CSCI-1680
Network Layer:
Inter-domain Routing

Nick DeMarinis

Based partly on lecture notes by Rachit Agarwal, Rodrigo Fonseca, Jennifer Rexford, Rob Sherwood, David Mazières, Phil Levis, John Jannotti
Administrivia

• IP: Due next Monday (10/24)
• PLQ grades: you should have received a bunch of emails from me last night
• HW3: Out today/tomorrow
Today

Exterior routing: how and why?

- Border Gateway Protocol (BGP)
- Today: Mechanics
- Thursday: Policies and Implications
Why Inter vs. Intra

For extension routing policy is most important.
Why Inter vs. Intra

- BGP allows for administrative input on routing policies
Why Inter vs. Intra

- BGP allows for administrative input on routing policies
- BGP is a policy control and information hiding protocol
  - intra == trusted, inter == untrusted
  - Different policies by different ASs
  - Different costs by different ASs
AS Relationships

Policies are defined by relationships between ASes

Example from Kurose and Ross, 5th Ed
AS Relationships

Policies are defined by relationships between ASes

- **Provider**: HIGHLY CONNECTED NETWORK, ROUTES TRAFFIC FOR CUSTOMERS.

Example from Kurose and Ross, 5th Ed
AS Relationships

Policies are defined by relationships between ASes

- Provider
- **Customer**: 
  
  PAY $ PROVIDER(S) FOR CONNECTIVITY
Policies are defined by relationships between ASes

- Provider
- Customer
- Peers: * intern connection between providers *

Example from Kurose and Ross, 5th Ed
Choice of Routing Algorithm
WHY NOT ONE OF THE ALGORITHMS WE ALREADY KNOW?

CHALLENGE
- SCALE
- AUTONOMY/POLICY
  - DON'T WANT TO SHARE ALL ROUTES.

LINK-STATE: SHARE ALL ROUTES W/ ALL NODES.
  - CAN'T SHARE W/ EVERYONE (SCALE)
  - CAN'T ENFORCE POLICY.

DISTANCE-VECTOR
- A BIT MORE PRIVACY,
  BUT NO WAY TO DO POLICY.
  - COUNT TO INFINITY
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

Given path info, can select route based on policy.
BGP: How it works

• Abstract each AS to a single node
• Destinations are CIDR prefixes
BGP: How it works

- Abstract each AS to a single node
- Destinations are CIDR prefixes
- Exchange prefix reachability with neighbors
  - E.g., “I can reach prefix 128.148.0.0/16 through ASes 44444 3356 14325 11078”
  - May choose to not advertise some paths to some neighbors
BGP: How it works

- Abstract each AS to a single node
- Destinations are CIDR prefixes
- Exchange prefix reachability with neighbors
  - E.g., “I can reach prefix 128.148.0.0/16 through ASes 44444 3356 14325 11078”
  - May choose to not advertise some paths to some neighbors
- Select a single path by routing policy

Pick one of many advertisements to use in your FIB table.
BGP: How it works

- Abstract each AS to a single node
- Destinations are CIDR prefixes
- Exchange prefix reachability with neighbors
  - E.g., “I can reach prefix 128.148.0.0/16 through ASes 44444 3356 14325 11078”
  - May choose to not advertise some paths to some neighbors
- Select a single path by routing policy
- Critical: learn many paths, propagate one
  - Add your ASN to advertised path
BGP Implications

- Explicit AS Path == Loop free (most of the time)
- Not all ASs know all paths
- Reachability not guaranteed
  - Decentralized combination of policies
- AS abstraction -> loss of efficiency
- Scaling
  - 74K ASs
  - 959K+ prefixes
  - ASs with one prefix: 25K
  - Most prefixes by one AS: 10008 (Uninet S.A. de C.V., MX)

Source: cidr-report 18Oct2022
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
  – NOW: Russia taking over Ukraine’s traffic
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 Example

Only 1 Router Per AS (for now)

AS 1
1.2.0.0/16

AS 2
1.2.0.0/16: AS 2 1

AS 3

AS 4

AS 5
1.2.0.0/16: AS 5 1

AS 2 PROPAGATES THE PREFIX FROM AS 1, ADDS ITSELF TO THE ADVERTISED PATH.
BGP Example

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

- **BGP speakers**: nodes that communicates with other ASes over BGP
- Speakers connect over TCP on port 179
- Exact protocol details are out of scope for this class; most important messages have type UPDATE
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, ...
  - Extensible: can add new types of attributes
Example

- **NLRI**: 128.148.0.0/16
- **AS-Path**: ASN 44444 3356 14325 11078
- **Next Hop IP**
- **Various knobs for traffic engineering:**
  - Metric, weight, LocalPath, MED, Communities
  - Lots of voodoo
Demo: AS11078
IP PREFIXES | ROUTE AGGREGATION

138.16.0.0/16 138.16.0.0/17

IDEA: ALLOCATE SMALLER NETWORKS FROM THIS PREFIX.

FOR EXAMPLE, COULD DIVIDE INTO TWO NETWORKS

1. 138.16.0.0/17
2. 138.16.128.0/17

IDEA: AS3 COMBINES, OR AGGREGATES PREFIXES FOR ITS CUSTOMERS

⇒ USE HIERARCHY OF ADDRESSES

IN PRACTICE... NOT SO EASY