The ISP Column
A monthly column on things Internet

September 2012
Geoff Huston

Network Service Models and the Internet

In recent times we've covered a lot of ground in terms of the evolution of telecommunications services, riding on the back of the runaway success of the Internet. We've taken the computer and applied a series of transformational changes in computing power and size, battery technology and radio systems to create a surprising result. We've managed to put advanced computation power in a form factor that fits in the palm of my hand, and couple it with a communications capability that can manage data flows of tens if not hundreds of megabits per second. All in a device that has as few as two buttons! And we done this at such scale that the manufacturing cost of these devices is now down to tens of dollars per unit. The Internet is not just at the center of today's mass market consumer service enterprise, the Internet is now at the heart of many aspects of our lives. It's not just the current fads of the social networking tools, but so much more. How we work, how we buy and sell, even what we buy and sell, how we are entertained, how democracies function, even how our societies are structured, and so much more, are all activities that are now mediated by the Internet.

But a few clouds that have strayed into this otherwise sunny story of technological wonder. Perhaps the largest of these clouds is that the underlying fabric of the Internet, the network's numbering plan, is now fracturing. We've run out of IP addresses in the Asia-Pacific region, and the same fate awaits Europe and the Middle East in the coming weeks. At the same time, the intended solution, namely the transition to a version of the IP protocol with a massively larger number space, IPv6, is still progressing at an uncomfortably slow pace. While the numbers look like a typical "up and to the right" Internet data series, the vertical axis tells a somewhat different story. The overall deployment of IPv6 in today's Internet currently encompasses under 1% of the total user base of the Internet, and its possible that the actions of the open competitive market in Internet-based service provision won't necessarily add any significant further impetus to this necessary transition.

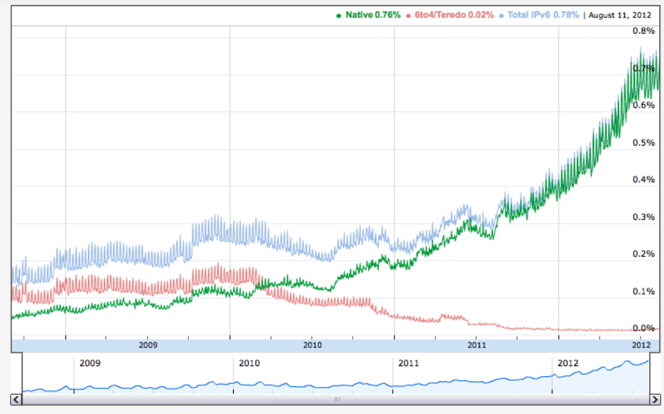

Figure 1 – IPv6 Preference in the Internet
[http:// http://www.google.com/ipv6/statistics.html#tab=ipv6-adoption]

We've gone through a number of phases of explanation for this apparently anomalous success-disaster situation for the Internet. Initially, we formed the idea that the slow adoption of IPv6 was due to a lack of widely appreciated knowledge about the imminent demise of IPv4 and the need to transition the network to IPv6. We thought that the appropriate response would be a concerted effort at information dissemination and awareness raising across the industry, and that's exactly what we did. But the response, as measured in terms of additional impetus with the uptake of IPv6 in the Internet, was not exactly overwhelming.

We then searched for a different reason as to why this IPv6 transition appeared to be stalling. There was the thought that this was not so much a technical issue as a business or a market-based issue, and there was the idea that a better understanding of the operation of markets and the interplay between markets and various forms of public sector initiatives could assist in creating a stronger impetus for IPv6 in the service market. The efforts at stimulation of the market to supply IPv6 goods and services through public sector IPv6 purchase programs has not managed to create a "tipping point" for adoption of IPv6.

There has been the idea that the realization of IPv4 exhaustion would focus our thinking and bring some collective urgency to our actions. But while IPv4 address exhaustion n the Asia Pacific region in 2011 has created some immediate interest in IPv4 address extension mechanisms, the overall numbers on IPv6 adoption have stubbornly remained under 1% of the internet's 2 billion user base.

Why is this? How can we deliberately drive this prodigious network into the somewhat perverse outcomes that break to basic end-to-end IP architecture by attempting to continue to overload the IPv4 network with more and more connected devices? What strange perversity allows us to refuse to embrace a transition to a technology than can easily sustain the connection needs of the entire silicon industry for many decades to come and instead chose a path that represents the general imposition of additional cost and inefficiency?

Perhaps there is something more fundamental going on here that reaches into the architectural foundations of the Internet that may explain, to some extent, this evident reluctance of critical parts of this industry to truly engage with this IPv6 transition and get moving.

Telephony Network Intelligence

Compared to today's "smart" phone, a basic telephone handset was a remarkably basic instrument. The entire telephone service was constructed with a model of a generic interface device that was little more than a speaker, a microphone, a bell and a pulse generator. The service model of the telephone, including the call initiation function of dialing and ringing, the real time synchronous channel support to support bi-directional speech, all forms of digital/analogue conversion and of course the call accounting function, were essentially all functions of the network itself, not the handset. While the network was constructed as a real time switching network, essentially supporting a model of switching timeslots within each of the network's switching elements, the service model of the network was a "full service" model.

The capital investment in the telecommunications service was therefore an investment in the network; in the transmission, switching and accounting functions. Building these networks was an expensive undertaking in terms of the magnitude of capital required. By the end of the twentieth century the equipment required to support synchronous time switching included high precision atomic time sources, a hierarchy of time division switches to support the dynamic creation of edge-to-edge synchronous virtual circuits and a network of transmission resources that supported synchronous digital signaling . Of course while these switching units were highly sophisticated items of technology, most of this investment capital in the telephone network was absorbed by the last mile of the network, or the so-called "local loop".

While the financial models to operate these networks varied from operator to operator, it could be argued that there was little in the way of direct incremental cost in supporting a "call" across such a network, but there is a significant opportunity or displacement cost. These networks have a fixed capacity, and the requirements for supporting a "call" are inelastic. When a timeslot is being used by one call this slot is unavailable for use by any other call.

Telephony Tariffs

When constructing a retail tariff structure for telephony a number of models were used. One model was a "subscription model", where, for a fixed fee, a subscriber could make an unlimited number of calls. In other words the operator's costs in constructing and operating the network were recouped equally from all the subscribers to the network, and no transaction-based charges were levied upon the subscriber. This model works exceptionally well where the network's capacity to service calls is of the same order as the peak call demand that is placed on the network. In other words where the capacity of the network is such that the marginal opportunity or displacement cost to support each call is negligible there is no efficiency gain in imposing a transactional tariff on the user. In the United States' telephone network, for example, a common tariff structure was that the monthly telephone service charge also allowed the subscriber to make an unlimited number of local calls.

Another model in widespread use in telephony was the use of a smaller fixed service change and a per transaction charge for each call that was made. Here a subscriber is charged a fee for each call (or "transaction") that is initiated by the subscriber. The components to determine the charge for an individual transaction include the duration of the call, the distance between the two end parties of the call, the time of day and day of the week. This allowed a network operator to create an economically efficient model of exploitation of an underlying common resource of fixed capacity. This model of per-call accounting was widespread, used not only by some operators in local call zones, and more widely by telephone service operators in long distance and international calls. This model allowed the operator to generate revenue, and recoup its costs, from those users who made use of the service, and, through use of the pricing function, the network operator could moderate peak demand for the resource to match available capacity.

This per-transaction service model of telephony was available to the operator of the telephone service simply because the entire function of providing the telephone service was a network-based service. The network was aware of who initiated the transaction, who 'terminated' the transaction, how long the transaction lasted, and what carriers were involved in supporting the transaction. Initially this transactional service model was seen as a fair way to fairly allocate the not inconsiderable costs of the construction and operation of the network to those who actually made use of used the network, and allocate these costs in proportion to the relative level of use, though I suspect that this fair cost allocation model disappeared may decades ago as these per-transaction service tariffs because less cost-based and more based on monopoly rentals.

IP Network Minimalism

The Internet is different. Indeed, the Internet is about as different to telephony as one could possibly imagine.

The architecture of the Internet assumes that a network transaction is a transaction between computers. In this architecture the computers are highly capable signal processors and the network is essentially a simple packet conduit. The network is handed "datagrams", which the network is expected to deliver most of the time. However within the architecture the network may fail to deliver the packets, reorder the packets or even corrupt the content of the packets. The network is under no constraint as to the amount of time it takes to deliver the packet. In essence the expectations that the architecture imposes on the network is about as minimal as possible. Similarly, the information that the edge-connected computers now expose to the network is also very limited. To illustrate this its useful to look at the fields that the internet protocol exposes to the network.

In IPv4 the fields of the internet protocol header are a small set, as shown in Figure 2. An IP packet header exposes the protocol version, the header length, the total length of the IP packet, packet fragmentation control fields, type of service fields, a hop counter, a header checksum and the source and destination addresses. In practice, the type of service field is unused, and the length and checksum fields have information that is also contained in the data link frame header. What's left is the protocol version protocol field, packet length, the fragmentation control fields, a hop counter and the source and destination addresses. Of these, the packet length, fragmentation control, hop counter and destination address are the fields used by the network to forward the packet to its ultimate destination.

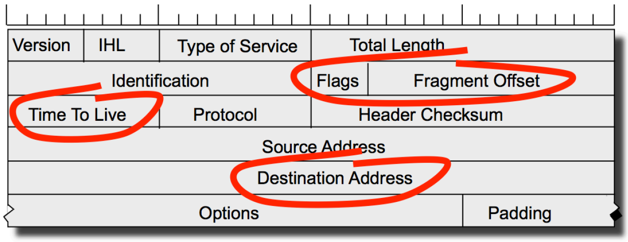

Figure 2 – The IPv4 Packet Header

In IPv6 this minimal approach was further exercised with the removal of the fragmentation control fields and the checksum fields (Figure 3). Arguably, the traffic class and flow label are unused, leaving only the protocol version, payload length, a hop counter and the source and destination addresses exposed to the network. In IPv6 the minimal network level information is now reduced to the packet length, the hop counter and the destination address.

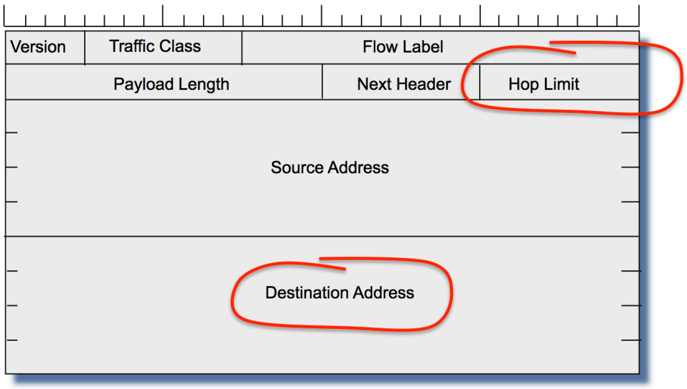

Figure 3 – the IPv6 Packet Header

These fields represent the totality of the amount of information that the internet protocol intentionally exposes to the network. There are no transaction identifiers, no call initiation or call teardown signals, or even any reliable indication of relative priority of the packets. All the network needs to "see" in each carried packet is a hop counter, a packet length and a destination address. Within this model the actions of each switching element are simple:

```
for each received packet:
        decrement the hop counter
                if the counter value is zero then discard the packet
        look up the packet's destination address in a local table
                if the lookup fails then discard the packet
        look up the output queue from the located table entry
                if the queue is full discard the packet
                if the packet is too large for the outbound interface then
                        fragment the packet to fit, if permitted (IPv4)
                        or discard the packet (IPv6).
        queue the packet for onward transmission
```

The Internet Service Model

What happened to "transactions" in this service model? What happened to network state? What happened to resource management within the network? What happened to all of the elements of network-based communications services? The simple answer is that the architecture of the Internet it's none of the underlying network's business. From a network perspective, IP has thrown all of that away!

In the context of the Internet service architecture a "transaction" is now merely an attribute of the application that is run on the end systems, and the underlying network is simply unaware of these transactions. All the network sees is IP packets, and each packet does not identify to the network any form of compound or multi-packet transaction.

Since a transaction is not directly visible to the IP network operator, then this implies that any effort for a IP service provider to use a transactional service tariff model becomes an exercise in frustration, given that there are no such network-visible interactions that could be used to create a transactional service model. In the absence of a network-based transactional access model the Internet Service Provider (ISP) has typically used an access-based service model as the basis of the IP service tariff. Some basic differentiation is provided by the ability to apply price differentials to different access bandwidths, but this is a relatively coarse form of market segmentation. Finer levels of transactional based prices, such as pricing a video stream, or even pricing a single web page fetch are not an inherent feature of such an access-based tariff structure.

The consequence for ISPs here is that within a single network access bandwidth class this service model does not differentiate between heavy and light users, and is insensitive to the services operated across the network and insensitive to the average and peak loads imposed by these services. Like the flat rate local telephone access model, the Internet pricing model is typically a flat rate model that takes no account of individual network transactions. The ISP's service delivery costs are in effect equally apportioned across the ISP's user base.

Interestingly, this has been a positive feature for the Internet. With no marginal incremental costs for usage of the network, users are in effect encouraged to use the Internet. In the same vein suppliers are also encouraged to use the Internet, as they can deliver goods and services to their customer base within imposing additional costs to the users. For example, we've seen Microsoft and Apple move towards a software distribution model that is retiring the use of physical media, and moving to an all-digital Internet-based service model to support their user base. We've also seen other forms of service provision where the access-based tariff model has enabled services that would otherwise not be viable - here Netflix is a good example of such services that have been enabled by this flat rate tariff structure.

The other side effect of this shift in the architecture of the Internet is that it has placed the carriage provider - the network operator - into the role of a commodity utility. Without any ability to distinguish between various transactions, because the packets themselves give little in terms of reliable information, the carriage role is an undistinguished commodity utility function. The consequent set of competitive pressures in a market that is not strongly differentiated ultimately weans out all but the most efficient of providers from the service provider market, as long as competitive interests can be bought to bear on these market segments.

Invariably, consumers value the services that a network enables, rather than the network itself. In pushing the transaction out of the network to the application, the Internet's architecture also pushed value out of the network as well. Given that a service in the Internet model is an interaction between applications running between a application service provider and their client, its clear that the network operator is not a party to the service transaction. An ISP may also provide services to users, but its by no means an exclusive role, and others are also able to directly interact with customers and generate value through the provision of good and services, without the involvement of the underlying network operators.

The Internet's Content Business Model

This unbundling of the service provision function from the network has had some rather unexpected outcomes. The initial forays of providing content to users believed that this was no different to many retail models, where the content provider formed a set of relationships with a set of users. The direct translation of this model encountered a number of problems, not the least of which was reluctance on the part of each individual user to enter into a panoply of service and content relationships. When coupled with considerations of control of secondary redistribution of the original service, this created some formidable barriers to the emergence of a highly valuable market for content and services on the Internet.

However, as with many forms of mass market media, the advertising market provides some strong motivation. With a traditional print newspaper, the full cost of the production of the newspaper is often borne largely by advertisers rather than by the newspaper's readers. But newspaper advertising is a relatively crude exercise, in that the advertisement is visible to all readers, while it is of interest to a much small subset. The Internet provided the potential to customize the advertisement.

The greatest market value for advertisements is generated by those operations that gain the most information about their customers. These days it is a lot to do with knowledge of the consumer. It could be argued that Facebook's $1B purchase of Instagram was based on the observation that the combination of an individual's pictures and updates forms an amazingly rich set of real time information about the behaviour and preferences of each individual consumer. It could also be argued that Google's business model is similarly based on forming a comprehensive and accurate picture of each individual user's preferences, which is then sold to advertisers at a significant premium simply due to its tailored accuracy. And the mobile services are trying to merge the user's current location with the knowledge of their preferences to leverage even greater value.

These developments are heading in the direction of a multi-party service model, where the relationship between a content provider and a set of users providers the content provider to resell these users to third parties through advertising. This on-selling of users' profiles and preferences is now a very sophisticated and significant market. As reported in http://www.techi.com/2012/03/a-breakdown-of-googles-top-advertisers/, some 90% of Google's $37.9B income was derived from advertising revenue. The cost per click for "cheap car insurance" is reported in the same source to be $33.97!

The Plight of the Carrier

While the content market and the associated service plane is now an extraordinarily valuable activity, the same is not the case for the network operator. Their carriage function been reduced from complete service delivery management to a simple packet carrier without any residual visibility into the service plane of the network. Obviously, network carriers look at these developments with dismay. Their own traditional value-added market has been destroyed, and the former model where the telco owned everything from the handset onward has now been replaced by a new model that relegates them to a role similar to electricity or water reticulation, with no prospect of added value. The highly valuable service level transactions are effectively invisible to the internet's carriage service providers.

There is an evident line of thought in the carriage industry that appears to say: "If we could capture the notion of service level transaction in IP we could re-cast our service profile into a per-transaction profile, and if we can do that then we could have the opportunity to capture some proportion of the value of each transaction."

Short of traffic interception, could the network operators working at the internet level of the network protocol stack have a means to identify these service level transactions? The generic answer is "no", as we've already seen, but there are some other possibilities that could expose service level transactions to the network operator.

QoS to the Rescue?

The recent calls by the European Network and Telephone operators' group, ETNO, advocating the widespread adoption of IP Quality of Service (QoS) appear to have some context from this perspective of restoring transaction visibility to the IP carriage provider. The QoS model is one where an application undertakes a QoS "reservation" with the network. The network is supposed to respond with a commitment to reserve the necessary resources for use by this transaction. The application then uses this QoS channel for its transaction, and releases the reservation when the transaction is complete.

From the network operator’s perspective the QoS-enabled network is now being informed of individual transactions, identifying the end parties for the transaction, the nature of the transaction and its duration, as well as the resource consumption associated with the transaction. From this comes the possibility for the QoS IP network operator to move away from a now commonplace one-sided flat access tariff structure for IP services, and instead use a transactional service model that enables the network operator to impose transaction-based service fees on both parties to a network service if it so choses. It also interposes the network operator between the content provider and the consumer, permitting the network operator to mediate the content service and potentially convert this gateway role into a revenue stream.

Of course the major problem in this QoS model is that its based on a critical item of Internet mythology - the myth that inter-provider QoS exists on the Internet. As I’ve said elsewhere, QoS is not part of today’s Internet, and there is no visible prospect that it will be part of tomorrow’s Internet either! [http://www.potaroo.net/ispcol/2012-06/noqos.html]

Knotting up NATs

But QoS is not the only possible approach to exposing service level transactions to the carriage level IP network operator. Interestingly enough, the twin factors of the exhaustion of IPv4 addresses and the lack of uptake of IPv6 offers the IP network operator another window on what the user is doing, and, potentially, another means of controlling the quality of the user’s experience through isolating individual user level transactions at the network level.

When there are not enough addresses to assign each and every customer a unique IP address the ISP is forced to use private addresses and operate a Network Address Translator (NAT) within the carriage network.

However NATs are not stateless passive devices. A NAT records every TCP and UDP session from the user, as well as the port addresses used by the application, when it creates a binding from an internal IP address and port to an external IP address and port. That’s a new NAT binding that is created for every user transaction; every conversation, every web site, every streamed video and literally everything else. If you were to take a peek at the NAT logs that record this binding information you would find a rich stream of real time user data that show precisely what each user is doing on the network. Every service transaction is now visible at the network level. How big is the temptation for the IP network operator to peek at this carrier-operated NAT log and analyze what it means?

Potentially, this transaction data could be monetized, as it forms a real time data feed of every customer's use of the network. At the moment carriers feel that they are being compelled to purchase and install this NAT functionality because of the IPv4 address situation. I suspect that when they look at the business case for purchasing and deploying these Carrier Grade NAT devices there is a parallel business case that can be made to inspect the NAT logs and perhaps to either on-sell the data stream or analyze it themselves to inform themselves of their customer's behaviour. And, as already noted, there is already market evidence that such information about each individual user's activities can be worth significant sums.

But it need not necessarily be limited to a passive operation of stalking the user's online behaviour. NATs offer a way for the carriage operator to obtain real time feeds of customer behaviour without actively intruding themselves into the packet stream. The NAT neatly segments the customer’s traffic into distinct transactions which are directly visible to the NAT operator. If the carriage provider were adventurous enough it could bias the NAT port binding function to even make some content work "better" than other content, through either slowing down the binding function for certain external sites or rationing available ports to certain less preferred external sites. In effect NATs provide a number of exploitable levers of control for the carriage operator, bundled with a convenient excuse of "we had no choice but to deploy these NATs!"

Where Now?

In contrast, what does an investment in IPv6 offer the carriage provider? An admittedly very bleak response is that what is on offer with IPv6 is more of what has happened to the telecommunications carriage sector over the past 10 years, with not even the remote possibility of altering this situation. IPv6 certainly looks like forever, so if the carriers head down this path the future looks awfully bleak for those folk who are entirely unused to, and uncomfortable with, a commodity utility provider role.

So should we just throw up our hands at this juncture and allow the carriage providers free rein? Are NATs inevitable? Should we view the introduction of transactional service models in the Internet as a necessary part of its evolution? I'd like to think that these are inevitable developments for the Internet, and that there are other paths that could be followed here. The true value for the end consumer is not in the carriage of bits through the network, but in the access to communication and services that such bit carriage enables. What does that imply for the future for the carriage role? I suspect that the role of commodity utility operator is one possible future here.

This is not the first time an industry has transitioned from production of a small volume of highly valuable units to production of a massively larger volume of commodity goods, each of which have a far lower unit value, but with an aggregate total that is much larger. The computing industry's transition from mainframe computers to mass market consumer electronics is a good example of such a transformation. As many IT sector enterprises have shown, it is possible to make such transitions. IBM is perhaps a classic example of an enterprise that has managed a number of successful transformations that have enabled it to maintain relevance and value in a rapidly changing environment.

The models for electricity distribution have seen a similar form of evolution in the last century. In the 1920's in the UK electricity was a low volume premium product. The prices for electricity were such that to keep just 5 light bulbs running for one day in a household cost the equivalent of an average week's wages. The consequent years saw public intervention in the form of nationalization of power generation and distribution that transformed electricity supply into a commonly available and generally affordable commodity.

The challenge the Internet has posed for the carriage sector is not all that different. The old carriage business models of relatively low volume high value transaction-based telecommunication services of telephony and faxes find no resonance within the service model of the Internet. In the architecture of the Internet it's the applications that define the services, while the demands from the underlying carriage network have been reduced to a simple stateless datagram delivery service. Necessarily, the business models of carriage have to also change to adapt to this altered role, and one of the more fundamental changes is the dropping of the transaction-based model of the provision of telecommunications services for the carriage provider. What this implies for the Internet's carriage sector is perhaps as radical as the transformation of the electricity supply industry during the period of the construction of the national grid systems in the first half of the twentieth century.

The necessary change this implies is from a high value premium service provider dealing in individual transactions across the network to that of a high volume undistinguished commodity utility operator. The architectural concepts of a minimal undistinguished network carriage role and the repositioning of service management into end-to-end applications is an intrinsic part of the architecture of the Internet itself. It's not a universally acclaimed step, and certainly not one that is particularly popular in today's carriage industry, but if we want to see long term benefits from the use of the Internet in terms of positive economic outcomes and efficient exploitation of this technology in delivering goods and services, then it's a necessary step in the broader long term public interest.

Disclaimer

The views expressed are the author's and not those of APNIC, unless APNIC is specifically identified as the author of the communication. APNIC will not be legally responsible in contract, tort or otherwise for any statement made in this publication.

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