Accounting and Routing in the Internet

Introduction

There has been discussion of proposals to engage in the collection of traffic flow measurement information for monitoring and to support charging and accounting. Some suggest that network interconnection points, such as the Border Gateway Protocol (BGP) connections between networks operated by administrations and operating agencies authorized by national governments, would be suitable points to collect detailed information about the traffic flows and usage patterns which could then be used to charge data traffic back to its source.

BGP, which is an Internet routing protocol, has been proposed as a delivery vehicle for statistical and charge accounting information required to bill for the routing of Internet traffic for specific applications such as voice over IP. Additional discussions have built on these suggestions to see whether it would be possible for the source of a traffic flow to control the path it takes through the Internet, and for the receiver of traffic to determine what path it took.

This note offers a technical perspective on Internet routing and accounting, intended to inform the discussion.

Types of routing protocols used in the Internet

Internet routing, in its various forms, derives from a branch of mathematics called Graph Theory, and specifically from a theorem by Bellman, Ford, and Moore. They described an algorithm by which one might start from a node in a network and successively add pairs of arcs and new connected nodes until every node has been connected to the graph. We implement the Bellman-Ford-Moore algorithm in our routing protocols in two ways: a distributed algorithm which we call "distance vector" (also called "Bellman Ford") and a centralized algorithm which we call "link state" (also known as "shortest path first").

In distance vector protocols, each router applies a local policy to its route table and announces the resulting set of routes to its neighbors. The neighboring routers integrate the information into their own route tables, and then announce their new route tables to their neighbors. In this way, route advertisements propagate outward from the destination routers like ripples in a pond. A router receiving a route advertisement assumes that it can send datagrams intended for the advertised route back down the path that the route advertisement took. Examples of distance vector protocols include RIPv2, RIPng, and the Border Gateway Protocol (BGP).

In link state protocols, each router within a network announces a list of its directly connected routers and address prefixes. “A network” may be a defined portion of a network, such as an OSPF Area. Each announcement is distributed throughout the network so that each router has a complete set of these lists. When a router notices a change in the available routing information, it calculates routes to each remote router, and to each address prefix those routers are announcing. Examples of link state protocols include OSPFv2, OSPFv3, and IS-IS, which was originally written for CLNP and has been extended to route IPv4 and IPv6.

Most routing protocols are used within networks, and are called "Interior Gateway Protocols" or IGPs. IGPs attempt to find the best route between any two places, and usually calculate symmetric routes; the sequence of routers and interconnections from A to B is exactly the reverse of the route from B to A. However, one protocol, BGP, is used between networks, and has a very different objective: it attempts to find routes that conform to a network administration's policies, and (by the way) also work reasonably well for the delivery of traffic. The routes derived by BGP are often asymmetric: datagrams from A to B may take a completely different path than those flowing from B to A.

Inter-AS policy routing: the Border Gateway Protocol

To discuss BGP, we must introduce some terms. An **Autonomous System** (**AS**) is a network or set of networks under common administration treated as a unit by the network administrator and, as a consequence, by remote networks. A single company may operate one or more Autonomous Systems. Note that this definition is not geopolitical; ASes operated by multinational companies often transcend national borders. An AS is identified by an **AS Number**, which is allocated by its **Regional Internet Registry (RIR)**. A **BGP route announcement**, which in other distance vector protocols is merely a list of prefixes and their metrics, consists of a set of prefixes along with a set of attributes. One of these attributes is the "**AS Path**": the list of AS numbers representing the ASes that the BGP announcements traversed to reach this location. In other words, this is the list of Autonomous Systems the traffic may follow to reach the destination (keeping in mind that the routing announcements and the routed traffic go in different directions).

Each AS uses routing policies, which implement commercial routing contracts, to determine how it behaves. The routing policies fall into two broad categories: *announcement policies* describe what an AS is willing to say to another AS; *filtering policies* describe the subset of information announced by a neighbor that an AS will use. All BGP routers within an AS apply the same routing policies.

Announcement policies

There are two broad types of contracts between ASes. Service providers offer either **transit** (“go via me to the entire Internet”) or **peering** (“go via me to my customers”) contracts. Announcement policies in a peering contract inform the peer of the prefixes of the AS's own customers. Announcement policies in a transit contract are asymmetric; the upstream AS advertises its entire route database (often aggregated) to its customer, and the customer advertises its own prefixes to the upstream AS.

Filtering policies

If an announcement policy implements a contract, a filtering policy enforces it. An AS receiving announcements from its neighbors filters them, ignoring routes that would cause routing problems or that don't conform to the operator’s policy. For example, an AS will usually ignore received announcements of routes to its own customers (because it prefers to serve its own customers directly), routes that include loops (because they would never successfully deliver datagrams), improper routes (because they would "black-hole" traffic), and other routes at its discretion.

A government network that has a policy of not letting its national data leave the country would filter any route advertisement whose AS Path starts and ends within the country but which contains an AS that is not entirely within the country. In that way, the government network might choose a less efficient route, or might find itself with no route at all to a destination within its own country, but would be sure of honoring its geopolitical policy.

A class of filtering policy being standardized in the **IETF** requires verification of an announcement; the RIRs maintain a database of public keys for their member ASes, and associate the prefixes they allocate with the Autonomous System numbers authorized to announce them (and which are therefore found in the origin of an AS Path). Routers may then filter routes contained in announcements where the prefixes or some subsets of the prefixes are not properly signed using the indicated key. An incorrectly signed announcement indicates that an unauthorized AS is announcing routes inappropriately perhaps because of a configuration error, or maybe for a malicious reason.

Integration of routing data

A BGP Router often has ongoing sessions with a number of neighbors, and chooses among the routes that its neighbors announce to it by first filtering them according to its filtering policies, and then accepting the routes that give it the shortest AS Path to any given prefix. It may also use other attributes advertised by the neighbor, or add attributes itself, that color that decision. Furthermore, it may have internal policies regarding some routes, designed to block attack traffic or other communications from using its network. The route table resulting from the integration of routing information from several neighbors with that of the IGP is called the "Routing Information Base", or **RIB**, and an optimized version of that used in data forwarding is called the "Forwarding Information Base" or **FIB**.

What a router knows, and what it doesn’t know

Since Internet routing is "to a destination", the result is that any given router has a pre-calculated decision table regarding how *it* might forward traffic – it knows what it itself will do.

A BGP router does not, however, know how data will be routed *to or through it;* it can at most infer a few possibilities. This is because inter-AS routing in the Internet is asymmetrical and based on locally-enforced policies.

Generally speaking, if a packet is forwarded to one AS, and that AS must forward it to another, the AS will route traffic using as little of its own resources as possible. This results in a behavior called "hot potato routing"; packets bounce quickly from AS to AS until they arrive at the AS serving the destination, which carries the packet as far as it needs to go. Since any two edge ASes generally have different upstream providers, traffic from A to B will primarily use B’s provider and will get to it by a path selected by A, and traffic from B to A will primarily use A’s provider and will get to it using a path selected by B. In addition, costs change; the contract between two ASes may, for example, stipulate that the cost of transit service will be one amount for a certain amount of traffic over a stated time interval, and another amount if more is sent during that interval. In that case, the AS might send traffic up to that threshold, and then change routing to a second AS whose transit cost is lower than the new charge imposed by the first AS. Alternatively, it might distribute traffic to a remote destination across several "next hop" ASes. In addition, there can be time-of-day charging, resulting in different routing decisions at different times of day.

Between issues of this type and the fact of occasional failures in a complex global network, Internet routing changes constantly, and the changes are often not obvious in a given router’s route table; each of the options is "one of the possible routes", but different routes are in actual use at different times. The one thing that can be said for sure is that companies will do their best to honor their service level agreements with their customers at the lowest possible cost to themselves.

Accounting Traffic in the Internet today

Business model changes

In the past, voice traffic was transported over a dedicated voice infrastructure, and the data network infrastructure was established in parallel so that voice and data traffic did not interfere with each other. Traditional voice accounting and performance functions are standardized within **SS7** (Common Channel Signaling System No. 7), the global standard for telecommunications, defined by the ITU-T. The success of data networks led to the development of techniques to encapsulate voice traffic in IP packets, and thus Voice over IP (VoIP) was born.

During the initial phase of VoIP, the **Public Switched Telephone Network** (**PSTN**) switches remained in the network and took full control over the voice calls. Instead of a dedicated voice trunk between the PSTN switches, gateways to an IP data network now interconnected them by encapsulating voice streams in IP packets. Accounting records and performance statistics were still gathered by the PSTN switches, as defined by SS7. In addition, the gateways provided further details from an IP transport point of view.

The next step was the development of voice technologies that no longer required PSTN components. Instead, a software switch or IP telephony server delivered the PSTN switch functionality. Alternatively, peer-to-peer applications (such as Skype for voice over the Internet) connect the IP phones without central call control. **H.323** (ITU-T) and **Session Initiation Protocol** (**SIP**) became the standards for voice call signaling and control in an IP network. Relevant data sources for accounting and performance purposes became IP telephony applications as well as voice gateways and network elements.

It should be noted, however, that carrying voice traffic over an IP network is not the same as carrying VoIP on the Internet. It is now common practice for telephony operators to use IP within their own private and dedicated voice networks, but they may also allow data traffic within those networks and may connect them to the Internet. On the other hand, many applications available to end users allow them to run voice calls over the Internet from their IP phones or their personal computers.

Disrupted Charging Model

With the deployment of the Internet, the fundamental paradigm of communication changed. Communication was not limited to voice or video, as it had been in the PSTN and the ISDN; any application that uses IP datagrams communicates, and those communications are more commonly exchanges of files or interactive data exchanges. As a result, with the move from SS7 to VoIP, the business model changed: Instead of applying the legacy SS7 paradigm "Don't forward traffic if you can't bill it.", the voice traffic is now "just another series of packets" or "just another session", and the accounting of communication must not only consider voice, but any session or flow of traffic using the network.

In data networks, end users pay a monthly flat fee for Internet connectivity. They may use their allowed bandwidth for email, web browsing, or VoIP, and there is no difference in pricing. But as voice traffic moves from the traditional telephony service to VoIP, there is reduction in the direct revenue for the traditional telcos (Telecommunications Companies) unless they are able to carry data traffic and charge effectively for it.

Many countries are concerned about the impact of services like Skype on their telephone-originated revenue. Telephone revenue is important for inward investment in telecommunications infrastructure and in some cases also serves to supplement general taxation income. Pricing agreements for domestic telephone calls are purely a national matter and are controlled by regulators. Pricing for international calls is subject to complex multi-party agreements, and charging is often asymmetrical with calls to one country from another costing more than calls in the opposite direction.

The charging model for VoIP services is significantly different: for VoIP-to-VoIP calls, both parties pay just their local fee for Internet connectivity, while for VoIP-to-telephone calls the caller may pay an additional fee roughly equivalent to the cost of a local call in the receiver's country. Furthermore, the fee for a VoIP-to-telephone call will be paid to the VoIP company (such as Skype) operating outside the caller's home country.

If one focuses on voice, the loss of traditional telephony revenues raises the question of detecting VoIP traffic and either charging extra for it or blocking it. Such charges or blocks might be imposed at national borders, and charges might be levied on the VoIP service provider or on the local user. But given that the fundamental communication paradigm has changed, that would seem to miss the point; why require complex new measures to tax a slender proportion of the whole traffic flow when the fundamental and growing business is in Internet data communications as a whole?

To understand the economics here, one must also understand the costs. These include, at least, the cost of equipment (a one-time charge), the cost of right of way (which may be leased or purchased), and the cost of bandwidth (which is usually a monthly charge). Import tariffs charged by governments drive the cost of equipment up, land use permits can be very expensive, and recovery of sunk cost in undersea cables can be exorbitant. Another thing is the basic cost of doing business; it is estimated that in the PSTN, the implied costs in per-minute charging for telephone calls is $0.70 for every dollar of telephone company revenue. That is the reason telephone companies have moved from metering calls to selling blocks of minutes per month.

Implications for Law Enforcement

Related to this general topic is the issue of law enforcement access to communications. For much of the 20th century, law enforcement access to communications consisted of PSTN Wiretap, which includes the ability for law enforcement to access billing records or require real time call reporting (in the US, referred to as “Pen Register” and “Trap and Trace”), and also includes the ability to capture and record actual conversations.

In the Internet, in which voice is just another application, narrowing the lens to voice makes little sense; if a criminal knows that voice can be tapped but instant messaging or electronic mail cannot, he will use the untapped communication mechanisms. Hence, in the Internet, intercept standards focus on the capture of IP datagrams, and data retention specifications focus on information available from IPFIX in some circumstances, and more generally from application (electronic mail, the web, and instant messaging) log files.

Law enforcement has asked, and answered, the same question: in an Internet world, a focus on voice is an anachronism.

IP accounting in the Internet: NetFlow and IPFIX

In the late-1990’s, Cisco's proprietary **NetFlow** protocol, which built on concepts developed at UCSD, became the de facto IP accounting standard throughout the industry. The basic output of NetFlow is a **flow record** exported from a device such as a router on which the NetFlow services are enabled. An exporter monitors packets entering a network interface, and creates flow records that describe the sessions that the packets are part of. These flow records are exported to a collector, which uses the information for network monitoring, capacity planning, security analysis, and, in some situations, billing.

IPFIX, which stands for “IP Flow Information eXport,” is an IETF effort to standardize an export protocol similar to NetFlow - specifically, a protocol that exports flow-related information. The IPFIX protocol specifications are largely based on the NetFlow version 9 export protocol. The IPFIX protocol, which is a flexible protocol based on templates, can choose from a long series of information elements for its export: either well-known ones, as registered by the Internet Assigned Numbers Authority (IANA), or enterprise-specific ones.

A typical flow record could be composed of:

- source IP address (identifies the system initiating the connection),
- destination IP address (identifies the system it is communicating with),
- protocol, destination port (identify the protocol used),
- application (the result of the Deep Packet Inspection, identifying the real application, such as Skype),
- the flow start and end times
- number of packets and bytes sent
- input interface of the monitoring router

In other words, the IPFIX exports information about the Who (the communicating systems), the What (protocol and application), the When (timing), the Where (router and interface location), and How Much (the data volume). A typical session between a pair of systems A and B is reflected in a pair of flow records, one from A to B and another from B to A.

General considerations for traffic measurement and options for international internet connectivity

There are further proposals that administrations take appropriate measures nationally to ensure that parties (including operating agencies authorized by nation states) involved in the provision of international Internet connections negotiate and agree to bilateral commercial arrangements, or other arrangements as agreed between administrations, enabling direct international Internet connections that take into account the possible need for compensation between them for the value of elements such as traffic flow, number of routes, geographical coverage and cost of international transmission, and the possible application of network externalities, amongst others. For more information, look here at a supplement to the ITU-T’s D.50 recommendation.

These approaches assume that an unmodified routing protocol (BGP) can deliver traffic statistics (which BGP doesn't do) on a hard-to-detect application such as Skype. It also assumes that if a Skype call (or other traffic stream) crosses a specific country, that country could be identified and receive some revenue for this call (exactly like in the SS7 world).

In simple language, the proposals are based on the assumption that BGP could be used to deliver traffic statistics to report on the use of an application such as Skype. It is implied that using these mechanisms, if a Skype call crosses a specific country, that country could be identified and receive some revenue for this call (as would be enabled by SS7).

However, BGP doesn't currently deliver traffic statistics; it is a routing protocol, not a measurement and accounting protocol. Also, applications running on top of RTP (the transport used by H.323 and SIP voice and video) are particularly hard to detect: one of the reasons is that they have RTP has no assigned UDP port number.

As discussed in the next section: this model simply will not work!

It will simply not work!

Reason 1: BGP ASes are not countries

Even if it were possible to collect information about the ASes through which the Skype traffic passes (perhaps using the BGP AS-PATH information for the associated routes) there is a big problem: BGP ASes routinely transcend national boundaries. As previously described, an AS is a network operator's administrative domain, and such domains are often organized across national boundaries. So it is not always possible to determine which countries the traffic traverses. This is a major roadblock to national billing.

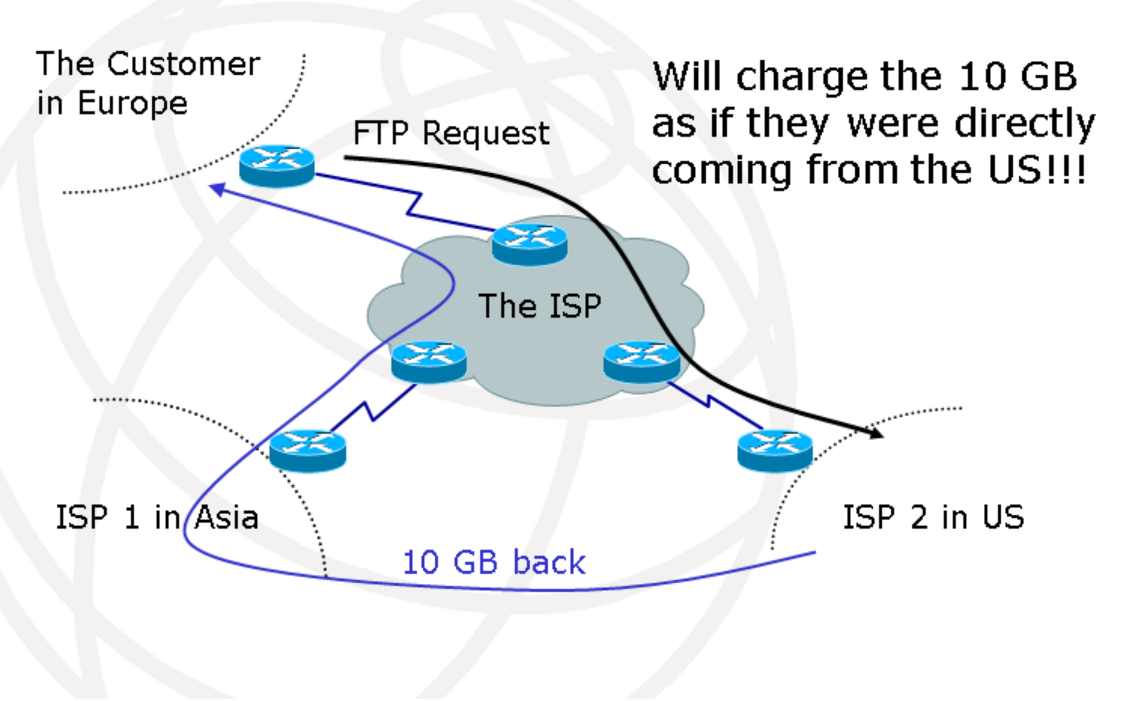
Reason 2: which traffic direction to bill?

If an end user pushes some traffic (like a huge video), he should pay for this. However, what if an end user watches some YouTube videos, should the content provider paid for this? Maybe not, because it is the end user not the content provider who benefits from the viewing the content. Or maybe the rule is that the traffic initiator should pay the bill, like in the old telephony. These questions become even more important when the traffic volume is dramatically different in different directions. What if a user logs into a remote server, and initiates a download from his own local server: would the user pay for the small request, while the remote server would pay for pushing the content? As you see, it’s not obvious who should be charged for the traffic. It should also be remembered that the destination and source already pay for connectivity to the Internet, and that that payment will have an element of bandwidth and data-usage associated with the pricing. What is for sure is that, to cover all the different scenarios, the destination and source paths must be identified.

Reason 3: BGP traffic might be asymmetric

As we learned in the BGP section, BGP policies will depend on an AS's own commercial interests. Illustrating the hot potato routing with an example below, it could happen that the traffic from a customer in Europe accessing a server in the USA would be sent directly to the US, while the return traffic would come back via Asia. In BGP, if the routes are stable (which is absolutely not a given), the operator would know the route the traffic will take to reach the destination: this is the destination AS-PATH. However, there is no way for the same operator to know the route the return traffic took: although the source AS-PATH may provide an approximation it is not a guarantee. The source AS-PATH is a lookup in the BGP table for the source address; In other words, this is the path the operator would use ***to reach*** the source address, and not the path the traffic took when it came ***from*** the source address. In conclusion, source sensitive billing is not possible with an asymmetric protocol such as BGP.

Issue: BGP Asymmetry Problem



Will charge the 10 GB as if they were directly coming from the US!!!

Reason 4: applications are hard to detect

Skype traffic flows are hard to detect. Indeed, some information needs to be correlated between different packets, imposing some stateful requirements on the Deep Packet Inspection (**DPI**) system. For Skype detection, the DPI system is required to analyze up to 6 consequent packets in the flow in order to determine, with a certain level of confidence, that the flow carries a Skype call. Since it is discoverable only by fairly sophisticated heuristics, we're talking about specialized detection and measurement equipment, equipment that due to algorithmic complexity is limited in throughput rate and which has to look at the actual data flow rather than at an offline summary of it. Such equipment is typically very expensive to the point of negating the benefits of charging.

On small routers typically used in a branch office, it is possible to use DPI, in conjunction with IPFIX, and export flow records containing: source IP address, destination IP address, protocol, source port, destination port, and application. However, on the routers running BGP, which are big routers whose primary goal is to route packets as fast as possible, it is not practical to include a DPI engine.

On the other hand, only the router running BGP contains the required destination AS-PATH.

Reason 5: Hiding from DPI

Some applications don’t want to be discovered by DPI engines, and constantly change their specifications. That is, the pattern of information that a DPI engine uses to identify a flow as belonging to an application can be easily changed by the application making it hard to detect and forcing the DPI engines to evolve constantly. The ultimate obfuscation is encryption, which completely hides the packet payload, and as a consequence makes it impossible to identify the application.

Reason 6: IPFIX and sampling

Sampling is the process of selecting only a fraction of the traffic passing across an interface for flow reporting. As a matter of practicality, sampling may be the only option on the high speed interfaces of BGP routers because inspecting all packets at line rate is prohibitively expensive. Typically, sampling rates of 100 or 500 (only 1 out of 100, or 1 out of 500 packets are selected) are used in the BGP routers in the Internet today. IPFIX exports the flow records generated from the sampled packets, along with the information about the sampling rate. Sampling gives good statistical results to generalize information about traffic flowing through a router, but sampling introduces some approximations regarding billing (by multiplying the number of bytes and packets by the sampling rate) and cannot be used to detect and charge for individual flows.

Reason 7: Only the BGP routes will be accounted

While BGP is the normal protocol for exchanging routes between ASes, routes can also be manually added to the routing table (static routes). Obviously, an accounting system exclusively based on the BGP protocol would fail to account the traffic following the static routes.

What if we could really make work?

The previous section provides a number of issues that make per-flow billing in the Internet hard or impossible, but let’s assume just for a minute, that we find solutions to all these problems: what would the Internet look like? If a country receives some money for traffic traversing its infrastructure, there will be an incentive to attract more traffic. The model breaks when each country wants to attract the traffic. Indeed, each country will advertise, with BGP announcements, that the rest of the world (any other BGP AS) is available via its own network. That is, instead of a shortest path paradigm for traffic routing, we will arrive at an Internet where traffic is routed through as many countries as possible! In the end, it will result in an unstable Internet, where the user quality of experience (**QoE**) will be bad if the content is not local.

So what is the solution?

There is no magic solution to solve the issues of Internet access cost for developing countries. Certainly it is not advisable to try to solve a socio-economic problem with a simple change in technology.

However, small steps are possible. Recently, a new Internet Exchange Point (**IXP**) was launched in Kinshasa, Democratic Republic of the Congo. The Kinshasa IXP (KINIX) was funded through the Internet Society's Community Grants Program and is managed by the Democratic Republic of Congo ISP Association (ISPA-DRC), as part of its DRC-IX project, which aims to establish IXPs in Kinshasa, Lubumbashi, and Goma. KINIX will serve as a catalyst for innovation and development of Internet services and applications in the Democratic Republic of Congo, and will support Government efforts to implement E-government services and lower the cost of developing local hosting and application development. The presence of KINIX will improve local Internet resilience by eliminating the dependence on international connectivity for local Internet services and Internet-based communications.

The Internet Society's Africa Interconnection and Traffic Exchange program has been actively supporting the development of IXPs and regional interconnection in the region. The program aims to have 80% of Internet traffic exchanged in Africa by 2020, keeping local traffic local. This objective has been boosted by the appointment of the Internet Society to implement the African Union's African Internet Exchange System (AXIS) program.

Maybe increased connectivity between ASes through IXPs will magnify the benefits of Internet connectivity, will attract more local online business, and provide a significant economic stimulus as larger percentages of the population are able to get on line. Perhaps these benefits will go some way to offset the lost revenues from the declining legacy telephone systems. What is the for sure is that using DPI to monitor and charge VoIP calls in the same old way … will simply not work!

Fred Baker, Adrian Farrel, and Benoit Claise