Inter-Domain Routing: BGP

Brad Karp
UCL Computer Science
(drawn mostly from lecture notes
by Hari Balakrishnan and Nick Feamster, MIT)

CS 3035/GZ01
6th December 2011
Outline

• Context: Inter-Domain Routing
• Relationships between ASes
• Enforcing Policy, not Optimality
• BGP Design Goals
• BGP Protocol
• eBGP and iBGP
• BGP Route Attributes
• Synthesis: Policy through Route Attributes
Context: Inter-Domain Routing

- So far, have studied intra-domain routing
  - Domain: group of routers owned by a single entity, typically numbering at most 100s
  - Distance Vector, Link State protocols: types of Interior Gateway Protocol (IGP)

- Today’s topic: inter-domain routing
  - Routing protocol that binds domains together into global Internet
  - Border Gateway Protocol (BGP): type of Exterior Gateway Protocol (EGP)
Context: Why Another Routing Protocol?

• **Scaling challenge:**
  – millions of hosts on global Internet
  – ultra-naïve approach: use DV or LS routing, each 32-bit host address is a destination
  – naïve approach: use DV or LS routing, each subnet’s address prefix (i.e., Ethernet broadcast domain) is a destination
  – DV and LS cannot scale to these levels
    • prohibitive message complexity for LS flooding
    • loops and slow convergence for DV
    • Keeping routes current costs traffic proportional to product of number of nodes and rate of topological change
Context: Scaling Beyond the Domain

• **Address allocation challenge:**
  – Each host on Internet must have unique 32-bit IP address
  – How to enforce global uniqueness?
  – Onerous to consult central authority for each new host

• **Hierarchical addressing:** solves scaling and address allocation challenges
Context: Hierarchical Addressing

• Divide 32-bit IP address **hierarchically**
  – e.g., 128.16.64.200 is **host at UCL**
  – e.g., 128.16.64 prefix is **UCL CS dept**
  – e.g., 128.16 prefix is **all of UCL**
  – destination is a **prefix**
  – writing prefixes:
    • 128.16/16 means “high 16 bits of 128.16.x.y”
    • netmask 255.255.0.0 means “to find prefix of 32-bit address, bit-wise AND 255.255.0.0 with it”
  – prefixes need not be multiples of 8 bits long
Hierarchical Addressing: Pro

• Routing protocols generally incur cost that increases with number of destinations
  – Hierarchical addresses aggregate
  – Outside UCL, single prefix 128.16 can represent thousands of hosts on UCL network
  – End result: “reduces” number of destinations in global Internet routing system

• Centralized address allocation easier for smaller user/host population
  – Hierarchical addresses assure global uniqueness with only local coordination
  – Inside UCL, local authority can allocate low-order 16 bits of host IP addresses under 128.16 prefix
  – End result: decentralized unique address allocation
Hierarchical Addressing: Con

- Inherent loss of information from global routing protocol \(\rightarrow\) less optimal routes
  - Nodes outside UCL know nothing about UCL internal topology
  - UCL host in Antarctica has 128.16 prefix \(\rightarrow\) all traffic to it must be routed via London

- Host addresses indicate both host identity and network attachment point
  - Suppose move my UCL laptop to Berkeley
  - IP address must change to Berkeley one, so aggregates under Berkeley IP prefix!
A routing domain is called an **Autonomous System (AS)**

Each AS known by a **unique 16-bit number**

IGPs (e.g., DV, LS) route among **individual subnets**

EGPs (e.g., BGP) route among **ASes**

AS owns **one or handful of address prefixes**; allocates addresses under those prefixes

AS typically a **commercial entity or other organization**

ASes often **competitors** (e.g., different ISPs)
Global Internet Routing: Naïve View

- Find globally shortest paths
- Dense connectivity with many redundant paths
- Route traffic cooperatively onto lightly loaded paths
Global Internet Routing: Naïve View

- Find globally shortest paths
- Dense connectivity with many redundant paths
- Route traffic cooperatively onto lightly loaded paths

No correspondence to reality!
Global Internet Routing, Socialist Style

- Multiple, interconnected ISPs
- **ISPs all equal:**
  - in how connected they are to other ISPs
  - in geographic extent of their networks
Global Internet Routing, Socialist Style

• Multiple, interconnected ISPs

Little correspondence to reality!

- in how connected they are to other ISPs
- in geographic extent of their networks
Global Internet Routing: Capitalist Style

- Tiers of ISPs:
  - Tier 3: local geographically, end customers
  - Tier 2: regional geographically
  - Tier 1: global geographically, ISP customers, no default routes

- Each ISP an AS, runs own IGP internally
- AS operator sets policies for how to route to others, how to let others route to his AS
Global Internet Routing: Capitalist Style

- Tiers of ISPs:
  - Tier 3: local geographically, end customers
  - Tier 2: regional geographically
  - Tier 1: global geographically, ISP customers, no default routes

- Each ISP an AS, runs own IGP internally
- AS operator sets policies for how to route to others, how to let others route to his AS

Reality!
Outline

• **Context:** Inter-Domain Routing
• **Relationships between ASes**
• **Enforcing Policy, not Optimality**
• **BGP Design Goals**
• **BGP Protocol**
• **eBGP and iBGP**
• **BGP Route Attributes**
• **Synthesis:** Policy through Route Attributes
AS-AS Relationships: Customers and Providers

- Smaller ASes (corporations, universities) typically purchase connectivity from ISPs
- Regional ISPs typically purchase connectivity from global ISPs
- Each such connection has two roles:
  - **Customer**: smaller AS paying for connectivity
  - **Provider**: larger AS being paid for connectivity
- Other possibility: **ISP-to-ISP connection**
AS-AS Relationship: Transit

- Provider-Customer AS-AS connections: **transit**
- Provider allows customer to route to (nearly) all destinations in its routing tables
- Transit nearly always involves payment from customer to provider
AS-AS Relationship: Peering

- **Peering**: two ASes (usually ISPs) mutually allow one another to route to some of the destinations in their routing tables

- Typically these are their **own customers** (whom they provide transit)

- By contract, but **usually no money changes hands**, so long as traffic ratio is narrower than, e.g., 4:1
Financial Motives: Peering and Transit

- Peering relationship often between competing ISPs
- **Incentives to peer:**
  - Typically, two ISPs notice their own direct customers originate a lot of traffic for the other
  - Each can avoid paying transit costs to others for this traffic; shunt it directly to one another
  - Often better performance (shorter latency, lower loss rate) as avoid transit via another provider
  - Easier than stealing one another’s customers
- Tier 1s must typically peer with one another to build complete, global routing tables
Financial Motives: Peering and Transit (cont’d)

- Disincentives to peer:
  - Economic disincentive: transit lets ISP charge customer; peering typically doesn’t
  - Contracts must be renegotiated often
  - Need to agree on how to handle asymmetric traffic loads between peers
Outline

- Context: Inter-Domain Routing
- Relationships between ASes
- Enforcing Policy, not Optimality
- BGP Design Goals
- BGP Protocol
- eBGP and iBGP
- BGP Route Attributes
- Synthesis: Policy through Route Attributes
The Meaning of Advertising Routes

- When AS A advertises a route for destination D to AS B, it effectively offers to forward all traffic from AS B to D
- Forwarding traffic costs bandwidth
- ASes strongly motivated to control which routes they advertise
  - no one wants to forward packets without being compensated to do so
  - e.g., when peering, only let neighboring AS send to specific own customer destinations enumerated peering contract
Advertising Routes for Transit Customers

• ISP motivated to advertise routes to its own customers to its transit providers
  – Customers paying to be reachable from global Internet
  – More traffic to customer, faster link customer must buy

• If ISP hears route for its own customer from multiple neighbors, should favor advertisement from own customer
Routes Heard from Providers

• If ISP hears routes from its provider (via a transit relationship), to whom does it advertise them?
  – Not to ISPs with peering relationships; they don’t pay, so no motivation to provide transit service for them!
  – To own customers, who pay to be able to reach global Internet
Example: Routes Heard from Providers

- ISP P announces route to $C'_P$, own customer, to X
- X doesn’t announce $C'_P$ to Y or Z; no revenue from peering
- X announces $C'_P$ to $C_i$; they’re paying to be able to reach everywhere
Routes Advertised to Peers

• Which routes should an ISP advertise to ASes with whom it has peering relationships?
  – Routes for all own downstream transit customers
  – Routes to ISP’s own addresses
  – Not routes heard from upstream transit provider of ISP; peer might route via ISP for those destinations, but doesn’t pay
  – Not routes heard from other peering relationships (same reason!)
Example: Routes Advertised to Peers

- **ISP X** announces $C_i$ to Y and Z
- **ISP X** doesn’t announce routes heard from ISP P to Y or Z
- **ISP X** doesn’t announce routes heard from ISP Y to ISP Z, or vice-versa
Route Export: Summary

- ISPs typically provide **selective transit**
  - Full transit (export of all routes) for own transit customers in both directions
  - Some transit (export of routes between mutual customers) across peering relationship
  - Transit only for transit customers (export of routes to customers) to providers

- These decisions about what routes to advertise motivated by **policy (money)**, not by optimality (e.g., shortest paths)
Route Import

- Router may hear many routes to same destination network
- **Identity** of advertiser very important
- Suppose router hears advertisement to own transit customer from other AS
  - Shouldn’t route via other AS; longer path!
  - Customer routes higher priority than routes to same destination advertised by providers or peers
- Routes heard over peering higher priority than provider routes
  - Peering is free; you pay provider to forward via them
- customer > peer > provider
Outline

- Context: Inter-Domain Routing
- Relationships between ASes
- Enforcing Policy, not Optimality
- **BGP Design Goals**
- BGP Protocol
- eBGP and iBGP
- BGP Route Attributes
- Synthesis: Policy through Route Attributes
Border Gateway Protocol (BGP): Design Goals

- Scalability in number of ASes
- Support for policy-based routing
  - tagging of routes with attributes
  - filtering of routes
- Cooperation under competitive pressure
  - BGP designed to run on successor to NSFnet, the former single, government-run backbone
BGP Protocol

- BGP runs over TCP, port 179
- Router connects to other router, sends OPEN message
- Both routers exchange all active routes in their tables (possibly minutes, depending on routing table sizes)
- In steady state, two main message types:
  - announcements: changes to existing routes or new routes
  - withdrawals: retraction of previously advertised route
- No periodic announcements needed; TCP provides reliable delivery
BGP Protocol (cont’d)

• BGP doesn’t chiefly aim to compute shortest paths (or minimize other metric, as do DV, LS)
• Chief purpose of BGP is to announce reachability, and enable policy-based routing
• BGP announcement:
  – IP prefix: [Attribute 0] [Attribute 1] [...]