The ISP Column
An occasional column on things Internet

Where's the Money?

- Internet Interconnection and Financial Settlements

Geoff Huston
APNIC
January 2005

It is often the case that it takes a significant amount of effort to produce seemingly simple outcomes. While we now take the net for granted and see nothing terribly spectacular behind the click of a mouse to bring up a web page, the mechanics behind this action are no less extraordinary than they ever were, no matter how much we now see this as the Internet simply doing what it was designed to do.

While the technical engineering aspects of the Internet are impressive, the business side of the network is no less impressive. In this article we will look at the manner in which Internet Service Providers (ISPs) interact from the perspective of their financial dealings, and take a look at some of the business pressures that are defining the structure of today's Internet industry.

Service Supply Systems and Cost Distribution

Any large multi-provider distributed service sector has to address the issue of cost distribution at some stage in its evolution. Cost distribution is the means by which various providers can participate in the delivery of a compound service to a customer who purchases the service from a single provider, and where providers can each be fairly compensated for their costs in an equitable structure of interprovider financial settlement.

As an example, when an airline ticket is purchased from one air service provider, various other providers and service enterprises may play a role in the delivery of the service. The customer does not separately pay the service fee of each airport baggage handler, caterer, air terminal or other form of service. The customer's original fare, paid to the travel service provider, is further distributed to those other providers who incurred cost in providing components of the total service. These costs are incurred through sets of service contracts, and are the subject of various forms of interprovider financial settlements, all of which are invisible to the customer. Posting an international letter (remember them?), or making a telephone call also may involve a number of service providers working in concert to complete the service transaction, each of whom need to be compensated for their contribution.

The Internet is in a very similar situation. Across the world there are some tens of thousands of constituent networks who must interconnect in one fashion or another to provide comprehensive end-to-end service to each client. In supporting a data transaction between two clients, the two parties often are not clients of the same network or the same ISP. Indeed, the two ISPs often do not directly interconnect, and one or more additional ISPs must act in a transit provider role in order to complete the network transaction.

So we can now pose the Internet's business engineering question with a little more clarity. Within the Internet environment, how do all the parties to various network transactions who incur cost in supporting third party transactions receive compensation? Or, more simply, what is the cost distribution model for the Internet?

Lets examine the basis for Internet interprovider cost distribution models and then look at the business models currently used in the interprovider Internet environment. This area commonly is termed "financial settlements", a term the Internet has borrowed from the telephony industry.

Financial Settlement Models

Financial settlements have been a continual topic of discussion within the domain of Internet interconnection. To look at the Internet settlement environment, let's first look at the use of interprovider financial settlements within the international telephony service industry. Then, we will look at the application of these generic principles to the Internet environment.

Within the traditional telephony model, interprovider peering takes place within one of three general models: Bilateral Settlement, Sender Keep All and Transit Fees.

- Bilateral Accounting Rate Settlements

  One of the better understood, or at any rate one of the most common, models of inter-provider interconnection is the bilateral call accounting settlement model used in the telephone industry.

  This model can be described in a simple two party transaction: Alice, a customer of service provider (A) calls Bob, a customer of a second service provider (B). In this model Alice pays her provider A for the call, and Bob incurs no cost from B in receiving the call. Provider A has received all the revenue for the call, while B, in terminating the call, has performed an equal service to that of A. To some level of approximation both A and B have performed equivalent functions in supporting Alice's call to Bob, but only Alice's provider has received revenue for the call. The system corrects itself only if A compensates B for its costs in terminating Alice's call.

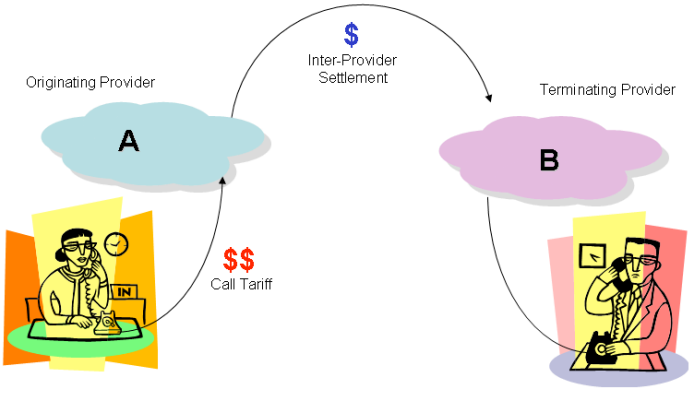

Figure 1 – Two Party Call

In practice, for each accounting period, A will tally the number of minutes of calls originated by A and terminated by B, and also tally the number of minutes of calls originated by B and terminated by A. The net difference is the settlement minutes. If A is a net originator of calls, then A will pay B an amount equal the settlement minutes multiplied by the agreed call accounting rate. If A is a net terminator of call minutes, then B will pay A the settlement minutes multiplied by the call accounting rate.

The call accounting rate is the financial charging rate that has been bilaterally agreed upon between the two parties. Because both parties can charge each other using the same accounting currency, the ultimate financial settlement is based on the net outcome of the two sets of call-minute transactions. The financial settlements provide the accounting balance intended to ensure equity of cost distribution in supporting the costs of the calls made between the two providers.

This financial settlement system has an extensive legacy in international telephony. The call accounting rates were set to a rate that was intended to cover the network costs associated with supporting a call.

- Sender Keeps All Settlement Model

The second financial model is that of Sender Keeps All (SKA), in which each service provider invoices its originating client's user for the end-to-end services, but no financial settlement is made across the inter-provider interconnection structure. Within the bilateral settlement model, SKA can be regarded as a boundary case of bilateral settlements, where both parties simply deem the outcome of the call accounting process to be absolutely equal, and consequently no financial settlement is payable by either party as an outcome of the interconnection.

- Unilateral Transit Fees

The third model is that of unilateral transit fees, in which the parties align into the roles of provider and customer, where the provider charges the customer for services provided. This arrangement is used in the telephony environment as the basis of the long-distance/local access interconnection arrangements, where the long distance provider assumes the role of service provider to the local access operator.

Again, this case can be viewed as a boundary case of a general bilateral settlement model, where in this case the parties agree to apply call accounting in only one direction, rather than bilaterally.

Settlement Models - Topics and Trends

The bilateral call accounting telephony settlement model is by no means stable, and currently, significant pressure is being placed on the international accounting arrangements to move away from bilaterally negotiated uniform call accounting rates to separately negotiated rates for calls in each direction of a bilateral interconnection. Simultaneously, communications deregulation within many national environments is changing the transit fee model as local providers extend their network into the long-distance area and commence interconnection arrangements with similar entities. Criticism also has been directed at the bilaterally negotiated settlement rates, because of the observation that in many cases the accounting rates are not cost-based rates but are based on a desire to create a revenue stream from accounting settlements.

In theory the accounting rate system is intended to create a fair and balanced outcome, where settlement payment ensures that neither provider can leverage off the other and indirectly sell services provided by the other party without fair compensation.

While this sounds fine in theory, in practice this has not been the case. In recent years the generally used international call accounting rates have little relationship to the actual network costs of a call. Costs have declined sharply while the call accounting rates have remained high. The resulting high margins, based on the difference between the accounting rate share and the network operating costs, have created an incentive for service providers to game the system in various ways. For example, by unilaterally raising the international call tariffs, a service provide will hope to generate higher levels of incoming call minutes, resulting in a net cash inflow through call accounting settlements.

> As a side note, the Federal Communications Commission of the United States (FCC) asserted that U.S. telephone operators paid out some $5.6 billion in settlement rates in 1996, and the FCC voiced the view that accounting rates had shifted into areas of non-cost-based settings, rather than working as a simple cost distribution mechanism. A similar view was voiced by the European Union at the time that the accounting rate settlement system was no longer producing efficient outcomes in terms of consumer prices for telephony services.

Call settlement rates do not readily encompass network interconnections that include one of more transit providers in addition to the originating and terminating providers. Also, experience in the past decade has shown that the benefits of lowering accounting rates are not automatically passed on to the consumer.

There is one important attribute of the telephone call accounting settlement model that is important to the Internet, namely that there remains a widespread expectation that the same settlement structures, and the same cost distribution outcomes can be supported in the inter-provider environment of the Internet. There is the expectation that, through use of such a cost-based financial settlement structure, large and small providers could co-exist competitively with a neutral interconnection regime that would not allow the larger players to shut out smaller players and not allow smaller players undue leverage from the investments made by the larger players.

As ideal as all this sounds, the Internet operator business environment has proved highly resilient to the introduction of such forms of inter-provider settlement arrangements. Are the incumbent service providers desperately attempting to ward off competitive new entrants? Is this because the legacy telcos are attempting to use their vastly greater market power to chase smaller competitors out of their markets in an attempt to establish a monopoly position with respect to this service? After some two decades of experience in the deregulation of the telephone service industry, where the focal point of the activity has generally been concerned with the dismantling of the entrenched monopoly position of the legacy incumbent operators, such viewpoints remain a sentimental favourite for many, particularly for those in the regulatory sector.

However, rather than blaming a recalcitrant service provider industry for being unwilling to work under a call accounting model for inter-provider financial settlements, there is another more likely explanation here. This view is that the differences between the telephony model and the Internet are quite fundamental and have repercussions that extend all the way through to the business models for provider interconnection. It may well be that the service provider industry already achieved whatever level of inter-provider equilibrium it is going to achieve given these quite fundamental differences in the service model.

These differences include:

- Single-Service vs Multi-Service Networks

  The telephony system is a single service network that supports a single core service: human conversations. There is a single model of a network transaction, namely a "call", and a uniform model of the states to establish, maintain and tear down a "call". Each network provides the same technical service and, in theory at any rate, each provider's service is directly substitutable for any other. The call is a undistinguished commodity, and it is possible to establish a benchmark accounting rate for the cost of supporting a call with a service provider's network.

  There is no analogous "service" unit in the Internet model. The basic transaction is an unreliable packet delivery service, which has no direct corresponding service level artifact. Services in the Internet architecture are not defined by the network, but are defined by the end systems. At a packet level network services are not always directly substitutable, and the parameters of the packet delivery service, such as sustainable peak delivery rate, delay, loss and jitter characteristics differ from service provider to service provider. It is not obvious how a benchmark accounting rate could be established given that there is no common service characteristic that is being provided by the network, nor is there a single service transaction model being supported by the networks.

- Network Transaction Unit - "Calls" and "Packets"

  The conventional means of supporting a call within a telephony network uses a time-switched architecture where the network supports a switching state that supports an edge-to-edge synchronous circuit. The service of supporting a "call" can be seen as a network-based activity

  While it might be argued that an Internet TCP session has much in common with a call, this concept of an originating TCP call-minute is not readily identified within the packet forwarding fabric in the interior of the network, and accordingly this is not a viable settlement unit. Unlike a telephony call, no concept of state initiation exists 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 TCP session, and therefore, no means is readily available to the operator to identify that a call has been initiated, and by which party. Of course the use of User Datagram Protocol (UDP), and various forms of tunneling traffic, also confound any such TCP call-minute accounting mechanism.

  On a more general level, the Internet's packet-based architecture includes no network state as a pre-requisite to support a particular service. Indeed the network's state, as a collection of packet forwarding directives, is one which is not directly related to the flow of data packets. The implication is that the network is entirely unaware of the existence of service transactions.

- Reliable and Unreliable Network Services

  When a packet is passed across an interconnection from one 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 transit path, may drop the packet for quite legitimate reasons, and will remain within the protocol specification in so doing. Indeed, the TCP protocol uses packet drop as a rate-control signal. If the packet is used as the accounting unit in a general cost distribution environment, should the provider who receives and subsequently drops the packet be able to offset an accounting credit for the interconnection? The logical response is that packet based accounting should apply only to successfully delivered packets, but such an accounting structure is highly challenging to implement accurately within the Internet environment.

- Symmetric and Asymmetric Network Paths

  Circuit-switched networks have symmetric paths, where both traffic flows (forward and reverse) flow on the same network path, and if the path traverses a provider interconnection than all the traffic associated with the service passes across that interconnection.

  There is no such assurance of symmetry in the Internet, and as the density of interconnection increases the likelihood of asymmetric paths increases. This implies that while the forward path for an end-to-end transaction traverses a particular inter-provider interconnection, the reverse path may use a different path and a different inter-provider interconnection.

Settlement Models for the Internet

With this in mind we can now turn our attention to the Internet and look at the forms of interconnection and associated financial arrangements that are used in this environment.

Where a wholesale or retail service agreement is in place between two ISPs, one ISP is, in effect, a customer of the other ISP. In this relationship, the customer ISP (downstream ISP) is purchasing transit and connectivity services from the supplier ISP (upstream ISP). The downstream ISP resells this service to its clients. The upstream ISP must announce the downstream ISP's routes to all other customers and other egress points of the ISP's networks to honor the service contract to the downstream ISP customer. This is very similar to the transit fee model.

However, given two ISPs who interconnect, the decision as to which party should assume the upstream provider role and which party should assume the downstream customer role is not always immediately obvious to either party, or even to an outside observer. Greater geographic coverage may be the discriminator here that allows the customer/provider determination. However, this factor is not the only possible one within the scope of the determination of respective roles. One ISP may host significant content and may observe that access to this content adds value to the other party's network, a factor that may be used as an offset against a more conventional customer relationship. In a similar vein, an ISP with a very large client population within a limited geographic locality may see this large client base as an offsetting factor with its provider.

An objective and stable determination of which ISP should be the provider and which should be the client is not always possible. In many contexts, the question is inappropriate, given that for some traffic classes the respective roles of provider and client may swap over. The question often is rephrased along the lines of, "Can two providers interconnect without the implicit requirement to cast one as the provider and the other as the client?" Exploration of some concepts of how the question could possibly be answered is illustrative of the problem space here.

Packet Cost Accounting

One potential accounting model is based on the observation that a packet incurs cost as it is passed through the network. For a small interval of time, the packet occupies the entire transmission capacity of each circuit over which it passes. Similarly, for a brief interval of time, the packet is exclusively occupying the switching fabric of the router. The more routers the packet passes through, and the greater the number and distance of transmission hops the packet traverses, the greater the incurred cost in carrying the packet.

A potential settlement model could be constructed from this observation. The strawman model is that whenever a packet is passed across a network boundary, the packet is effectively sold to the next provider. The sale price increases as the packet transits through the network, accumulating value in direct proportion to the distance the packet traverses within the network. Each boundary packet sale price reflects the previous sale price, plus the value added in transiting the ISP's infrastructure. Ultimately, the packet is sold to the destination client.

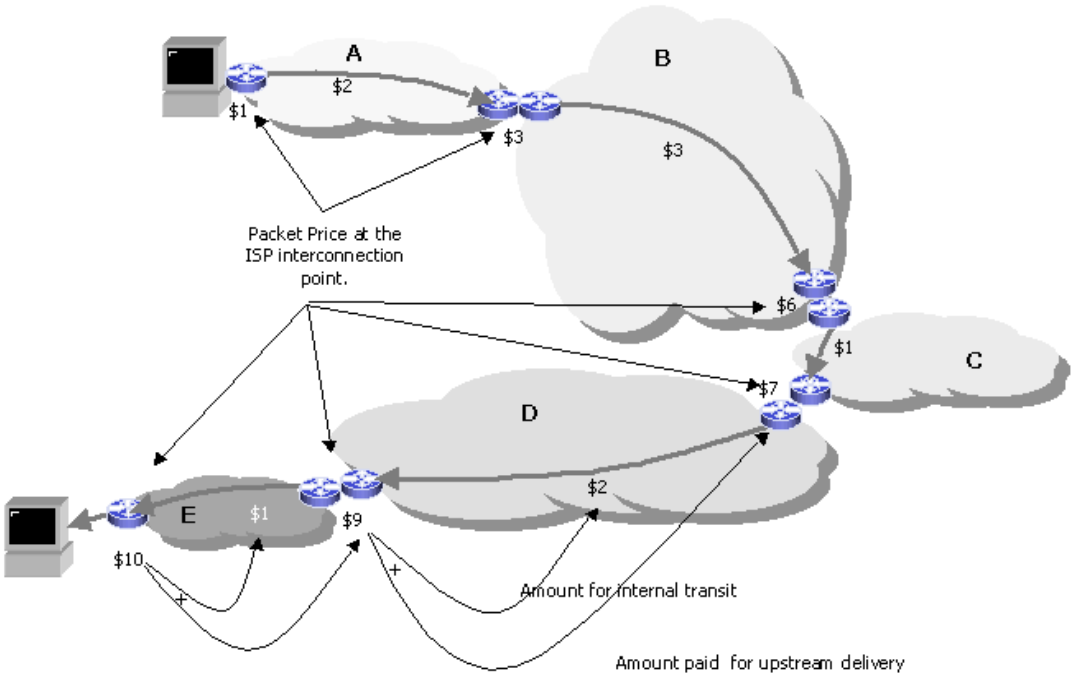

Figure 2 – Packet Cost Transfer Model

As with many strawman models, this one has numerous critical weaknesses, but let's look at the strengths first. An ISP gains revenue from a packet only when delivered on egress from the network, rather than in network ingress. Accordingly, a strong economic incentive exists to accept packets that will not be dropped in transit within the ISP, given that the transmission of the packet generates revenue to the ISP only on successful delivery of the packet to the next hop ISP or to the destination client. This factor places strong pressure on the ISP to maintain quality in the network, because dropped packets imply foregone revenue on local transmission. Because the packet was already purchased from the previous provider in the path, packet loss also implies financial loss. Strong pressure also is exerted to price the local transit function at a commodity price level, rather than attempt to undertake opportunistic pricing. If the chosen transit price is too great, the downstream provider has the opportunity to extend its network to reach the next upstream provider in the path, resulting in bypassing the original upstream ISP and purchasing the packets directly from the next hop upstream source. Accordingly, this model of per-packet pricing, using a settlement model of egress packet accounting, and locally applied value increments to a cumulative per-packet price, based on incremental per-hop transmission costs, does allow for some level of reasonable stability and cost distribution in the interprovider settlement environment.

However, the weaknesses of this packet accounting appear to overweigh any positive aspects of this approach. Firstly, reliable packet delivery cannot be guaranteed for all packets and some level of packet drop is inevitable, irrespective of average traffic loads in the network. To minimize the liability from accepting undeliverable packets each ISP is forced to maintain a complete routing table and accept only packets for which a specific

route is held. More critical is the issue that the mechanism is open to abuse, as packets that are generated by one provider can be transmitted across the provider interface, which in turn results in revenue being generated for the originating provider. Per-packet accounting within the core of the network is a significant refinement of existing technology. Within a strict implementation of this model, packets require the concept of an attached value that ISPs augment on an ingress-to-egress basis. Finding a reliable way in which to store this associated per packet value yet to do so in a fashion that has no impact on existing semantics of packet transmission and with negligible incremental cost is a challenging task. And of course the concept of DOS attacks and other forms of traffic abuse take on a new and even more malicious intent when individual packet delivery incurs additional cost for the recipient.

These traffic-based metrics exhibit critical weaknesses because of their inability to resist abuse and the likelihood of exacting an interprovider payment even when the traffic is not delivered to an ultimate destination. Of more concern is that this settlement regime has a strong implication in the retail tariff domain, where tariffs derived from delivered traffic volume and total transit distance is then one of the more robust ways that a retail provider can ensure that there is an effective match between the interprovider payments and the retail revenue streams. Given that there is no intrinsic match of distance, and therefore cost, to any particular end-to- end network transaction, such a retail tariff mechanism would meet with strong consumer resistance.

TCP Session Accounting

Does an alternative settlement structure that can address these weaknesses exist? One approach is to perform significantly greater levels of analysis of the traffic as it transits a boundary between a client and the provider, or between two providers, and to adopt financial settlement measures that match the derived type of traffic being observed. As an example, the network boundary could detect the initial TCP SYN handshake, and all subsequent packets within the TCP session could be accounted against the session initiator, while UDP traffic could be accounted against the UDP source. Such detailed accounting of traffic passed across a provider boundary could allow for a potential settlement structure based on duration or volume.

Although such settlement schemes are limited more by imagination in the abstract, the technical considerations tend to rule this approach out of serious consideration. For a client-facing access router to detect a TCP flow and correctly identify the TCP session initiator requires the router to correctly identify the initial SYN handshake, the opening packet, and then record all in-sequence subsequent packets within this TCP flow against this accounting element. This identification process is often impossible to perform within the network at an interprovider boundary as the outcome of the routing configuration may be an asymmetric traffic path, so that a single interprovider boundary may see only traffic passing in a single direction.

However, the greatest problem with this, or any other traffic accounting settlement model, is the diversity of retail pricing structures that exist within the Internet today. Some ISPs use pricing based on received volume, some on sent volume, some on a mix of sent and received volume, and some use pricing based on the access capacity, irrespective of volume. This discussion leads to the critical question when considering financial settlements: Given that the end client is paying the local ISP for comprehensive Internet connectivity, when a client's packet is passed from one ISP to another at an interconnection point, where is the revenue for the packet? Is the revenue model one in which the packet sender pays or one in which the packet receiver pays? The packet egress model described here assumes a uniform retail model in which the receiver pays for Internet packets. The TCP session model assumes the session initiator pays for the entire traffic flow. This uniformity of retail pricing is simply not mirrored within the retail environment of the Internet today. Although this session-based settlement model does attempt to promote a quality environment with fair carriage pricing, it cannot address the fundamental issue of financial settlements.

Internet Settlement Structures

For a financial settlement structure to be viable and stable, the settlement structure must be a uniform abstraction of a relatively uniform retail tariff structure. This conclusion is critically important to the entire Internet financial settlement debate.

The financial structure of interconnection must be an abstraction of the retail models used by the two ISPs. If a uniform "sender pays" retail model is used, the party originating the packet pays the first ISP a tariff to deliver the packet to its destination within the second ISP; then the first ISP is in a position to fund the second ISP to complete the delivery through an interconnection mechanism. If, on the other hand, its a "receiver pays" model is used, in which the receiver of the packet funds its carriage from the sender, then the receiving ISP funds the originating ISP for every packet received from the originating ISP, and the per packet funding will be variable based on the distance the packet has travelled to reach this particular interconnection.

If no uniform retail model is used then when a packet is passed from one provider to the other no understanding exists about which party receives the revenue for the carriage of the packet and accordingly, which party settles with the other party for the cost incurred in transmission of the packet.

The answer to these issues within the Internet environment has been to commonly adopt just two models of interaction. These models sit at the extreme ends of the business spectrum, where one is a customer/provider relationship, and the other is a peering relationship without any form of financial settlement, or SKA. These models approximately correspond to the second and third models described previously from traditional models of interconnection within the communications industry.

No Settlement and No Interconnection

Examining the option of complete autonomy of operation, without any form of interaction with other local or regional ISPs, is instructive within this examination of settlement options. One scenario for a group of ISPs is that a mutually acceptable interconnection relationship cannot be negotiated, and all ISPs in the group operate disconnected network domains with dedicated upstream connections to third party ISPs and no interconnection. The outcome of such a situation is that third-party connectivity would take place, with transit traffic flowing between the local ISPs being exchanged within the domain of a mutually connected third-party ISP (or via transit across a set of third-party ISPs). For example, for an Asian country, this situation could result in traffic between two local entities, both located within the same country, being passed across the Pacific, routed across numerous network domains within the United States, and then passed back across the Pacific. Not only is this scenario inefficient in terms of resource utilization, but this structure also adds a significant cost to the operation of the ISPs, a cost that ultimately is passed to the consumer in higher prices for Internet traffic.

This situation is not entirely theoretic - the Internet has seen such arrangements appear in the past; and these situations are still apparent in parts of today's Internet. Such arrangements have arisen, in general, as the outcome of an inability of a group of ISPs to negotiate a stable local peering structure.

However, such positions of no interconnection have proved to be relatively short-lived because of the high cost of operating international transit environments, the instability of the significantly lengthened interconnection paths, and the unwillingness of foreign third-party ISPs to act (often unwittingly) as agents for domestic interconnection in the longer term. As a result of these factors, such off-shore connectivity structures generally have been augmented with domestic interconnection arrangements.

The resultant general operating environment of the Internet is that effective isolation is not in the best interests of the ISP, nor is isolation in the interests of other ISPs or the consumers of the ISPs' services. In the interests

of a common desire to undertake rational and cost-effective use of communications resources, each national (or regional) collection of ISPs acts to ensure local interconnectivity between such ISPs. A consequent priority is to reach mutually acceptable ISP interconnection arrangements.

Sender Keeps All

Sender Keeps All (SKA) peering arrangements are those in which traffic is exchanged between two or more ISPs without mutual charge (an interconnection arrangement with no financial settlement). Within a national structure, typically the marginal cost of international traffic transfer to and from the rest of the Internet is significantly higher than domestic traffic transfer. In these cases, any SKA peering is likely to relate to only domestic traffic, and international transit would be provided either by a separate agreement or independently by each party.

This SKA peering model is most stable where the parties involved perceive equal benefit from the interconnection. This interconnection model generally is used in the context of interconnection or with providers with approximate equal dimension, as in peering regional providers with other regional providers, national providers with other national providers, and so on. The parties themselves do not have to agree on what that value or dimension may be in absolute terms. Each party makes an independent assessment of the value of the interconnection, in terms of the perceived size and value of the ISP and the value of the other ISP. If both parties reach the conclusion that in their terms a net balance of value is achieved, then the interconnection is on a stable basis. If one party believes that it is larger than the other and SKA interconnection would result in leverage of its investment by the smaller party, then an SKA interconnection is unstable.

The essential criterion for a stable SKA peering structure is perceived equality in the peering relationship. This criterion can be achieved in many ways, including the use of entry threshold pricing into the peering environment or the use of peering criteria, such as the specification of ISP network infrastructure or network level of service and coverage areas as eligibility for peering.

A typical feature of the SKA peering environment is to define an SKA peering in terms of traffic peering at the client level only. This definition forces each peering ISP to be self-sufficient in the provision of transit services and ISP infrastructure services that would not be provided across a peering point. This process may not result in the most efficient or effective Internet infrastructure, but it does create a level of approximate parity and reduces the risks of leverage within the interconnection. In this model, each ISP presents at each interconnection or exchange only those routes associated with the ISP's customers and accepts only traffic from peering ISPs at the interconnection or exchange directed to such customers. The ISP does not accept transit traffic destined to other remote exchange locations, nor to upstream ISPs, nor traffic directed to the ISP's infrastructure services. Equally, the ISP does not accept traffic that is destined to peering ISPs, from upstream transit providers. The business model here is that clients of an ISP are contracting the ISP to present their routes to all other customers of the ISP, to the upstream providers of the ISP, and to all exchange points where the ISP has a presence. The particular tariff model chosen by the ISP in servicing the customers is not material to this interconnection model. Traffic passed to a peer ISP at the exchange becomes the responsibility of the peer ISP to pass to its customers at its cost.

Another means of generating equity within an SKA peering is to peer only within the terms of a defined locality. In this model, an ISP would present routes to an SKA peer in which the routes correspond to customers located at a particular access POP, or a regional cluster of access POPs. The SKA peer's ability to leverage advantage from the greater level of investment (assuming that the other party is the smaller party) is now no longer a factor, because the smaller ISP sees only those parts of the larger ISP that sit within a well-defined local or regional zone.

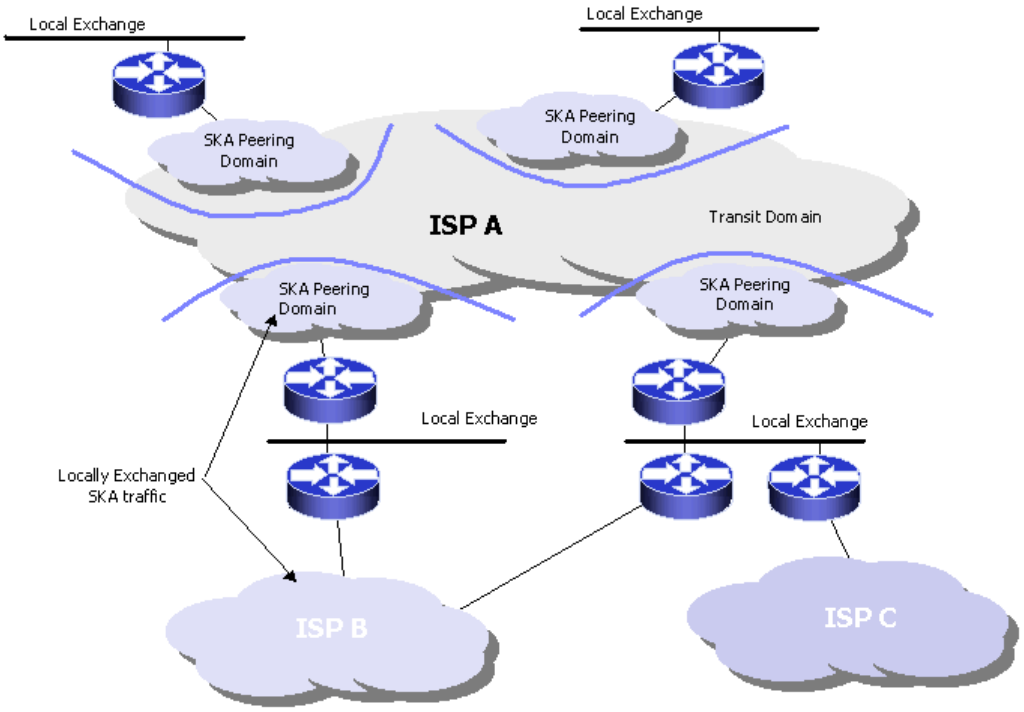

Figure 3 – Locality SKA Peering

A common outcome of widespread use of SKA interconnections is a provider structure that is segregated into two domains - a set of transit ISPs, whose predominate investment direction is in terms of high-capacity carriage infrastructure and high-capacity switching systems, and a collection of local access ISPs, whose predominate investment direction is in service infrastructure supporting a strong retail focus. Local ISPs participate at exchanges and announce local routes at the exchange on an SKA basis of interconnection with peer ISPs. Such ISPs are strongly motivated to prefer to use all routes presented at the exchange within such peering sessions, because the ISP is not charged any transit cost for the traffic under an SKA settlement structure. The exchange does not provide comprehensive connectivity to the ISP, and this connectivity needs to be complemented with a separate purchase of transit services. In this role, the local ISP becomes a client of one or more transit ISPs explicitly for the purpose of access to transit connectivity services.

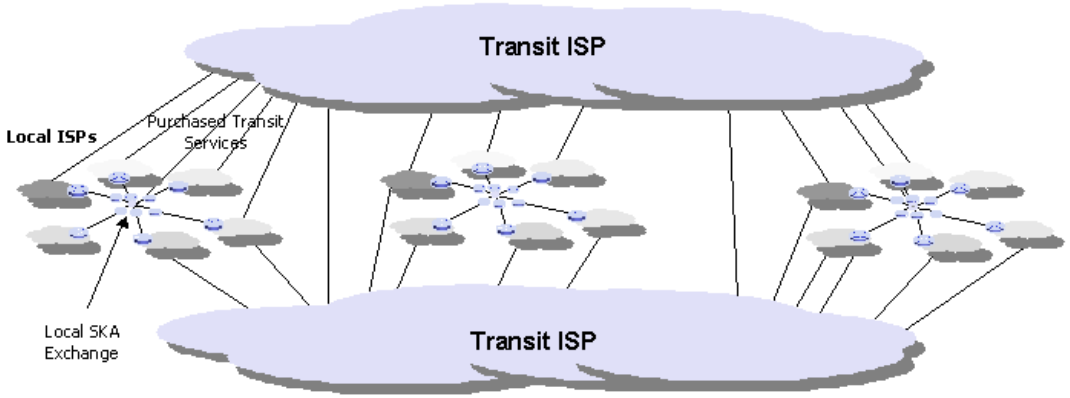

Figure 4 – ISP Structure of local and Transit Operators

In this model, the transit ISP must have established a position of broad-ranging connectivity, with a well-established and significant market share of the wholesale transit business. A transit ISP also must be able to present customer routes at a carefully selected set of major exchange locations and have some ability to exchange traffic with all other transit ISPs. This latter requirement has typically been implemented using private interconnection structures, and the associated settlements often are negotiated bilaterally. These settlements possibly may include some element of financial settlement.

Negotiated Financial Settlement

The alternative to SKA and provider/client role selection is the adoption of a financial settlement structure. The settlement structure is based on both parties effectively selling services to each other across the interconnection point, with the financial settlement undertaking the task of balancing the relative amounts.

The simplest form of undertaking this settlement is to measure the volume of traffic being passed in each direction across the interconnection and to use a single accounting rate for all traffic. At the end of each accounting period, the two ISPs would financially settle based on the agreed accounting rate applied to the net traffic flow. Which way the money should flow in relationship to traffic flow is not immediately obvious. One model assumes that the originating provider should be funding the terminating provider to deliver the traffic, and therefore, money should flow in the same direction as traffic. The reverse model assumes that the overall majority of traffic, is traffic generated in response to an action of the receiver, such as web page retrieval or the downloading of software. Therefore, the total network cost should be imposed on the discretionary user, so that the terminating provider should fund the originating provider. This latter model has some degree of supportive evidence, in that a larger provider often provides more traffic to a smaller attached provider than it receives from that provider. Observation of bilateral traffic flow statistics tends to support this, indicating that traffic-received volumes typically coincide with the relative interconnection benefit to the two providers.

The accounting rate can be negotiated to be any amount. There is a caveat on this ability to set an arbitrary accounting rate, because where an accounting rate is not cost-based, business instability issues arise. For greater stability, the agreed settlement traffic unit accounting rate would have to match the average marginal cost of transit traffic in both ISP networks for the settlement to be attractive to both parties. Refinements to this approach can be introduced, although they are accompanied by significant expenditure on traffic monitoring and accounting systems.

The refinements are intended to address the somewhat arbitrary determination of financial settlement based on the receiver or the sender. One way is to undertake flow-based accounting, in which the cost accounting for the volume of all packets associated with a TCP flow is directed to the initiator of the TCP session. Here, the cost accounting for all packets of a UDP flow is directed to the UDP receiver. The session-based accounting is significantly more complex than simple volume accounting, and such operational complexity would be reflected in the cost of undertaking such a form of accounting. However, asymmetric paths are a common feature of the inter-AS environment, so that it may not always be possible to see both sides of a TCP conversation and perform an accurate determination of the session initiator.

Another refinement is to use a different rate for each provider, where the base rate is adjusted by some agreed size factor to ensure that the larger provider is not unduly financially exposed by the arrangement. The adjustment factor can be the number of Points of Presence, the range of the network, the volume carried on the network, the number of routes advertised to the peer, or any other metric related to the ISP's investment and market share profile. Alternatively, a relative adjustment factor can simply be a number, without any basis in a network metric, to which both parties agree.

Of course, such a relative traffic volume balance is not very robust either, and the metric is one that is vulnerable to abuse. The capability to adjust the relative traffic balance comes from the direct relationship between the routes advertised and the volume of traffic received. To reduce the amount of traffic received, the ISP reduces the number of routes advertised to the corresponding peer. Increasing the number of routes, and at the same time increasing the number of specific routes, increases the amount of received traffic. When there is a rich mesh of connectivity, the primary objective of routing policy is no longer that of supporting basic connectivity, but instead the primary objective is to maximize the financial return to the operator. If the ISP is paying for an "upstream" ISP service, the motivation is to minimize the cost of this contract, either by maximizing the amount of traffic covered under a fixed cost, or minimizing the cost by minimizing the traffic exchanged with the upstream ISP. Where there is a financially settled interconnection, the ISP will be motivated to configure its routing policies to maximize its revenue from such an arrangement. And of course an ISP will always prefer to use customer routes wherever possible, as a basic means of maximizing revenue into the operation.

Of greater concern is the ability to abuse the interconnection arrangements. One party can generate and then direct large volumes of traffic to the other party. Although overt abuse of the arrangements is often easy to detect, greed is a wonderful stimulant to ingenuity, and more subtle forms of abuse of this arrangement are always possible. To address this, both parties would typically indicate in an interconnection agreement their undertaking not to indulge in such forms of deliberate abuse. Notwithstanding such undertakings by the two providers, third parties can still abuse the interconnection in various ways. Loose source routing can generate traffic flows that pass across the interconnection in either direction. The ability to remotely trigger traffic flows through source address spoofing is possible, even where loose source routing is disabled. This window of financial vulnerability is far wider than many ISPs are comfortable with, because it opens the provider to a significant liability over which it has a limited ability to detect and control. Consequently, financial settlement structures based on traffic flow metrics are not a commonly deployed mechanism, because they introduce significant financial risks to the ISP interconnection environment.

The Settlement Debate

The issue of Internet settlements, and associated financial models of settlement, has occupied the attention of a large number of ISPs, traditional communications carriers, public regulators, and many other interested bodies for many years now. Despite these concentrated levels of attention and analysis, the Internet interconnection environment remains one where there are no soundly based models of financial settlement in wide-spread use today.

It is useful to look further into this matter, and pose the question: "Why has the Internet managed to pose such a seemingly intractable challenge to the ISP industry?"

The prime reason is likely to be found within the commonly adopted retail model of ISP services. The tariff for an ISP retail service does not implicitly cover the provision of an Internet transmission service from the client to all other Internet-connected hosts. In other words, the Internet service, as retailed to the client, is not a comprehensive end-to-end service.

In a simple model of the operation of the Internet, each ISP owns and operates some local network infrastructure, and may choose to purchase services from one or more upstream service providers. The service domain offered to the clients of this network specifically encompasses an Internet subdomain limited to the periphery of the ISP network together with the periphery of the contracted upstream provider's service domain. This is a recursive domain definition, in that the upstream provider in turn may have purchased services from an upstream provider at the next tier, and so on. After the client's traffic leaves this service domain, the ISP ceases to directly, or indirectly, fund the carriage of the client's traffic, and the funding burden passes over to a funding chain linked to the receiver's retail service. For example, when traffic is passed from an ISP client to a

client of another provider, the ISP funds the traffic as it transits through the ISP and indirectly funds the cost of carriage through any upstream provider's network. When the traffic leaves the provider's network, to be passed to either a different client, another ISP, or to a peer provider, the sender's ISP ceases to fund the further carriage of the traffic. In other words, these scenarios illustrate the common theme that the retail base of the Internet is not an end-to-end tariff base. The sender of the traffic does not fund the first hop ISP for the total costs of carriage through the Internet to the traffic's destination, nor does the ultimate receiver pay the last hop ISP for these costs. The ISP retail pricing structure reflects an implicit division of cost between the two parties, and there is no consequent structural requirement for interprovider financial balancing between the originating ISP and the terminating ISP.

An initial reaction to this partial path service model would be to wonder why the Internet works at all, given that no single party funds the carriage of traffic on the complete path from sender to receiver. Surely this would imply that once the traffic had passed beyond the sending ISP's service funded domain the traffic should be discarded as unfunded traffic? The reason why this is not the case is that the receiver implicitly assumes funding responsibility for the traffic at this handover point, and the second part of the complete carriage path is funded by the receiver. In an abstract sense, the entire set of connectivity paths within the Internet can be viewed as a collection of bilaterally funded path pairs, where the sender funds the initial path component and the receiver funds the second terminating path component. This underscores the original observation that the generally adopted retail model of Internet services is not one of end-to-end service delivery, but instead one of partial path service, with no residual retail price component covering any form of complete path service.

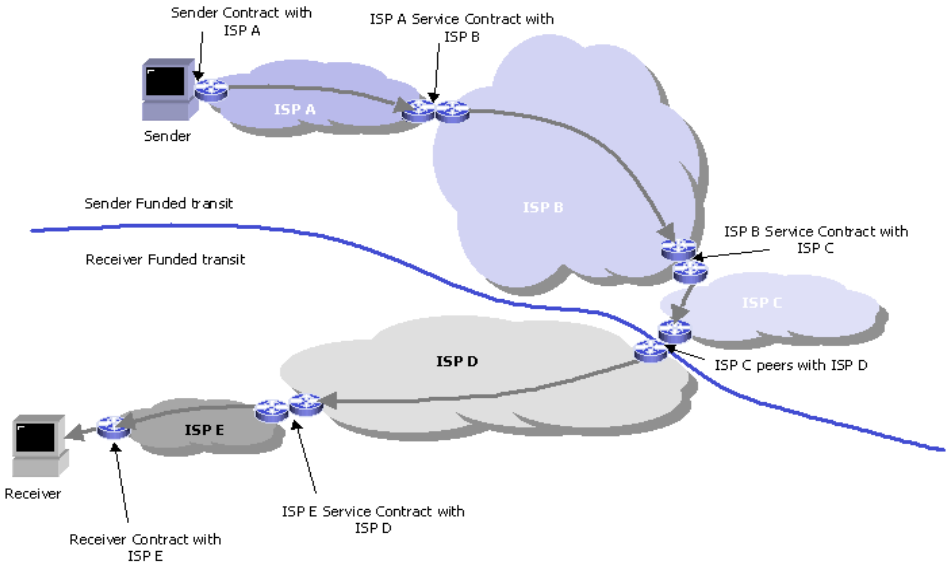

Figure 5 – Partial Path Funding in the Internet

Financial settlement models typically are derived from a different set of initial premises than those described here. The typical starting point is that the retail offering is a comprehensive end-to-end service, and that the originating service provider utilizes the services of other providers to complete the delivery of all components of the retailed service. The originating service provider then undertakes some form of financial settlement with those providers who have undertaken some form of an operational role in providing these service elements. This cost-distributed business structure allows both small and large providers to operate with some degree of financial stability, which in turn allows a competitive open service market to thrive. Through the operation of open competition, the consumer gains the ultimate price and service benefit of cost-efficient retail services.

The characteristics of the Internet environment tend to create a different business environment to that of a balanced cost distribution structure. Here there is a clear delineation between a customer/provider relationship

and a peer relationship, with no stable middle ground of a financially settled inter-ISP bilateral relationship. An ISP customer is one that assumes the role of a customer of one or a number of upstream providers, with an associated flow of funding from the customer to the upstream provider, whereas an ISP upstream service provider views the downstream provider as a customer. An ISP peer relationship is where the two ISPs execute a peering arrangement, where traffic is exchanged between the two providers without any consequent financial settlement, and such peering interactions are only stable while both providers perceive some degree of parity in the arrangement; for example, when the two providers present to the peering point Internet domains of approximate equality in market coverage and market share. An ISP may have multiple simultaneous relationships, being a customer in some cases, an upstream provider in others, and a peer in others. In general, the relationships are unique within an ISP pairing, and efforts to support a paired relationship which encompasses elements of both peering and customer/provider pose significant technical and business challenges.

The most natural business outcome of any business environment is for each provider to attempt to optimize its business position. For an ISP, this optimization is not simply a case of a competitive impetus to achieve cost efficiency in the ISP's internal service operation, because the realization of cost efficiencies within the service provider's network does not result in any substantial change in the provider's financial position with respect to upstream costs or peering positioning. The ISP's path toward business optimization includes a strong component of increasing the size and scope of the service provider operation, so that the benefits of providing funded upstream services to customers can be maximized, and non-financially settled peering can be negotiated with other larger providers.

Regulation and Interconnection

Currently the Internet admits only a small set of possible inter-provider interconnection options: when two providers interconnect either one assumes the role of provider and the other of their customer, or the two providers agree on an SKA arrangement. Other arrangements tend to be re-labelled variations on these themes, such as "paid peering" which is a variation of the provider / customer transit model, or "regional peering" where a provider may only present a subset of their complete route set to its SKA peers in a particular regional interconnection arrangement.

Without the adoption of a settlement regime that supports some form of cost distribution among Internet providers, there are serious structural problems in supporting a highly diverse and well populated provider industry sector. These problems are exacerbated by the additional observation that the Internet transmission and retail markets both admit significant economies of scale of operation. The combination of these two factors leads to the economic conclusion that the Internet market is not a long term sustainable open competitive market that is capable of supporting a wide diversity of players both large and small. This aggregation is already well underway in today's Internet, and direction of the Internet market will be forged through the tension between this aggregation pressure and various national and international public policy objectives that relate to the Internet industry.

Under such circumstances the natural market outcome is that of continuing aggregation of providers, where smaller players have to distinguish themselves through the servicing of various forms of niche opportunities and only larger players have the capability to service the commodity mass market.

Various forms of regulatory constraints in a number of regulatory regimes have attempted to even the playing field and provide sustainable conditions for smaller ISPs in the larger commodity markets. These regulatory measures have involved the imposition of pricing constraints on trunk and last mile access transmission facilities, the opening up of local switching offices for third party equipment access, and even regulation of

wholesale tariffs in an endeavour to ensure that smaller players have access to an equitable cost base for their operation.

Efforts to extend this regulatory activity to the area of regulated interconnection and peering have been investigated by various national regimes, but with little in the form of regulatory outcomes to date. Attempts to impose measures of mandatory domestic SKA peering at nominated exchange points tend to create an environment where there is a disincentive for larger players to aggressively invest in further infrastructure given the ability for smaller players to leverage this investment to their advantage without making comparable investments of their own. This has the risk of leading to an excessively fragmented serviced industry where natural economies of scale are not realized, and the consumer base is exposed to an inefficient supply industry which inherently imposes price premiums at the retail service level. The issue here is the supply of Internet services is not an end in itself - the objective is to ensure an efficient and effective service industry that provides the necessary foundation for other economic activities that can themselves leverage the capabilities of the Internet.

The problem stated here is not in the installation of transmission infrastructure, nor is it in the retailing of Internet services. The problem faced by the Internet industry is in ensuring that each provider of infrastructure is fairly compensated when its infrastructure is used. In essence, the problem is how to distribute the revenue gained from the retail sale of Internet access and services to the providers of carriage infrastructure. While continued growth has effectively masked these problems for the past decade, after market saturation occurs and growth tapers off, these issues of financial settlement between the various Internet industry players will shape the future of the entire global ISP industry.

While it is not completely clear that the deregulated open market nature of the Internet can sustain a diverse, efficient and effective service provider industry, it is also unclear what form of regulatory constraints or intervention are appropriate, if any.

One school of thought is that the Internet continues to operate are unprecedented levels of efficiency and the vibrant competition at all levels of the industry continues to produce outcomes of price and performance to the consumer that were never seen in the regulated telephony world. There is competition in the supply of all forms of Internet services, from local access to global transit, and prices continue to efficiently reflect the cost of supply. There is no widespread evidence of any form of deliberate market distortion, such as price gouging, dumping, hoarding, or cartel behavior. This school of thought poses the rhetorical question: "What precisely is broken in this environment that requires regulatory intervention at a national, regional or international level?"

Others see it differently. They point to the net revenue flow to the so-called "Tier One" providers who have no requirement to purchase transit from any other provider. They point to the current form of international transmission provisioning where the smaller provider now needs to fund all the transmission necessary to meet at the provider's access point, rather than an historical model of carefully managed cost-sharing. They point to the prevailing conditions at national regional and international levels where the smaller players are forced into various niche markets, while the larger players continue to increase their total market size in the larger mass markets. They see the potential for cartel-like behavior of the larger players, and the possibility of the imposition of monopoly rentals on services where all competition has been eliminated. They see such outcomes as undesireable, and see a definite role for some form of regulatory involvement at a national and international level to mitigate such risks and ensure that there remains a fair and open competitive environment in the supply of Internet services.

It is certainly the case that the Internet has to date achieved unprecedented outcomes in terms of cost effectiveness of the service, in terms of speed of deployment and in terms of the ability to rapidly bring new technologies to bear. This is an extremely efficient supply chain. There is a strong risk that regulatory involvement, if applied inappropriately, will trigger structural inefficiencies that ultimately will be reflected at

the consumer level in higher prices and inferior services. Competition is not an end in itself, nor is regulatory impost. The challenge here is to foster the conditions that allow the Internet to be a productive and efficient platform for all. That, for me, appears to be at the heart of the challenge of the Information Society.

Disclaimer

About the Author

Geoff Huston B.Sc., M.Sc., has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. He is author of a number of Internet-related books, and has been active in the Internet Engineering Task Force for many years.

www.potaroo.net