Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.829 Fall 2002 — **Wide-Area Unicast Internet Routing** — March 2002

Overview. This lecture looks at routing issues in the Internet at large, focusing on how Internet Service Providers exchange routes with each other. We look at peering and transit relationships between providers and discuss BGP4, the current wide-area Internet routing protocol.

These notes were originally prepared for 6.829 Computer Networks in Fall 2001. They have been slightly revised for 6.033 Computer Systems Engineering in Spring 2002.

1 Introduction

This goal of this lecture is to give you a good sense of the reality of wide-area Internet routing. We will look at how Internet Service Providers exchange routing information (and packets) between each other, and how the way in which they buy service from and sell service to each other and their customers influences the technical research agenda of Internet routing in the real-world.

An abstract, highly idealized view of the Internet is shown in Figure **??**, where end-hosts hook up to routers, which hook up with other routers to form a nice connected graph of essentially "peer" routers that cooperate nicely using routing protocols that exchange "shortest-path" or similar information and provide global connectivity. The same view posits that the graph induced by the routers and their links has a large amount of redundancy and the Internet's routing algorithms are designed to rapidly detect faults and problems in the routing substrate and route around them. Some would even posit that the same routing protocols today perform load-sensitive routing to dynamically shed load away from congested paths on to less-loaded paths.

While at a high-enough level there are some vague elements of truth in the above description, this abstraction is actually quite misleading. It's actually a myth, or perhaps wishful thinking, that much of this happens! The real story of the Internet routing infrastructure is that the Internet service is provided by a large number of commercial enterprises, generally in competition with each other. Cooperation, required for global connectivity, is generally at odds with the need to be a profitable commercial enterprise, which often occurs at the expense of one's competitors—the same people with whom one needs to cooperate. How this is achieved in practice (although there's lots of room for improvement), and how we might improve things, is an interesting and revealing study of how good technical research can be shaped and challenged by commercial realities.

A second pass at developing a good picture of the Internet routing substrate is shown in Figure **??**, which depicts a group of Internet Service Providers (ISP's) somehow cooperating to provide global connectivity to end-customers. This picture is closer to the truth, but the main thing it's missing is that not all ISP's are created equal. Some are bigger and more "connected" than others, and still others have global reachability in their routing tables. There are names given to these "small," "large," and "really huge" ISP's: *Tier-3 ISP's* are ones that have a small number of usually localized (in geography) end-customers; *Tier-2 ISP's* generally have regional scope (*e.g.,* state-wide, region-wide, or non-US country-wide), while *Tier-1 ISP's*, of which there are a handful, have global scope in the sense that their routing tables actually have routes to all currently reachable Internet prefixes (*i.e.,* they have no default routes). This organization is shown in Figure **??**.

The current wide-area routing protocol, which exchanges *reachability information* about routeable IP-address prefixes between routers at the boundary between ISP's, is *BGP-4* (for "Border Gateway Protocol, Version 4"). More precisely, the wide-area routing architecture is divided into *autonomous systems* (AS's) that exchange reachability information. An AS is owned and administered by a single commercial entity, and implements some set of policies in deciding how to route its packets to the rest of the Internet, and how to export its routes (its own, those of its customers, and other routes it may have learned from other AS's) to other AS's. Each AS is identified by a unique 16-bit number.

Within an AS, an entirely different routing protocol operates. These routing protocols are called *Interior Gateway Protocols*, or IGP's, and include protocols like RIP, OSPF, IS-IS, and IGRP. (This makes protocols like BGP-4 "Exterior Gateway Protocols" or EGP's.) The key difference between BGP-4 and IGPs is that the former is concerned with providing *reachability information* and facilitating *routing policy* implementation, in a *scalable* manner, whereas the latter are typically concerned with optimizing a path metric. Scalability is typically not a major concern in the design of IGP's (or at least, it's safe to say that all known IGP's don't scale as well as BGP-4 does).

The rest of this lecture is in two parts: first, we will look at inter-AS relationships (transit and peering); then, we will study some salient features of BGP-4. We don't have time to survey IGP's in this lecture, but you should be familiar with the more well-known ones like RIP and OSPF (or at least with distance-vector and link-state protocols). To learn more about IGP's if you're not familiar with them, read a standard networking textbook (*e.g.,* Peterson & Davie, Kurose & Ross, Tanenbaum) or a book on routing protocols (*e.g.,* Huitema).

2 Inter-AS Relationships: Transit and Peering

Consider the picture shown in Figure **??**. It shows an ISP, with AS number X, directly connected to a *provider* (from whom it buys Internet service) and a few *customers* (to whom it sells Internet service). In addition, the figure shows two other ISP's to whom it is directly connected, with whom X exchanges routing information via BGP.

There are two prevalent forms of AS-AS interconnection. The first form is *transit*, wherein one ISP (the "provider" P in Figure **??**) provides access to all (or most) destinations in its routing tables. Transit almost always is meaningful in an inter-AS relationship where financial settlement is involved; the provider charges its customers for Internet access, in return for forwarding packets on behalf of customers to destinations (and in the opposite direction in many cases). Another example of a transit relationship in Figure **??** is between X and its customers (the C_i's).

The second prevalent form is called *peering*. Here, two AS's (typically ISP's) provide mutual access to a subset of each others' routing tables. The subset of interest here is their own transit customers (and the ISP's own internal addresses). Like transit, peering is a business deal, but it may not involve financial settlement. While paid peering is not unheard of, in many cases they are reciprocal agreements. As long as the traffic ratio between the concerned AS's is not highly asymmetric (*e.g.,* 4:1 is a commonly believed and quoted ratio), there's usually no financial settlement. Peering deals are almost always under NDA and held quite confidential. (Paid peering arrangements are apparently common in some parts of the world; Norton mentions some examples.)

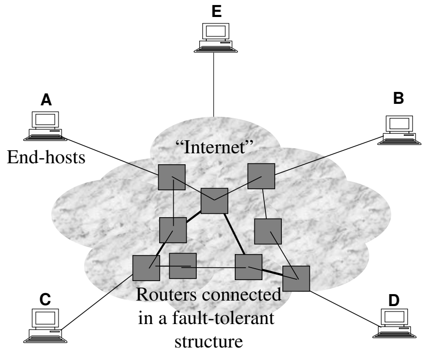

Figure 1: This is a rather misleading abstraction of the Internet routing layer.

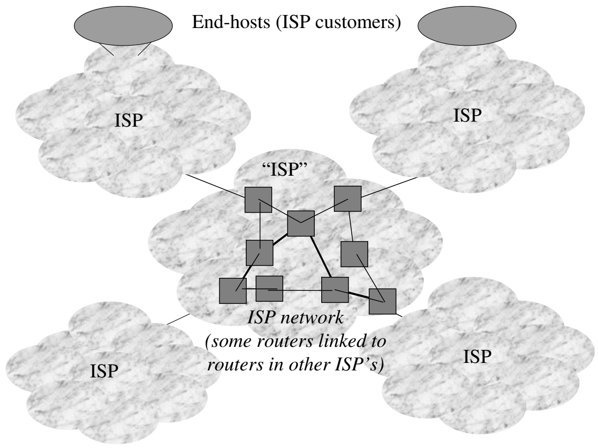

Figure 2: The Internet is actually composed of many competing Internet Service Providers (ISP’s) that cooperate to provide global connectivity. This picture suggests that all ISP’s are “equal,” which isn’t actually true.

2.1 Peering v. Transit

A key point to note about peering relationships is that they are often between business competitors. The common reason for peering is the observation by each party that a non-trivial fraction of the traffic emanating from each one is destined for the other’s direct transit customers. Of course, the best thing for each of the ISP’s to try to do would be to wean away the other’s customers, but this may be hard to do. The next best thing, which would be in their mutual interest, would be to avoid paying transit costs to *their* respective providers, but instead set up a transit-free link between each other to forward packets for their direct customers. In addition, this has the advantage that this more direct path would lead to better end-to-end performance (in terms of latency, packet loss rate, and throughput) for their customers. It’s also worth noticing that a Tier1 ISP usually will find it essential to be involved in peering relationships with other ISP’s (especially other Tier1 ISP’s) to obtain global routing information in a default-free manner.

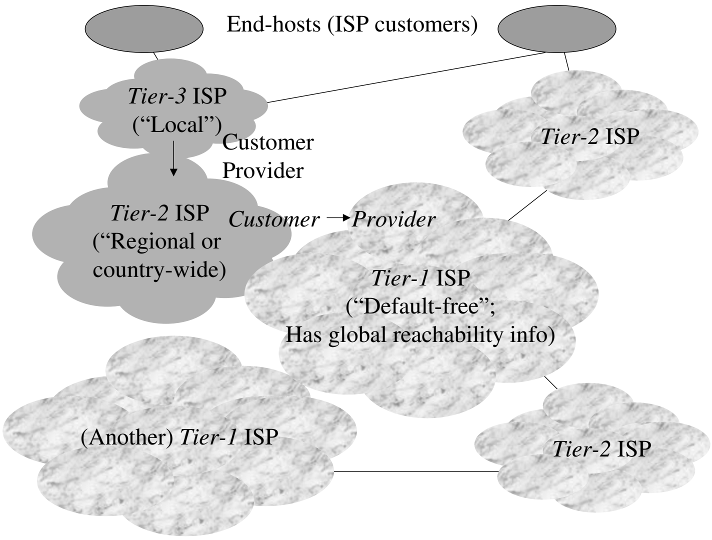

Figure 3: A more accurate picture of the wide-area Internet routing infrastructure, with various types of ISP's defined by their respective reach. *Tier-1* ISP's have "default-free" routing tables (i.e., they don't have any default routes), and typically have global reachability information. There are a handful of these today (about five or so).

Balancing these potential benefits are some forces against peering. Transit relationships generate revenue; peering relationships usually don't. Peering relationships typically need to be renegotiated often, and asymmetric traffic ratios require care to handle in a way that's mutually satisfactory. Above all, these relationships are often between competitors vying for the same customer base.

2.2 Exporting Routes: Route Filtering

Each AS (ISP) needs to make decisions on which routes to export to its neighboring ISP's using BGP. The reason for this is that no ISP wants to act as transit for packets that it isn't somehow making money on. Observe that in general packets flow in the opposite direction to the (best) route advertisement for any destination, which means that an AS should be careful of what routes it advertises. An advertisement for any destination means that some other AS that hears the advertisement may believe that the place where the advertisement came from is a good place to send packets for any destination corresponding to the advertisement.

Transit customer routes. To an ISP, its customer routes are likely the most important, since the view it provides to its customers is the sense that *all* potential senders in the Internet can reach them. This means that it is in the ISP's interest to advertise routes to its transit customers to as many other connected AS's as possible.

Transit provider routes. Does an ISP want to provide *transit* to the routes exported by its provider to it? Most likely not, since the ISP isn't making any money on providing such transit facilities. An example of this is shown in Figure **??**, where C'_P is a customer of P, and P has exported a route to C'_P to X. It isn't in X's interest to advertise this route to everyone, e.g., to other ISP's with whom X has a peering relationship. An important exception to this, of course, is

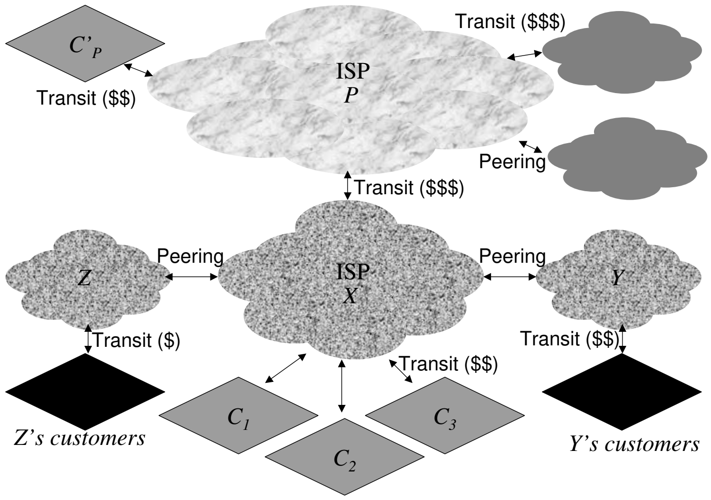

Figure 4: Inter-AS relationships; transit and peering.

X's transit customers who are paying X for service—the service X provides its customers C_i's is that they can reach any location on the Internet via X, so it makes sense for X to export as many routes to X as possible.

Peer routes. It usually makes sense for an ISP to export only selected routes from its routing tables to other peering ISP's. It obviously makes sense to export routes to all of ones transit customers. It also makes sense to export routes to addresses within an ISP. However, it does not make sense to export an ISP's transit provider routes to other peering ISP's, since that may cause a peering ISP to use the advertising ISP to reach a destination advertised by a transit provider. This would expend ISP resources but not cause any money to reach it.

The same situation applies to routes learned from other peering relationships. Consider ISP Z in Figure **??**, with its own transit customers. It doesn't make sense for X to advertise routes to Z's customers to another peering ISP (Y), since X doesn't make any money on Y using X to get packets to Z's customers!

These arguments show that most ISP's end up providing *selective transit*: typically full transit capabilities for their own transit customers in both directions; some transit (between mutual customers) in a peering relationship; and transit only for one's transit customers (and ISP-internal addresses) to one's providers.

The discussion so far may make it sound like BGP is the only way in which to exchange reachability information between an ISP and its customers or between two AS's. This is not true—a large fraction of end-customers (typically customers who don't provide large amounts of further transit and/or aren't ISP's) do not run BGP sessions with their providers. This is because BGP is complicated to configure, administer, and manage, and isn't very useful if the set of addresses in the customer is relatively unchanging. These customers interact with their providers via *static routes.* These routes are usually manually configured. Of course, information about customer address blocks will in general be exchanged by a provider using BGP to other AS's (ISP's) to achieve global reachability to the customer premises.

2.3 Importing Routes

The previous section described the issues considered by an AS (specifically, routers in an AS involved in BGP sessions with routers in other AS's) while deciding which routes to export. In a similar manner, when a router hears many possible routes to a destination network, it needs to decide which route to install in its forwarding tables.

This is a fairly involved process in BGP and requires a consideration of several attributes of the advertised routes. At this stage, we consider only one of the many things that a router needs to consider, but it's the most important consideration. It has to do with who advertised the route. Typically, when a router (*e.g.,* X in Figure **??**) hears advertisements to its transit customers from other AS's (*e.g.,* because the customer is multi-homed), it needs to ensure that packets to the customer do not traverse additional AS's unnecessarily. This usually means that customer routes are prioritized over routes to the same network advertised by providers or peers. Second, peer routes are likely more preferable to provider routes, since the purpose of peering was to exchange reachability information about mutual transit customers. These two observations imply that typically routes are imported in the following priority order:

$$customer > peer > provider$$

This rule (and many others like it) can be implemented in BGP using a special attribute that's locally maintained by routers in an AS, called the LOCAL PREF attribute. The first rule in route selection with BGP is to pick a route based on this attribute. It is only if this attribute is *not* set for a route, are other attributes of a route even considered. This doesn't imply, however, that most routes in practice are selected using the LOCAL PREF attribute; other attributes like the length of the AS path tend to be quite common.

3 BGP-4

We now turn to how reachability information is exchanged using BGP-4, and how routing policies like the ones explained in the previous section can be expressed and enforced.

The design of BGP, and its current version (4), was motivated by three important needs:

1. Scalability. The division of the Internet into separate routing domains, called autonomous systems (AS's), under independent administration, was done while the backbone of the then Internet was under the administration of the NSFNet. An important requirement for BGP was to ensure that the Internet routing infrastructure remained scalable as the number of connected networks increased.

2. Policy. The ability for each AS to implement and enforce various forms of routing policy was an important design goal. One of the consequences of this was the development of the BGP attribute structure for route announcements, and allowing route filtering.

3. Cooperation under competitive circumstances. BGP was designed in large part to handle the transition from the NSFNet to a situation where the "backbone" Internet infrastructure would no longer be run by a single administrative entity. Rather, routing in the Internet would be handled by a large number of mutually competing ISP's, who would (loosely) cooperate to provide global connectivity. This implies that the routing protocol should allow AS's to make purely local decisions on how to route packets, from among any set of choices.

In the old NSFNET, the backbone routers exchanged routing information over a tree topology, using a routing protocol called EGP. (While the modern use of the term EGP is to think of it as a family of exterior gateway protocols, its use in the context of NSFNET refers to the specific on used in that network.) Because the backbone routing information was exchanged over a tree, the routing protocol was relatively simple. However, the evolution away from a singly administered backbone made the NSFNET EGP obsolete and required a more sophisticated protocol, BGP.

3.1 The Protocol

As protocols go, the operation of BGP is quite straightforward. BGP-4 runs over TCP, on well-known port (179). To start participating in a BGP session with another router, a router sends an OPEN message after establishing a TCP connection to it on the BGP-4 port. After the OPEN is completed, both routers exchange their tables of all active routes (of course, applying all applicable route filtering rules). This process may take several minutes to complete, especially on routers that have a large number of active routes.

After this, there are two main types of messages on the BGP session. First, there are KEEPALIVE messages sent in both directions to check if the BGP session is still running. Second, there are *route updates* sent on the session. These updates only send any routing entries that have changed since the last update (or transmission of all active routes). There are two kinds of updates. The first are *announcements*, which are changes to existing routes or new routes. The second are *withdrawals*, which are messages that inform the receiver that the named routes no longer exist. This usually happens when some previously announced route can no longer be used. Because BGP uses TCP, which provides reliable and in-order delivery, routes do not need to be periodically announced unless they change. However, the absence of the KEEP ALIVE messages allows a router to remove all routes from its tables that came from an external neighbor that no longer exists.

Unlike many IGP's, BGP does not simply optimize any metrics like shortest-paths or delays. Because its goals are to provide reachability information and enable routing policies, its announcements do not simply announce some metric like hop-count. Rather, they have the following format:

$$IP\ prefix : Attributes$$

where for each IP prefix announced there are one or more attributes that are announced as well. There are a substantial number of standardized attributes in BGP-4, and we'll look at some of them in more detail in the rest of this lecture. Recall that one BGP attribute has already been introduced to us, the LOCAL PREF attribute. It isn't an attribute that's disseminated with route announcements, but is an important attribute used locally while selecting a route for a destination.

There are two types of BGP sessions: *eBGP* sessions are between BGP-speaking routers in different AS's, while *iBGP* sessions are between BGP routers in the same AS. They serve different purposes, but use exactly the same protocol.

3.2 Inter-AS Conversations: eBGP

eBGP is the "standard" mode in which BGP is used, since after all BGP was designed to exchange network routing information between different AS's in the Internet. This is shown in Figure **??**, where the BGP routers implement route filtering rules and exchange a subset of their routes with routers in other AS's.

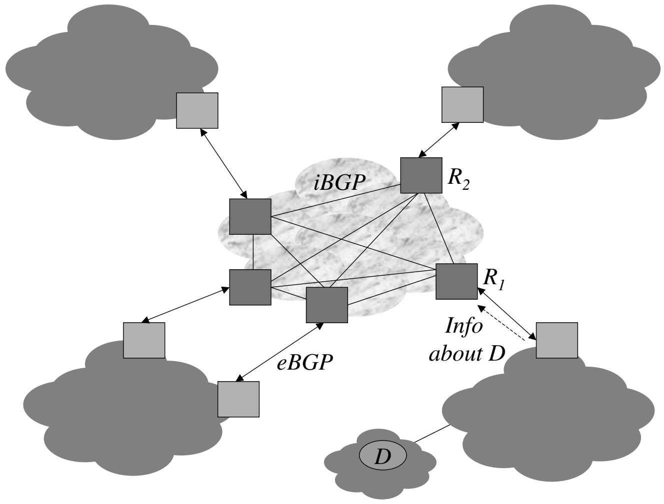

Figure 5: eBGP and iBGP.

3.3 Maintaining Intra-AS Consistency: iBGP

In general each AS will have more than one router that will participate in eBGP sessions with neighboring AS's. During this process, each router will obtain information about some subset of all the prefixes that the entire AS knows about. Two things need to be accomplished at this stage with the route announcements heard from different neighbors:

1. *Completeness.* One of the goals of BGP is to allow each AS to be treated as a single monolithic entity. This means that the several eBGP-speaking routes in the AS must exchange external route information so that they have a complete view of all external routes. For instance, consider Figure **??**, and prefix D. Router R_2 needs to know how to forward packets destined for D, but R_2 hasn't heard a direct announcement on any of its eBGP sessions for D.

   This calls for some kind of route information exchange *within* an AS. This is provided by iBGP sessions running in each AS.

2. *Consistency.*

   BGP attempts to achieve scalability by abstracting each AS into a monolithic entity, but this would be defeated if each eBGP-speaking router had an entirely different and arbitrary set of routes to a given destination. To first order, all routers in an AS should treat any packet destined for an external network in the same way, as far as the deciding which AS to forward the packet to next. Routers within an AS need a way to achieve route consistency for external routes, and a way to consistently make route announcements and withdrawals. This is provided by iBGP sessions running between the BGP-speaking routers.

An important question concerns the topology over which iBGP sessions should be run. One possibility is to use an arbitrary connected graph and "flood" updates of external routes to all BGP routers in an AS. This would require additional techniques to avoid routing loops. BGP solves this problem by simply setting up a *complete mesh* of iBGP sessions, where every BGP router maintains an iBGP session with every other BGP router in the AS. Flooding updates is now straightforward; simply send it to all your iBGP neighbors.

It is important to note that *iBGP is not an IGP* like RIP or OSPF, and it cannot be used to route packets between internal nodes. Rather, iBGP sessions provide a way by which routers inside an AS can use the same protocol (BGP) to exchange information for completeness. In fact, iBGP sessions and messages are themselves routed between the BGP routers in the AS via whatever IGP is being used in the AS! Like eBGP, iBGP also uses TCP.

One might wonder why iBGP is needed, and why one can't simply use whatever IGP is being used in the AS to also send BGP updates. There are several reasons this is inconvenient, but the most important ones have to do with the state model assumed by BGP and the fact that BGP announcements use a large (and rich) set of attributes not present in most IGP's. The first point bears some elaboration—whereas many IGP's rely on periodic route announcements to achieve route consistency in the presence of packet loss and link failures, BGP announcements aren't periodically repeated. Only the KEEP ALIVE messages are periodic. The second point about attribute translation implies that to preserve all the information about routes gleaned from eBGP sessions, it is best to run BGP sessions inside an AS as well.

The requirement that the iBGP routers be connected via a complete mesh limits scalability. As a result, two methods to handle this have arisen, both based on manual configuration into some kind of hierarchy. The first method is to use *route reflectors*, while the second sets up *confederations* of BGP routers. We won't discuss how these are done in this class.

3.4 Routes and Path Selection

We're now in a position to understand what the anatomy of a BGP route looks like and how route announcements (and withdrawals) allow a router to compute a forwarding table from all the routing information. This forwarding table typically has one chosen path in the form of the egress interface (port) on the router, corresponding to the next neighboring IP address, to send a packet destined for a prefix. Recall that each router implements the longest prefix match on each packet's destination IP address.

3.5 Exchanging Reachability: NEXT HOP Attribute

A BGP route announcement has a set of attributes associated with each announced prefix. One of them is the NEXT HOP attribute, which gives the IP address of the router to send the packet to. As the announcement propagates, the NEXT HOP field is changed, with each router replacing the current value with its own. While there are many ways to deal with this within as AS, the important point is that this field definitely changes when an AS boundary is crossed.

This information allows packet forwarding to occur, since packets flow in the opposite direction to the route announcements for each prefix.

3.5.1 Length of AS Paths: ASPATH Attribute

Another attribute that changes as a route annoucement traverses different AS's is the ASPATH attribute, which is a *vector* that lists all the AS's (in reverse order) that this route announcement

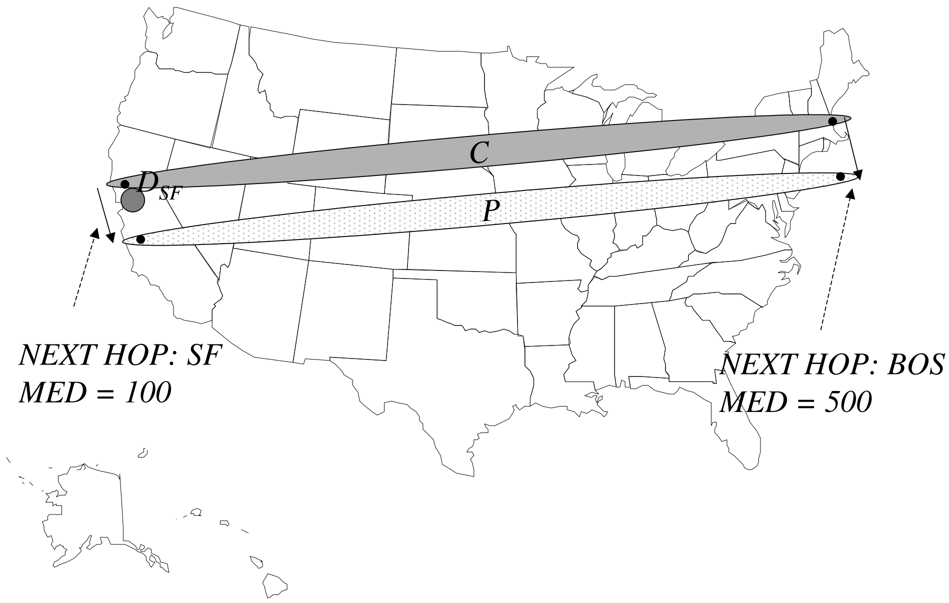

Figure 6: MED's are useful in many situations, *e.g.*, if C is a transit customer of P, to ensure that cross-country packets to C traverse P's (rather than C's wide-area network). However, if C and P are in a peering relationship, MED may (and often will) be ignored. In this example, the MED for D_SF is set to 100 at the SF exchange point, and 500 in Boston, so P can do the right thing if it wants to.

has been through. Upon crossing an AS boundary, the first router prepends the unique identifier of its own AS and propagates the announcement on (subject to its route filtering rules). This use of a "path vector"—a list of AS's per route—is the reason BGP is classified as a *path vector protocol*.

A path vector serves two purposes. The first is *loop avoidance*. Upon crossing an AS boundary, the router checks to see if its own AS identifier is already in the vector. If it is, then it discards the route announcement, since importing this route would simply cause a routing loop when packets are forwarded.

The second purpose of the path vector is to help pick a suitable path from among multiple choices. If no LOCAL PREF is present for a route, then the ASPATH length is used to decide on the route. Shorter ASPATH lengths are preferred to longer ones. However, it is important to remember that BGP isn't a strict shortest-ASPATH protocol (classical path vector protocols would pick shortest vectors), since it pays attention to routing policies. The LOCAL PREF attribute is always given priority over ASPATH. Many routes in practice, though, end up being picked according to shortest-ASPATH.

3.5.2 Choosing Between Multiple Exit Points: MED Attribute

There are many situations when two AS's are linked at multiple locations, and one of them may prefer a particular transit point over another. This situation can't be distinguished using LOCAL PREF (which decides which AS' announcement to import) or shortest ASPATH (since they would be equal). A BGP attribute called MED, for *multi-exit discriminator* is used for this.

It's best to understand MED using an example. Consider Figure **??** which shows a provider-customer relationship where both the provider P and customer C have national footprints. Cross-country bandwidth is a much more expensive resource than local bandwidth, and the customer would like the provider to incur the cost of cross-country transit for the customer's packets. Suppose we want to route packets from the east coast (Boston) destined for D_{SF} to traverse P's network and not C's. We want to prevent P from transiting the packet to C in Boston, which would force C to use its own resources and defeat the purpose of having P as its Internet provider.

A MED attribute allows an AS, in this case C, to tell another (P) how to choose between multiple NEXT HOP's for a prefix D_{SF}. Each router will pick the smallest MED from among multiple choices. No semantics are associated with how MED values are picked, but they must obviously be picked and announced consistently amongst the eBGP routers in an AS. In our example, a MED of 100 for the SF NEXT HOP for prefix D_{SF} and a MED of 500 for the BOS NEXT HOP for the same prefix accomplishes the desired goal.

An important point to realize about MED's is that they are usually ignored in AS-AS relationships that don't have some form of financial settlement (or explicit arrangement, in the absence of money). In particular, most peering arrangements ignore MED. This leads to a substantial amount of *asymmetric routes* in the wide-area Internet, as we'll see in the next lecture. For instance, if P and C were in a peering relationship in Figure **??**, cross-country packets going from C to P would traverse P's wide-area network, while cross-country packets from P to C would traverse C's wide-area network. Both P and C would be in a hurry to get rid of the packet from their own network, a form of routing sometimes called *hot-potato routing*. In contrast, a financial arrangement would provide an incentive to honor MED's and allow "cold-potato routing" to be enforced.

3.5.3 Putting It All Together

So far, we have seen the most important BGP attributes: LOCAL PREF, ASPATH, and MED. We are now in a position to discuss the set of rules that BGP routers in an AS use to select a route from among multiple choices.

These rules are shown in Table **??**, in priority order.

3.6 Failover and Scalability

BGP allows multiple links (and eBGP sessions) between two AS's, and this may used to provide some degree of fault tolerance and load balance. Overall, however, BGP wasn't designed for rapid fault detection and recovery, so these mechanisms are generally not particularly useful over short time scales. Furthermore, upon the detection of a fault, a router sends a withdrawal message to its neighbors. To avoid massive route oscillations, the further propagation of such route announcements is *damped*. Damping causes some delay (configurable using a timer) before problems can be detected and recovery initiated, and is a useful mechanism for scalability.

Priority	Rule	Remarks
1	LOCAL PREF	Highest LOCAL PREF (§**??**). *E.g.,* Prefer transit customer routes over peer and provider routes.
2	ASPATH	Shortest ASPATH length (§**??**) *Not* shortest number of Internet hops or delay.
3	ORIGIN	iBGP-*originated* preferred to eBGP-*originated*. Allows internally-originated routes to be selected over external ones.
4	MED	Lowest MED preferred (§**??**). May be ignored, esp. if no financial incentive involved.
5	eBGP > iBGP	Did AS learn route via eBGP (preferred) or iBGP? Note: this is different from #3 since it doesn't apply to internal routes.
6	IGP path	Smallest IGP path length to next hop. If all else equal so far, pick shortest internal path.
7	Router ID	Smallest router ID (IP address). A random (but unchanging) choice.

Table 1: How a BGP-speaking router selects routes.

With BGP, faults may take minutes to detect and it may take several minutes for routes to converge to a consistent state afterwards.

3.6.1 Multi-homing: Promise and Problems

Multi-homing typically refers to a technique by which a customer can exchange routes and packets over multiple distinct provider AS's. An example is shown in Figure **??**, which shows the topology and address blocks of the concerned parties. This example uses *provider-based addressing* for the customer, which allows the routing state in the Internet backbones to scale better because transit providers can aggregate address blocks across several customers into one or a small number of route announcements to their respective providers.

Today, multi-homing doesn't actually work while still preserving the scalability of the routing infrastructure. Figure **??** shows why. Here the customer (C) address block 10.0.0.0/16 needs to be advertised not only from provider P_2 to the rest of the Internet, but *also* from provider P_1. If P_1 didn't do so, then longest prefix matching would cause all packets to the customer to arrive via P_2's link, which would defeat the purpose of using P_2 only as a backup path.

Now, given that this route needs to be advertised on both paths, how does C ensure that both paths aren't used? One hack to achieve this is by *padding* the exported ASPATH attribute. On the path through P_1, the normal ASPATH is announced, while on the path through P_2, a longer path is advertised by padding it with C's AS number multiple times.

A good way to do extensive multi-homing without affecting routing scalability is a good open problem. In addition to the fact that customer routes must be advertised along multiple paths, effective multi-homing today is often not possible unless the customer has a large address block. To limit the size of their routing tables, many ISP's will not accept routing announcements for fewer than 8192 contiguous addresses (a "/19" netblock). Small companies, regardless of their fault-tolerance needs, do not often require such a large address block, and cannot effectively multi-home. Notice that provider-based addressing doesn't really work, since this requires handling two distinct

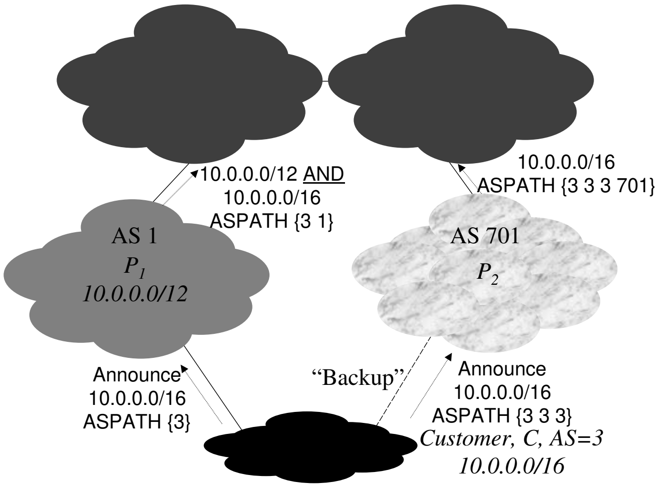

Figure 7: Customer C is multi-homed with providers P_1 and P_2 and uses provider-based addressing from P_1. C announces routes to itself on both P_1 and P_2, but to ensure that P_2 is only a backup, it uses a hack that pads the ASPATH attribute. However, notice that P_1 must announce (to its providers and peers) *explicit* routes on both its regular address block *and* on the customer block, for otherwise the path through P_2 would match based on longest prefix in the upstream AS's!

sets of addresses on its hosts. It is unclear how *on-going* connections (*e.g.*, long-running ssh tunnels, which are becoming increasingly common) on one address set can seamlessly switch on a failure in this model.

3.6.2 Convergence Problems

BGP does not always converge quickly after a fault is detected and routes withdrawn. Depending on the eBGP session topology between AS's, this could involve the investigation of many routes before route convergence occurs. The paper by Labovitz *et al.* from ACM SIGCOMM 2000 explains this in detail, and shows that under some conditions this could take a super-exponential number of steps.

In practice, it's been observed that wide-area routes are often (relative to what's needed for "mission-critical" applications) unavailable. Although extensive data is lacking, the observations summarized in Table **??** are worth noting.

3.7 Summary

BGP is actually a rather simple protocol, but its operation in practice is extremely complex. It has a large number of configuration parameters and allows for a rich set of attributes to be exchanged in route announcements. There are a number of open and interesting research problems in the area of wide-area routing, relating to policy, failover, scalability, and configuration—and understanding the behavior and performance of wide-area routing.

Researchers	Finding	Time-frame
Paxson	Serious routing pathology rate of 3.3%	1995
Labovitz *et al.*	10% of routes available less than 95% of the time	1997
Labovitz *et al.*	Less than 35% of routes available 99.99% of the time	1997
Labovitz *et al.*	40% of path outages take 30+ minutes to repair	2000
Chandra *et al.*	5% of faults last more than 2 hours, 45 minutes	2001
Andersen *et al.*	Between 0.23% and 7.7% of overlay "path-hours" experienced serious 30-minute problems in 16-node overlay	2001

Table 2: Internet path failure observations, as reported by several studies.

4 Summary

This lecture looked at issues in wide-area unicast Internet routing, focusing on real-world issues. We first looked at inter-AS relationships and dealt with transit and peering issues. We then discussed many salient features and quirks of BGP-4, the prevalent wide-area routing protocol today.