Routing protocols - BGP
==========

Outline
- Recap of Distance Vector algorithm
- Hierarchical routing - intra and inter domain
- Need for a new interdomain routing protocol, history of BGP.
- Path vector, AS, provider, customer, peer
- BGP route advertisement and route selection
- iBGP and e BGP
- Implementation details

==========

Distance vector routing: "tell about everyone to your neighbors".

- Every node exchanges a distance vector, a vector containing its estimate of distance to each destination, with its neighbors.
- Upon receiving a neighbor's distance vector, a node updates its distance vector by adding link cost to neighbor.
- If a better path is found through neighbor, it updates its best route.
- Distance vector has "count to infinity" problem. Simple solution is "split horizon" or "poisoned reverse".

Routing in the Internet is typically hierarchical.

- Typically in an organization (Autonomous System or AS), hosts are grouped into subnets based on physical proximity.
- A subnet is connected by a layer 2 technology (e.g., Ethernet) to its first hop IP router.
- Multiple IP routers are connected to each other by point-to-point links or Ethernet.
- All these routers in an organization run an "intra-domain" routing protocol or "Interior Gateway Protocol" (IGP) like OSPF (link state routing protocol) or RIP (distance vector protocol) to compute paths among themselves.
- For every router in an organization knows how to reach other internal IP destinations.

What about IP destinations outside the organization?

- There are special "border" routers at the edges of organizations that connect to other border routers.
- Internet Service Providers (ISPs) run a bunch of border routers that connect various organizations and other ISPs.
- These border routers run "inter-domain" routing protocols (BGP is the defacto standard today).
- These inter-domain routing protocols determine paths between organizations.
- The intra-domain and inter-domain routing protocols together fill up the forwarding tables in such a way that every IP router along the path can correctly route packets.

Why separate inter-domain and intra-domain?

- For scale.
  The internet routing tables will become very bulky if every IP router runs shortest path to every IP prefix. Instead, with the interdomain and intradomain separation, each set of protocols needs to handle lesser information.

- For policy.

Interdomain routing may not want to do simple shortest path, but complex policies based on business deals and trust, as we see below.

History of BGP.

Initially, when the Internet comprised of a small number of universities, the edge routers at these organizations just ran a simple routing protocol between them. These edge routers and few more routers added for connecting these (together called the Internet backbone) was managed by the US government for free. Soon, the internet expanded, and the internet backbone was commercialized. We now have Internet Service Providers (ISPs) that run a bunch of BGP routers (and intradomain routers to connect their organization) that connect various organizations. So the backbone of the internet is now composed of several ISPs. All the backbone routers have adopted the current version of BGP as the interdomain routing protocol since 1990s.

- BGP is a "path vector" protocol.
    - It is based on the distance vector philosophy, where you tell your neighbors about all the destinations you know of.
    - However, you don't just tell the cost, but you tell the entire path to the destination.
    - This addition of specifying the entire path avoids routing loops and counting to infinity (because if a neighbor is announcing a path through you already, you won't use that path).
    - Like DV protocols, there is a route announcement phase (where you exchange path vectors) and a best route computation phase (where you decide what your best path is).
    - These two phases have slight differences with general DV protocols (that do simple shortest path) in order to incorporate policy.

What is the granularity at which the path is specified? Do you list all the IP routers on the path?

- The internet paths are listed as sequence of Autonomous System Numbers(ASN).
- An AS is an organization that can be considered as one unit for the purpose of interdomain routing. Every AS has a unique AS number. For example, DAIICT is an AS. An ISP is an AS.
- A large organization that is spread across many locations can have different AS numbers for each part.
- Routers inside an AS run intradomain routing, and routers across ASes run interdomain routing.

So how does a path vector announcement in BGP look like?

- For every prefix, you list the AS path so far, and a few extra attributes for policy.
- For route selection, you pick the shortest AS path, along with some other decision criteria based on policy.

ASes are mainly of two types:

- End-user organizations
- ISPs that provide connectivity between these organizations.
- ISPs are classified into tiers 1, 2, and 3 (informally).

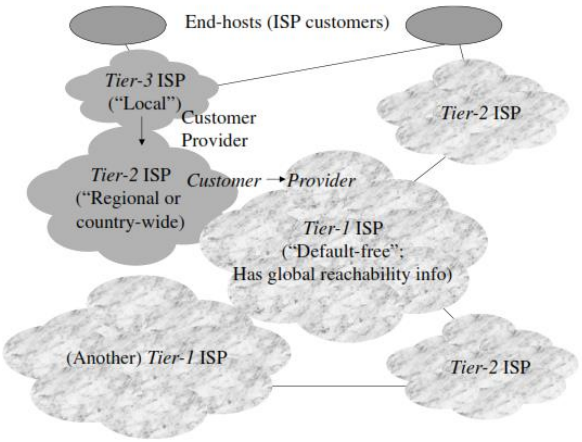

Fig.2: General organization of ISPs

AS to AS relationships are of two types: **transit and peering**.

Transit relationship exists between ISPs and their customers.

- A customer pays an ISP some money for Internet connectivity. This means that the ISP takes the responsibility of announcing the customer's IP address over BGP to the rest of the Internet.
    - That is how you get connected to the Internet through your ISP
- Advertising a route means to carry traffic, and traffic flows in the reverse direction of route advertisements.
- This means that traffic to the customer from other hosts in the Internet will flow via the ISP. Similarly, when the customer sends traffic, the ISP will have routes to several destinations, and will forward the customer's traffic onwards.
- "Buying service from an ISP" mean;s that the ISP is getting paid for announcing your routes, bringing you traffic, and sending your traffic. This type of a provider-customer relationship is also called a **transit relationship.**
    - In other words, a route advertisement from you to an ISP for a destination prefix is an agreement by ISP that it will forward packets sent by you destined for any destination in the prefix

Peering relationship

Consider two organizations that have lots of traffic to each other. Instead of both of them paying a provider to forward traffic to each other, they may seek to establish a connection directly, and forward each other's traffic directly. Such a relationship is called a **peering relationship**. Peering is usually between similar ASes that have roughly equal traffic to each other, and does not involve any payments. It is intended to carry traffic between peers by cutting out the middleman ISP.

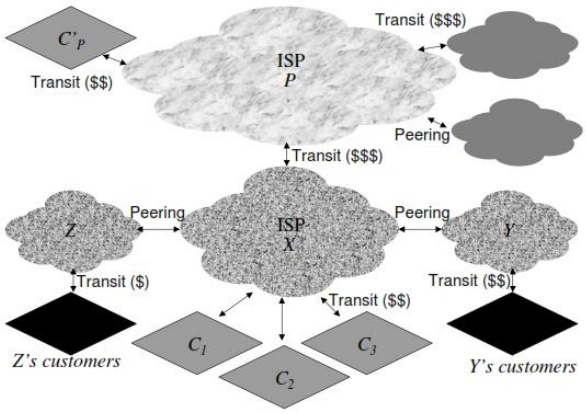

Fig.2: Transit and Peering relationship

- It shows an ISP, X, directly connected to a provider ISP P (from whom it buys Internet service) and a few customers C1, C2, C3 (to whom it sells Internet service: this is TRANSIT RELATIONSHIP
- It also shows two other ISPs Z and Y to whom it is directly connected, with whom X exchanges routing information via BGP: this is PEERING RELATIONSHIP since X, Y and Z do not pay anything to each other but agree to exchange/advertise routes based on some mutual contract

Is your ISP always guaranteed to be connected to everyone in the entire Internet, and have routes to every destination?

- Not in the case of smaller ISPs.
- Usually, the smaller ISPs buy service from bigger ISPs, which buy from bigger ISPs are so on. The ISPs at the top of the hierarchy are called Tier-1 ISPs. By definition, a Tier-1 ISP does not have any other ISP as its provider. All Tier-1 ISPs peer amongst themselves. The customers of Tier-1 ISPs are Tier-2 ISPs and so on.
- ASes or BGP routers which have BGP sessions between them are "neighbors" as far as the path vector routing protocol is considered. We will now revisit the question of what gets advertised to neighbors and how best path is computed.

Policy decision: when do you advertise a route?

- Two rules.
    - First, routes from customers and self-routes (routes to destinations within an ISP or organization) are announced to all neighbors - because you want to provide as much visibility as possible.
    - Two, routes about all destinations that you learn from your neighbors are announced only to your customers.

What is the logic behind these rules?

Why not announce peer or provider routes to other peers and providers?

Because you don't want to carry traffic on behalf of your peers and providers (you don't make any money from it). The only announcements that happen are intended to provide reachability to self and customers, and let customers and self-reach everyone.

No other type of connectivity suits business interests. Contrast this with an intradomain routing protocol whose goal is to provide shortest path connectivity between everyone.

Policy decision: which routes do you prefer during best route selection? Even before shortest AS hop count, you have a policy based rule.
Prefer customer routes > peer routes > provider routes. Why?

- If you use customer route, it means you will send traffic through customer, and customer pays for it.
- If you use peer routes, nothing lost nothing gained.
- If you use provider route to send traffic, you pay for it.
  So use the cheapest option.
- Typically, routes that come in on a link are marked with an attribute called "localpref" to indicate if it is a customer link / peer link / provider link. This attribute is checked first before seeing shortest AS hop coount.

How do typical routes look at AS level?
Typically, you have zero or more customer to provider links, then you hit a peering link or a tier-1 provider, then you go down zero or more provider to customer links to reach the other end point.

BGP is between border routers. How does this translate to forwarding table at the interior routers?

- BGP sessions are of two types: eBGP (external BGP) between border routers in different organizations, and iBGP (internal BGP) between routers in the same organization.
- That is, an organization may have more than one BGP routers (each potentially talking to several other BGP routers in different organizations).
- Each BGP router sends all the external routes it learns to all other internal BGP routers using BGP itself - this is called an iBGP session.

The main difference between iBGP and eBGP is as follows.

- Every BGP route for a prefix has a value called "next hop".
- In eBGP sessions, the next hop is simply the IP address of the other BGP endpoint, indicating that if this route is used, traffic should be sent to this next hop.
  - This next hop is updated on eBGP sessions.
- On the other hand, the next hop over iBGP sessions is simply set to the first BGP router that introduced the route into an organization.
  - That is, if BGP router A sends a BGP route to B over iBGP, which in turn sends it to C, then the next hop of the BGP route will be IP address of A and not that of B.
- So, all internal BGP routers will have BGP routes from all BGP routers (typically all BGP routers talk to each other), with next hop being the BGP router that introduced the route.
- The internal routers will also be running an intradomain IGP to populate routes to internal destinations.
- So when a packet comes in for the external destination, the internal router looks up the BGP routing table. It may find several routes from several border BGP routers. It will pick the closest border router based on the next hop.
- Then the path to the border router will be determined by the IGP / intradomain protocol. So the forwarding is done by combining the intradomain routing tables with the interdomain routing tables.

Some implementation details.

- BGP runs as a userspace process over TCP on a well known port (179).
- Two routers that want to become BGP neighbors (also called BGP peers) open a TCP connection between them.
- Then, they exchange the path vectors (and other information) for all prefixes. This is called the **full table dump.**
- From then on, only updates (announcements, withdrawals, any changes) are communicated.
- Periodically, the peers / neighbors also exchange heartbeat / keep alive messages to let the other end point know they are alive.
- The BGP routing table consists of prefixes followed by a set of attributes for each prefix.
- Here are some important attributes associated with each prefix:
    - Localpref: a local variable (specific to each link) which indicates whether the route came from a customer, provider, or peer.
    - AS path: sequence of ASes the route has traversed.
    - Next hop: the IP address of the next hop BGP router for this route. If this route is picked as the best route, traffic will be sent to this next hop.
        - For eBGP sessions, the next hop is the BGP peer at the other end of the BGP session.
        - For routes learned over iBGP, the next hop is the border BGP router that first introduced the route into the network.
    - Multi-exit discriminator (MED): When two ASes are connected at multiple points, the MED attribute is used to indicate a preference for which connection should be used to send packets.

    - BGP neighbors exchange their current best route with their peers subject to the policy we discussed earlier (announce customer and self-routes to everyone, and all best routes to customers).
    - When updates arrive from neighbors, a BGP router updates its best route for every prefix using the following criteria in this order.
        - Highest localpref (prefer customer > peer > provider)
        - Shortest AS path preferred
        - Lowest MED preferred
        - eBGP routes preferred over iBGP (if you are directly connected to external network, use that instead of going via your own network)
        - Lowest IGP path to next hop border router
        - Smallest router ID or IP address to break ties.