routing policies in ISP networks by Caesar and Rexford

## Import Policy: Local Preference

- Favor one path over another based on local policy
  - Override the influence of AS path length
  - Local admin policy given priority
  - Apply local policies to prefer a path




25

## Import Policy: Filtering

- Discard some route announcements
  - Detect configuration mistakes and attacks
- Examples on session to a customer
  - Discard route if prefix not owned by the customer
  - Does not want routing for that prefix via the peer

26

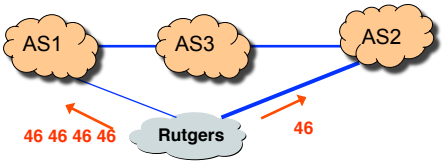
## Export Policy: Filtering

- Discard some route announcements
  - Limit propagation of routing information
- Not forwarding prefixes
  - Do not want others to use you as an intermediary for that prefix

27

## Export Policy: Attribute Manipulation

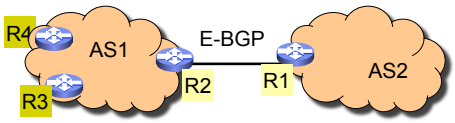
- Modify attributes of the active route
  - To influence the way other ASes behave
- Example: AS_PATH padding
  - Artificially inflate AS path length seen by others
  - Convince some ASes to send traffic another way
  - May not work always: AS2 may have a higher LOCAL_PREFERENCE




28

## Internal vs. External BGP

- Internal-BGP or i-BGP used to distribute routes within AS
- Egress routers use E-BGP or BGP
- R4 and R3 learn routes from R2 using i-BGP
- R1 and R2 talk e-BGP (different AS)
- R2, R4 and R2, R3 and R3, R4 taal i-BGP (same AS)



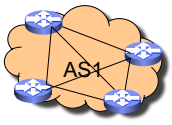

29

## Internal BGP (I-BGP)

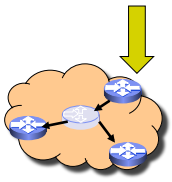
- Same messages as E-BGP
- Different rules about re-advertising prefixes:
  - Rule #1: Prefix learned from E-BGP can be advertised to I-BGP neighbor and vice-versa, but
  - Rule #2: Prefix learned from one I-BGP neighbor cannot be advertised to another I-BGP neighbor
  - Reason: no AS PATH within the same AS and thus danger of looping.
  - Means each I-BGP speaker must be connected directly with every other I-BGP within the same AS
  - Full MESH!!!

30

## Route reflectors




Mesh does not Scale
$O(N^2)$ sessions

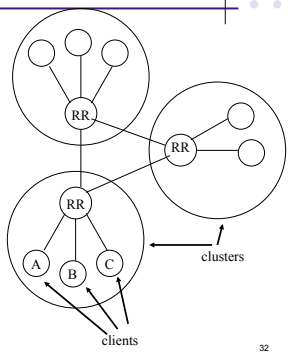


Only N-1 sessions
The RR only advertises best routes

31

## Route Reflectors

- Problem: requiring a full mesh of I-BGP sessions between all pairs of routers is hard to manage for large AS's.
- Solution:
  - group routers into clusters.
  - Assign a leader to each cluster, called a route reflector (RR).
  - Members of a cluster are called clients of the RR
- I-BGP Peering
  - clients peer only with their RR
  - RR's must be fully meshed




32

## Route Reflectors: Rule on Announcements

- If received from RR, reflect to clients
- If received from a client, reflect to RRs and clients
- If received from E-BGP, reflect to all - RRs and clients
- RR's reflect only the best route to a given prefix, not all announcements they receive.
  - helps size of routing table
  - sometimes clients don't need to carry full table

33

## Announcement loop

CISCO manual on BGP configuration

**Caution** Incorrectly setting BGP attributes for a route reflector can cause inconsistent routing, routing loops, or a loss of connectivity. Setting BGP attributes for a route reflector should be attempted only by an experienced network operator.

| Command | Purpose |
|---|---|
| Router(config-router)**# no bgp client-to-client reflection** | Disables client-to-client route reflection. |

RFC 4456- BGP Route Reflectors
When a route is reflected, it is possible through misconfiguration to form route re-distribution loops.

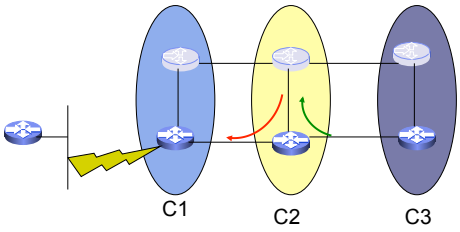
34

## Avoiding Loops with Route Reflectors

- Loops cannot be detected by traditional approach using AS-PATH because AS-PATH not modified within an AS.
- Announcements could leave a cluster and re-enter it.
- Two new attributes introduced:
  - ORIGINATOR_ID: router id of route's originator in AS
    *rule*: announcement discarded if returns to originator
  - CLUSTER_LIST: a sequence of cluster id's. set by RRs.
    *rule*: if an RR receives an update and the cluster list contains its cluster id, then update is discarded.

35

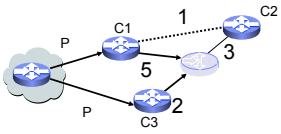
## Announcement loops prevention




36

## Route reflector vs Full mesh

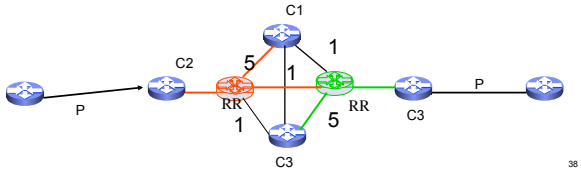
- In a full mesh, every router hears about every egress announcement
- Has complete visibility ; each router picks the shortest IGP path (among all routers announcing a prefix)
- Not so with RRs
- Who does C2 choose as the egress point?




37

## Forwarding loops

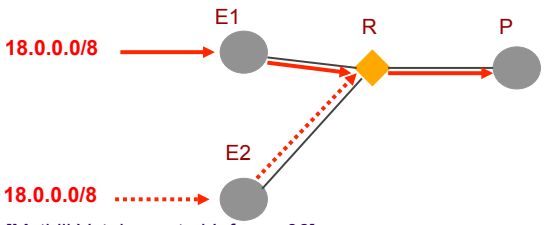
- Each router picks the shortest path among the routes it has heard
- Two different routers can consider each other has the intermediary to the shortest path
- Forwarding loop!!




38

## Key insight for emulating full-mesh

- For every BGP router P, every egress E
  - P and E have iBGP session, OR → if true for all P , what do we have?
  - P should be the client of a route reflector on the shortest path between P and E




39

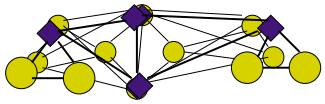
[Mythili Vutukuru et.al Infocom06]

## Can we do better than full mesh?

- Graph separator
- Choose a separator and make them route reflectors
- Connect RRs into a mesh
- Make members of connected components clients of all RRs in the separator
- Connect all members in each connected component into a mesh
- Recursively apply to the components

40

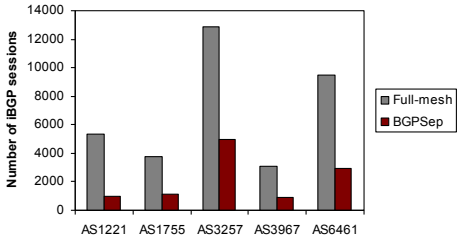
## Separator Algorithm



41

## Evaluation

- 2.5 to 5X fewer iBGP sessions on ISP topologies [Source: Rocketfuel]




42

## BGP misconfiguration

- Each BGP router individually configured
- Script or manually generated
- Access CISCO web site (Google key: BGP configuration CISCO)
- https://www.cisco.com/en/US/docs/ios/12_2/ip/configuration/guide/1cfbgp.html
- You will see in a lot of places, the following
- **Caution** If the authentication string is configured incorrectly, the BGP peering session will not be established. We recommend that you enter the authentication string carefully and verify that the peering session is established after authentication is configured.

Understanding BGP Misconfiguration SIGCOMM 2002 Ratul Mahajan et.al.

43

## BGP Misconfigurations

- Types of BGP misconfigurations
  - Origin misconfiguration
  - Export misconfiguration
- Types of Impact of BGP misconfigurations
- How to identify misconfigurations?
- Results
  - How often?
  - Impact level?
  - Cause?
  - How to avoid them?
- Conclusion

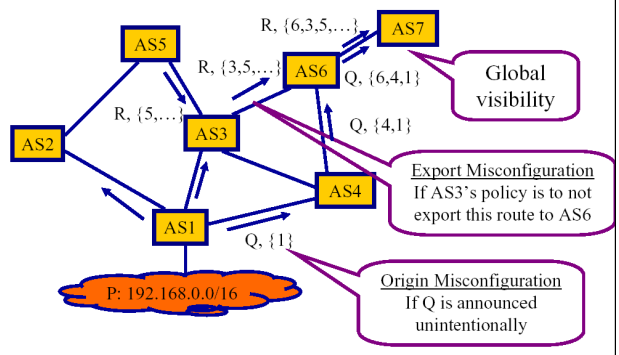
44

## BGP misconfigurations

- Configurations errors that result in unintended BGP route announcements or route suppressions
- defined by local operational practices, *not* global standards
  - No universally accepted list of "Dos & Don'ts"
  - Read CISCO/JUNIPER manual, but then who reads manuals!
- Misconfiguration: behavior unintended by the operator
  - Includes both slips (inadvertent errors) and mistakes (erroneous plan)
- Focus on two broad classes of globally visible faults
  - Origin misconfiguration
    - Announce a more specific prefix than necessary
    - Prefix hijacking
  - Export misconfiguration
    - AS-path in violation of one of the policies

45

## BGP Misconfiguration (2)



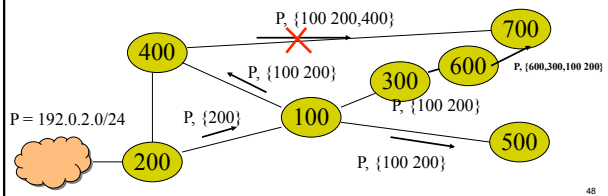

## Other Types of Misconfigurations

- Filter out routes that should be announced
  - Appears to users as failures
- Two links connect to a neighboring AS, misconfigure to use the less-preferred link
- Not easily identifiable from global BGP changes
  - Need information internal to ASes
  - OK to ignore → no significant impact on connectivity

47

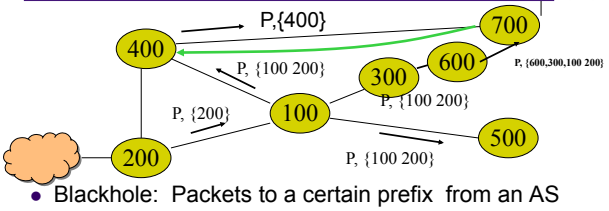
## Impacts of Misconfigurations

- Routing Load: unnecessary updates pressure already-loaded BGP speaking routers
- Connectivity Disruption: partially or globally
- Policy violation: carry unwanted traffic




48

## Prefix hijacking




- Blackhole: Packets to a certain prefix from an AS discarded
- detour: data sent to wrong destinations

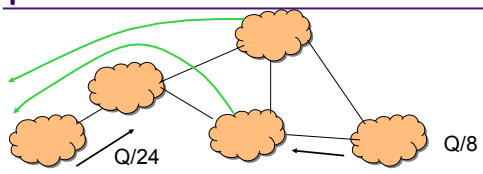


## Detecting prefix hijacking

- The real AS has no way of knowing
- AS does not keep track of all the ASes that it supposed to hear
- Bogus AS may still route to origin AS
  - Only performance degradation
- Diagnosing prefix hijacking need to be done externally
- Traceroute from specific hosts
  - Map ORIGIN IP address and AS number
- Inspect prefix announcements at various BGP routers



## Announcing a more specific prefix



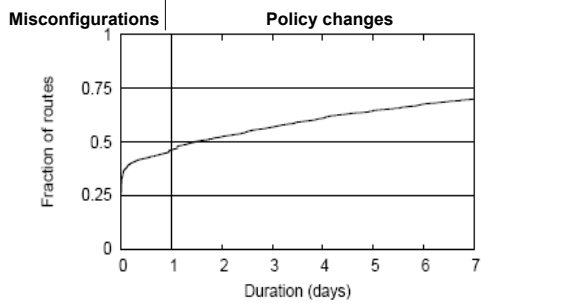

- All ASes will choose bogus route for the more specific prefix
- Traffic for that prefix acc to longest prefix match will be rerouted to BOGUS AS



## Methodology

- Data collected from a special BGP peer
- 23 peers in 19 different ASes
- How to distinguish between legitimate policy changes and misconfigurations
- Look at timescale
- Most policy changes (change to an existing prefix) lasts couple of days
- Changes that lasts < day due to misconfigurations
- Look for change signatures
- Pick signatures with shorter lifetimes

## Lifetime of New Routes




- New route: new prefix or existing prefix with new origin

## Signatures

- Signature 1
  - Short lived new route
  - Misconfiguration
- Signature 2
  - Short lived route disappearance
  - Failures
- Incident: A set of prefix signatures closely related in time from the same AS
- Connectivity incident: incidents that lead to connectivity problems
- Research: Mine signatures and apply ML techniques to classify various BGP events

54

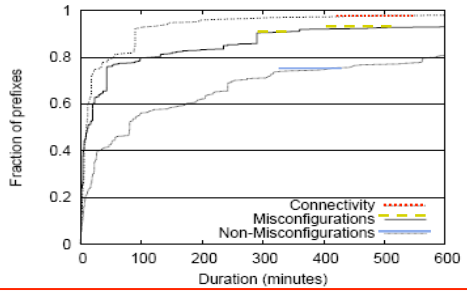
## Origin Misconfiguration Analysis

- Origin misconfiguration: accidentally inject routes for prefixes into global BGP tables

| | Old route | New route |
|---|---|---|
| Self deaggregation | a.b.0.0/16 X→Y→Z | a.b.c.0/24 X→Y→Z |
| Related Origin | a.b.0.0/16 X→Y→Z | a.b.0.0/16 X→Y<br>a.b.0.0/16 X→Y→Z→O<br>a.b.c.0/24 X→Y<br>a.b.c.0/24 X→Y→Z→O |
| Foreign Origin | a.b.0.0/16 X→Y→Z | a.b.0.0/16 X→Y→O<br>a.b.c.0/24 X→Y→O |

55

## How long do short-lived origin changes last?




Misconfigurations last shorter than non-misconfigurations, connectivity problems are detected/fixed sooner

56

## Results: Origin misconfiguration

| | Potential misconfigs per day | Email responses (% of potential) | Misconfigs (% of email) | Connectivity (% of misconfigs) |
|---|---|---|---|---|
| Prefixes | 605 | 352 (58%) | 339 (96%) | 13 (4%) |
| Incidents | 178 | 52 (29%) | 45 (86%) | 6 (13%) |

(200 long-lived new routes/day)
(vs. 1000 failures/day)

- Misconfiguration detection accuracy is high
- Large number of misconfigurations
  - Extrapolated estimate is 580 (605 * 0.96) prefixes per day
  - 3 in 4 new routes seen in a day result from misconfigs
- Most misconfigurations don't disrupt connectivity

## Export Misconfiguration

- Inadvertent export of a route in violation of the exporters policy

| | Potential misconfigs per day | Email responses (% of potential) | Misconfigs (% of email) | Connectivity |
|---|---|---|---|---|
| Paths | 96 | 64 (66%) | 61 (96%) | - |
| Incidents | 35 | 12 (36%) | 10 (86%) | - |

58

## Research

- Can you hear me now, it must be BGP, Nate Kushman,Srikanth Kandula and Dina Katabi, ACM SIGCOMM CCR, April 2007
- Impact of Hot-Potato Routing Changes in IP Networks. R. Teixeira,A. Shaikh, T. G. Griffin, J. Rexford. IEEE/ACM Transactions on Networking, vol. 16, no. 6, December, 2008. Pages 1295-1307.

59

## BGP convergence delay

- How long before a route change converges in the network
- Two Time factors
- Time to detect a failure
  - Keep-alive 60 seconds
  - Hold timer: 180 seconds
- On failure detection, throw away peer routes and announce changes

60

## Route change propagation

- New route announcement requires path exploration
- Path Path exploration is expensive
  - Large number of possible paths
  - Might have to explore (nearly) all of them
- Minimum Route Advertisement Interval
  - Minimum time between advertisement of routes for a given destination to a given neighbor
  - allows for combining multiple messages in one
  - Typical value of 30 seconds
- Convergence delay
  - (30 seconds) * (# of paths) + 180 seconds

61

## Route changes

- Link failures
- Reachability issues (router reboot)
- Session resets
- Lots of path changes
- How to deal with transient changes
- Do not want to be too quick
- At the same time not wait too long
- Strike a balance
- Route Flamp damping (RFD)

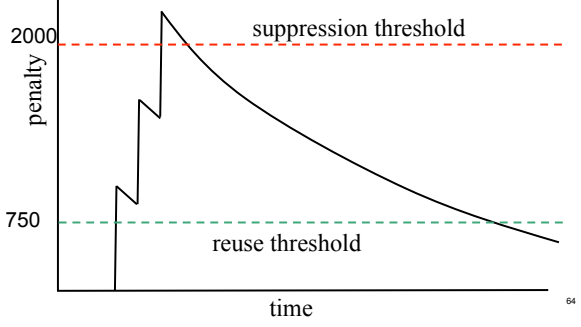
62

## Route flap damping

- First flap on a route (withdrawn & readvertised), asses penalty (1000 points), put the prefix in historical category
- Second flap (another 1000 points), do not advertise this route to others
- Penalty is decayed if it does not flap further
- Once the penalty falls below 750, the route is removed from dampened state

63

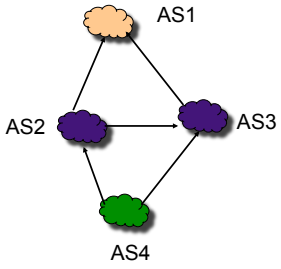
## Damping Penalty Function




64

## Effect of damping

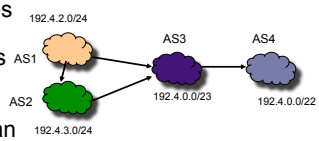
- If Route to AS1 flaps beyond suppression penalty
- AS2 will not advertise routes to AS1 via AS2
- AS4 will then stick to route AS1 via AS3




65

## Aggregation can help route flapping

- Specific-route changes can result in flapping
- But aggregated routes may not exhibit route flapping
- Hence aggregation can mask route flapping and reduce instability because it reduces the number of networks visible in the core Internet.




66

## Current Research

1. **A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance** F. Wang, Z. M. Mao, J. Wang, L. Gao, R. Bush ACM SIGCOMM 2006
- **Rationality and Traffic Attraction: Incentives for Honest Path Announcements in BGP** Sharon Goldberg (Princeton University); Shai Halevi (IBM T. J. Watson Research); Aaron D. Jaggard (Rutgers University); Vijay Ramachandran (Colgate University); Rebecca N. Wright (Rutgers University) ACM SIGCOMM 2008

67