NSWI184 – Řízení počítačových sítí

Přednáška šestá

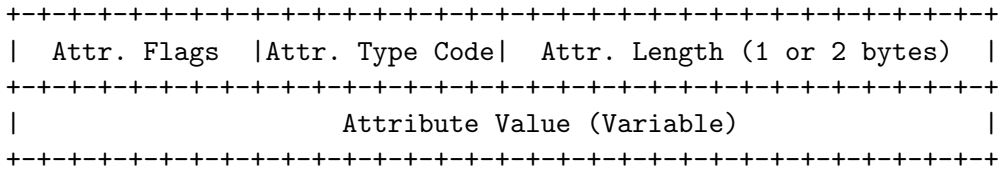
Ondřej Zajíček, Kateřina Kubecová

2025-11-05

BGP Path Attributes

- TLV (type, length, value) format
- Type consists of flags and type code
- Each type only once

Attribute Format



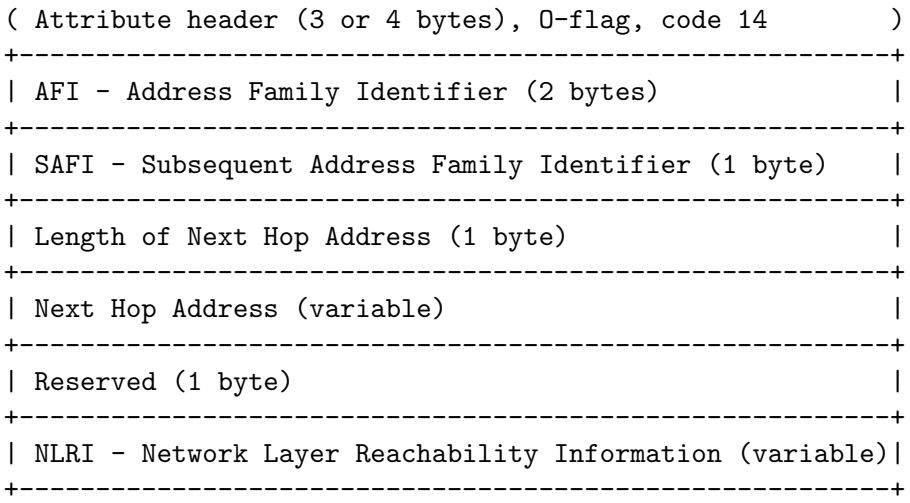
Attribute Flags (8 bits)

- Optional (O)
- Transitive (T)
- Partial (P)
- Extended Length (E)

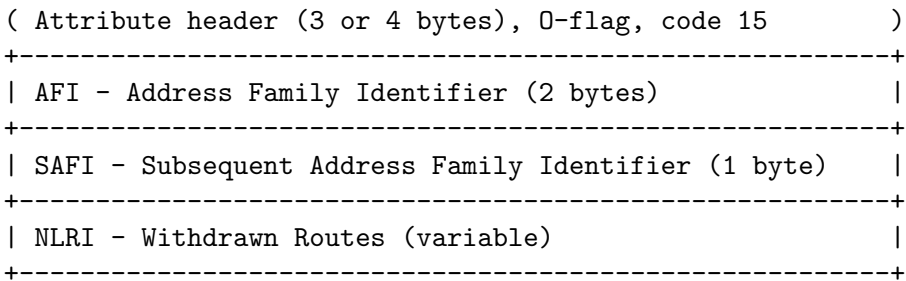
Multiprotocol BGP

- RFC 4760
- MP_REACH_NLRI / MP_UNREACH_NLRI attributes
- Replaces both NLRI and Next Hop attribute
- Multiple ‘streams’ in one BGP session
- Endpoint addresses vs. payload NLRI

MP_REACH_NLRI Attribute



MP_UNREACH_NLRI Attribute



IPv6 in Multiprotocol BGP

- RFC 2545
- NLRI is the same (prefix length, prefix)
- But what about BGP next hop?
- Global-scope IP needed to resolve IGP route
- Link-local IP preferred when used as immediate next hop
- Solution – send both!
- Next hop length 16 or 32 bytes
- Global-only (IBGP, multihop)
- Global + link-local (EBGP, direct)

Link-local BGP

- Non-standard extension
- Direct BGP session over link-local addresses
- No global-scope next hop, only link-local one
- Various encodings (LL, ::/LL, LL/LL)

BGP Route Refresh

- Import policy may reject some routes
- Import policy may change
- Need to re-evaluate all received routes
- ROUTE-REFRESH message (RFC 2918)
- Request for re-transmit of all routes (re-feed)
- Limited to specific AFI/SAFI
- Associated capability

BGP Enhanced Route Refresh

- Export policy may change, triggering re-feed
- Problem: missing demarcation for re-feed
- Re-feed have to be differential
- We may want just to send the new state

BGP Enhanced Route Refresh (continued)

- BoRR / EoRR messages (RFC 7313)
- Extension of ROUTE-REFRESH message
- Used in the other direction
- BoRR signals begin of re-feed
- EoRR signals end of re-feed
- Routes not mentioned during re-feed are implicitly withdrawn
- Limited to specific AFI/SAFI
- Associated capability

BGP Graceful Restart

- RFC 4724
- Maintain forwarding during restart of BGP control plane
- Support negotiated via capabilities
- Silent TCP reset $\rightarrow$ graceful restart
- NOTIFICATION message $\rightarrow$ regular restart
- Restarting server keeps data plane (FIB)
- Neighbors keep routes in routing table
- After restart, routers wait for full route exchange
- End of initial feed marked with End-of-RIB
- Timeout in case of something gone wrong

BGP Long-Lived Graceful Restart

- RFC 9494
- Extends graceful restart mechanism for longer periods
- Applies after regular graceful restart timeout
- Routes are still kept, but marked with LLGR_STALE community
- Stale routes are depreferenced in best path selection
- Removed after Long lived stale time
- Configured separately for each AFI/SAFI

BGP Add-Path

- RFC 7911
- Allows BGP speakers to advertise multiple paths for the same prefix
- Path Diversity: multiple backup paths available for faster convergence
- Load Balancing: enables traffic distribution across multiple paths
- Extends NLRI by 32bit Path Identifier to distinguish advertisements
- Negotiated via BGP capabilities, independent for each direction

Error Handling - Syntactic Errors

- RFC 7606
- Session reset
- AFI/SAFI disable
- Treat-as-withdraw
- Attribute discard

Error Handling - Semantic Errors

- AS path loop
- Route reflector loop
- Unresolvable next hop
- OTC / role mismatch
- Result: Path ineligible
- Optional peer-as-check

BGP Limits

- Attribute length: 64k
- Message length: 4k
- Extended message length (RFC 8564): 64k
- Options (capabilities) in OPEN: 255 bytes
- Extended options (RFC 9072): 64k
- But OPEN still max 4k

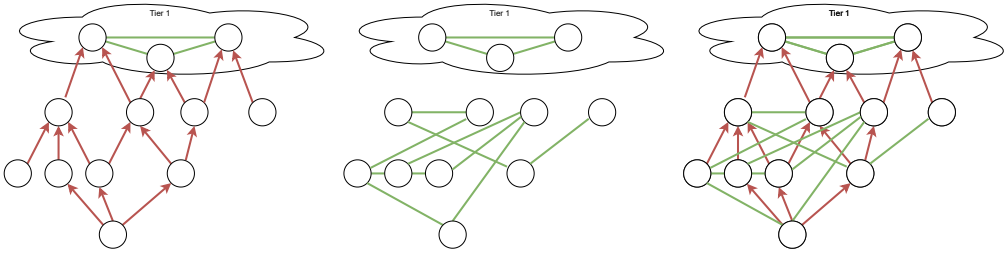
Internet Topology

- Global network of autonomous systems is not arbitrary graph
- It has hierarchical structure, but not simple one
- Many overlapping and interconnected hierarchies
- Transit connectivity - connectivty to whole internet

Roughly three kinds of relationship between ASes

- Upstream: provides you transit connectivity, provider
- Downstream: you provide them transit connectivity, customer
- Peer: provides you connectivity to them and their downstream, and vice versa

Example Internet Topology



Tier 1 Networks

- Top-level ISPs in the internet hierarchy
- No upstream providers
- Settlement-free peering with all other tier 1 networks
- Global backbone infrastructure
- Direct connections to all major internet exchanges

Internet Exchange Points (IXPs)

- Infrastructure where multiple networks interconnect
- Replace direct peering links with shared switched network
- Direct BGP sessions between border routers
- Or BGP sessions to route servers

Route Servers

- Replace mesh of BGP sessions in IXP with central point
- Facilitate BGP routing exchange between border routers
- Do not forward traffic, only exchange routing information

Filtering Policy

Depends on kind of relationship between ASes

- Upstream: import everything (except your network), export only you and your downstreams
- Downstream: import only expected networks, export everything
- Peer: import only expected networks, export only you and your downstreams

Route leaks: Routes exported contrary to the expected relationship, i.e., routes from an upstream or a peer exported to another upstream or peer.

BGP Roles

- RFC 9234
- Formalization of inter-AS relationships
- Five defined roles (provider, customer, peer, RS, RS-client)
- Prevents BGP session with mismatched roles
- Prevents route leaks with OTC attribute
- Attached when route is propagated downstream or to a peer
- Route with OTC attribute can be propagated only to customers

Internet Routing Registry (IRR) Databases

- Public databases of address ranges and routing policies of ASes
- Source of info which networks to expect from a peer or customer
- Routing Policy Specification Language (RFC 2622)
- Unfortunately not fully reliable

Resource Public Key Infrastructure (RPKI)

- Problem: Anyone can originate routes regardless of IP allocation
- RPKI: Resources are cryptographically signed by RIRs
- Chain of trust corresponding to how resources are distributed
- RPKI cache – collects and validates resource certificates
- Routers download records from RPKI cache and validate BGP routes
- RPKI-RTR protocol (RFC 6810)

Route Origin Authorizaton (ROA)

- ROA record: Authorize ASNs to originate routes
- (ASN, prefix, max-length)
- Routes validated against these records
- Result of ROA validation used in routing policy
- Prevents accidental leaks of internal routes
- Not bulletproof againts intentional attacks

ROA validation

- Find ROA records that cover route prefix
- Compare last ASN from AS_PATH to ROA ASN
- Compare route prefix length to max-length
- Any ROA matches prefix and passes checks $\rightarrow$ Valid
- Any ROA matches prefix but fails checks $\rightarrow$ Invalid
- No ROA matches prefix $\rightarrow$ Unknown

AS Provider Authorizaton (ASPA)

- ASPA record: Describe a valid set of AS providers
- (ASN, provider-ASN), authorized by ASN
- AS_PATH attribute validated against these records
- Result of ASPA verification used in routing policy
- Should be combined with ROA validation and peer ASN check