October 2024
Geoff Huston

The IPv6 Transition

I wrote an article in May 2022, asking "Are we there yet?" about the transition to IPv6. At the time I concluded the article on an optimistic note, observing that we may not be ending the transition just yet, but we are closing in. I thought at the time that we won't reach the end of this transition to IPv6 with a bang, but with a whimper. A couple of years later, I'd like to revise these conclusions with some different thoughts about where we are heading and why.

The state of the transition to IPv6 within the public Internet continues to confound us. RFC 2460, the first complete effort at a specification of the IPv6 protocol was published in December 1998, more than twenty-five years ago. The entire point of IPv6 was to specify a successor protocol to IPv4 due to the prospect of running out of IPv4 addresses. Yet we ran out of IPv4 addresses more than a decade ago while the Internet is largely sustained through the use of IPv4. This transition to IPv6 has been going on for 25 years now, and if there was any urgency to be instilled in the transition effort by the prospect and then the reality of IPv4 address exhaustion, then we've been living with exhaustion a very long time now, and we're inured to it. It's probably time to ask the question again: How much longer is this transition to IPv6 going to take?

At APNIC Labs we've been measuring the uptake of IPv6 for more than a decade now. We use a measurement approach that looks at the network from the perspective of the Internet's user base. What we measure is the proportion of users who can reach a published service when the only means to do so is by using IPv6. The data is gathered using a measurement script embedded in an online ad, and the ad placements are configured to sample a diverse collection of end users on an ongoing basis.

The IPv6 adoption report, showing our measurements of IPv6 adoption across the Internet's user base from 2014 to the present is shown in Figure 1.

On the one hand, Figure 1 is one of those classic "up and to the right" Internet curves which show continual growth in the adoption of IPv6. The problem is in the values in the scale of the Y-axis. The issue here that in 2024 we are only at a level where slightly more than one third of the Internet's user base can access an IPv6-only service. Everyone else is still in an IPv4-only Internet.

This seems to be a completely anomalous situation. It's been over a decade since the supply of "new" IPv4 addresses has been exhausted, and the Internet has not only been running on empty, but also being tasked to span an ever-increasing collection of connected devices without collapsing. In late 2024 its variously estimated that some 20 billion devices use the Internet, yet the Internet's IPv4 routing table

only encompasses some 3.03 billion unique IPv4 addresses. The original "end-to-end" architecture of the Internet assumed that every device was uniquely addressed with its own IP address, yet the Internet is now sharing each individual IPv4 address across an average of 7 devices, and apparently it all seems to be working! If "end-to-end" was the sustaining principle of the Internet architecture then as far as the users of IPv4-based access and services are concerned, then it's all over!

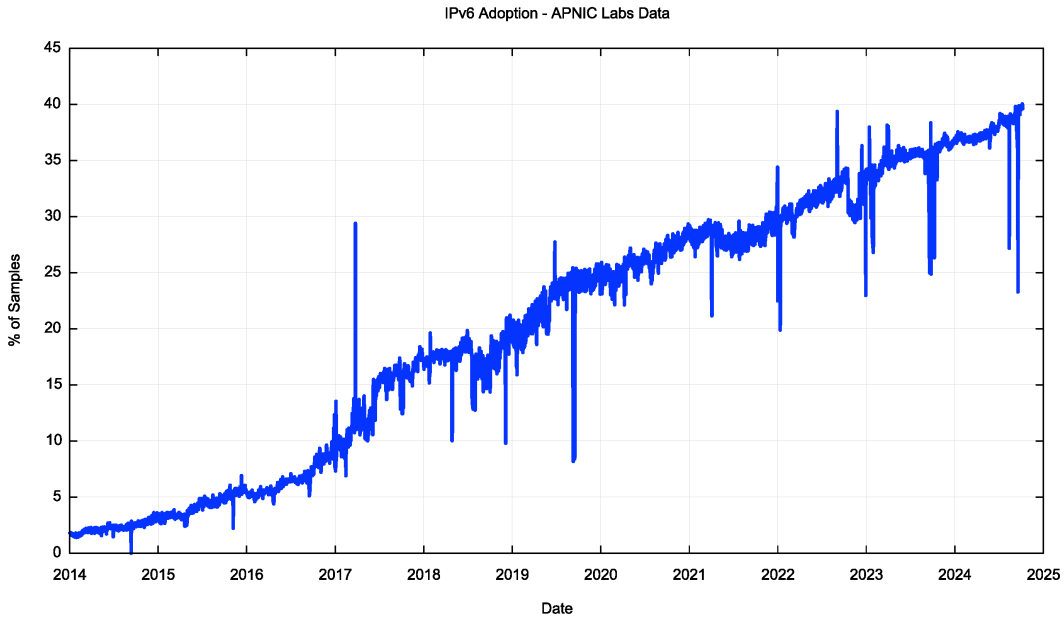


Figure 1 – IPv6 Adoption - 2014 to Now, APNIC Labs Data

IPv6 was meant to address these issues, and the 128-bit wide address fields in the protocol has sufficient address space to allow every connected device to use their own unique address. The design of IPv6 was intentionally very conservative. To a first level of approximation IPv6 is simply "IPv4 with bigger addresses". There are also some changes to fragmentation controls, changes to the address acquisition protocols (ARP vs Neighbour Discovery), and changes to the IP Options fields, but the upper-level transport protocols are unchanged. IPv6 was intended to be a largely invisible change to a single level in the protocol stack, and definitely not intended to be a massive shift to an entirely novel networking paradigm.

In the sense of representing a very modest incremental change to IPv4, the IPv6 design achieved its objective, but in so doing it necessarily provided little in the way of any marginal improvement in protocol use and performance. IPv6 was no faster, no more versatile, no more secure than IPv4. The major benefit of IPv6 was to mitigate the future risk of IPv4 address exhaustion. In terms of conventional market operations, many markets, including that of the Internet, apply a hefty discount factor to future risk. The result is that the level of motivation to undertake this transition is highly variable given that the expenditure to deploy this second protocol does not realise tangible benefits in terms of lower cost, greater revenue or greater market share. In a networking context where market-based coordination of individual actions is essential, this level of diversity of views of the value of running a dual stack network leads to reluctance on the part of individual actors and sluggish progress of the common outcome of the transition. As a result, there is no common sense of urgency.

To illustrate this, we can look at the time series shown in Figure 1 and ask the question: "If the growth trend of IPv6 adoption continues at its current rate, how long will it take for every device to be IPv6 capable?" This is the same as looking at a linear trend line placed over the data series used in Figure 1, looking for the date when this trend line reaches 100%. Using a least-squares best fit for this data set from January 2020 to the present day, and using a linear trend line, we can come up with Figure 2.

This exercise predicts that we'll see completion of this transition in late 2045, or some 20 years into the future. It must be noted that there is no deep modelling of the actions of various service providers, consumers, and network entities behind this prediction. The only assumption that drives this prediction

is that the forces that shaped the immediate recent past are unaltered when looking into the future. In other words, this exercise simply assumes that "tomorrow is going to be a lot like today."

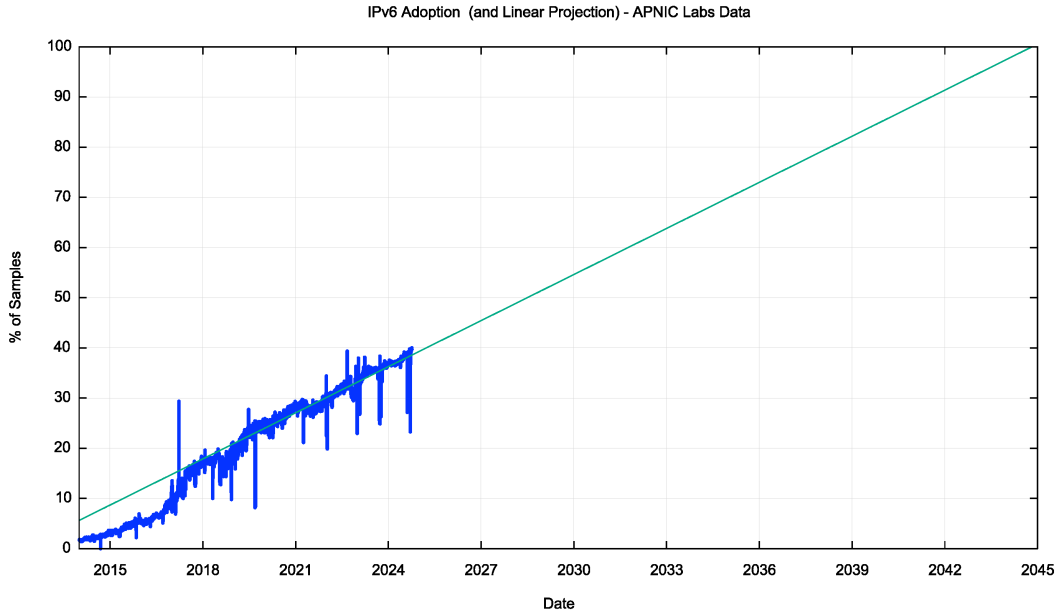


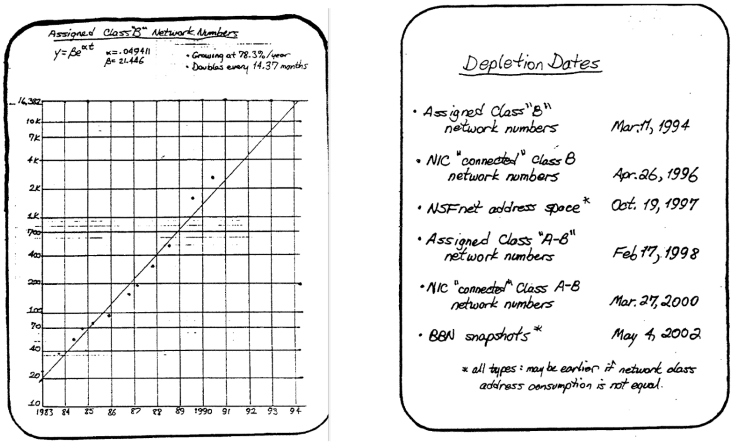
Figure 2 – IPv6 Adoption - Projection, APNIC Labs Data

The projected date in Figure 2 is less of a concern than the observation that this model predicts a continuation of this transition for a further two decades. If the entire concept of IPv6 was to restore a coherent address plan across the collection of Internet-connected devices, then having this model of coherent unique device addressing be placed in abeyance for a total of some 30 years, from around 2015 through to 2045, leads to question the role and value of such a unique device addressing framework in the first place! If we can operate a fully functional Internet without such a coherent end device address architecture for three decades, then why would we feel the need to restore address coherence at some point in the future? What's the point of IPv6 if it's not address coherence?

Something has gone very wrong with this IPv6 transition, and that's what I'd like to examine in this article.

A Little Bit of History

By 1990 it was clear that IP had a problem. It was still a tiny Internet at the time, but the growth patterns were exponential, doubling in size every 12 months. We were stressing out the pool of Class B IPv4 addresses and in the absence of any corrective measures this address pool would be fully depleted in 1994 (Figure 3).



"Internet Growth" Frank Solensky, Proc. IETF, Aug 1990

Figure 3 – IPv4 Depletion Predictions, Frank Solensky, 18th IETF Proceedings, August 1990

We were also placing pressure on the routing system at the time. The deployed routers in 1992 only had enough memory to support a further 12 to 18 months of routing growth. The combination of these routing and addressing pressures was collectively addressed in the IETF at the time under the umbrella of the ROAD effort (RFC 1380).

There was a collection of short-, medium- and longer-term responses that were adopted in the IETF to address the problem. In the short term, the IETF dispensed with the class-based IPv4 address plan and instead adopted a variably sized address prefix model. Routing protocols, including BGP, were quickly modified to support these classless address prefixes. Variably sized address prefixes added additional burdens to the address allocation process, and in the medium term the Internet community adopted the organisational measure of the Regional Internet Registry structure to allow each region to resource the increasingly detailed operation of address allocation and registry functions for their region. These measures increased the specificity of address allocations and provided the allocation process with a more exact alignment to determine adequate resource allocations that permitted a more diligent application of relatively conservative address allocation practices. These measures realized a significant increase in address utilization efficiency. The concept of "address sharing" using Network Address Translation (NATs) also gained some traction in the ISP word. Not only did this dramatically simplify the address administration processes in ISPs, NATs also played a major role in reducing the pressures on overall address consumption.

The adoption of these measures across the early 1990's pushed a two-year imminent crisis into a more manageable decade-long scenario of depletion. However, they were not considered to be a stable long-term response. It was thought at the time that an effective long-term response really needed to extend the 32-bt address field used in IPv4. At the time the transition from mainframe to laptop was well underway in the computing world and the prospect of further reductions in size and expansion of deployment in smaller embedded devices was clear at the time. An address space of 4 billion was just not large enough for what was likely to occur in the coming years in the computing world.

But in looking at a new network protocol with a vastly increased address space, there was no way that any such change would be backward compatible with the installed base of IPv4 systems. As a result, there were a few divergent schools of thought as to what to do. One approach was to jump streams and switch over to use the Connectionless Transport profile of the OSI protocol suite and adopt OSI NSAP addresses along the way. Another was to change as little as possible in IP except the size of the address fields. And there were a number of ideas being thrown about in the area of proposing significant changes to the IP model.

By 1994 the IETF had managed to settle on the minimal change approach, which was IPv6. The address field was expanded to 128 bits, a Flow ID field was introduced, fragmentation behaviour was altered and pushed into an optional header and ARP was replaced with multicast.

The bottom line was that IPv6 did not offer any new functionality that was not already present in IPv4. It did not introduce any significant changes to the operation of IP. It was just IP, with larger addresses.

Transition

While the design of IPv6 consumed a lot of attention at the time, the concept of transition of the network from IPv4 to IPv6 did not.

Given the runaway adoption of IPv4, there was a naive expectation that IPv6 would similarly just take off, and there was no need to give the transition much thought. In the first phase, we would expect to see applications, hosts and networks adding support for IPv6 in addition to IPv4, transforming the internet into a dual stack environment. In the second phase we could then phase out support for IPv4.

There were a number of problems with this plan. Perhaps the most serious of these was a resource allocation problem. The Internet was growing extremely quickly, and most of our effort was devoted to keeping pace with demand. More users, more capacity, larger servers, more content and services, more responsive services, more security, better defence. All of these shared a common theme: scale. We could either concentrate our resources on meeting the incessant demands of scaling, or we could work on IPv6 deployment. The short and medium-term measures that we had already taken had addressed the immediacy of the problems of address depletion, so in terms of priority, scaling was a for more important priority for the industry than IPv6 transition. Through the decade from 1995 to 2005 the case for IPv6 quietly slumbered in terms of mainstream industry attention. IPv4 addresses were still available, and the use of classless addressing (CIDR) and far more conservative address allocation practices had pushed the prospect of IPv4 address depletion out by more than a couple of decades. There were many more pressing operational and policy issues for the Internet that absorbed the industry's collection attention on the day.

However, this was merely a brief period of respite. The scaling problem accelerated by a whole new order of magnitude in the mid 2000's with the introduction of the iPhone and its brethren. All of a sudden this was not just a scale problem of the order of tens or even hundreds of millions of households and enterprises, but it transformed to a scale problem of billions of individuals and their personal devices and added mobility into the mix. As a taste of a near term future, the production scale of these "smart" devices quickly ramped up into annual volumes of hundreds of millions of units. The entire reason why IPv6 was a necessity was coming into fruition. But at this stage we were just not ready to deploy IPv6 in response. Instead, we rapidly increased our consumption of the remaining pools of IPv4 addresses and we supported the first wave of large-scale mobile services with IPv4. Dual stack was not even an option in the mobile world at the time. The rather bizarre economics of financing 3G infrastructure meant that dual stack infrastructure in a 3G platform was impractical, so IPv4 was used to support the first wave of mobile services. This quickly turned to IPv4 and NATs as the uptake of mobile services gathered momentum.

At the same time the decentralised nature of the Internet was hampering IPv6 transition efforts. What point was there in developing application support for IPv6 services if no host had integrated IPv6 into its network stack? What point was there in adding IPv6 to a host networking stack if no ISP was providing IPv6 support? And what point was there in an ISP in deploying IPv6 if no hosts and no applications would make use of it? In terms of IPv6 at this time, nothing happened.

The first efforts to try and break this impasse of mutual dependence was the operating system folk, and fully functional IPv6 stacks were added to the various flavours of Linux, Windows and MAC OS, as well as in the mobile host stacks of iOS and Android.

But even this was not enough to allow a transition to achieve critical momentum. It could be argued that this situation made the IPv6 situation worse and set back the transition by some years. The problem was that with IPv6-enabled hosts there was some desire to use IPv6. However, these hosts were isolated "islands" of IPv6 sitting in an ocean of IPv4. The concentration of the transition effort then fixated on various tunnelling methods to tunnel IPv6 packets through the IPv4 networks (Figure 4). While this can be performed in a manual manner when you have control over both tunnel endpoints, this was not that useful an approach. What we wanted was an automated tunnelling mechanism that took care of all these details.

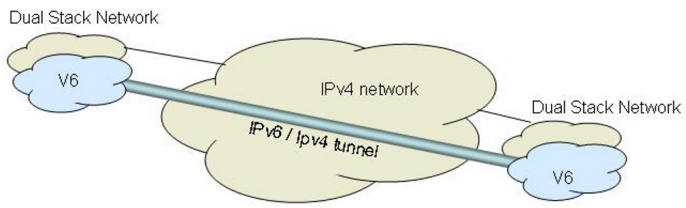


Figure 4 – Phase 1 of the IPv6 Transition

The first such approach that gathered some momentum was 6to4. The first problem with 6to4 was that is required public IPv4 addresses, so it could not provide services to IPv6 hosts that were behind a NAT. The more critical problem was that firewalls had no idea how to handle these 6to4 packets, and the default action when in doubt is to deny access. So 6to4 connections encountered an average of a 20% - 30% failure rate in the public Internet, which made it all but unusable as a mainstream service. The NAT traversal issue was also a problem, so a second auto-tunnel mechanism was devised that performed NAT sensing and traversal. This mechanism, Teredo, was even worse in terms of failure rates, and some 40% of Teredo connection attempts were observed to fail.

Not only were these Phase 1 IPv6 transition tools extremely poor performers, as they were so unreliable, but even when they worked the connection was both fragile and slower than IPv4. The result was perhaps predictable, even if unfair. It was not just the transition mechanisms that were viewed with disfavour, but IPv6 itself also attracted some opprobrium.

Up until around 2011 IPv6 was largely ignored as a result in the mainstream of the public Internet. A small number of service providers tried to deploy IPv6, but in each case they found themselves with a unique set of challenges that they and their vendors had to solve, and without a rich set of content and services on IPv6, then the value of the entire exercise was highly dubious! So, nothing much happened.

Movement at last!

It wasn't until the central IPv4 address pool managed by the IANA was depleted at the start of 2011, and the first RIR, APNIC, ran down on its general allocation pool in April of that year, that the ISP industry started to pay some more focussed attention to this IPv6 transition.

At around the same time the mobile industry commenced their transition into 4G services. The essential difference between 3G and 4G was the removal of the PPP tunnel through the radio access network from the gateway to the device and its replacement by an IP environment. This allowed a 4G mobile operator to support a dual stack environment without an additional cost component, and this was a major enabler for IPv6. Mapping IPv4 into IPv6 (or the reverse) is fragile and inefficient for service providers as compared to native dual stack. In the six-year period, from 2012 to the start of 2018 the level of IPv6 deployment rose from 0.5% to 17.4%. At this stage IPv6 was no longer predominately tunnelled, as many networks supported IPv6 in native node (Figure 5).

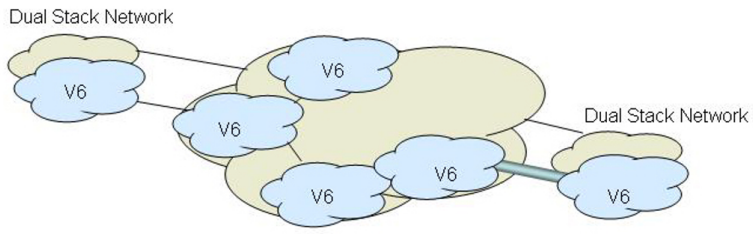


Figure 5 – Phase 2 of the IPv6 Transition

The problem here was that we were late with this phase of the transition. The intention of this transition was to complete the work and equip every network and host with IPv6 before we ran out of IPv4 addresses (Figure 6).

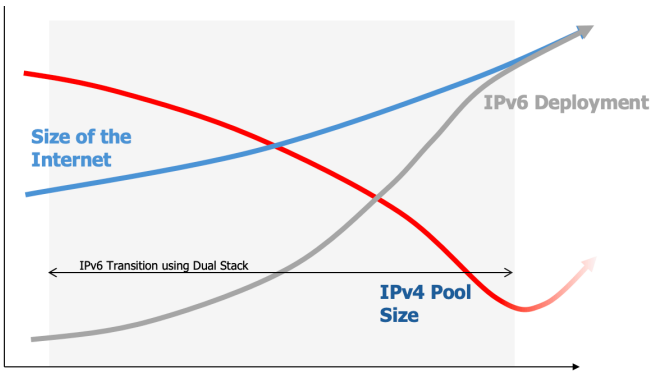


Figure 6 – The IPv6 Transition Plan

Where we had got to by 2012 was a far more challenging position. The pools of available IPv4 address space were rapidly depleting and the regional address policy communities were introducing highly conservative address allocation practices to eke out the remaining address pools. At the same time the amount of IPv6 uptake was minimal. The transition plan for IPv6 was largely broken (Figure 7).

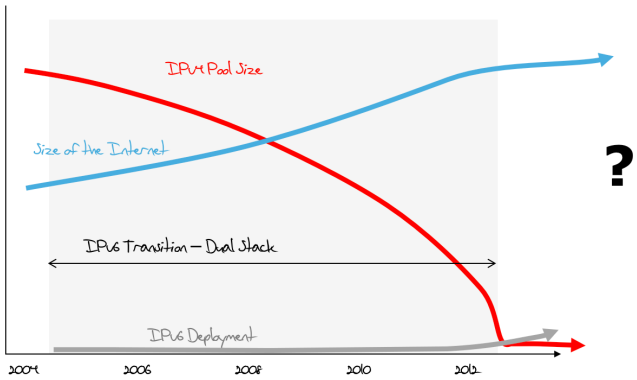


Figure 7 – The IPv6 Transition Plan in 2012

NATs and Address Scarcity Pressures

At this point there was no choice for the Internet, and to sustain growth in the IPv4 network while we were waiting for IPv6 to gather momentum we turned to NATs. NATs were a challenging subject for the IETF. The entire concept of coherent end-to-end communications was to eschew active middleware in the network, such as NATs. NATs created a point of disruption in this model, creating a critical dependency upon network elements. They removed elements of network flexibility from the network and at the same time reduced the set of transport options to TCP and UDP.

The IETF resisted any efforts to standardise the behaviour of NATs, fearing perhaps that standard specifications of NAT behaviour would bestow a legitimacy on the use of NATs, an outcome that that a number of IETF participants were very keen to avoid. This aversion did not reduce the level of impetus behind NAT deployment. We had run out of IPv4 addresses and IPv6 was still a distant prospect, so NATs were the most convenient solution. What this action did achieve was to create a large variance of NAT behaviours in various implementations, particularly with respect to UDP behaviours. This has exacted a cost in software complexity where an application needs to dynamically discover the type of NAT (or NATs) in the network path if it wants to perform anything more complex than a simple two-party TCP connection.

Despite these issues NATs were a low friction response to IPv4 address depletion where individual deployment could be undertaken without incurring external dependencies. On the other hand, deployment of IPv6 was dependant on other networks and servers also deploying IPv6. NATs made

highly efficient use of address space for clients, as not only could a NAT use the 16-bit source port field, but by time-sharing the NAT binding, NATs achieved an even greater level of address efficiency. A major reason why we've been able to sustain an Internet with 10's of billions of connected devices is through the widespread use of NATs.

Server architectures were also changing. The introduction of TLS (Transport Layer Security) into the web server world included a point in TLS session establishment where the client informs the server platform the name of the service that it intended to connect to. Not only did this allow TLS to validate the authenticity of the service point, but this also allowed a server platform to host an extremely large collection of services from a single platform (and a single platform IP address) and perform individual service selection via this TLS Server Name Indication (SNI). The result is that server platforms perform service selection by name-based distinguishers (DNS names) in the session handshake, allowing a single server platform to serve large numbers of individual servers. The implications of the widespread use of NATs and the use of server sharing in service platforms has taken the pressure off the entire IPv4 address environment.

One of the best ways to illustrate the changing picture of address scarcity pressure in IPv4 is to look at the market price of address transfers over the past decade. Scarcity pressure is reflected in the market price. A time series of the price of traded IPv4 addresses is shown in Figure 8.

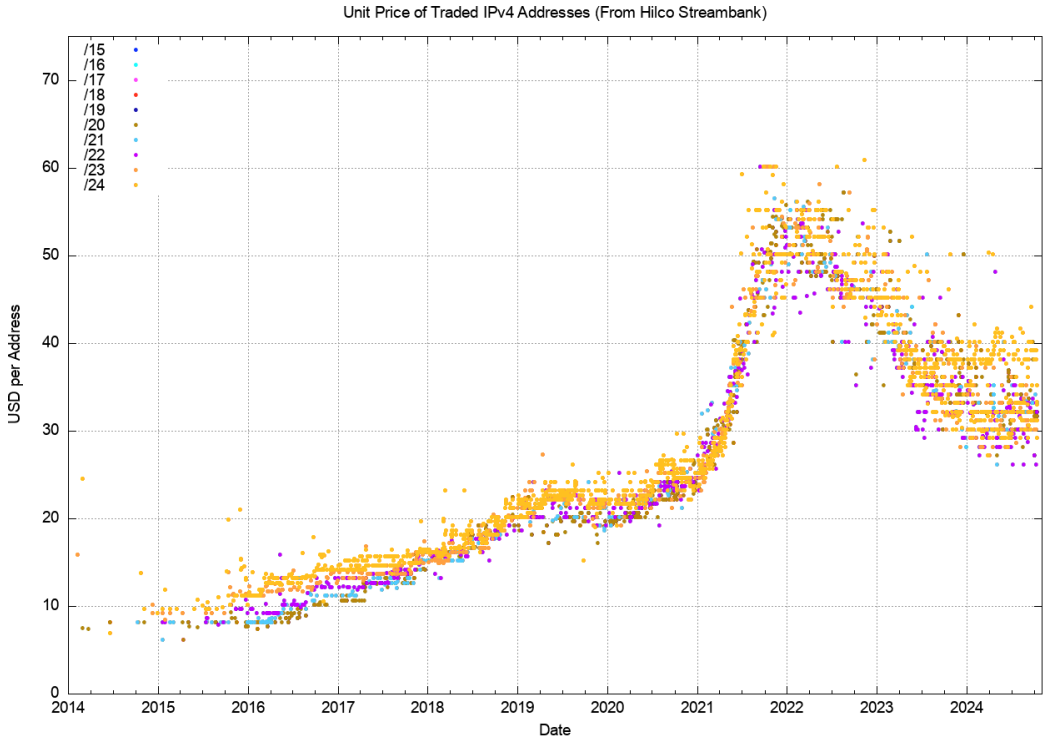


Figure 8 – Market Prices of IPv4 Address Transfers (Data from Hilco Streambank)

The period of the Covid outbreak coincided with a rapid price escalation over 2021, but the price has since declined to between $30 to $40 per address, and this price, admittedly over a $16 range from $26 to $42 per address, has been stable across 2024. This price data indicates that IPv4 addresses are still in demand in 2024, but the level of demand appears to have equilibrated against available levels of supply, implying that there is no scarcity premium in evidence in the address market in 2024. This data points to the combination of the efficacy of NATs in extending the efficiency of IPv4 addresses by making use of the 16 bits of port address space plus the additional benefits of using shared address pools.

However, it's not just IPv4 that has alleviated the scarcity pressure for IPv4 addresses. Figure 1 indicates that over the past decade the level of IPv6 adoption has risen to encompass some 40% of the user base of the Internet. Most applications, including browsers, support *Happy Eyeballs*, which is a shorthand

notation for preferring to use IPv6 over IPv4 if both protocols are available for use in support of a service transaction. As network providers roll out IPv6 support, the pressure on their IPv4 address pools for NAT use is relieved due to the applications' preference to use IPv6 where available.

How much longer?

Now that we are somewhere in the middle of this transition, then the question is: How much longer is this transition is going to take?

This seems like a simple question, but it does need a little more elucidation. What is the "end state" when we can declare the transition to be over? When will this transition be "complete"? Is it the time when there is no more IPv4-based traffic on the internet? Or is it the time when there is no requirement for IPv4 in public services on the Internet? Or do we mean the point when IPv6-only services are viable? Or perhaps we should look at the market for IPv4 addresses and define the endpoint of this transition at the time when the price of IPv4 addresses completely collapses? Perhaps we can take a more pragmatic position here and rather than looking for completion as the point when the Internet is completely bereft of all use of IPv4 addresses and their use, we could define "completion" as the point when use of IPv4 is no longer necessary. This would imply that when a service provider can operate a viable Internet service using only IPv6 and having no supported IPv4 access mechanisms at all, then we would've completed this transition.

What does this imply? Certainly, the ISP needs to provide IPv6. But as well all the connected edge networks and the hosts in these networks need to support IPv6. After all, the ISP has no IPv4 services at this point of completion of the transition. It also implies that all the services used by the clients of this ISP must be accessible over IPv6. Yes, this includes all the popular cloud services and cloud platforms, all the content streamers and all the content distribution platforms. It also includes specialised platforms such as Slack, Xero, Atlassian and similar. The data published at Internet Society's Pulse page reports that only some 47% of the top 1000 web sites are reachable over IPv6, and clearly a lot of service platforms have work to do, and this will take more time.

When we look at the IPv6 adoption data for the United States there is another somewhat curious anomaly (Figure 9).

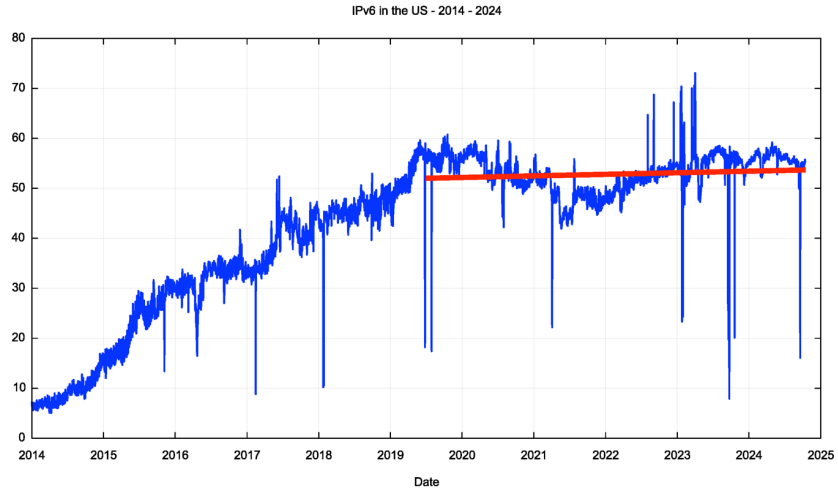


Figure 9 – IPv6 Adoption in the US - 2014 to Now, APNIC Labs Data

The data shows that the level of IPv6 use in the US has remained constant since mid-2019. Why is there no further momentum to continue with the transition to IPv6 in this part of the Internet? I would offer the explanation that the root cause is a fundamental change in the architecture of the Internet.

Changes to the Internet Architecture

The major change to the Internet's architecture is a shift away from a strict address-based architecture. Clients no longer need the use of a persistent unique public IP address in order to communicate with servers and services. And servers no longer need to use a persistent unique public IP address to provide clients with access to the service or content. Address scarcity takes on an entirely different dimension when unique public addresses are not required to number every client and every distinct service.

Some of the clues that show the implications of this architectural shift are evident when you look at the changes in the internal economy of the Internet. The original model of IP was a network protocol that allowed attached devices to communicate with each other. The network providers supplied the critical resource to allow clients to consume content and access services. At the time the costs of the network service dominated the entire cost of the operation of the Internet, and in the network domain distance was the dominant cost factor. Network providers who provided distance services (so-called "transit providers") were the dominant providers. Little wonder that we spent a lot of our time working through the issues of interconnection of network service providers, customer/provider relationships and various forms of peering and exchanges. The Internet Service Providers were in effect brokers in the rationing of the scarce resource of distance capacity. This was a classic network economy (Figure 10).

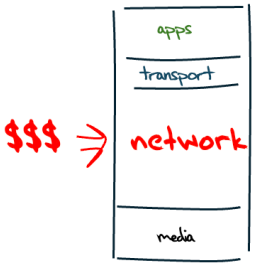


Figure 10 – The Classic Network Economy

For many years the demand for communications services outstripped available inventory, and price was used as a distribution function to moderate demand against available capacity. However, all of this changed due to the effects of Moore's Law consistently changing the cost of computing and communications.

The most obvious change has been in the count of transistors in a single integrated circuit. Figure 11 shows the transistor count over time since 1970.

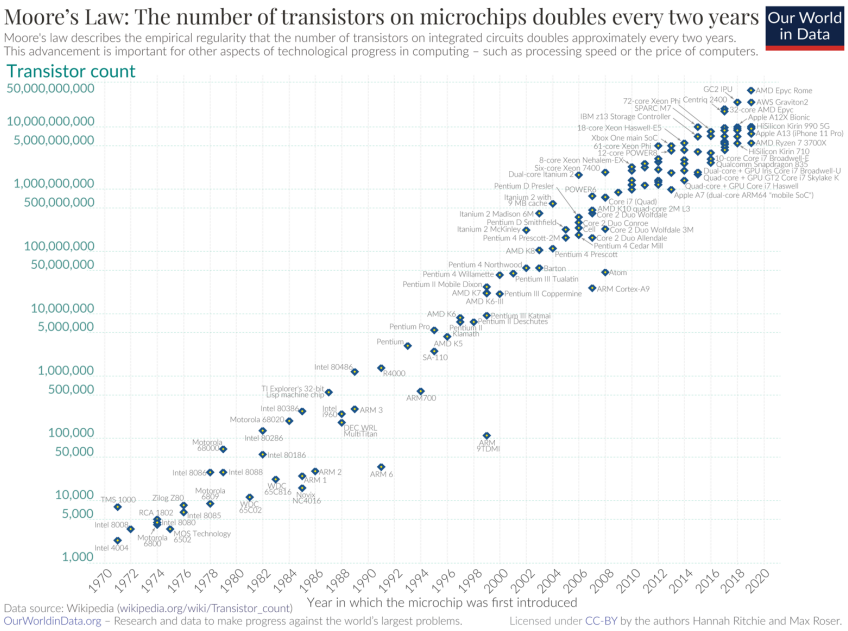


Figure 11 – Transistor count over time – from https://assets.ourworldindata.org/uploads/2020/11/Transistor-Count-over-time.png

The latest production chips in 2024 are the Apple M3, a 3nm chip with up to 92 billion transistors. With perhaps the possible exception of powering AI infrastructure, these days processing capability is an abundant and cheap resource.

This continual refinement of integrated circuit production techniques has an impact on the size and unit cost of storage (Figure 12). While the speed of memory has been relatively constant far more than a decade, the unit cost of storage has been dropping exponentially for many decades. Storage is also an abundant resource.

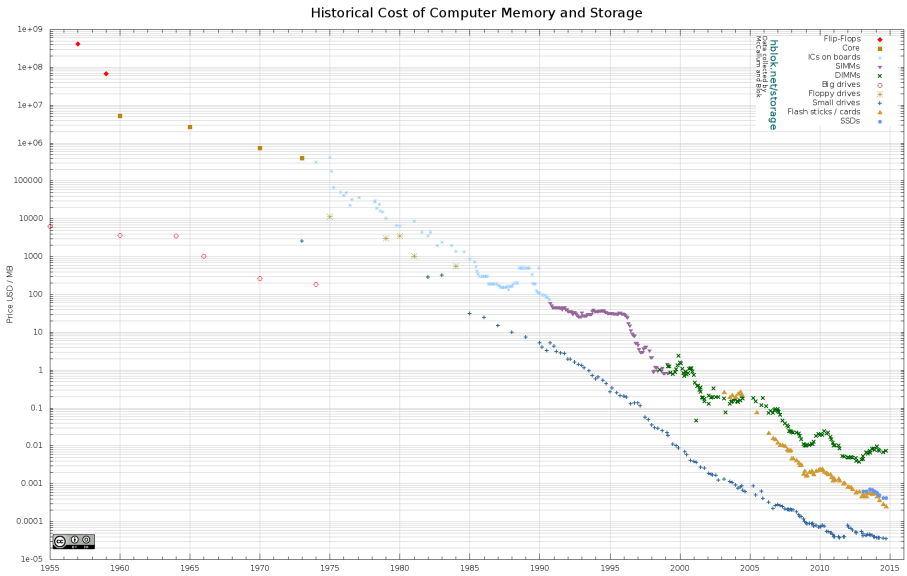


Figure 12 – Computer Memory and Storage unit costs over time – from http://aiimpacts.org/wp-content/uploads/2015/07/storage_memory_prices_large-_hblok.net_.png

These changes in the capabilities of processing have also had a profound impact on communications costs and capacities. The constraining factor in fibre communications systems is the capabilities of the digital signal processors and the modulators. As silicon capabilities improve, it's possible to improve the signal processing capabilities of transmitters and receivers, which allows for a greater capacity per wavelength on a fibre circuit (Figure 13).

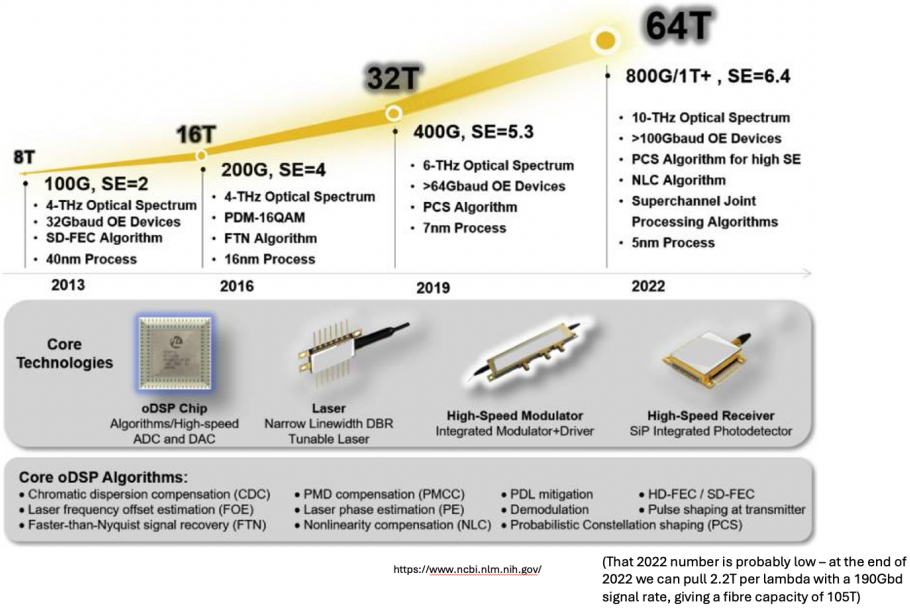


Figure 13 – Fibre Capacity over time - from https://www.ncbi.nlm.nih.gov/

This change from scarcity to abundance in processing, storage and transmission capacity has had a profound impact on the service model of the Internet. The model has changed from an *on-demand* pull to a *just-in-case* model of pre-provisioning. These days we load replicas of content and services close to the edge of the network where the users are located and attempt to deliver as much of the content and service as possible from these edge points of presence to the users in the adjacent access networks. These changes in the underlying costs of processing and storage have provided the impetus for the expansion of various forms of content distribution networks (CDNs) which now serve almost the entirety of Internet content and services. In so doing, we've been able to eliminate the factor of distance from the network and most network transactions occur over short spans.

The overall result of these changes is the elimination of distance in pushing content and services to clients. We are able to exploit the potential capacity in 5G mobile networks without the inefficiencies of operating the transport protocol over a high delay connection. Today's access networks operate with greater aggregate capacity, and the close proximity of service delivery platform and client allow transport protocols to make use of this capacity, as transport sessions that operate over a low latency connection are also far more efficient. Service interactions across shorter distances using higher capacity circuitry results in a much faster Internet!

As well as "bigger" and "faster," this environment of abundant communications, processing and storage capacity is operating in an industry when there are significant economies of scale. And much of this environment is funded by capitalising a collective asset that is infeasible to capitalise individually, namely the advertisement market. The result of these changes is that a former luxury service accessible to just a few has been transformed into an affordable mass-market commodity service available to all.

However, it's more than just bigger, faster and cheaper. This shift into abundance of basic inputs for the digital environment has shifted the economics of the Internet as well. The role of the network as the arbiter of the scarce resource of communication capability has dissipated. In response, the economic focus of the Internet economy has shifted up the protocol stack to the level of applications and services (Figure 14).

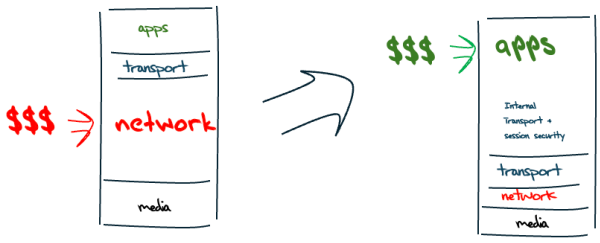


Figure 14 – The Transformation of the Network Economy

Now let's return to the situation of the transition to IPv6. It is left to networks and network operators to make the investments to switch to a dual stack platform initially (and then ultimately to remove support for IPv4). But this change is really not visible, or even crucial, to the content or service world. If IPv4 and NATs perform the carriage function adequately, then there is no motivation for the content and service operators to pay a network a premium to have a dual stack platform.

It's domain names that operate as service identifiers, and its domain names that underpin the users' tests of authenticity of the online service, and it's the DNS that increasingly is used to steer users to the "best" service delivery point for content or service. From this perspective addresses, IPv4 or IPv6, are not the critical resource for a service and its users. The "currency" of this form of CDN networking is names.

So where are we in 2024? Today's public Internet is largely a service delivery network using CDNs to push content and service as close to the user as possible. The multiplexing of multiple services onto underlying service platforms is an application-level function tied largely to TLS and service selection

using SNI field of the TLS handshake. We use the DNS to perform "closest match" service platform selection. It's the objective of a CDN to directly attach to the access networks where its users are located, and the result is a BGP routing table inside the CDN with an average AS Path Length that is intended to converge to 1! From this respect the DNS has supplanted the role of routing! We may not route "names" in today's Internet, but it is certainly operating in a way that is largely isomorphic to such a named data network.

There are a few additional implications of this architectural change for the Internet. TLS, like it or not (and there is much to criticise about the robustness of TLS), is the sole underpinning of authenticity in the Internet. DNSSEC has not gathered much momentum to date. DNSSEC is too complex, too fragile and just too slow to use for the majority of services and their users. Some value its benefits highly enough that they are prepared to live with its shortcomings, but that's not the case for most name holders and most users, and no amount of passionate exhortations about DNSSEC will change this! It supports the view that it's not the mapping of a name to an IP address that's critical. What's critical is that the named service is able to demonstrate that it operated by the owner of the name. Secondly, RPKI, the framework for securing information being passed in the BGP routing protocol, is really not all that useful in a service network where there is no routing!

The implication of these observations is that the transition to IPv6 is progressing very slowly not because this industry is chronically stupid or short-sighted. There is something else going on here. IPv6 alone is not critical to a large set of end user service delivery environments. We've been able to take a 1980's address-based architecture and scale it more than a billion-fold by altering the core reliance on distinguisher tokens from addresses to names. There was no real lasting benefit in trying to leap across to just another 1980's address-based architecture (with only a few annoyingly stupid differences, apart from longer addresses!).

Where is this heading in the longer term? We are pushing everything out of the network and over to applications. Transmission infrastructure is becoming an abundant commodity. Network sharing technology (multiplexing) is decreasingly relevant. We have so much network and computing resources that we no longer have to bring consumers to service delivery points. Instead, we are bringing services towards consumers and using the content frameworks to replicate servers and services With so much computing and storage the application is becoming the service, rather than just a window to a remotely operated service.

If that's the case, then will networks matter any more? The last couple of decades have seen us stripping out network-centric functionality and replacing this will an undistinguished commodity packet transport medium. Its fast and cheap, but it's up to applications to overlay this common basic service with its own requirements. As we push these additional functions out to the edge and ultimately off the network altogether, we are left with simple dumb pipes!

> You could argue that this is nothing new, and it's a continuation of the disruption that the Internet itself bought to bear on the predecessor telephone network infrastructure. The Internet architecture shifted functionality out of the core of the network and replaced the network service of synchronous real-time end-to-end virtual circuits with an extremely basic data packet delivery service where networks were permitted to drop, duplicate, reorder and re-time these packets in flight across the network.
>
> It was left to the control functions that were embedded in the attached devices (such as the TCP protocol, for example) to create a functional reliable end-to-end communications service model. Internet hosts only valued a network to the level of a basic (and imperfect) packet delivery

service. A network's clients were unwilling to pay a price premium for network-level services that were already being provided by the edge devices.

The result is a diminished network, dramatically reduced both in role and in value. This diminished role impairs the operators of networks to raise additional revenue through augmented services, whether it's through variable service responses through Quality of Service responses or even as basic as IPv6 protocol support.

At this point it's useful to ask: What "defines" the Internet? Is the classic response, namely: "A common shared transmission fabric, a common suite of protocols and a common protocol address pool." still relevant these days? Or is today's network more like: "A disparate collection of services that share common referential mechanisms using a common name space?"

When we think about what's important to the Internet these days, is the choice of endpoint protocol addressing really important? Is universal unique endpoint addressing a 1980's concept whose time has come and gone? If network transactions are localised, then what is the residual role of unique global endpoint addressing for clients or services? And if we cannot find a role for unique endpoint addressing, then why should we bother? Who decides when to drop this concept? Is this a market function, so that a network that uses local addressing can operate from an even lower cost base gains a competitive market edge? Or are carriage services so cheap already that the relative benefit in discarding the last vestiges of unique global addresses so small that it's just not worth bothering about?

And while we are pondering such questions, what is the role of referential frameworks in networks? Without a common referential space then how do we usefully communicate? What do we mean by "common" when we think about referential frameworks? How can we join the 'fuzzy' human language spaces with the tightly constrained deterministic computer-based symbol spaces?

Certainly, there is much to think about here!

And where does this leave the transition to IPv6? I suspect that the dual stack world we're in is a world we will be stuck in for quite some time. There seems to be no appetite to resolve this situation by completing the transition any time soon, and absolutely no desire to back out and revert to a IPv4-only network. This is where we are, caught in a partial state of transition to IPv6 that is taking on an unfortunate air of permanence! And as the preponderance of value in this environment continues to move up the protocol stack into service, content and today generative content in the guise of AI, there is little continued capacity to place collective attention on questions that have been left unresolved for decades.

It may well be that the question of when will this IPv6 transition end is a question that engenders decreasing levels of interest and attention in line with the larger picture of decreasing relative economic value of the answer! Silicon abundance has enabled a few select content and service operators to privatise much of the former public communications platform, and in so doing they have managed to shrink the public Internet to a set of margins at the edges. That implies that the answer to the IPv6 transition question may soon be: "Who cares anyway?"

Disclaimer

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

Author

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

www.potaroo.net