BGP Traffic Engineering

Andy Davidson
CTO @Allegro Networks
Director @ LONAP, IXLeeds

andy.davidson@allegro.net
AFPIF 2013, Casablanca, Morocco
3rd September 2013
Why do Traffic Engineering?

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

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 ASN**s
You need data

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 – Incompatible with Netflow, typically implemented by switch vendors.
• Jflow – Essentially Netflow on Junipers
Enabling Netflow (example)

```
ip route-cache flow
Enables Netflow on an Interface

ip flow-export version X origin-as
Defines Netflow options

ip flow-export destination <ip> <port>
Defines the collector address

ip flow-export source loopback0
For consistent source IP addressing
```
6500/7600 sup720 Netflow

mls netflow interface
mls flow ip interface-full
mls flow ipv6 interface-full
mls nde sender
ip flow-capture mac-addresses
ip flow-export version 9 origin-as
ip flow-export destination 192.0.2.100 5500 vrf vrf-netflow
ip flow-top-talkers

interface GigabitEthernet1/1
  ip flow ingress

Order that you enter configuration matters.

With special thanks to Nick Hilliard of INEX for this config
XR Flexible Netflow

flow exporter-map fem-default
  version v9
    options interface-table timeout 300
    options sampler-table timeout 300

transport udp 5500
source Loopback0
destination 192.0.2.100

flow monitor-map fmm-ipv4
  record ipv4
  exporter fem-default
  cache entries 1000000

sampler-map sm-flow-default
  random 1 out-of 100

interface TenGigE0/0/2/2
  flow ipv4 monitor fmm-ipv4 sampler sm-flow-default ingress

router bgp 65533
  address-family ipv4 unicast
    bgp attribute-download

With special thanks to Nick Hilliard of INEX for this config
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...
<table>
<thead>
<tr>
<th>IP Address</th>
<th>Time and date</th>
<th>Amount of Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>90.155.42.152</td>
<td>[01/Sep/2013:19:09:50 +0100]</td>
<td>304 189</td>
</tr>
<tr>
<td>90.155.42.152</td>
<td>[01/Sep/2013:19:10:10 +0100]</td>
<td>200 212</td>
</tr>
<tr>
<td>90.155.42.152</td>
<td>[01/Sep/2013:19:10:29 +0100]</td>
<td>6782</td>
</tr>
<tr>
<td>90.155.42.152</td>
<td>[01/Sep/2013:19:11:02 +0100]</td>
<td>200 5216</td>
</tr>
<tr>
<td>90.155.42.152</td>
<td>[01/Sep/2013:19:11:33 +0100]</td>
<td>200 5216</td>
</tr>
<tr>
<td>90.155.42.152</td>
<td>[01/Sep/2013:19:12:18 +0100]</td>
<td>304 189</td>
</tr>
<tr>
<td>90.155.42.152</td>
<td>[01/Sep/2013:19:13:13 +0100]</td>
<td>200 7525</td>
</tr>
<tr>
<td>90.155.42.152</td>
<td>[01/Sep/2013:19:13:31 +0100]</td>
<td>304 189</td>
</tr>
</tbody>
</table>
Other ways to get “data”

• Wild Guess
  – Your instinct is better than you think?
  – Content networks will talk to eyeballs
  – Eyeball networks will talk to content
  – Confirm with top talkers, etc.
  – But you should use Netflow. 😊
    • Early “quick wins” may provide hard data
    • Hard data provides stronger business case
Data tells you

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

• Your **top traffic originators** or **destinations**
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.
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.
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?
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?
If you lose RT1, how will traffic volume via PP3 and EX1 be delivered?
Localpref – blunt hammer

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

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
Configuration Example

Larger flows are in ASNs
Listed in as-path 30 and 40

Deterministic exits configured
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
If you understand your top flows, you will cope with traffic growth and failures.
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.
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

- 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 prepending 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>
2.4 **Restrict Route Propagation**

A customer network may use these communities to restrict propagation of its routes to AS3320 peers. However, the well known community NOPEER should be employed instead of these where appropriate.

<table>
<thead>
<tr>
<th>Community Value</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>65010 : X</td>
<td>No Export to AS X</td>
<td>Do not advertise route(s) to named AS3320 peer (ASN=X)</td>
</tr>
<tr>
<td>65010 : 65001</td>
<td>No Export by Class: Peer</td>
<td>Do not advertise route(s) to AS3320 peers.</td>
</tr>
<tr>
<td>65010 : 65002</td>
<td>No Export by Class: Upstream</td>
<td>Do not advertise route(s) to AS3320 upstream.</td>
</tr>
<tr>
<td>65010 : 65003</td>
<td>No Export by Class: Peer &amp; Upstream</td>
<td>Do not advertise route(s) to AS3320 peers and upstream.</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.

```
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.
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!
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
Aggregate transit & peering capacity

• Buy transit/peering capacity through a reseller who can offer many providers on a single link
  – Different providers presented on separate VLANs
  – Failures in a transit or peering will result in traffic shifting to another provider on same link
  – Access to multiple providers on single commit?
  – Not available everywhere, but Allegro offer this in London

• Does not replace need for backup to reseller
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.davidson@allegro.net
Email me to request a copy of this presentation!

Feedback and introduction to peering co-ordinators welcome

Twitter: @andyd
+44 161 200 1610 (Manchester, UK office hours)