BGP Peering Strategy & Data

Andy Davidson
CTO @ Asteroid International
Director @ LONAP, IXLeeds, Euro-IX

andy@asteroidhq.com

APRICOT 2017, HCMC, Vietnam
March 2017
Why do Traffic Engineering?

- Manage your capacity demands
- Ensure service quality
- Recover from Failures
- Manage service/circuit costs
- Handle traffic growth

James Cridland http://www.flickr.com/photos/jamescridland/
Life starts out very simply, “send traffic to peers if possible, then transit providers”

But what about when your network grows?

What about when your traffic grows?

What if you add more cities/POPs/exchanges?
Real examples

• Circuits with *cost difference* > $100/Mbit
• Regional networks - *poor local peering*
• Circuit failure causing *congestion*
• Changing *customer demand/behaviour*
  – Increased quality expectation
  – New high bandwidth services such as video
Internal network TE

- **Simple** compared with Interdomain TE

- You administrate both sides
  - You know the **price** of all paths
  - The IGP knows the **capacity** of all paths
  - IGP protocols let you map price, capacity to shape routing using **cost**.
Inter-domain TE

• You do NOT control both sides
  – Path vector protocols hide metric, capacity, cost
  – Simplicity of BGP protocol imposes limitations
  – **Volume of traffic** matters, not # of routes

• However, large volume of traffic is usually with a **small number of other ASNs**
You need data

AS-Stats
Manuel Kasper
https://neon1.net/as-stats/

.. But more on this later
Netflow

- **Export** information about packets routed through your network
- Normally **sampled**
- Sent to a **collector** over UDP
- A variety of commercial and open-source tools sort and display these **flow records**.
Different Flow protocols

• Netflow – Designed by Cisco in ‘90s, published as a standard (v9 is RFC3954 and supports IPv6)
• IPFIX (RFC5101) Based on Netflow 9, 2008
• sFlow – Nice protocol but incompatible with Netflow, typically implemented on L2 switch.
• Jflow – Essentially Netflow on Junipers
Other ways to get data

• Log file analysis
  – Useful before you have a network, for working out the benefit of building a network/peering.
  – Best for ‘single service’ networks
    • DNS providers have DNS logs with time & IP
    • Web providers have web logs with time & IP
    • Hosted email providers have mail logs...
Data tells you

• Your traffic **direction**
  – Mainly inbound
  – Mainly outbound
  – Balanced

• Your **top traffic originators** or **destinations**
As-stats

- https://neon1.net/as-stats/
- Open source
- Quick to setup, simple to use
- Resource intensive
Who are my key peers?
(or potential peers)

Top originators of traffic to me, top consumers of my content

Chart colour relates to an interface on the edge of my network
Historical data

New peering added, traffic growth!
= $$$$$$ for my network!

Bonus please!
Inmon sflow-rt

- Produces realtime traffic graphs
- “What is the situation right now”
- Useful to check peering config has taken effect
- Also useful in abuse mitigation
Realtime aggregate data

<table>
<thead>
<tr>
<th>bgpsources</th>
<th>bgpsourcepeers</th>
<th>bps</th>
<th>bgpsourcepeers/ = 41230'</th>
</tr>
</thead>
<tbody>
<tr>
<td>15169</td>
<td>15169</td>
<td>714</td>
<td></td>
</tr>
<tr>
<td>6185</td>
<td>6185</td>
<td>299</td>
<td></td>
</tr>
<tr>
<td>20950</td>
<td>20950</td>
<td>392</td>
<td></td>
</tr>
<tr>
<td>32904</td>
<td>32904</td>
<td>392</td>
<td></td>
</tr>
<tr>
<td>9090</td>
<td>9090</td>
<td>290</td>
<td></td>
</tr>
<tr>
<td>2900</td>
<td>2900</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td>54996</td>
<td>54996</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>3356</td>
<td>3356</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>178</td>
<td>178</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>16509</td>
<td>16509</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>19879</td>
<td>19879</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>5481</td>
<td>5481</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>36408</td>
<td>36408</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>2900</td>
<td>2900</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>13414</td>
<td>13414</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>8708</td>
<td>8708</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>59999</td>
<td>59999</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>1273</td>
<td>1273</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>20940</td>
<td>20940</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>15322</td>
<td>15322</td>
<td>122</td>
<td></td>
</tr>
</tbody>
</table>
Realtime transit analysis

<table>
<thead>
<tr>
<th>bgpsources,bgpsourcepeeras</th>
<th>bps</th>
<th>bgpsourcepeeras='1273'</th>
</tr>
</thead>
<tbody>
<tr>
<td>bgpsources</td>
<td></td>
<td>bps</td>
</tr>
<tr>
<td>bgpsourcepeeras</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1273</td>
<td></td>
</tr>
<tr>
<td>3356</td>
<td></td>
<td></td>
</tr>
<tr>
<td>54994</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4872</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5483</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16623</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4812</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20940</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9506</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12878</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9304</td>
<td></td>
<td></td>
</tr>
<tr>
<td>56046</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32934</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3320</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8764</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17816</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17921</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Export from Sflow-RT

- RESTful export into logstash/influxdb/grafana for historical data

pmacctd

- http://www.pmacct.net
- Open Source
- High performance, high scale, powerful
- Most flexible, most configuration required

- Collector → own reports
1) Configure a collector

sfacctd_port: 2100
sfacctd_as: sflow
sfacctd_renormalize: true
!
plugins: print[testprint]
!
aggregate[testprint]: in_iface, out_iface, proto, peer_src_ip, peer_dst_ip, peer_dst_as, peer_src_as, src_as, dst_as
!
print_output_file[testprint]: /path/to/spool/blabla-$peer_src_ip-%Y%m%d-%H%M.txt
print_output[testprint]: csv
print_output_separator[testprint]: ;
print_refresh_time[testprint]: 60
print_history[testprint]: 1m
print_history_roundoff[testprint]: m

Use the ASN data from the router if it exists, no need to setup BGP flow export
2) Get a report

SRC_AS; DST_AS; PEER_SRC_AS; PEER_DST_AS; PEER_SRC_IP; PEER_DST_IP; IN_IFACE; OUT_IFACE; PROTOCOL; PACKETS; BYTES
41230; 224; 41230; 2603; x.x.x.253; x.x.x.246; 3; 4; tcp; 2048; 151552
41230; 15169; 41230; 15169; x.x.x.253; x.x.x.246; 3; 4; tcp; 10240; 880640
41230; 50247; 41230; 24724; x.x.x.253; x.x.x.246; 3; 4; tcp; 2048; 167936
41230; 9269; 41230; 1273; x.x.x.253; x.x.x.237; 3; 3; tcp; 2048; 135168
41230; 3356; 41230; 1273; x.x.x.253; x.x.x.237; 3; 3; tcp; 32768; 2375680
41230; 209; 41230; 1273; x.x.x.253; x.x.x.237; 3; 3; udp; 2048; 2940928
20940; 0; 20940; 0; x.x.x.253; x.x.x.246; 3; 4; tcp; 43008; 65458176

Red line represents a single flow with Google on behalf of a user
Keeping historical data

plugins: mysql[5mins], mysql[hourly]

sql_optimize_clauses: true
sql_dont_try_update: true
sql_multi_values: 1024000

sql_history_roundoff[5mins]: m
sql_history[5mins]: 5m
sql_refresh_time[5mins]: 300
sql_table[5mins]: acct_bgp_5mins

sql_history_roundoff[hourly]: h
sql_history[hourly]: 1h
sql_refresh_time[hourly]: 3600
sql_table[hourly]: acct_bgp_1hr

plugin_buffer_size: 10240
plugin_pipe_size: 1024000
aggregate: tag, src_as, dst_as, peer_src_as, peer_dst_as, peer_src_ip, peer_dst_ip, local_pref, as_path
Example by pmacct author Paolo Lucente

```sql
mysql> SELECT * FROM int_tm-20130803_1400 LIMIT 10;
|
+-----------------+-----------------+-----------------+-----------------+-------------------------------+-------------------+-----------|
| iface_in | peer_ip_src | peer_ip_dst | peer_dst_as | stamp_inserted | bytes |
+-----------------+-----------------+-----------------+-----------------+-------------------------------+-------------------+-----------|
| 212 | 10.0.0.107 | 10.0.0.3 | 65000 | 03-08-2013 14:00 | 859 |
| 212 | 10.0.0.107 | 10.0.0.253 | 65001 | 03-08-2013 14:00 | 5358 |
| 212 | 10.0.0.107 | 10.0.0.234 | 65002 | 03-08-2013 14:00 | 6181 |
| 212 | 10.0.0.107 | 10.0.0.251 | 65003 | 03-08-2013 14:00 | 27002 |
| 205 | 10.0.0.107 | 10.0.0.233 | 65004 | 03-08-2013 14:00 | 1200 |
| 258 | 10.0.0.107 | 10.0.0.240 | 65005 | 03-08-2013 14:00 | 560 |
| 212 | 10.0.0.107 | 10.0.0.252 | 65006 | 03-08-2013 14:00 | 62682 |
| 212 | 10.0.0.107 | 10.0.0.234 | 65007 | 03-08-2013 14:00 | 3843 |
| 212 | 10.0.0.107 | 10.0.0.17 | 65008 | 03-08-2013 14:00 | 21074 |
| 205 | 10.0.0.107 | 10.0.0.254 | 65009 | 03-08-2013 14:00 | 2023 |
```

- Export into reports, web interface, spreadsheet
- Multiple back ends supported, including time series databases
- Very flexible approach but needs more setup time
Kentik

- https://www.kentik.com
- Hosted solution
- Zero configuration, zero equipment needed
- Point Netflow at their collector and reports follow
Traffic by Source ASN
Breakdown by region (US)
Top flows per interface
Outbound vs Inbound

- Outbound heavy networks
  - Somewhat easier life

- Inbound heavy networks
  - You must trick the Best Path Selection methods of networks sending you traffic.
  - Their config change will move your traffic.
Outbound Heavy Traffic Engineering

• BGP Best Path Selection Algorithm
• Traffic engineering is about ‘tricking’ this process
• Affects traffic in outbound direction
  – Local Preference
  – AS PATH length
  – Lowest Origin Type
  – Lowest MED
  – Prefer eBGP paths
  – Lowest IGP Metric
  – Oldest route
Mainly outbound, single POP

• Localpref
  – A **hammer** – blunt tool, inflexible.. But it is a tool.
  – “Generally” prefer to send traffic to customers, then peers, then transits.
  – Manage top ‘n’ networks, so that there is a **preferred path**, and a **failure path**, with capacity on both circuits.
Localpref – blunt hammer

10.0.0.0/8 Localpref 100 via 100 123
10.0.0.0/8 Localpref 500 via 300 200 200 200 200 123

Which link will you prefer?
AS123 here is trying to shape inbound traffic via AS100. Why?
   Higher capacity link?
   More reliable?

What should you do?

Answer: It depends on the volume of traffic, cost of capacity, value of traffic
Mainly outbound – Many POPs

• Use hot potato routing to best effect
  – Nearest exit routing
  – Understand who your top traffic sinks are and peer at all POPs
  – Ignore MEDs from others – unless you want to carry the traffic on your backbone
AS_PATH prepending (outbound)

• Make BGP prefix paths "appear" longer via less preferred circuits

• “BGP path for 10.0.0.0/8 654_789 is congested”

• 123 456 789
• 654 789

• 100 100 100 654 789

• Will not vary inbound packet route (though this route might not be congested!)
Inbound traffic engineering

• Much harder
  – Trick others’ Best Path calculations
  – You do not administrate origin party router

• But remember...
  – Largest flows come from a small number of networks
  – Content networks want to deliver traffic to you as well as possible!
Selective Announcements

- **Shortest prefix**
- Local Preference
- AS PATH length
- Lowest Origin Type
- Lowest MED
- Prefer eBGP paths
- Lowest IGP Metric
- Oldest route

Prefix length considered before BGP.

10.0.0.0/16 vs 10.0.0.0/17 & 10.128.0.0/17
Problem of Selective Announcements

• Often filtered
• Considered rude – might lead to depeering
• Never announce ‘globally’
...But can be used to great effect

- To the same peer or transit provider, announce aggregate and regional pfx

Add NO_EXPORT community

Use with permission
AS_PATH prepending (inbound)

• Signal preferred path by growing AS_PATH on less preferred paths
• Marginal effect which *degrades quickly*
• Signal backup link to a single AS, but load-balancing capacity is much harder
• May not be heard at ‘distant’ ASNs
• Another ‘blunt’ tool, but can move some traffic.
2.5 AS Path Prepending

AS path prepping is a common way of making routes less attractive since AS path length is usually one of the BGP path selection criteria. A customer network may use these communities to selectively request AS3320 to insert additional copies of the AS number 3320 when propagating the customer routes to neighbors.

<table>
<thead>
<tr>
<th>Community Value</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>65012 : X</td>
<td>AS Prepend 2x to AS X</td>
<td>Prepend 3320 two times to named peer (ASN=X)</td>
</tr>
<tr>
<td>65013 : X</td>
<td>AS Prepend 3x to AS X</td>
<td>Prepend 3320 three times to named peer (ASN=X)</td>
</tr>
<tr>
<td>6501n : 65001</td>
<td>AS Prepend by Class: Peer</td>
<td>Prepend 3320 n times to peers. n=2 or 3.</td>
</tr>
<tr>
<td>6501n : 65002</td>
<td>AS Prepend by Class: Upstream</td>
<td>Prepend 3320 n times to upstream.</td>
</tr>
<tr>
<td>6501n : 65003</td>
<td>AS Prepend by Class: Peer &amp; Upstream</td>
<td>Prepend 3320 n times to peers and upstream.</td>
</tr>
<tr>
<td>6501n : 65004</td>
<td>AS Prepend by Class: Customer</td>
<td>Prepend 3320 n times to customers.</td>
</tr>
<tr>
<td>6501n : 65005</td>
<td>AS Prepend by Class: Customer &amp; Peer</td>
<td>Prepend 3320 n times to customers and peers.</td>
</tr>
<tr>
<td>6501n : 65006</td>
<td>AS Prepend by Class: Customer &amp; Upstream</td>
<td>Prepend 3320 n times to customers and upstream.</td>
</tr>
<tr>
<td>6501n : 65007</td>
<td>AS Prepend by Class: All</td>
<td>Prepend 3320 n times to all AS3320 neighbors.</td>
</tr>
</tbody>
</table>
MEDs

- **Lowest** MED wins.
  - Opposite of Nearest Exit routing, “carry traffic to me”
  - Only works to the same peer in multiple regions
  - Copy IGP metric to MED
  - Normally subject to negotiation

- Sometimes honoured, often when network traffic is latency or loss sensitive.
MEDs are often filtered

• Many networks set MED to 0 when they learn prefixes, so that hot potato routing will override MED.

```plaintext
route-map peers-in permit 10
    set local-preference 200
    set metric 0
```
Origin changing

- IGP
- EGP
- Incomplete

Highest priority

route-map PEERS permit 10
set origin igp

route-route-map TRANSIT permit 10
set origin incomplete

Often peers set to ‘igp’ or ‘egp’ statically on routers to nullify effects of Origin changing.
Inbound – what does work well?

• Overprovisioning
• Peer with top networks **widely** (buy options!)
  – Failure of single link will not break adjacency
  – Failures can be handled in predictable ways
• Build **relationships**
• **Constantly monitor and manage**
• If you care about your traffic, let it go. 😊
  – Playing games with peering hurts your customers’ traffic

• Affecting distant ASNs is very hard – a region may only see a single next-hop ASN.
Deterministic routing

- Local Preference
- AS PATH length
- Lowest Origin Type
- Lowest MED
- Prefer eBGP paths
- Lowest IGP Metric
- Oldest route

Top flows should leave your network via deterministic means, and not left to BGP Best Path selection (or to chance).

If you are relying on oldest route to make the decision, you risk traffic taking unpredictable routes.

However, oldest routes do break the ‘flapping sessions’ problem. You need to monitor and manage your top flows constantly.
What does “manage relationships” mean?

• Go back to your data
  – Collect and share information with peering co-ordinators at forums like this
  – You will stand out if you know exactly how much traffic you will exchange at peak with a peer
  – Protect your peer’s interests
    • Discuss mutual points of interconnection that suit both
    • Respond to abuse complaints promptly
    • Use contacts to reach other peering co-ordinators
    • Respond promptly to BGP session down/flapping
    • List your network on PeeringDB!
Publish in IRRDB

- Publish your circuit and adjacency information in IRRDB
- Make it possible for other networks to figure out how you are *intending* your traffic to be routed
AS2 is your largest flow - via PP2 - maybe needs a second private peer backup on RT2?
AS2 is your largest flow - via PP2 - maybe needs a second private peer backup on RT2? AS1 via PP1, configure a backup over EX1 or EX2 for deterministic routing?
AS2 is your largest flow - via PP2 - maybe needs a second private peer backup on RT2? AS1 via PP1, configure a backup over EX1 or EX2 for deterministic routing? Can you move larger peers behind EX1 and EX2 onto private peering?
AS2 is your largest flow - via PP2 - maybe needs a second private peer backup on RT2? AS1 via PP1, configure a backup over EX1 or EX2 for deterministic routing? Can you move larger peers behind EX1 and EX2 onto private peering? If there is an exchange failure, where will the traffic go? How big a flow should you care about?
AS2 is your largest flow - via PP2 - maybe needs a second private peer backup on RT2?
AS1 via PP1, configure a backup over EX1 or EX2 for deterministic routing?
Can you move larger peers behind EX1 and EX2 onto private peering?
If there is an exchange failure, where will the traffic go? How big a flow should you care about?
If you lose RT2, how will traffic to PP3 and traffic volume via EX2 be delivered?
Buying transit in a smart way

• **Buying from a well peered transit provider:**
  – Can improve quality for the reasons discussed
  – Hides capacity problems from you automatically

• **Buying from your top traffic destination**
  – If your business relies on the traffic quality, it may make sense to pay
  – **Data** may help you negotiate good terms
Dealing with a “no” to peering

• Paid peering is one option
  – Often more expensive than full IP transit
  – “Once a customer, never a peer”

• Pay for other services in return for peering
  – Transport for example

• Peer around the problem
  – Try to peer directly with downstream customers
  – Try to sell directly to downstream customers
    • If you are better peered, you can sell based on quality
Constantly manage

• Peering on the Internet changes every day.
• Capacity on the Internet grows every day.
• Small networks become large.
• Large networks become larger (consolidation)
• A “bad” path might become good overnight
Questions?

---

Andy Davidson
andy@nosignal.org

Email me to request a copy of this presentation!

Feedback and introduction to peering co-ordinators welcome

Twitter: @andyd