NSWI184 – Řízení počítačových sítí

Přednáška pátá

Ondřej Zajíček, Kateřina Kubecová

2025-10-29

Introduction to BGP

- **Border Gateway Protocol**
- RFC 4271 (BGPv4)
- Path vector routing protocol
- Exchange routing information on Internet
- Key characteristic: Policy-based routing

Path Vector Routing

- Similar to distance vector routing
- Keep whole path instead of just metric
- Simple loop avoidance
- No counting to infinity

BGP Types

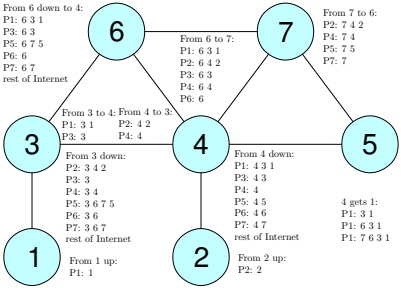
- External BGP (EBGP): Between different ASes
- Internal BGP (IBGP): Within the same AS
- Different rules and behaviors

Autonomous System

- Collection of networks under single administrative domain
- Presents unified routing policy to the internet
- Implements consistent routing decisions across its networks
- Internally connected
- Identified by Autonomous System Number (ASN)
- 2-byte ASNs: 1–65534 (64512–65534 private)
- 4-byte ASNs: RFC 6793 (4.2G–($2^{32} - 2$) private)
- Assigned by IANA / RIRs (Regional Internet Registries)
- Private ASNs: For internal use only

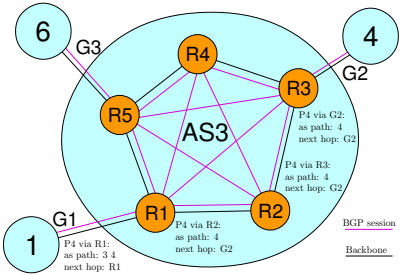
External BGP

By abstracting each autonomous system as a single node, EBGP performs inter-AS routing in the same manner as RIP or Babel performs routing between individual routers.



Internal BGP

- Within AS, IGP (OSPF) is used for local routing
- Within AS, IBGP is used to distribute global BGP routes to all border routers
- Full mesh: Each BGP speaker has session with all other BGP speakers in AS



BGP Connection

- Single TCP session, port 179
- Manually configured, no neighbor detection
- Direct connection (EBGP) or multi-hop (IBGP)
- Usually symmetric, both sides try to connect
- (passive mode / collision detection)
- Neighbor also called ‘peer’

BGP Basic Operation - Session

- Establish session
- Initial route updates (feed)
- End-of-RIB (RFC 4724)
- Incremental route updates
- Advertised routes valid until withdrawn
- Session termination withdraws everything

BGP Basic Operation - Router

- Receive routes (Adj-RIB-In)
- Apply import policy / filter
- Best path selection (Local-RIB)
- Apply export policy / filter
- Advertise routes (Adj-RIB-Out)

BGP Messages

1. OPEN: Initial message, parameter negotiation
2. UPDATE: Advertise or withdraw routes
3. NOTIFICATION: Terminating message
4. KEEPALIVE: Ensure liveness
5. ROUTE-REFRESH: Ask for re-feed (RFC 2918)

BGP State Machine

- Idle: Neither accepting nor initiating connection
- Connect: TCP connection initiated, waiting for TCP establishment
- Active: Ready to accept TCP connection, not trying to connect
- OpenSent: Connection established, OPEN sent, waiting for OPEN
- OpenConfirm: OPEN messages exchanged, waiting for KEEPALIVE
- Established: Normal BGP operation, routes are advertised

BGP Timers

- ConnectRetryTimer (120 s) - Retry TCP connect attempt
- KeepaliveTimer ($1/3 \times$ HT) - Periodic sending of KEEPALIVE
- HoldTimer (90 s) - Waiting for KEEPALIVE
- SendHoldTimer ($2 \times$ HT) - Last message sent (RFC 9687)
- (MinASOriginationIntervalTimer)
- (MinRouteAdvertisementIntervalTimer)

BGP Message Header

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                                                               |
+                                                               +
|                           Marker                              |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Length               |       Type     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

BGP Session Security

- Based on TCP extensions
- MD5 authentication
- TCP-AO authentication
- TTL security

BGP OPEN Message

- Initial handshake
- Exchange of ASNs and Router IDs
- Negotiation of HoldTimer
- Negotiation of capabilities

BGP Capabilities

- Negotiate optional features during session establishment
- Protocol extensibility with backwards compatibility
- Unsupported capabilities are ignored
- Each capability has 1-byte code and variable-length data

Common capabilities:

Code	Capability	Description
1	Multiprotocol	Support for non-IPv4 address families
2	Route Refresh	Dynamic route table refresh without session reset
64	Graceful Restart	Maintain forwarding during BGP restart
65	4-byte ASN	Support for 32-bit AS numbers
69	ADD-PATH	Advertise multiple paths for same prefix

BGP KEEPALIVE Message

- Send periodically by KeepaliveTimer
- Expected by HoldTimer
- Error code 4: Hold Timer Expired

BGP NOTIFICATION Message

- Send to terminate BGP session
- Error code and subcode, optional data
- Error code 6: Cease (non-error termination)
- Data are often part of message containing error
- Text message for administrative shutdown (RFC 9003)

BGP UPDATE Message

```
+-----------------------------------------------------+
|   Withdrawn Routes Length (2 octets)                |
+-----------------------------------------------------+
|   Withdrawn Routes (variable)                       |
+-----------------------------------------------------+
|   Total Path Attribute Length (2 octets)            |
+-----------------------------------------------------+
|   Path Attributes (variable)                        |
+-----------------------------------------------------+
|   Network Layer Reachability Information (variable) |
+-----------------------------------------------------+
```

- Withdrawn routes (NLRI)
- Advertised routes (NLRI and path attributes)

BGP Route

- Key and value pair
- Key: NLRI (e.g. IP prefix)
- Value: List of path attributes
- UPDATE messages add, remove or update key-value mapping

BGP NLRI

- Network Layer Reachability Information
- Describes route destinations
- Multiple types of NLRI, identified by AFI/SAFI pairs
- Basic type IPv4-Unicast (1/1):
  - Prefix length (1 byte)
  - Prefix (0-4 bytes)
  - E.g. 192.168.1.0/24 $\rightarrow$ [24, 192, 168, 1]
- Other types require Multiprotocol capability

BGP AFI / SAFI

Address Family Identifiers:

- AFI 1: IPv4
- AFI 2: IPv6
- AFI 25: Layer 2

Subsequent Address Family Identifiers:

- SAFI 1: Unicast
- SAFI 2: Multicast
- SAFI 4: Unicast with MPLS labels
- SAFI 70: Ethernet VPN
- SAFI 128: MPLS-labeled L3VPNs
- SAFI 133: Flow Specification rules

BGP Path Attributes

- Additional information for routes
- Key for best path selection and policy filtering
- Some well-known, other optional
- Some mandatory, other discretionary
- Some transitive, other non-transitive
- Some EBGP-only, some IBGP-only
- Each has defined rules for origination and propagation

BGP Well-Known Attributes

- AS_PATH: Sequence of AS numbers
- NEXT_HOP: IP address of next hop
- LOCAL_PREF: Local preference
- MULTI_EXIT_DISC: Preference of entry points into AS
- (ORIGIN, ATOMIC_AGGREGATE, AGGREGATOR): Mostly irrelevant

AS Path Attribute

- AS_PATH / bgp_path
- Sequence of AS numbers encoding path from origin AS
- Newest is first, oldest (origin AS) is last
- Loop detection: Reject if own ASN in path
- Path selection: Default / primary metric
- Prepending: Artificially increase path length
- Mandatory

BGP Next Hop Attribute

- NEXT_HOP / bgp_next_hop
- IP address of BGP next hop
- Not to be confused with immediate next hop
- Resolve in IGP routing table to get immediate next hop
- EBGP: Usually changed to IP of advertising router
- IBGP: Usually forwarded unmodified
- Mandatory

Local Preference Attribute

- LOCAL_PREF / bgp_local_pref
- Local preference for given route
- 32-bit number, higher is better
- Assigned by local policy to EBGP routes
- Mandatory on IBGP, forbidden on EBGP

Multi-Exit Discriminator (MED) Attribute

- MULTI_EXIT_DISC / bgp_med
- Suggest preferred entry point to AS
- 32-bit number, lower is better
- Assigned by local policy to routes sent to EBGP
- Forwarded only to neighboring ASes
- Comparison only between paths from same AS
- Causes non-transitivity in path selection
- Sometimes implemented in non-compliant way

BGP Best Path Selection

1. Local Preference (higher better)
2. AS Path length (shorter better)
3. (Origin attribute)
4. MED (lower better, only between paths from same AS)
5. Prefer EBGP route over IBGP route
6. IGP metric to BGP Next Hop
7. Disambiguation (Neighbor router ID, IP address)