CSCI-1680
Network Layer:
Inter-domain Routing

Rodrigo Fonseca
Today

- **Last time: Intra-Domain Routing (IGP)**
  - RIP distance vector
  - OSPF link state
- **Inter-Domain Routing (EGP)**
  - Border Gateway Protocol
  - Path-vector routing protocol
Why Inter vs. Intra

• Why not just use OSPF everywhere?
  – E.g., hierarchies of OSPF areas?
  – Hint: scaling is not the only limitation

• BGP is a policy control and information hiding protocol
  – intra == trusted, inter == untrusted
  – Different policies by different ASs
  – Different costs by different ASs
Types of ASs

- **Local Traffic** – source or destination in local AS
- **Transit Traffic** – passes through an AS
- **Stub AS**
  - Connects to only a single other AS
- **Multihomed AS**
  - Connects to multiple ASs
  - Carries no transit traffic
- **Transit AS**
  - Connects to multiple ASs and carries transit traffic
AS Relationships

- How to prevent X from forwarding transit between B and C?
- How to avoid transit between CBA?
  - B: BAZ -> X
  - B: BAZ -> C? (=> Y: CBAZ and Y:CAZ)

Example from Kurose and Ross, 5th Ed
Choice of Routing Algorithm

• Constraints
  – Scaling
  – Autonomy (policy and privacy)

• Link-state?
  – Requires sharing of complete information
  – Information exchange does not scale
  – Can’t express policy

• Distance Vector?
  – Scales and retains privacy
  – Can’t implement policy
  – Can’t avoid loops if shortest path not taken
  – Count-to-infinity
Path Vector Protocol

• **Distance vector algorithm with extra information**
  – For each route, store the complete path (ASs)
  – No extra computation, just extra storage (and traffic)

• **Advantages**
  – Can make policy choices based on set of ASs in path
  – Can easily avoid loops
BGP - High Level

• Single EGP protocol in use today
• Abstract each AS to a single node
• Destinations are CIDR prefixes
• Exchange prefix reachability with all neighbors
  – E.g., “I can reach prefix 128.148.0.0/16 through ASes 44444 3356 14325 11078”
• Select a single path by routing policy
• Critical: learn many paths, propagate one
  – Add your ASN to advertised path
Why study BGP?

• Critical protocol: makes the Internet run
  – Only widely deployed EGP

• Active area of problems!
  – Efficiency
  – Cogent vs. Level3: Internet Partition
  – Spammers use prefix hijacking
  – Pakistan accidentally took down YouTube
  – Egypt disconnected for 5 days
BGP Example

AS 1
1.2.0.0/16

Only 1 Router
Per AS (for now)
BGP Example

Only 1 Router Per AS (for now)
BGP Example

Only 1 Router Per AS (for now)
BGP Example

Only 1 Router Per AS (for now)
BGP Example

AS 1
1.2.0.0/16

AS 2

AS 3

AS 4

AS 5

Only 1 Router Per AS (for now)
BGP Protocol Details

• **Separate roles of speakers and gateways**
  – Speakers talk BGP with other ASs
  – Gateways are routes that border other ASs
  – Can have more gateways than speakers
  – Speakers know how to reach gateways

• **Speakers connect over TCP on port 179**
  – Bidirectional exchange over long-lived connection
BGP Implications

• **Explicit AS Path == Loop free**
  – Except under churn, IGP/EGP mismatch

• **Reachability not guaranteed**
  – Decentralized combination of policies

• **Not all ASs know all paths**

• **AS abstraction -> loss of efficiency**

• **Scaling**
  – 37K ASs
  – 350K+ prefixes
  – ASs with one prefix: 15664
  – Most prefixes by one AS: 3686 (AS6389, BellSouth)
Integrating EGP and IGP

• **Stub ASs**
  – Border router clear choice for default route
  – Inject into IGP: “any unknown route to border router”

• **Inject specific prefixes in IGP**
  – E.g., Provider injects routes to customer prefix

• **Backbone networks**
  – Too many prefixes for IGP
  – Run internal version of BGP, iBGP
  – All routers learn mappings: Prefix -> Border Router
  – Use IGP to learn: Border Router -> Next Hop
iBGP

AS 1
1.2.0.0/16

AS 2

AS 3

AS 4

AS 5

Only 1 Router Per AS (for now)
iBGP keeps AS consistent
Multiple Peering Points!
BGP Messages

• Base protocol has four message types
  – OPEN – Initialize connection. Identifies peers and must be first message in each direction
  – UPDATE – Announce routing changes (most important message)
  – NOTIFICATION – Announce error when closing connection
  – KEEPALIVE – Make sure peer is alive

• Extensions can define more message types
  – E.g., ROUTE-REFRESH [RFC 2918]
Anatomy of an UPDATE

- Withdrawn routes: list of withdrawn IP prefixes
- Network Layer Reachability Information (NLRI)
  - List of prefixes to which path attributes apply
- Path attributes
  - ORIGIN, AS_PATH, NEXT_HOP, MULTI-EXIT-DISC, LOCAL_PREF, ATOMIC_AGGREGATE, AGGREGATOR, ...
  - Each attribute has 1-byte type, 1-byte flags, length, content
  - Can introduce new types of path attribute – e.g., AS4_PATH for 32-bit AS numbers
Example

- **NLRI**: 128.148.0.0/16
- **AS Path**: ASN 44444 3356 14325 11078
- **Next Hop IP**: same as in RIPv2
- **Knobs for traffic engineering**:  
  - Metric, weight, LocalPath, MED, Communities  
  - Lots of voodoo
BGP State

• BGP speaker conceptually maintains 3 sets of state
  
  • Adj-RIB-In  
    – “Adjacent Routing Information Base, Incoming”  
    – Unprocessed routes learned from other BGP speakers  
  
  • Loc-RIB  
    – Contains routes from Adj-RIB-In selected by policy  
    – First hop of route must be reachable by IGP or static route  
  
  • Adj-RIB-Out  
    – Subset of Loc-RIB to be advertised to peer speakers
Demo

• **Route views project:** [http://www.routeviews.org](http://www.routeviews.org)
  – telnet route-views.linx.routeviews.org
  – show ip bgp 128.148.0.0/16 longer-prefixes

• **All paths are learned internally** (iBGP)

• **Not a production device**
Next class

- BGP Policy Routing and Security