Internet peering and settlements

Geoff Huston, APNIC Chief Scientist

The business world today features many complex global service activities that involve multiple interconnected service providers. Customers normally expect to execute a single paid transaction with one service provider, but many service providers may assist in the delivery of the service. These contributory service providers seek compensation for their efforts from the initial provider. However, within a system of interdependent providers a service provider may undertake both roles of primary and contributory provider, depending of the context of each individual customer transaction.

In a system where there are many mutual service provision transactions it is common to see a balance of services rather than individual transaction payments between providers. Inter-provider financial settlements are also commonly used as a means of reconciling residual imbalances in the accounting of such mutual service provision tasks. In this article I'd like to describe how this has been applied to the Internet, and look at the Internet's approach to interconnection and financial settlements.

For example, today airlines have codeshare agreements where a customer may purchase a ticket from one airline while the flight is operated by another airline. Given that one provider has received the revenue for the service, and another provider has incurred the costs of providing the service, there is a need to pass some form of payment from one provider to the other. Rather than undertaking a separate inter-provider financial transaction for each codeshare journey, airlines can use a more efficient arrangement that uses inter-provider settlements. Each airline retains the original ticket sales revenue, and accumulates amounts in credit and debit from the execution of the codeshare flights between providers. They then settle any residual imbalance in the mutual service account with a single financial transaction at the end of each settlement period.

The Internet operates in a similar fashion: its services are provided by some 40,000 constituent network service providers that must not only interconnect with one another, but also execute a set of inter-provider arrangements to ensure that each service provider is duly compensated for their respective efforts in providing end-to-end services to the network's clients.

The telephony model

To look at the Internet interconnection and settlement structure, it is useful to look at its immediate predecessor, which is the telephony inter-provider financial settlement model.

The retail model for telephony appears largely to have been borrowed from the postal service, where the caller pays its local carrier for the entire cost of the call, while the called party pays nothing to receive the call. When both the caller and the called party are connected to the same carrier this is quite straightforward, and the carrier charges the caller for the entire cost of the call. When we take the same model and apply it to international phone calls, the common intent is to preserve the same simple retail model: the caller pays the cost of the call. Given this now involves the telephone carriers undertaking mutual service provision activities, the telephone industry devised the concept of inter-carrier call accounting financial settlements to redress any residual imbalances in mutual service provision.

Within the framework of this interconnection model, two national carriers interconnect at an agreed handover point. As part of this interconnection they establish a call minute settlement rate, which is the rate one carrier bills the other for the residual imbalance of terminated calls incoming from the originating carrier's network, as the calls pass through a handover point into the terminating carrier's network.

The originating provider receives a payment from the caller for the entire call, and accumulates a debit to the terminating provider for the termination costs of the call. The terminating provider receives no payment from the called party, and accumulates a credit from the originating provider for the same call termination cost. A periodic financial payment from one provider to the other allows the two providers to "settle" the net of these debit and credit accounts, thereby providing a form of equity of cost distribution in meeting the costs of the calls made between the two providers.

Calls are measured in units of call minutes in the interconnection domain, so the debit and credit accounts are also measured in call minutes. Where there is equity of call accounting rates between two providers, bilateral inter-provider financial settlements are used in accordance with the originating call minute imbalance, in which the provider hosting the greater number of originating call minutes pays the other party according to a bilaterally negotiated rate per residual call minute imbalance as the mechanism of cost distribution between the two providers.

It's notable that the general bilateral telephony settlement model does not admit multi-party transit arrangements. Such arrangements are handled using further forms of inter-carrier agreements, where a carrier may hand-off call requests to a third party carrier at some mutually agreed call minute rate, and similarly a carrier may engage another carrier to act as its call terminating carrier for an agreed share of the call termination settlement fees.

Because the telephony model includes local monopolies, there is no inherent market-based capability that prevents a carrier setting its call termination settlement rates to a level that is in excess of its actual call termination costs. The resultant distortions and economic inefficiencies in the inter-carrier domain have acted as a powerful driver behind the increasing interest from some high volume calling party

carriers in Voice over IP (VoIP) solutions that bypass these call accounting settlements. In such situations the originating carrier uses VoIP trunking instead of call handover at the inter-carrier interconnection point and terminates the call request within the terminating carrier's network as a regular internal call. This allows the originating carrier to avoid the call termination settlement rates.

The end customers of these VoIP trunk services benefit from a lower priced service offering, adding an associated commercial pressure on the terminating carrier to remove the monopoly rental component from its call termination settlement rates. It would be a highly retrograde step to see a new wave of international regulation that entrenches these inefficient distortions of monopoly rentals in the telephone sector by attempting to prevent, by regulation, the use of alternate voice trunking solutions, such as VoIP, in the international telephony domain.

Another by-product of this monopoly-based distortion in the inter-carrier settlement rates is the emergence of a lobby group comprising the current beneficiaries of these arrangements who, perhaps understandably, are highly motivated to see these arrangements continue and potentially extend to encompass the Internet. This contemplated extension of interconnection regulations from telephony into the Internet environment again represents a retrograde step in terms of introducing regulatory distortions and significant inefficiencies into what is at present a functional and efficient Internet interconnection market.

Internet considerations

There are a number of important technical differences that exist between the telephony and Internet models of carrier interconnection. These differences are fundamental and have confounded all attempts to cleanly map telephony interconnection models into the Internet environment. The most critical of these differences are described as follows:

There is no "call"

Unlike a telephony call, there is no concept of state initiation in an IP network to pass a call request through a network and lock down a network transit path in response to a call response. The network undergoes no state change in response to a packet being passed through the network. Therefore, no means is readily available to the carriage service operator to identify that a call has been initiated, and by which party.

Packets may be dropped

When a packet is passed across an interconnection from one carriage service provider to another, no firm guarantee is given by the second provider that the packet will definitely be delivered to the destination. The second provider, or subsequent providers in the packet's transit path, may drop the packet for quite legitimate reasons, and will remain within the Internet Protocol specification in so doing. Indeed, the TCP protocol uses packet drop as a rate-control signal, which is necessary for its efficient operation.

The broader implication here is that the quality of the packet transit service one carrier may anticipate from its adjacent carrier peer when passing packets across an interconnection is inherently undefined.

Packet paths are not necessarily symmetric

The inter-carrier path a packet takes from one client to another is not necessarily the same path that a packet takes in the reverse direction. This path asymmetry means there is no direct analogy to the virtual circuit model used in telephony. When two clients, A and B, exchange traffic, a carrier may only see traffic flowing in the direction from B to A and not see any traffic from A to B. This does not indicate there is no such traffic, but that the traffic flowing in the opposite direction uses a different inter-carrier transit path through the network.

In a hypothetical case of traffic flow-based inter-carrier service accounting, when a carrier sees just one half of a traffic flow, it's unclear how a carrier can reliably determine whether it should claim a service debit or a credit from its adjacent carriers in passing the traffic towards its intended destination.

End-to-end resource management

In the telephone world, the establishment of a virtual circuit to support a voice call represents an exclusive claim on a unit of capacity in the network that excludes all other potential users of that capacity. In the Internet, there is no concept of resource exclusion. In other words, the Internet does not use a network-based management regime to allocate its resources to support individual transactions.

The end-to-end architecture of the Internet places the responsibility for resource management and allocation on the collection of end systems that generate traffic at any point in time. In this architecture it's conventional to see traffic flows across the network being regulated by the Internet's transport protocol, TCP. Each traffic session adapts to attempt to make the most efficient use of the entire set of network resources, while at the same time attempting to sustain a stable state where each individual traffic flow claims an equal volume of network resources for itself.

Internet interconnection arrangements

The retail model used by the Internet is not one of "sender pays," or "receiver pays". The retail service is not one where either the sender or the receiver funds the entire end-to-end transit of a packet through the Internet as part of the retail tariff structure for Internet services. Indeed, given that the end-to-end transmission of a packet is not even an assured outcome in the Internet architecture, such a tariff model

would not match the nature of the service provided by the underlying IP network. The Internet retail model is one where a customer contracts with a carriage service provider for an access service. This is a customer/provider relationship, where the customer funds its carriage service provider for all packets that are sent or received by the customer.

The translation of this retail service model into the inter-carrier interconnection environment preserves this particular form of customer/provider relationship at the retail edge of the network, translating it into the context of two interconnecting carriers. A carrier that is a customer of another carrier pays for an access service, where the customer carrier funds all traffic to and from its provider (or upstream transit) carrier.

Who is the provider and who is the customer in such an arrangement is not pre-determined by any objective or regulatory determination. Each carrier assesses its value and the value that the other carrier is bringing to the proposed interconnection.

If one carrier believes it brings the greater value to the interconnection, then it would naturally only contemplate the interconnection as the provider and the other carrier as its customer. If the other carrier reaches a similar conclusion that the first party is providing a greater value to the interconnection, then they would likely proceed to the next step of negotiation of service terms and conditions between provider and customer that recognizes the extent of the difference in relative value. Sometimes this takes the form of a conventional wholesale relationship, while at other times more creative labels are used for much the same form of customer relationship, such as *paid peering.*

If both carriers assess their relative value to be greater than the other, and both would assume the role of provider in an interconnection, the mismatch in relative value perceptions would imply that any attempt to execute an interconnection between these two carriers would not be stable. Such a failure to directly interconnect will not necessarily partition the network. A more typical outcome in such a case is that any traffic exchanged between customers of these two carriers would be passed through indirect transit arrangements.

There is another possible outcome of this self-assessment of perception of value in an interconnection, where both parties see approximately equal mutual benefit in interconnecting. Here a "Sender Keep All" (SKA) relationship is appropriate, where the parties exchange traffic in both directions but do not exchange any funds in either direction. These SKA arrangements are typically referred to as peer relationships.

In all of these relationships, the parties themselves do not have to agree on what that measured value or scope may be in absolute terms. Each party makes an independent assessment of the value of the interconnection; both in terms of the perceived size and value it brings to the interconnection and the value of the assets that the other party brings. If both parties reach the conclusion in their respective terms that a net balance of value is achieved, the SKA interconnection is a stable one. If one party believes it brings a greater value to the interconnection than the other, then any SKA interconnection would result in leverage of its investment by the smaller party, and an SKA interconnection would be unstable.

These two forms of interconnection, namely the customer/provider relationship and the SKA peer relationship, form the basis of the entire set of connections that collectively support a coherent and fully connected Internet.

An outcome of this interconnection model is that the service providers' options for business optimization include a strong incentive to increase the size, scope, and efficiency of their operations within the consumer and wholesale market space. That way, non-financially settled SKA peering can be negotiated with larger providers, thereby reducing the additional outlays required to purchase upstream transit services to complement these peering connections. This in turn results in a highly interconnected Internet that improves both the performance and operational costs of the service offering, both of which translate to improved consumer offerings in a highly competitive service industry.

Market-based interconnection

In many ways interconnection in the Internet can be seen as the operation of an open market within the broader framework of a generally deregulated industry. Each party brings assets to the market place and attempts to reach mutually satisfactory arrangements with other actors in the same market. Each party is attempting to reach precisely the same outcome, namely comprehensive connectivity, in a maximally efficient manner to minimize its expenditure while meeting its requirements for connectivity and capacity.

More than twenty years of experience in operating this market-based framework for Internet connectivity has shown that the Internet is capable of scaling from a few dozen service providers to currently over 40,000 component networks. At the same time it has proved to be capable of sustaining its objective, namely that of universal connectivity across the entire Internet's interconnection domain. This has all been achieved without the imposition of overriding regulatory impost and with ever-increasing performance service offerings for the end consumer at pricing levels that continue to fall over time. The evidence from this experience points to a conclusion that this market operates in an efficient manner, and this efficiency directly translates into cost efficiencies in the retail market for Internet services.

The ultimate beneficiary of this form of market-based Internet connection is the end consumer who, for the price of a local access service across the last mile, gains access to the entire world of the Internet.

As Chief Scientist at APNIC, Geoff Huston heads APNIC Labs, conducting leading research on Internet infrastructure, standards, and operations. During the 1990s, Huston was instrumental in the establishment of the Internet in Australia, first through the academic network, AARNet, and then through Telstra's national Internet services.