BGP Opaque AFI

Petr Lapukhov,
Facebook
petr@fb.com

Use case (3)

- BGP is getting more widespread in DC
  - Used for routing, programming, monitoring
- 3rd party apps run on network devices
  - BGP already available
- Apps need to coordinate/exchange data
- Examples:
  - Discovery
  - Share some state (e.g. link state)
  - Ad-hoc resource allocation

Why use BGP? (3)

- Re-use existing transport system
- Already transports non-routing data
  - Effectively “broadcast state”
  - E.g.: flow-spec, L2 VPN discovery
- BGP filtering is flexible
  - E.g. communities, ORF, etc
- Best-path selection still performed

Implementation: new AFI (4)

- "Opaque data" AF
- New SAFI named "Key-Value binding"
  - Key = NLRI
  - Value = new optional non-transitive attribute
- To announce a binding send:
  - MP_REACH_NLRI + [OPAQUE_VALUE_ATTR]
  - (attribute associated with NLRI)
- To remove a binding, send
  - MP_UNREACH_NLRI

Client (on-box) API

- **Option 1**
  - A thread in client maintains BGP sessions
    - Opaque AFI + other AFIs enabled if needed
  - Client originates UPDATE messages
  - Bonus: state change notifications
- **Option 2**
  - Standalone BGP injector, e.g. ExaBGP
  - Some 3rd party API is used:
  - REST, Thrift, ZMQ, text -> file

Challenges

- Key + Value size limited to ~4K
  - *draft-ietf-idr-bgp-extended-messages*
- Ideally, Value should be in NEXT_HOP
  - Size is limited to 256 bytes ☹
- UPDATE packing
  - One MP_REACH_NLRI per UPDATE
  - … Due to OPAQUE_VALUE attribute

Challenges (cont.)

- Key name collisions
  - Intentionally kept out of scope
  - UUID scheme could be used
  - ASN + Originator-ID could be used

Next steps

- Encode both Key and Value in NLRI field
  - No optional attribute needed
  - Value portion treated as “attribute” (not as a key in RIB)
- Withdraw only announces key string, not value (keyed in RIB)
- Add new VPN-Key-Value SAFI to support RD for enforcing key uniqueness