February 2022
Geoff Huston

What's an "Address"?

I work at APNIC, the Regional Internet Registry that serves the Asia Pacific Region. APNIC provides common infrastructure services for the region that support the unique assignment of IP address blocks to Internet network operators within our region through the operation of an address registry. In short, IP addresses are what we do. So, when there are discussions in technology circles about evolving the Internet's address model in varying ways, it's a topic where we at APNIC have a deep interest! Currently, there are discussions in the IETF's Internet Area on the topic of architectural evolution of the Internet and its implications for the changing role of IP addresses, and I'd like to share some of my thoughts on this topic here.

Identity, Location and Reachability

Addresses in various guises are a consistent attribute in most communications systems. It's what we used to put on the envelope of a letter to tell the postal service where we wanted the letter delivered, when we used to write letters to each other. It's what we used to associate with a telephone handset so that we could initiate a conversation, when we had a telephone network. And it's what we attach to every IP packet that moves through the Internet. Addresses, and the way we used them, are a vital component of the architecture of any network, and the Internet is no exception here.

The consistent attributes within any addressing scheme are the aspects of *identity*, *location* and *reachability*.

Identity is associated with concepts of uniqueness. In theory, a useful addressing scheme assigns a unique address to each entity. If it doesn't, then we need to resort to additional measures to resolve conflicts where multiple entities share the same address. A telephone number is a good example here, in that the desired property of the telephone number is that it is associated with a single connected handset (This was a useful starting point, although in this world of mobile devices, various forms of call redirection, group pickup and similar, we seem to have spent a lot of effort to support exceptions to this generic principle in those times when telephones really mattered in our society.

IP addresses were assigned a similar role in the Internet architecture, although there were some subtle changes with the Internet as compared to other computer network protocols. Strictly speaking, IP addresses do not identify the computer. They identify the point of attachment of a computer (or IP protocol stack) to a network. If a computer has multiple active network interfaces, then it will have multiple IP addresses. IP packets that have a particular destination IP address will, in theory, be presented to the host's local protocol stack on the interface with the matching IP address. In theory, the outgoing packets are similarly constrained, in that the source IP address attached to a packet will reflect the network interface used by the host to pass that packet into the network.

However, identity without *location* not so useful for a network. Knowing that some entity exists (or has an address) but not knowing where it might be located within the network is not exactly helpful in the context of a network, so most addressing systems attempt to perform form of association between addresses as an identification token and addresses as a location attribute. In the telephone world the E.164 number often used a progressive number code to identify the country, the local and even the city. Two telephone numbers with a common prefix were considered to be in the same local call zone, and

making calls between them generally attracted a lower tariff. A simplified system was adopted in the IP world and IP addresses are divided into a network part and a host part. All hosts attached to the same network share a common network prefix in their IP addresses, and their addresses vary only in the host part. In IPv4 we subsequently allowed this network identifier demarcation point to be variable, while in IPv6 we reverted to a fixed model for the most part, where the host identifier occupied the least significant 64 bits while the network part is the high order 64 bits of the address. So, to the extent that every IP address is the concatenation of a network identifier and host identifier, then IP addresses include a location field in the form of a network identifier.

Finally, addresses can contain some form of *reachability* information, indicating how the network can facilitate a communication from one connected entity to another. This can be used in the context of setting a network state that creates from form of virtual circuit between the two parties, allowing subsequent transactions to reference this circuit state. In stateless packet-switched networks the address information is loaded into every active network element, allowing each packet to be passed through the network in a manner that is not necessarily dependant on the handling of any previous packets.

The classical IP address architecture uses a single token set, the IP address to support all three functions. An IP address is intended to function as an endpoint identifier, a locator and the way that a packet can inform the network's packet forwarding decisions to pass the packet towards its intended destination.

This combination of roles into a single address has been variously described as a key element of the Internet's architecture that enabled significant cost efficiencies in the network platform and as a case of semantic overload where the requirements of one function are in juxtaposition to another (such as the desire to disambiguate location and identity in mobile network platforms, for example).

Architectural Implications of Addresses as Identity Tokens

This overloaded semantic of the implicit association of an IP address with both an identity and a location within the network was perhaps not that important a distinction in the Internet's architecture until mobile networks started to be incredibly valuable for the industry. The efforts to create an "all-IP" mobile framework that allowed a mobile device to roam into a network and attach to it, while still maintaining its identity as provided by its "home" IP address failed comprehensively in the marketplace. Neither Mobile IPv4 or Mobile IPv6 got anywhere in terms of support, from platform vendors or from network operators, mobile or otherwise.

The solution was far more pragmatic and involved some obvious trade-offs. A device can maintain its IP address within the same mobile network using circuit switching and mobile handoffs within the access part of the mobile network. Sessions can remain open and the handoff between access points is seamless at the IP level. Roaming to different providers is not seamless, and it invariably requires the assignment of a new IP address set to the device and upper-level associations need to be rebuilt. By and large consumers have accepted these limitations, and there is little residual demand for a totally seamless mobile handoff with full roaming. For those who have such a need, the solution lies in tunnelling and various forms of VPN solutions.

The deeper implication of an addresses as an identity token was the unique association of an IP address and an attached device. This was perfectly reasonable in the time of large mainframe shared computers, but the cracks started to show as computing platforms become smaller and more customised to individual use with the personal computer, and it was really struggling when the massive numbers of personal use mobile devices that were subsequently bought to bear on the Internet. The concept of a 32-bit address space with a numeric capacity of 2,294,967,296 addresses seemed like a vast space in 1980s, but a crammed and woefully inadequate space by the 2000's. But the cracks in the Internet's address architecture were visible many years earlier in the late 1980's with the leap of the Internet out of the research world into world of commerce.

Re-casting Address Semantics

It's clear in retrospect that we commenced down the road of re-casting the role of IP addresses at about the same time as we started down the track of consumer deployment via dial-up ISPs.

A customer created a connection over a phone call, and as part of the session establishment the customer presented their login credentials to the provider at the other end of the call and was provided with an IP address to use in the ensuing session.

The client did not have a pre-assigned IP address range for their network. The connection provided them with a single IP address for the duration of the session. In this way, we introduced the notion of time-shared addresses, where the IP address you got was temporal for the life of the connection.

We subsequently introduced the notion of port-based address sharing via Network Address Translators, where this single address was shared among a collection of local devices. I personally don't believe that the prospects of impending scarcity in IPv4 addresses were the predominate driving factors behind the adoption of this model. I think it was a more mundane issue of cost transfer from the network provider to the client. From the dial-up ISP's perspective we headed down the shared address path for IP addressing because it was cheaper for the ISP industry at the time to do so. The ISP's address pool requirements scaled with the number of simultaneous connections, a number that was potentially far smaller than the number of the ISP's clients, and the entire set of tasks associated with address management was able to be managed with the ISP's Access, Authentication and Accounting tools.

The subtle change here was the incorporation of intermittently connected devices into the Internet model. Obviously, these devices would be poor candidates as service hosts, as they would be inaccessible for much of the time, so we started to think of the Internet in terms of a segmentation of devices into *clients* and s*ervers*. Clients initiated connections to servers. They did not initiate connections to other clients. Accordingly, clients had no requirement for a persistent form of network identity other than their access credentials. And if that's the case then we could share IP addresses across multiple clients.

At around the same time, in the early 90's we were heading down the "running out of addresses" path which ultimately led to the design of IPv6.

It is useful here to contrast the differences in these two approaches of deliberative planning and incremental piecemeal responses. The IPv6 design effort was an IETF-led command-and-control effort that attempted to anticipate future industry needs and produce a technology that would meet these needs. On the other hand, the industry was being driven by two imperatives, one was unchecked levels of escalating demand that completely swamped any efforts to satiate it (we were building as fast as we could, not as fast as consumers wanted), and the other was the continuing imperative to strip cost out of the service. Time-sharing addresses was a pragmatic and cheap way of responding to this situation that could be performed by individual network operators without necessarily coordinating their actions with any other party. It appears that NATs did not gain in popularity because a prudent and far-sighted industry was wanting to defer the costs of a large-scale protocol transition. They gathered momentum in deployment because they represented a cheaper approach to the service provider.

The important point here is that this recasting of the role of addresses was not a particularly deliberative process at the time. It was a consequence of an industry response to the prevailing conditions, and in a deregulated market-based activity that's the only driving factor out there.

IPv6

IPv6 was a return of the original address semantics for the Internet architecture, with a few slight tweaks along the way. In the manner of the EUI-48 Ethernet MAC addresses, IPv6's 128 bits had ample space for devote 64 bits for a potentially unique individual host interface address. Accordingly, the host identifier part of the IP address was not just unique in the context of the network but was intended to be unique in the same way that individual MAC addresses are intended to be unique.

However, at the time the Internet was grappling with more than just the prospect of IPv4 address exhaustion at some ill-defined point in the future. The more immediate problem was how to scale up the network's infrastructure to cope with escalating demand. Not only were the number of connected clients increasing at exponentially growing rates, but the scale of service demand per customer was growing at similarly explosive rates. The result was a billion-fold expansion of the network across the 15 years from 1995 to 2010. Impressive as this outcome has been, the scope of this activity has been pointedly focused, and there was little appetite to add the large-scale deployment of IPv6 to this already challenging agenda. It was also the case that industry was already very comfortable with a bifurcated model of a network architecture that divided up the network into clients and servers. Servers might need a stable IP address in order to terminate connections, but clients were quite comfortable in working without a static IP address, and, by implication, clients (and servers) were quite comfortable in foregoing a fixed address-based identity.

The service architecture was also changing, and applications included the identity of the server in the initial exchange, including, for example, in the Service Name Indication field in Transport Layer Security, and including the service name in HTTP transactions. The implication is that specific service identity was no longer uniquely bound to an individual IP address. Many services could be located on a service host located behind a single IP address, and a service could be replicated on multiple service hosts each with their own IP address.

The result was that while IPv6 had sufficient space in the address field to allow network endpoints and services to be uniquely associated with an IPv6 address, there was no longer any need to do so, nor were there cost efficiencies to be realised. The reason why the transition to IPv6 has taken more than twenty years so far lies in the lack of a compelling proposition that applies to all parts of this industry. However, the Internet itself has not been standing still all this time. Other fundamental changes have also been taking place and these changes have profound implications on the role of IP addresses.

The Internet of Content

Today's Internet is an Internet that is dominated by Content Distribution Networks (CDNs) and their associated "clouds". If you include video streaming data then I've heard figures of between 70% to as high as 90% of all delivered data to end customers being video streaming (No, unfortunately I have not seen solid public data to confirm these whispered mutterings - I wish I did!). It also appears that yesterday's enterprise network demand is morphing into the same CDNs as cloud services. The inexorable demands of scaling and a hiatus of in-office work during this Covid pandemic has hastened the shift of enterprise computing into various forms of clouds as well. The shared public network and its infrastructure is being marginalised (and privatised).

Another way to look at these changes is to look at the change in the role of the network. If you looked at the Internet as a service delivery platform, then the role of the network was to teleport customers to the portal of their desired service. At one point in the late 90's one of the denser points of network infrastructure was located at Microsoft's facilities in Seattle, where distributions of Microsoft Windows were being published online, and the world's Windows users (and there were many at the time) were being presented to these servers via the network. However, this model of networking assumed a basic fallacy that computation and storage were expensive and network transmission was abundant and cheap. In fact, computation and storage is abundant and inexpensive in today's world. CDNs exploit today's environment of abundance to eliminate distance in communications. By bringing replicated content and service right under the noses of consumers we eliminate the cost and performance issues of accessing remote services. Shorter networks to connect clients to services are faster and cheaper networks. The overall service outcomes are faster, cheaper, and generally more resilient in this CDN world.

The public Internet world is now in the throes of the death of transit and the size of the public Internet has shrunk to the diameter of last mile access network. Why? Ultimately the answer is "It's cheaper this way!".

The Changing Role of Addresses

So why do we even need unique addressing anymore?

Surely, all I need to do is to distinguish myself from the other clients in the service cone of my local CDN. Why do I need to use an address that distinguishes me from the other billions of client endpoints that are not my local neighbours? Is it for the few residual applications that have not yet been sucked into the CDN world?

The issue here is that uniqueness costs. Why should we spend a disproportionate amount of resource to support a function used by a residual trace amount of traffic? Sooner or later network operators will cut out that cost and just not do it anymore. As the CDN's continue to exploit the abundance of computing and storage the current shift of more points of presence positioned ever closer to the end clients will continue and the essential value of global uniqueness in the addressing infrastructure of the network will be increasingly difficult to quantify.

There is also a second factor in these changes today's network architecture, that of sunk cost. Nobody wants to pay to upgrade existing common infrastructure. Nobody. So those who want to change the network need to build around, over and tunnel through existing infrastructure. In a deregulated world where piecemeal uncoordinated actions predominate, the level of coordination and orchestration required to uplift common shared infrastructure is simply impossible. We say to ourselves that we outgrew Flag Days on the Internet many decades ago, and that's true, but at times we appear not to understand precisely what that implies about today. We have built an application-based superstructure of encapsulated tunnels in the network that neatly circumvents the entire question of infrastructure renewal. Whereas IPv6 has been head-butting against the sunk cost of existing infrastructure for more than two decades then dramatic rise of QUIC, BBR, SVCB and HTTPS, and similar application-level technologies attests to the application world's extreme distaste to engage with the existing infrastructure.

Perhaps the question is not about the transition to IPv6 anymore. The case can be made that this is largely a question whose answer really has little of industry relevance to offer any more. The question is more about the service and application world and the way applications create their own environment in a manner that is as disengaged and isolated from the underlying infrastructure as possible. This is a pretty exact replay of the way IP's stateless packet-based hop-by-hop forwarding treated the circuit-switched telephone infrastructure a few decades ago! What goes around, comes around!

These days we appear to be treating IP addresses as no more than temporary ephemeral session tokens and we've passed all the heavy lifting of service identification over to the name system. These days you and I could be accessing the same service, yet we could be using entirely different IP addresses to do so. Or I could be accessing the same service at different times and be using different addresses each time. I find it somewhat ironic that we see increasing moves to pull in IP addresses as part of the set of personally identifying information in some regulatory regimes, yet what the larger network sees of end clients is a temporary NAT-managed binding to a public address that may be shared by hundreds if not thousands of others. IPv6's use of privacy addressing achieves a similar outcome in a different way.

And even that characterisation is changing. QUIC's use of a session token inside the encrypted envelope even makes the binding of an IP address to a single session fluid, as the same QUIC session can be address agile within a single session.

So perhaps an address these days is just an ephemeral transport token that distinguishes one packet flow from another, and really has little more in the way of semantic meaning. In this world what's holding the Internet together is not IP addresses and routing: it's the DNS and the common name system.

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre.

Author

Geoff Huston AM, B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region.

www.potaroo.net