Lecture 9: Interdomain routing

Anirudh Sivaraman

2018/10/15

Last lecture, we looked at how packets find their way from one router to another *within* an autonomous system (AS) (also known as a domain[1]). Formally, an AS is a network of routers that is controlled by a single administrative entity, such as a university, an enterprise company, an Internet Service Provider (ISP), or a content provider such as Google, Netflix, or Facebook. With this definition of an AS or domain, what we looked at last class can be termed *intradomain* routing.

This lecture we'll look at *interdomain routing*:[2] how packets find their way from one router to another across domains.

Interdomain routing is important because it is really what makes the Internet an internet, i.e., a network of interconnected networks. The interconnected networks are networks run by ASes or domains. Without interdomain routing, as an NYU student, you would be restricted to accessing web sites within NYU alone, and Google/Facebook/Netflix would have no way of delivering content to you.

Keeping with the Internet's original goals of low-effort interconnection between existing networks, the goal of interdomain routing is to ensure *reachability*, i.e., to ensure that there is *some* path—however poor that path may be—to get from one host on the Internet to another. This is in contrast to intradomain routing that we looked at last lecture. With intradomain routing, the goal was to find a good if not the best path, for some definition of best: lowest latency, highest bottleneck link capacity, etc. This is possible in intradomain routing because the routers are all under a single administrative entity who can optimize routing on the network as it sees fit.

After DNS and private IP addresses, this separation of routing into interdomain and intradomain routing is another example of the classic computer systems trick of hierarchy. Concretely, an interdomain routing protocol needs to scale enough to handle routing between domains in the world, while an intradomain routing protocol needs to scale enough to handle routing between all routers within a domain. Neither protocol needs to scale enough to handle the product of the number of routers and the number of domains in the world—a far more daunting task.

1 The Border Gateway Protocol

For the purpose of interdomain routing, there are two kinds of routers: border routers and internal routers. Border routers sit at the boundary of a domain and have a direct connection to another border router sitting at the boundary of an adjacent domain. An internal router is any router within a domain that is not a border router.

The predominant protocol for interdomain routing is the Border Gateway Protocol (BGP). This protocol is responsible for (1) discovering information regarding routing paths between domains and (2) propagating this information to the internal routers within a domain. In more detail, the first step is responsible for figuring out a sequence of domains that gets a packet from a source domain to a destination domain. The second step is responsible for telling an internal router in the source domain which border router it needs to send packets to so that the packets can eventually find their way to the destination. We'll discuss each of these two steps below.

[1] This use of the term domain refers to an independent entity in the Internet. It is distinct from the domain in the domain name system.

[2] We could have also called it inter AS routing to avoid confusion with domains in DNS. However, the term interdomain routing is so widely used that it would cause more confusion to invent a new term for it.

1.1 Exchanging information between border routers

For the next two subsections, let's imagine a really simple topology where three domains are arranged in straight line:

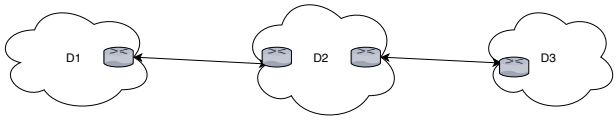

Figure 1: Simple topology with 3 domains (Routers here represent edge routers)

Each domain in the topology above has a border router that connects it to the next domain. The bidirectional arrows indicate that packets can flow in either direction.

The eBGP (for external BGP) protocol is responsible for discovering paths that cross multiple domains. A border router BR_{D1} at the border of one domain $D1$ tells the border router BR_{D2} at an adjacent domain $D2$ the following: any router/host within $D2$ can get to a particular set of IP addresses within $D1$ (the formal term for a set of IP addresses is a prefix) using the domain $D1$. Now, let's assume that $D2$ is, in turn, connected to another domain $D3$. Then $D2$'s border router at the boundary to $D3$ will tell $D3's$ border router two things: (1) $D3$ can get to prefixes within $D2$ using $D2$, and (2) $D3$ can get to prefixes with $D1$ using the sequence of domains $D2, D1$.

eBGP is what is called a *path-vector protocol*. Here each border router prepends its own domain name at the beginning of a sequence of domain names (called a path vector) before advertising it to the next border router. This is in contrast to a distance vector protocol that advertises distances to different destinations to its neighbors. eBGP is a path-vector protocol because knowledge of the full path (as opposed to just the distance of the path) is important: a domain may choose not to use or advertise a path to a destination if it passes through a particular domain for competitive reasons.

1.2 Propagating this information to the internal routers

Once the border routers know how to get to other border routers, the iBGP (internal BGP) protocol handles the propagation of this interdomain information to routers within each domain. Conceptually, the following sequence of events occur when propagating interdomain information within a domain.

1. A border router, say the one in domain $D3$, learns of a path to prefixes in another domain $D1$ through a sequence of domains $D2, D1$.
2. This border router in domain $D3$ (let's call it BR) will use iBGP to tell all its internal routers that they can reach prefixes in $D1$ by going through BR.
3. The intradomain routing protocol running inside $D3$ would have already found a path between any pair of routers (whether these are internal or border routers) inside $D3$.
4. Using the information from the intradomain routing protocol, any internal router within $D3$ knows the next internal router it needs to send packets to so that packets eventually end up at $D3$'s border router (BR), which will eventually get packets to $D1$ through $D2$.

2 Choosing between alternative paths

Internet topologies typically have redundancy built into them: there are multiple paths between two domains. This means a domain is often faced with a choice between two different paths. The simplest example of this is multihoming, where one domain (e.g., a campus network) connects to the Internet using two different ISPs for fault tolerance and better load balancing. For any given destination domain, the two ISPs may have

different paths that get to the same destination domain. In turn, this means that the campus network has two alternatives to pick from to get to a destination domain. How does it choose?

The BGP protocol doesn't mandate any particular way of picking between alternative paths. This is left to the administrative entity in charge of running a particular domain. Formally, this is accomplished by assigning each path a numeric attribute called a *local preference*. How this attribute is assigned is completely up to the domain. If two paths have the same local preference value, then the path with a smaller number of domains on it is chosen.

Note that neither the local preference nor minimizing the length of the path necessarily corresponds to improving performance. Why is this? The local preference can choose paths for entirely business-driven reasons that have nothing to do with optimizing packet latency. For instance, the phenomenon of *boomerang routing* has been observed in parts of Africa where traffic between two parts of Africa is routed through Europe! This is because many African ISPs choose to optimize the common case (an African client communicating with a European server) to the detriment of the uncommon one (client and server both in Africa). As a result, they may prefer to route their traffic through a European ISP, with whom they have a business relationship, rather than a local ISP with whom they may not have such a relationship [1].

Similarly, minimize the length of the path as measured by the number of domains on the path may not minimize packet latency either. This is because this metric captures nothing about the internal topology of each domain—and hence the internal latency required to transit each domain. Interdomain routing is full of such paradoxes and provides a fertile playground for Internet measurement research, which seeks to observe and explain such measurements.

3 Relationships between domains

We already briefly alluded to business relationships between domains when we discussed boomerang routing. We'll now make this more concrete and explain how business relationships affect the local preference attribute when choosing between different routing paths.

Simplifying considerably, there are two kinds of relationships between domains: *transit* and *peering*. In transit relationships, a *provider* domain provides Internet service for a *customer* domain. This is common when a regional ISP provides Internet service to many different local ISPs. It is typically associated with the customer domain paying the provider domain for Internet service. It might also involve the customer domain agreeing to provide a minimum amount of traffic to the transit domain to make it worth the provider's time and money. In peering relationships, two domains agree to carry traffic for each other that is destined to the other domain's customers. Peering relationships can either involve no payment (settlement-free peering) or, more rarely, could involve each party paying the other for carrying traffic on its behalf.

Let's compare peering and transit relationships. In transit relationships, there is a notion of a hierarchy, where a provider offers service to a customer. In peering the two entities are more-or-less equal. However, peering relationships suffer from the risk that one entity thinks the other is taking unfair advantage of the peering relationship. For instance, this happens when the traffic volume in one direction of peering is much higher than the other. As a result, both peering and transit relationships are revisted frequently depending on the availability of other ISPs to connect to the Internet and changes in traffic characteristics. In summary, peering and transit relationships are crucial to the Internet. In the absence of such peering and transit relationships, the Internet would be relegated to isolated islands and would not be a network of networks.

To understand these relationships graphically, you can think of the Internet's domain-level topology to a first approximation as following a tree-like structure. The upper levels of the tree consist of domains called *transit providers*, because they carry transit traffic on behalf of their customers. The lower levels of the tree (e.g., residential, campus, local ISP networks, content provider networks) are called *stub networks*; these typically generate traffic for the transit networks. Transit relationships occur up and down the tree; peering relationships occur horizontally within the same level of the tree.

3.1 Business relationships and local preferences

The types of relationships between domains affect the local preference for different paths in two different situations: (1) what paths the domain uses to get to destinations on the Internet and (2) what paths the domain chooses to advertise to other domains.

For instance, let's look at a multihomed setting, where a campus C connects to 2 ISPs A and B. Here, C is in a transit relationship with both A and B. This means that C can use either A or B to allow its own customers to get to the Internet. However, even though it is technically feasible, it would make no economic sense for C to carry traffic for A's other customers (say from another stub network N connected to A) from A to B or B to A. Hence, C will only advertise a path to an IP address prefix within N to its own customers. It will not advertise this path to A or B because then it will be carrying someone else's traffic.

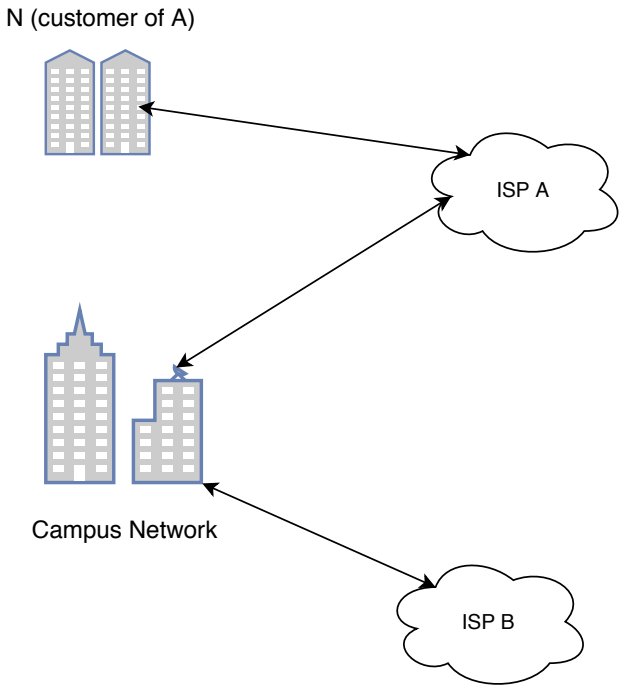

Figure 2: Campus connected to two ISPs

As another example, let's go back to the boomerang routing example in Africa discussed earlier. Let's say an ISP in Kenya wants its customers to connect to servers in a South African ISP. The Kenyan ISP could be forcing its customers to take a circuituos path through Europe for one of the two reasons below: (1) even though there is a South African ISP that could carry traffic and get to the South African server faster, the South African ISP does not advertise a path to its servers to the Kenyan ISPs, or (2) the South African ISP does advertise a path to its servers, but the Kenyan ISP chooses not to use it. The local preference attribute controls both the selective advertisement and the selective usage of paths. In this case, the underlying reason for both selective advertisement and usage might be the lack of a peering relationship between the Kenyan and South African ISPs. They might find it more economically benefical to use their money to negotiate a relationship with European ISPs instead because most traffic from Africa is destined to addresses outside of Africa.[3]

[3]We can only guess these relationships—another area of active research. The true nature of these relationships is typically confidential.

4 Historical notes: the flattening of the Internet

Historically, the Internet's domain-level topology was structured as a multi-level tree with multiple tiers of Internet Service Providers: global, regional, and local with peering agreements between them and transit agreements with their customers (clients and content providers such as Google/Facebook/Microsoft).

Over time, as content providers have grown in size and tried to expand their geographic presence, this hierarchy has become flatter. Content providers like Google now routinely enter into peering agreements with other ISPs, so much so that the distinction between an ISP and a content provider is blurring and the Internet's hierarchy is flattening into two levels: clients (mobile phones, desktops, laptops) and content providers/ISPs. The rationale for this is that by directly connecting with a local ISP and entering into a peering agreement with them, a content provider such as Google can bring their servers closer to the client, thereby improving client-perceived performance [2].

Hence, even though Google/Netflix/Facebook/YouTube/Yahoo are headquartered in the U.S., they have servers and local networks distributed around the world, allowing them to serve their customers from geographically proximate locations.

5 Historical notes: the YouTube outage

BGP wasn't explicitly designed with security in mind. This means a border router's advertisements are not authenticated. This has led to some high-profile incidents where a border router's incorrect advertisements are propagated without any checks to the rest of the Internet.

One notable example is the use of BGP to censor Internet access in some countries. This is done by having a regional ISP advertise a blackhole path to some Internet prefix to its customers. Instead of actually connecting customers to that Internet prefix, this blackhole path leads to a different IP address that either drops packets silently or notifies the customer that they have been blocked.

In 2008, Pakistan Telecom decided to block access to YouTube for all its customers. But, instead of selectively advertising this blackhole path to its own customers alone, it advertised this path to other domains as well. To make matters worse, its neighboring domains also continued to advertise this path without filtering it out because the lack of authentication in BGP makes it hard to know if a path is genuine or not. Soon enough, the entire Internet was sending traffic destined for YouTube to this blackhole.

Typically, these problems are resolved quickly by manual intervention, but this has provided quite some impetus for developing secure versions of BGP.

References

[1] Arpit Gupta, Matt Calder, Nick Feamster, Marshini Chetty, Enrico Calandro, and Ethan Katz-Bassett. Peering at the Internet's Frontier: A First Look at ISP Interconnectivity in Africa. In *Proceedings of the 15th International Conference on Passive and Active Measurement - Volume 8362*, PAM 2014, 2014.

[2] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. Internet inter-domain traffic. *ACM SIGCOMM Computer Communication Review*, 41(4):75–86, 2011.