Best Practices for Network Interconnections

Tom Scholl, AT&T Labs

NANOG 43 – June 3rd 2008

Why agree upon how to interconnect/peer?

- What IP addresses do we want to use?
  - Who provides the subnet? Can you do /31s?
- What MTU do we set?
- MD5? GTSM? BFD?
- BGP Timer Complications
  - Some vendors don’t like very low values
  - Some routers cant handle very low values
- How do we grow and add additional links?
  - Multihop? Multipath? Link Agg?

What happens when things aren’t applied consistently?

- Troubleshooting issues
  - Circuit ID / NOC contact not in the router config?
  - BGP neighbor description should have details
- MTU Problems
  - Black holes
- Incorrect policy application
  - Prefix-Leak
  - Incorrect local-preference application
  - AS-Path filters not being updated
  - Prefix-lists not being updated

So why have a presentation on this?

Because most operators will run into the same dilemmas and issues.

Not everyone will implement the same solution and in some cases, people don’t know what may be a better solution.

Most network operators do not openly share router configurations (though, if you are tossing out gear, please wipe the NVRAM first…).

Most vendors (sorry folks) are really not clued into how things work from an Internet peering point of view. Yes, there are some clueful people at each vendor, but not all account teams or business units know what goes on.

Routing Policy

- Community Based Filtering

  Use BGP communities to identify routes. Such markings can be used to specify:

  -Geographic Origination (NYC, Chicago, etc)

  -Relationship (Peer, Transit, Customer)

  -Traffic-Engineering Capability (Prepend, Set MED, etc)

Routing Policy

- Community Based Filtering

  Why use communities? Because IP prefix-filters don't scale.

  - IP Prefix-Filters require frequent updates
  - IP Prefix-Filters can be *long* (think router config size limits)

  Where to apply community based filters?

  Towards:

  -Peers

  -Customers

  -Transit

Routing Policy

- Community Based Filtering
  - Where should you insert communities?

    Routes from:

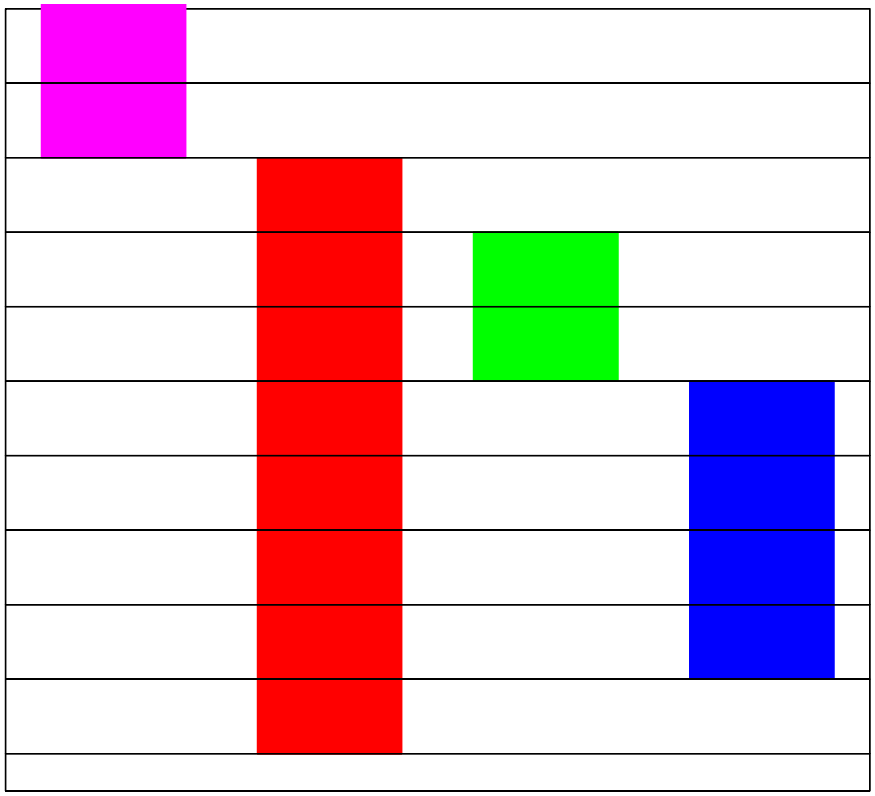
    - Peers
      - Need to mark this route so we know to send it to customers but not to peers or transit.
    - Customers
      - Need to advertise these routes to transit, peers and other customers.
    - Transit
      - Need to advertise these routes to customers only (not peers or other transit providers).

Local Preference

- Local-Preference is one of the most key policy control devices within your ASN
  - Only applies to outbound traffic (routes received)
- Typical deployments specify a tiered model of preference (which reflects economic factors)
  - 1 – Trust your infrastructure over anyone else
  - 2 – Trust your customers desires
  - 3 – Prefer peers over transit
    - Except in cases where there are issues with peers networks, latency or economic reasons to use transit.

Local Preference

- Example:



Exchanging MEDs

- How does it work?
  - MED (Multi Exit Discriminator) is a numeric value associated with a route. This value can be used as a tie-breaker in selecting a BGP path.
  - Many networks use MEDs that are derived from distance or some other administrative network.
  - Typically agreements to do MED’s are bidirectional between peers. Though some peers transmit MED values in “best effort” to help inform peers of desires.

Exchanging MEDs

- Pros:
  - You can influence your peers ability to select which peering point to select (providing best-exit routing instead of nearest-exit).
- Cons:
  - Every network uses their own scheme (IGP costs, static MED values) when utilizing MEDs. Some operators operate on an order of magnitude different from others.
  - Results in you having to trust the behavior of a peer network (think of a peer attempting to traffic-engineer within your network to workaround their own internal congestion issues).

Exchanging MEDs

- Common behaviors from Customer to Transit:
  - It is *expected* that customers have the ability to set MED to their transit providers.
  - It would also be nice if transit providers sent MED to their customers, too. This is not always done.
- Common behaviors between settlement free peers:
  - While MED may be transmitted, typically it is rewritten by the remote party.
  - MED exchange should be pre-negotiated before session turnup so both sides can validate their expectation.
  - Understanding the values used within a MED scheme is helpful when attempting to rewrite MED.

BGP Communities

- BGP Communities are good for:
  - Identification of a route
    - Where did it come from?
    - Who did it come from?
  - Transmitting a desire with a route:
    - Black-Hole
    - Reset BGP local-preference to X value
    - No-Export
    - No-Advertise
    - No-Export to a specific peer
    - Prepend
    - Prepend to a specific peer X amount of times

BGP Communities
-Accepting and Transiting them

- Not everyone accepts and transits them
  - Not all networks agree as to how to handle routes with communities:
    - Customer communities applied are sometimes stripped when being transmitted to peers, transit, customer.
    - “Internal” communities are sometimes stripped to peers, transit, customer and even public looking-glasses/route-servers.
    - To prevent a variety of issues, community filters may be applied to restrict what communities can be transmitted.
  - There is no expectation that a community you set will be seen outside of the AS domain you are communicating with.

BGP Communities
-Accepting and Transiting them

- It would be nice if everyone would agree to keep the community set intact.
  - This of course, may lead to excessively long communities applied on a given route.
  - Some vendors do not like lots of communities. Some routers may:
    - Reject the route
    - Refuse to append on additional communities
    - Bounce the session
- Networks *need* to sanitize BGP communities received in order to prevent bad behaviors.

BGP Communities
-Protecting Yourself

- Apply community list filters to prevent a customer or peer from transmitting a route with your internally defined communities.
  - Remember to use regular expressions to make sure you are writing the proper filter.
  - Also ensure that you are not permitting a customer to impact your transit providers.

BGP Communities
-Honoring Communities

- There is no expectation between peers that communities will be accepted. Though it would be nice.
  - Don't expect:
    - NO-EXPORT or NO-ADVERTISE to work
    - NO-EXPORT to work one day and work the next
      - "Hey look at all that traffic…"
    - A peer to transit your communities to peers or customers
    - Your peers to explain what a community means on sent route
  - Do expect:
    - Your peers to send you a variety of communities

BGP Communities
-De-Aggregates and MEDs

- De-Aggregate (subset of a larger Internet announced route) and MEDs can be combined to provide some routing enhancements.
  - This may improve end-user performance as both peers could pin-point the best exit point to reach some portion of your network.
  - By using NO-EXPORT, you reduce the amount of bloat in the Internet routing table by only permitting the deaggregates into that peers network.
  - This can also *hurt* you if that peer no longer accepts NO-EXPORT.

Security
-Maximum Prefix on Peers

- A simple mechanism to prevent routing table leaks and your accidental acceptance.
- Has some side effects:
  - Requires manual intervention to reset
    - There are knobs to periodically reset, but isnt that like looking down a barrel more than once?
  - Sudden BGP session shutdown
    - BGP convergence is slooooow
    - Drastic traffic rerouting
- Still, probably the best way we have to stop an accidental leak today.

Security
-Maximum Prefix on Peers

- The problem with maximum-prefix is:
  - Prefix threshold limits change over time (peer acquisitions, mergers, etc).
    - Networks neglect to remind peers to update max-pfx.
    - Peers ignore requests to update max-pfx.
  - Recovering from a max-pfx leak is not easy
    - Standard NOC contacts may not be permitted to touch peering routers
- Much of this has been discussed in previous postings and presentations.
  - Why not have maximum-prefix that “ignores” prefixes that exceeded the limit?

Security
-Maximum Prefix on Peers

- Remember, Maximum-Prefix operates differently across vendors!
  - Cisco applies it to prefixes received, regardless of acceptance
  - Juniper applies it based upon routes received and accepted
- Much of this has been discussed in previous postings and presentations.
  - Why not have maximum-prefix that “ignores” prefixes received after the peer has exceeded the limit?
  - Why not have maximum-prefix ***outbound***?

Security
-Prefix-Filtering

- Good on customers
  - Can be fed from a variety of databases/formats
  - Strongly recommended
- Mixed opinions on peers
  - Good for filtering *some* routes

Security
-Prefix-Filtering

- What are those “some” routes?
  - Known “Special Use” networks
    - RFC3330
  - RFC1918 address space
    - Mixed - some networks allow customers to advertise RFC1918 space with no-export (think cablemodem CPE management).
    - In general a good idea to filter this (easy target, range doesn’t change).
  - Unallocated IANA space
    - A good idea, however it *must* be kept up-to-date as IANA allocates out pieces several times a year.
    - Failure to update the filter will result in a lot of pain (and probably some NANOG postings…)

Security
-Prefix-Filtering

- Prefix-Length range filtering
  - Good practice to filter /0 to /7, as no one is using them.
  - Most networks deny /25 to /32 (since the Internet generally will refuse them)
  - Some networks permit /25 to /27 and apply NO-EXPORT for customers who have additional requirements (load-balancing, migration, etc)

Security
-AS-Path Filtering

- Filtering customers to only use what they’re allowed
  - Updating AS-Path filters frequently may not be fun
    - New downstream customers
    - How do you validate what is an authorized ASN?
  - Does however provide a “chain of evidence” of what ASNs a route traversed
- Filtering known private ASN ranges
  - Easy target to apply on customers & peers

Security
-AS-Path Filtering

- AS-Path filtering can have unintended consequences
  - Maybe people prepend ASNs for traffic-engineering
    - Exploiting BGP loop-prevention
  - Inbound AS-Path filtering on peers & transit can burn you when people apply the above scenario
- Filtering Unallocated ASNs
  - Prevent accepting routes from parties hijacking unallocated ASNs
  - Some Caveats:
    - IANA allocates new ASN ranges yearly
    - Keeping this filter up to date may not be easy
      - What happens if you forget to update it…

Security
-AS-Path Filtering

- Want to stop an easy accidental prefix leak to your peers?
  - Apply an ***<u>outbound</u>*** AS-path filter that includes popular/large network ASNs.
  - Would save a lot of people when they accidentally leak their transit routes to peers.

Security
-MD5

- The MD5 Story:

Not many people used it to begin with.

A security “issue” arises, many people start enabling and requesting MD5 support in a short period of time.

People begin to get suspicious.

Word spreads, MD5 is now mandatory on many sessions due to the above.

Security
-MD5

- Why the concern?
  - Additional overhead for the router to examine the MD5 hash in the TCP options may require additional resources.
    - Think router CPU/resource attacks.
- Is it needed?
  - It has benefits, but it only validates the remote neighbor. It is not meant to “secure” the BGP session from attack.
- Should you do it?
  - Make your own decisions, but do not rely upon this as a sole method to protect your infrastructure.
  - There are other options out there to authenticate your neighbor.

Security
-GTSM

- An alternative to easily validate a BGP neighbor
  - Basically impossible to spoof/manipulate TTL from beyond the defined scope.
  - Less administrative overhead than MD5.
  - But, not all hardware supports it (properly).
- But it is not a security mechanism
  - Once again, it does not “secure” the BGP session.
  - Routers are always vulnerable to attacks with or without these features.

Peering Configuration
-Link Bundling

- Peering traffic generally grows over time.
- N x Interface link bundling is a simple way to grow the peering interconnection.
- Implementations vary depending on the network and the hardware.

Peering Configuration
-Link Bundling

- 1) eBGP Multihop + Static Routes to Loopbacks

- Both routers peer with each others loopbacks.
- Since the loopback is not reachable, both routers static route the loopback to each interface.
- Static routes are pointed as an equal-cost multipath across all the interfaces.
- The router will recursively lookup the loopback to distribute traffic across all the interfaces.

Peering Configuration
-Link Bundling

- Pros:
  - Simple and has worked for years.

- Cons:
  - Static route mis-configuration can land you in trouble.

Peering Configuration
-Link Bundling

```
interface GigabitEthernet1/0
 ip address 192.168.1.0 255.255.255.254
interface GigabitEthernet2/0
 ip address 192.168.2.0 255.255.255.254


ip route 10.0.0.1 255.255.255.255 192.168.1.1
ip route 10.0.0.1 255.255.255.255 192.168.2.1
```

Assume 192.168.0.0/16 is something we announce internally as an aggregate.

What happens when GigabitEthernet1/0 goes down? Both static routes will still remain valid. Some of the traffic may be redirected back towards your Infrastructure.

Peering Configuration
-Link Bundling

- Things to remember?
  - Include the interface name as the next-hop

```
ip route 10.0.0.1 255.255.255.255 GigabitEthernet1/0 192.168.1.1
ip route 10.0.0.1 255.255.255.255 GigabitEthernet2/0 192.168.2.1
```

  - Make sure the next-hop IP address is correct ☺
  - Not all vendors load balance traffic correctly with static ECMPs
  - Some vendors have ECMP limitations (max 6 next-hops for static routes)

Peering Configuration
-Link Bundling

- 2) eBGP Multipath

- One BGP session for each peering interface.
- After a certain point in the BGP best path decision process, a route can be considered valid for multipath.
- The router would then consider multiple next-hops as valid and install into the FIB.

Peering Configuration -Link Bundling

- Pro:
  - One BGP session per interface can make troubleshooting easier (think of when working with a bad circuit).
- Con:
  - Some vendors implement this as a global command (applied on all neighbors), rather than a per-neighbor or per-group configuration.
    - This could lead to issues in the hardware (FIB)
  - No way to set multipath on specific prefixes.

Peering Configuration
-Link Bundling

- 3) Link Aggregation

- Aggregate links as a logical interface at Layer-2 and have a single Layer-3 IP interface.
- Don’t have to worry about the RIB having to perform L3 ECMP.

Peering Configuration
-Link Bundling

- Pro:
  - One BGP session, one IP interface.
- Con:
  - Not all hardware supports link aggregation or all the associated protocols (LACP).
  - Some vendors may have a poor ability to load balance across multiple interfaces.

Peering Configuration
-BFD

- Bidirectional Forwarding Detection
  - Plenty of presentation and documents exist on the protocol.
- Basically, a rapid method to detect failure of a neighbor by using fast keepalive messages.
- Why would I need this?
  - Default BGP keepalives result in 180 second detection time.
  - If there isn’t an interface failure, you are going to rely upon BGP keepalives to detect the failure.

Peering Configuration
-BFD

- So why not just set BGP keepalives low?
  - Because sometimes BGP keepalives are missed
    - When routers are busy, sometimes they forget sending things..
    - All BGP messages go directly to the route processor.
      - BFD hellos *typically* are done on a distributed linecard.
    - Setting low timers of 3 seconds (9 second dead timer) may sound nice, but during churn can result in dropping various BGP sessions.
    - Even values of 15/45 have been found to be troublesome with some peers.

Peering Configuration
-BFD

- BFD can operate at sub-second intervals.
- Vendors do have caveats regarding:
  - Number of BGP neighbors per-linecard, per-router.
  - What hello/dead timer values are used.

While some vendors may tout how cool BFD is, remember 3x300ms (900ms) seems to be the most realistic setting anyone will commit to.

(Barely legal sub-second convergence?)

Peering Configuration
-BFD

- BFD operates in two methods on vendors:
  - Distributed hello process on applicable linecards
  - Centralized hello process on a route-processor
- Each of the above implementations has its own pros and cons.
  - What are you trying to detect? RP failure or LC failure?
- BFD is not so pretty when operating with Link Aggregation
  - BFD packets will be sent on one member only, not all.
  - Do you want a single link failure on a Nx4 bundle to shutdown everyone?

Peering Configuration
-MTU

- Plenty of things have already been said/presented regarding MTU.
- Operating > 1500 bytes at an Internet exchange shared fabric is asking for trouble.
  - Not all members can accept > 1500
- However, agreeing upon MTU on a private interconnect is fairly easy. Except:
  - Make sure you use the right numbers and factor in overhead, FCS, etc.

Peering Configuration
-Public Exchanges

- Things to remember when turning up on a public Internet exchange:
  - Turn **OFF** Cisco Discovery Protocol (CDP)
  - Turn **OFF** Proxy ARP
  - Do **NOT** redistribute the public internet exchange prefix within your network
    - If you do, **please** do not export it outside your ASN
  - Do **NOT** enable OSPF/ISIS/EIGRP/RIP on your peering interface
    - Though if you do, perhaps someone should build an adjacency and sees what happens ☺

Peering Configuration
-Public Exchanges

- Turn **OFF** DHCP
- Turn **OFF** BOOTP Server
- Turn **OFF** IP Redirects
- Do **NOT** rewrite BGP next-hop to point to a non-peer
- Do **NOT** static route to members on the exchange
- If peering with someone else on the exchange and you are transmitting another peers routes, please **reset** BGP next-hop to point to you.

- Yes, people have done all of the above.

Peering Configuration -Public Exchanges

- What about MAC filtering?
  - Would stop non-peers from sending traffic to you.
  - Requires the IX to provide a list of MAC/IP addresses, which requires periodic updating.
  - Some hardware cannot do MAC filtering (or at least, without some impact)
  - An alternative would be to perhaps request the IX operator to control what ports can talk to each other…
    - But does that sound a bit too much like PeerMaker?
    - Why do I need to contact the IX to turn-up a peer?
    - Then again, if the interface to the IX was a simple working website, maybe that’s not so bad?

Peering Configuration -Netflow

- Netflow is a great tool to determine where traffic is going.
- Some things to remember:
  - Disregard what ASN (peer-as or origin-as) is embedded in a netflow packet. Remember, the router populates the ASN based upon what the RIB indicates as the best path.
  - Use internally stored BGP RIB snapshots to associate prefix with one or multiple ASNs.
  - Rely upon SNMP IfIndex to determine where traffic originated from.

Peering Configuration
-QoS

- QoS honoring & marking behavior differs from network to network.
- In general, expect peering traffic to be treated as best effort in your peers networks.
  - Its natural for a service provider to prefer their own traffic (VPN, IPTV, etc) over Internet traffic.
- It's best to maintain the original TOS/DSCP bits as traffic traverses the network.
  - How you treat this traffic internally, is another story.

Peering Configuration -Monitoring

- SNMP poll all peering interfaces.
  - As well as your infrastructure to determine if the counters match up.
- Not so easy to generate per-peer graphs on a public exchange:
  - MAC-Accounting can solve this.
    - It can also detect when a non-peer is sending you traffic.
- A large assortment of free tools exist out there to do the above.
    - Torrus (www.torrus.org) will automatically discover BGP peers and support MAC accounting.

End

- Please read:
  - Ren Provo’s presentation on BCP on peering/NOC communication
  - BGP Tutorials – Great start if you’re new
  - Richard Steenbergen’s BGP communities presentation
- IETF BCP work underway regarding peering interconnections.
  - Individuals involved:
    - Jon Nistor (TorIX)
    - Marco Rodrigues (Juniper)
    - Me

Send questions, comments, complaints to:

Tom Scholl, AT&T Labs

tom.scholl@att.com