ADD-PATH OVERVIEW

Dave Ward, John Scudder

NANOG 48, February 23, 2010

PROBLEM STATEMENT

BGP has implicit withdraw semantics

- On a peering session, an advertisement of a given prefix replaces any previous announcement of that prefix
  - If the prefix completely goes away, then it's explicitly withdrawn

BGP scaling techniques are widely used

- Route reflectors, confederations

Combined, these result in data hiding

- Available backup routes are hidden
- May be good for scaling… but problematic in other ways

USE CASES

Fast convergence, robustness and graceful shutdown schemes that require backup paths

- Because backup paths get “eaten” by route reflectors

Stability and correctness schemes that require additional paths

- For example fixes for MED oscillation or MED misrouting

Multipath schemes that require multiple next hops

And, implicit withdraw alone is potentially a problem for some types of inter-AS backup schemes

This is not an exhaustive list! Just examples.

SOLUTION SPACE

Problem space has two parts

- Implicit withdraw
- Scaling techniques (RRs, Confeds)

Implies solution can attack either (or both)

Add-path attacks implicit withdraw

- Because applicability is not limited by deployment scenario
  - Goal: general tool, not point solution
- Orthogonal to any changes to scaling techniques
  - So, can potentially be combined

ADD-PATH IN A NUTSHELL

Add a path identifier as part of the NLRI

- Very similar to Route Distinguisher in RFC 2547/4364 VPNs, but applicable to all address families

ADD-PATH IN DETAIL — CAPABILITY EXCHANGE

Peers exchange add-path capability

```
+------------------------------------------------+
| Address Family Identifier (2 octets)           |
+------------------------------------------------+
| Subsequent Address Family Identifier (1 octet) |
+------------------------------------------------+
| Send/Receive (1 octet)                         |
+------------------------------------------------+
```

- For each AFI/SAFI on the session, indicates whether to use add-path for receive, transmit, or both
- Implications:
  - Can choose to use add-path for only certain address families
  - Can choose to use add-path for only certain peerings, in selected direction

ADD-PATH IN DETAIL —
NLRI ENCODING

Each NLRI that is using the new encoding gets a Path Identifier

- Example, RFC 4271 (BGP-4, IPv4 prefix) looks like this:

```
 +--------------------------------+
 | Path Identifier (4 octets)     |
+--------------------------------+
| Length (1 octet)               |
+--------------------------------+
| Prefix (variable)              |
+--------------------------------+
```

- Path Identifier can be used to prevent a route announcement from implicitly withdrawing a previous one

ADD-PATH IN DETAIL — PATH IDENTIFIER USAGE

Path Identifier is chosen locally

- Only unique to a peering session
- Typically, automatically generated by implementation — no configuration involved

Example prefix encoding

- Normal BGP IPv4 route is identified by prefix: 10/8
- With add-path, identified by prefix and Path ID: (10/8, ID=1) is different from (10/8, ID=2)

REMINDER — BEST-EXTERNAL

Advertise best EBGP path into IBGP even if not using it as overall best

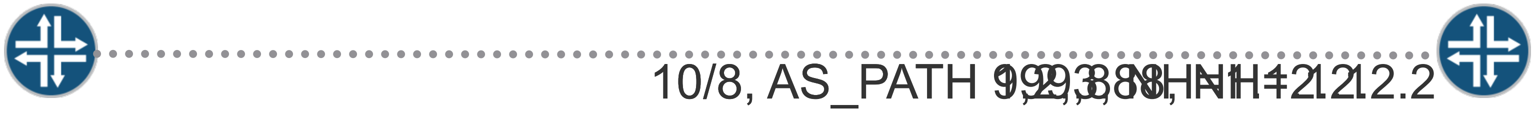
Analogous rules for route reflectors

- Advertise best client route to non-clients
- Advertise best non-client route to clients
- Requires full meshing of clients if used on reflector towards clients

Potentially useful on border routers even if add-path used within the AS

OPERATION

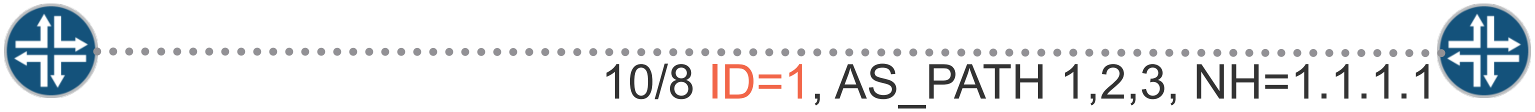
Conventional BGP



10/8, AS_PATH 1,2,3, NH=1.1.1.1

10/8, AS_PATH 999,888, NH=2.2.2.2

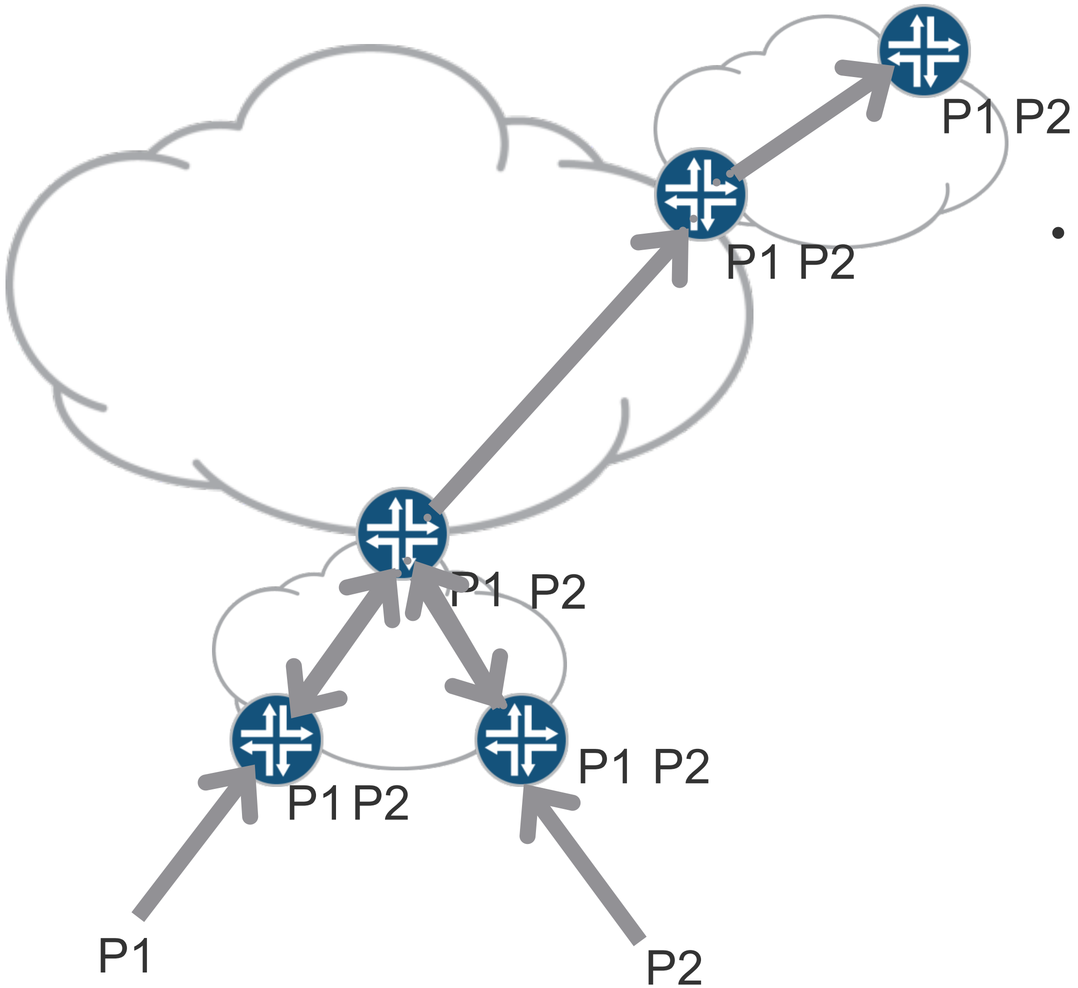
Add-Path



10/8 ID=1, AS_PATH 1,2,3, NH=1.1.1.1

10/8 ID=2, AS_PATH 999,888, NH=2.2.2.2

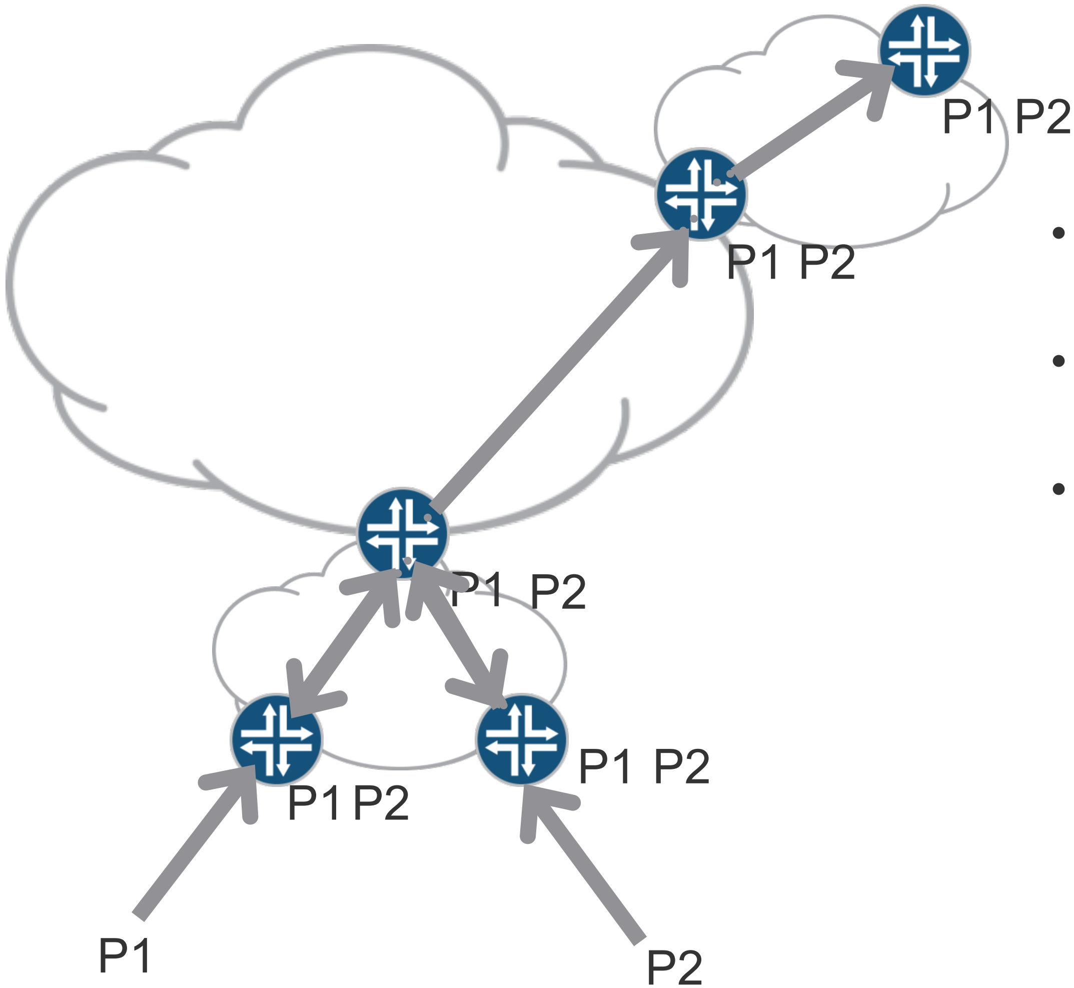
OPERATION — CONVENTIONAL BGP



- Worst-case failure of P2 takes five rounds to repair and causes EBGP route flap

Note: only single reflector per POP shown for simplicity

OPERATION — ADD-PATH



- PEs request add-path from RR
- PEs use best-external towards RR
- RRs use add-path second-best mode towards each other and PEs

Note: only single reflector per POP shown for simplicity

MEMORY OVERHEAD BACK-OF-ENVELOPE

Obvious: Additional paths → Memory overhead

Less obvious: Most overhead is at route reflectors

- Assume a configuration where RRs send best and second-best
- At worst, 2x on PEs (existing best path, plus second-best)
  - But PE sees at worst one full routing table from each of its RRs to begin with… typically two RRs
  - Most RRs see more routes than this today… implies PE can take it (assuming similar control plane hardware on PE and RR)
- On RRs, also 2x
  - RR also sees at worst one full routing table from each of its peer RRs… but typically, more peer RRs
  - Fortunately, RRs are easiest to scale up using larger (including outboard) control plane hardware

FURTHER NOTES ON MEMORY

Number of paths to be advertised is under operator control

- Fine tuning is possible, and advised!

In deployments that we’ve shown, no impact on global Internet routing

- Because add-path only used on IBGP

Overhead is purely control plane, not forwarding plane

- Unless you want some flavor of fast reroute in which case, some FIB overhead is inevitable (but payoff is good)

DEPLOYMENT CONSIDERATIONS

Path selection consistency is important

- Doubly so in traditional IP networks

Analysis shows selection to be consistent when border routers don’t advertise more than one path

- See draft-pmohapat-idr-fast-conn-restore-00

SOME NOTES ON SCALING

Memory is one scaling axis

- A deep route reflection hierarchy minimizes memory utilization
- But converges like a dog, relatively speaking

Convergence/restoration is another

- A flat IBGP mesh (with best-external) converges well
- But hides no routes at all

Ideally, find the “sweet spot” between the two

- Add-path enables tuning between the two extremes

CONCLUSION

Powerful tool with broad applicability

Clear benefits for

- Intra-domain deployment
- Fast restoration
- Stability

Other uses not yet explored

REFERENCES AND RELATED WORK

draft-ietf-idr-add-paths-02.txt

draft-pmohapat-idr-fast-conn-restore-00

draft-walton-bgp-route-oscillation-stop-02

draft-ietf-idr-best-external-00.txt

draft-vvds-add-paths-analysis-00

draft-ietf-grow-bgp-graceful-shutdown-requirements-01 everywhere