October 2021 Geoff Huston

# Fifty Years On

When did the Internet begin? It all gets a bit hazy after so many years, but by the early 1970's research work in packet switched networks was well underway and while it wasn't running TCP at the time (the flag day when the ARPANET switched over to use TCP was not until 1 January 1983) but was there was the base datagram internet protocol running in the early research ARPA network in the US. Given that this is now around 50 years ago, and given that so much has happened in this last 50 years then what does the next 50 years have in store? This was the question posed in a recent workshop hosted by IBM Research, on "Future of Computer Communications Networks" (https://ibm.biz/2021-fccn-conference), and I was invited to present at this workshop. I'd like to share my thoughts on this rather challenging topic, based on a presentation I made to this workshop (https://www.potaroo.net/presentations/2021-10-13-ibm-futures.pdf).

Luckily for me we were not asked to muse about the future of computers and computing over this same period, as the time span is long enough to think well beyond silicon-based structures and muse in the rather diverse directions of quantum physics and biological substrates for computation. At that point I find myself with few insights to offer in that space. However, thankfully this workshop was a more focused brief on "the nature and requirements of computer communications networks that will be needed by society 50 years from now". I guess that the supposition here is that society as we know it will survive largely unscathed, which in times of significant and fundamental societal change is always a dicey proposition, but putting that aside, let's look specifically at this question of the evolution of computer communications.

In addressing this question, I found myself wondering what we would've thought in response to this same question were it posed in 1971. When we look at the musings at the time, such as Kubrick's vision made a few years earlier in 1968 in the move "2001: A Space Odyssey", parts of the predictions about communications technologies seem quite prescient given the benefits of hindsight, while other aspects are just way off the mark. It illustrates the constant issue with such musing about the future: Predicting the future is easy. The tough bit is getting it right!

## Fifty Years On: 1921

But maybe we are not looking back far enough. What if we start this prediction exercise to ask the same question in the context of 1921, or 100 years ago. What might we have said about the public communications requirements in 1921 about our future needs 50 or even 100 years hence? At the time in 1921 the telephone was still a recent invention, and it was priced as a business tool rather than a consumer commodity. Even the telegram was expensive for everyday consumers, and the bulk of communications volume was the postage service. The nineteenth century penny postage system had changed much of the communications landscape for the world at the time, as letters had become accessible and affordable for many and assumed the role of the mainstream communications medium. What would a 50 year prediction of the future look like at that time? Clearly, the telephone was gaining momentum, and like electrification, the concept of an affordable, ubiquitous service for all was behind the effective monopoly that the US Congress granted to AT&T through the Kingsbury Commitment of 1913. Equally, by 1921 the concept of radio as a medium for communication was gathering momentum. It seems likely that prognostications as to the future communications needs over the ensuring 50 years would've been based around telephony and radio as key technologies. As indeed was the case in the ensuing decades.

What would such predictions have missed? I'd guess that the rise and fall of the fax would've been missed, and perhaps the massive obsession with television would also have been missed. Considering the enormous costs in deploying the telephone network and the scale of technology involved, the concept of making and electronic facsimile of a movie and transmitting it as a radio broadcast would seem to involve a huge technology shift that was unlikely to occur within in 50 years. Would a prediction of 1921 miss the rise of computers and the emergence of digital environments? Again, that's likely.

## Fifty Years On: 1971

Moving forward by 50 years, what if we had posed this same question in 1971? The computing environment was being transformed yet again at that time. The monolithic "mainframe" computing environment was being challenged by so-called "mini-computers" The notion of a small number of shared computing utilities was based on the inordinate cost of these devices, the esoteric use cases, and the need to defray these costs over multiple users and uses. For all the other information processing tasks we had various forms of clerical labour and filing clerks! Mini-computing challenged this concept of the highly expensive mainframe computer used for only the most esoteric and detailed tasks and not only brought down the cost of computing but changed the model of access. These were smaller scale device that could be used for a single purple or by a single user. It was no accident that Unix, a single user operating system platform used on a PDP-7 mini-computer came out of Bell Labs at that time (equally, it was no accident that Unix was a deliberate play on the term Multics, a time-sharing operating system for multiple concurrent users) So I guess that was evident at the time was that computers would continue to push further in the market space by building computers with a smaller physical form factor, yet with sufficient capability to perform useful work. What would've been harder to predict at that time was the rise of computers as a consumer product. The early offerings, such as the IMSAI 8080 or the Altair 8800 still looked like scientific computers. Over in the consumer space our collective fascination was still absorbed by pocket calculators, and their efforts to introduce a programming capability into the calculator. It was not obvious at the time that the pocket calculator market would enjoy only a fleeting moment in the mainstream, and the rather clunky Apple II was the true progenitor of the evolution of the computer industry.

At this time, 1971, it was time to also think about the needs of computer-based communication in a more focussed fashion. The telephone industry was in the process of undertaking revolutionary transformation of its internal technology, moving from frequency division multiplexing to digitisation and time division multiplexing. This offered a dramatic shift in the cost efficiency of telephony infrastructure. When we thought about the needs of computer communications, we had a split vision at the time. Local networks that connected peripheral devices to a common mainframe central computer were being deployed and these networks used dedicated infrastructure. The concepts of connecting these devices together were not seen as forming a market that was anywhere near the scale and value of telephony and therefore it was likely that computers would continue to ride across existing telephone infrastructure for the foreseeable future. At best it was thought that these computers could potentially interface into the telephone network at the point of the telephone network's digital switching infrastructure. From this came the envisaged paradigm of computer communications that mimicked telephone transactions with dynamic virtual circuits, such as the X.25 packet switched networks and the later promotion of ISDN as the telephone industry's twisted and perverted collective vision of what consumer "broadband" was mean to be. This created a split vision for computer communications, with the local networks advancing along a path of "always on, always connected" model of computer communications using dedicated transmission infrastructure, and longer distance networks straddling the capabilities of the telephone network with a model of discrete self-contained transactions as the driving paradigm using a shared transmission infrastructure.

What happened since 1971 to shape the world of today? Firstly, Moore's Law has been truly prodigious over these 50 years. In the 1980s the network was merely the transmission fabric for computers and the unit of this transmission was the packet. The network itself did very little, and most of the functionality was embedded in these multi-use mainframe computers. However, in the 1990's the momentum behind computers as a consumer product not only gathered pace but overwhelmed the industry, and the personal computer became a mandatory piece of office equipment for every workstation and increasingly for every home as well. But these dedicated devices were computers that were switched on at the start of the workday and switched off when the user went home. They were not the multi-purpose always on, always at work, mini models of the mainframe computers that they were displacing, but were more in the line of a smart peripheral device. At the hub of many computing environment of the 1990's was still a common shared storage and large-scale information processing resource. In the computing world we were making the distinction between the mainframe and the constellation of personal computers that surrounded them. Computer communications networks also made this distinction, and unlike the telephone networks that viewed every subscriber in the same terms (it was essentially a "true" peer-to-peer" network), computer networks started to think about a network architecture that made a fundamental distinction between "clients" and "servers". Computer networks started to amalgamate some of the essential services of a network, such as a common name service, and a routing system, into this enlarged concept of the network, while "clients" were consumers of the services provided by the network. In a sense the 1990's was a transformation of the computer network from the paradigm of telephony to the paradigm of broadcast television.

However, this change in the model of networking to client / server systems also created a more fundamental set of challenges in the networking environment. In the vertically bundled world of telephony the capacity of the network was largely determined by the deployment of telephone handsets, and therefore network provisioning was a deterministic process completely under the control of the telephone network operator. In the unbundled world of the emerging client/server model of the Internet of the 1990's the capacity requirements of the network were determined by the actions of the consumer market, and the coupling of consumer demand and network service became a function of the Internet market itself. This meant that by the 2000's there was a scramble to scale up the services provided within the server side of the network. The rapacious demands of all those devices being purchased by consumers were not matched by a commensurate level of investment in scaling the service infrastructure and the capacity of the connecting network. The pricing signals did not exist and the rise of "flat rate" access tariffs for network services exacerbated the issue. More consumer demand was not accompanied by more revenue which, in turn, meant that more infrastructure was funded by increasing the debt levels of the service and infrastructure provider. We had shifted the parameters of the communications infrastructure away from that of a tightly coupled economy where growth in use patterns translated directly into additional revue for infrastructure providers that, in turn, provided capital for more infrastructure to be built. In this new uncoupled economic model, only more users generated more revenue, and the escalating level of use could only be funded with the buildout of more infrastructure by the continued entry of additional, presumably low usage intensity, new users. If this sounds a lot like a huge pyramid scheme, you'd be right. That was the ISP industry of the late 1990's!

This environment created a feedback loop that amplified demand for service infrastructure, and it wasn't only the financial models that were under stress. The growth was such that the technology models were also under stress. Popular services hosted on a single platform were totally overwhelmed, and the network infrastructure that connected these services was also totally overwhelmed. The solution was to change the technology of service infrastructure and we started to make use of server farms and data centres, exchanges and gateways, and the hierarchical structuring of service providers into 'tiers'. We experimented once more with virtual circuits in the form of MPLS and VPNs and other related forms of network partitioning, and because these efforts to pace the capacity of the service realm tended to lag the demand from the client population we experimented with various forms of "quality of service" to perform selective rationing of those network resources that were under contention.

Perhaps the most fundamental change by the 2000's was the emergence of content distribution networks. Rather than bringing back all the clients to a single service delivery point (I recall Microsoft trying to service all online updates to Windows from their server farm located in Seattle, which was a challenge in both computing and communications terms) we turned to the model of replicating the service closer to the service's clients. In this way the client demand was expressed only within the access networks, while the network's interior was used to feed the updates to the edge service closer to the Internet had discovered edge-based distribution mechanisms that bought the service closer to the user, rather than the previous communications model that bought to the user to the service.

And this was just in time because with the advent of Apple's iPhone in 2007 a massive shift in the demand curve took place. The industry was forced to confront an increase in demand that appeared to be three to four orders of magnitude larger than that of the tethered personal computer. Kilobits per second just didn't

do it. Customers wanted multiple megabits to complete the immersive environment that was created on their mobile devices.

The last 50 years has seen an evolution in networking infrastructure. We've taken the packet-focussed Ethernet model and pushed it into high-speed long-distance infrastructure. We haven't constructed SDH circuit fabric for decades, and these days the packet switches of the Internet connect directly to the transmission fabric. Yet in all these transitions we still operating these packets using the Internet Protocol. Why and how has this happened? For me, the true genius of the Internet Protocol was to separate the application and content service environment from the characteristics of the underlying transmission fabric. Each time we invented a new transmission technology we could just map the Internet Protocol into it, and then allow the entire installed base of IP-capable devices to use this new transmission technology seamlessly. From point-to-point serial lines to common bus Ethernet systems to ring systems such as FDDI and DQDB and radio systems, each time we've been able to quickly integrate these technologies at the IP level with no change to the application or service environment. This has not only preserved the value of the investment in the Internet across successive generations of communications technologies but increased its value in line with every expansion of the Internet's use and users.

## Fifty Years On: 2021

This now allows us, at last, to look at the next 50 years in communications technologies. Now 50 years, as we have seen, is a long time in some ways, but in many other ways it's not that long. The transformations that occur across multiple centuries often shed every trace of the former state and every aspect of the "new" environment is completely novel. But I don't think that this has been the case for a 50-year prediction. Much of today's world was conceivable in 1971, or earlier. The transformation of mobile telephones into these "smart" devices was a clear trend in the early 1970's. The transformation of computing with the progressive refinement of silicon processing to make processors with billions of individual gates with incredibly small power consumption and extremely high clock speed did not entail a fundamental re-think of what a computer was internally. The designs may have shrunk, but their logic has been largely constant. The point is that the seeds of the factors that became dominant some fifty years later were evident in the world of 1971, and the same line of thought asserts that the seeds of the dominant factors of our world 50 years hence in this communications environment are probably with us today. Perhaps the issue here is that these are not the only seeds of ideas that we have today, and the real challenge lies in distinguishing the significant from the merely distracting.

So maybe it's pointless to try and paint a detailed picture of the computer communications environment 50 years hence. But if we brush over the details, then we can look at the driving factors that will shape that future, and select these factors based on the driving factors that have shaped our current world.

### What's driving change today?

### Bigger

When we stopped operating vertically integrated providers and used market forces to loosely couple supply and demand we managed to unleash waves of dramatic escalation in demand. We viewed telephony communications using a language of multiples of kilobits per second. Today our units of the same conversations are measured not in megabits or gigabits per second, but terabits per second. For example, the Google Echo cable, announced in March 2021, linking the US with Singapore transmitting the Pacific will be constructed with 12 fibre pairs, each with a design capacity of 12Tb/s. Yes, that's an aggregate cable capacity of 144Tb/s. We are building larger capacity transmission systems using photonics amplifiers, wavelength multiplexing and phase/amplitude/polarisation modulation to extract significant improvements in cable capacity.

Moore's low may have been prodigious, but frankly the consumer device industry has scaled at a far more rapacious rate. We appear to have sold some 1.4 billion mobile Internet devices in 2020, and have done this volume, and higher every year since 2015. Massive volumes and massive capability fuels more immersive content and services. How do we serve content to all these clients? We have become expert at server and

content aggregation, and these days Content Distribution Networks are dedicated to servicing these clients at a scale and speed that matches the capacity of these last mile access networks.

### Faster

At the same time as we are building bigger networks, both in terms of the number of connected clients and in the volume of data moved by the network, we want this data to be pushed through the network at ever faster rates.

We have been deploying very high-capacity mobile edge networks and even 3G now looks unacceptably slow for many consumers. The industry is being pushed into deployment of 5G systems that can deliver data to an endpoint at a boasted peak speed of 20Gb/s. Now this may be a "downhill, wind on your back, no-one else around" measurement, but it belies a reasonable consumer expectation that these mobile networks can now deliver 100's of Mb/s to connected devices. In the wired world DSL technology, and more generically guiding a digital signal over a legacy telco twisted copper pair conductor is largely irrelevant these days and continue use of legacy copper infrastructure access technology only survives in those countries where the national communications infrastructure program was taken over by a hopelessly incompetent and corrupted political process, such as Australia. Elsewhere we are rewiring our wired environment with fibre, and here the language of a unit of wired service is moving away from megabits to gigabits

But speed is not just the speed of the transmission system but the speed of the transition itself. Here, the immutable laws of physics come into play and there is an unavoidable signal propagation delay between sender and receiver. If "faster" is more than brute force volume but also "responsiveness" of the system to the client, then we want it both. We want both low latency and high capacity, and the only way we can achieve this is to reduce the "packet miles" for every transaction. If we serve content and services from the edge, then the unavoidable latency between the two parties drops dramatically. The system becomes more "responsive" because the protocol c conversation is faster.

But it's not just moving services closer to clients that makes a faster network. We've been studying the at times complex protocol dance between client and the network to transform a "click" to a visible response. We working to increase the efficiency of the protocols to generate a transaction outcome with a smaller number of exchanges between client and server. That translates to a more responsive network that feels faster to use.

### Better

This is a more abstract quality, but if "better" means "more trustworthy" and "better privacy" then it appears that we are making headway at last! The use of HTTPS, or encrypted content sessions is close to ubiquitous in today's web service environment. We've been working on sealing ups the last open porthole in TLS by using encrypted Server Name Indication in the Client Hello message. We are even taking this a step further with the approaches proposed in Oblivious DNS and Oblivious HTTPS to isolate the combination of the identity of the client and the transaction being performed so that no party on the network, not even the server itself, has a priori knowledge of this coupling of identity and transaction.

The content, application and platform sectors have all taken up the privacy agenda with enthusiasm, and the question of the extent to which networks are implicitly trustable or not really does not matter anymore. This question of trust includes the payload, the transaction metadata, such as DNS queries, and even the control parameters of the transport protocol. All network infrastructure is regarded as untrusted!

I suspect that this is an irrevocable step and the previous levels of implicit trust between services, applications, and content and the underlying platform and network frameworks are gone forever. Once it was demonstrated that this level of trust was being abused in all kinds of insidious ways then the applications and service environment responded by taking all necessary steps to seal over every point of potential exposure and data leakage. There's no coming back from this stance. The concept of internal paranoia across the levels of the protocol stack, where each level of the stack exposes only the functionally minimal set of items of information to the other layers that are required to complete the requested transaction and protects everything else, is now firmly entrenched in the operating model of network design and operation.

### Cheaper

We appear to be transitioning into an environment of abundant communications and computing capability. At the same time these systems have significant economies of scale. For example, the shift in transmission systems to improve the carriage capacity of a cable system a millionfold has not resulted in a millionfold increase in the price of the cable system, and in some cases the capital and operating cost of the larger system has in fact declined over the years. The result is that the cost per bit per unit of distance has plummeted as a result.

This abundance has also led to a decline in per-transaction tariffs. While it was feasible to charge a penny for a letter to be passed into the penny post, or change per minute for a phone call, the unit cost of a network transaction is generally so small that it is infeasible to generate a cost-based transactional tariff model of digital services.

It goes further than just the reduction in cost, however. Some of these services are funded indirectly and to the consumer they operate without. For example, a search on Google's search engine happens without any user tariff. It's free to the user. Obviously, this service is indirectly funded through advertising revenue. This advertising revenue is made because Google has assembled a rich profile of users and sells this profile information to advertisers through their management of advertising campaigns. Interestingly, if I, as a user, tried to sell my own individual profile to advertisers, the exercise would fail. Individually, I'm not a viable market, but when I am aggregated together with a few billion or so of my fellow Internet users collectively we represent a very valuable market. So valuable that it funds the search system and still makes money. It can be argued that much of the service environment is funded by service providers capitalising a collective asset that is infeasible to capitalise individually. The outcome is transformational in so far as a former luxury service that was accessible to just a privileged few who could assemble a team of dedicated researchers has been transformed into an affordable mass-market commodity service that is available to all.

### Bigger, Faster, Better and Cheaper

It was often said that it was impossible to meet all these objectives at once. Somehow the digital service platform has been able to deliver across all of these parameters. How has it done this?

The way in which we build service platforms to meet ever-larger load and ever-declining cost parameters is not just by building bigger networks, but by changing the way in which clients access these services. We've largely stopped pushing content and transactions all the way across a network and instead we serve from the edge.

Serving for the edge slashes packet miles which in turn slashes network costs and lifts the responsiveness factor which lifts speed. These seem to be the driving factors for the next few decades.

This is not a more ornate, more functional, more "intelligent" network. This is definitely not "New IP" or anything close. In fact, I would argue that these factors represent the complete antithesis of these attributes! This continues along the same theme as that of the Internet as compared to the predecessor telephone network. By pushing functions out of the network, we strip out common cost elements and push them out to the connected devices, where the computing industry is clearly responding with more capable devices that can readily undertake such functions. By pushing services out to the edge of the network we further marginalise the role of a common shared network in providing digital services.

For me, these factors appear to be the dominant factors that that will drive the next 50 years of evolution in computer communications and digital services.

### Some Issues to Think About

If these are the important drivers that have got us to where we are today, then it seems completely reasonable to believe that they will continue to exert pressure on the future directions of the digital environment. It is highly likely that we will continue to realise to further improvements in the communications infrastructure, in terms of constructing bigger, faster, better and cheaper networking systems.

But I suspect that there are several other seeds in today's environment that tease out some interesting questions in the coming years. Let's look at a just a few of these related topics.

### Addresses

The first is the nature of "addressing". The Internet borrowed from the telephone network where every endpoint was uniquely numbered and identified. The issue we experienced with the Internet was that initial estimate of 4 billion unique addresses proved to be inadequate and we were forced to embark on a protracted, costly and as yet still incomplete transition to a new version of the protocol that has a far larger address space. However, more than twenty years after we embarked on this technology transition the process is still incomplete. The fact that we have still not completed this transition, nor indeed have we got anywhere close to completion, and the observation that there is still no common sense of urgency about this transition raises the obvious question of why we think this transition is important to complete in any case.

Is this concept of universal endpoint identification via protocol addressing nothing more than a 1980's networking concept whose time has come and gone? Is this absolute addressing of endpoints a property of a network infrastructure that just can't keep up with the demands for ever larger networks? Should we dispense with absolute endpoint addressing, and continue down the track we've been using with network address translation and protocols that are address agile, such as QUIC, and treat these addresses as ephemeral session tokens that allows the network to disambiguate traffic between concurrent active sessions and not much else? We still need to uniquely identify services and service delivery points, bit it that necessarily a network function or should this be an attribute of the service application itself?

### Names

We have been overloading the name space to compensate for the shortfalls of the IPv4 address space. We used to think of the DNS in the context of the Internet as an alias for addresses to allow human use of the Internet. We seem to be loading additional information into the name system, and rather than a simple mapping of a name to an address we now want to use the DNS as a service rendezvous function. Not only can the DNS tell us the IP address we should use to send IP packets to a named service, but the DNS can also tell us what transport protocol to use and what service encryption credentials should be used.

The change is that the DNS is changing from being a common attribute of the network into a collection of service specific functions that can tell each client how to access the named service. To rephrase this in more generic terms, are "names" a common attribute of the network's infrastructure? Or are they dynamic and relative attributes of a service that permits clients to rendezvous with the service?

### References

This second issue of names leads on to the question of referential frameworks. How does a client identify a service? How can a client pass this identity to others as a reference? The critical semantic distinction is that for an individual client it is sufficient to identify a service in terms that are self-referential, but to use it as a reference independent of the client's context requires more context. (For example, I might identify my local corner shop by saying "turn left out of my house and proceed to the next intersection" but this is not a useful algorithm for you to use from your location.)

In a network of densely replicated service delivery points, there is an additional consideration. How can a client rendezvous with the "best" instance of a service delivery point? Is it up to the client to work out what is the "best" service point from all these alternative instances? Or should the network make this call? Should the service itself make this decision? Should the reference be absolute, but the resolution of that reference lead to a set of parameters to perform a service transaction provide an outcome relative to the client? Who performs such a resolution function? The network? Or the service?

### **Two-Party Transactions**

Finally, if we are questioning the basics of the name and address infrastructure of networking, what about the nature of transactions? Does it still make sense to think of the model of computer transactions in terms of a two-party, synchronous, exchange of data? In what contexts would multi-party transactions make sense? Two-party synchronised transactions work for many forms of human communications, but are they necessarily the best template for computer communication?

Obviously, there are no clear answers to these questions at present, but I suspect that these basic questions about the role of these elements of names, addresses and references are a core part of the evolution of the architecture of computer networks.

## Longer Term Trends

Where is all this going? It seems that in order to build networks that are effective in terms of bigger, faster, better and cheaper, then we seem to be achieving this by passing more and more of the network's functions out of the interior of the network and shifting them reside in a duplicated manner closer to all of the edges of the network, residing in a set of locations that are adjacent to all clients. We appear to have transformed transmission and computation from a scarce and expensive resource into an abundant and cheap commodity and this implies that sharing common pooled resources is no longer an essential part of the picture of service delivery. We are amassing so much transmission, computation, and storage that we are no longer motivated to use a common network to carry clients to distant service delivery points. Instead, we are shifting these services towards the client using just-in-case pre-provisioning for the service and the internal network is now used to support this service replication to synchronise all these edge service delivery points.

This, in turn, heralds a more significant change where the application is no longer a window to a remotely operated service, but the application is becoming the service itself. I guess that if the desire to position a service ever closer to the client ends in the question of why provision the service at a network point adjacent to the client if we could directly provision the service on the client's device?

This leads to a final couple of questions that I'd like to pose about the next 50 years in the communications realm.

### At the end of all this, will shared networks still matter?

What we are observing is a trend to strip out cost and function from the network and instead load them onto the end device. This has given us lower costs, higher speed and far greater agility in service provision. So, when do we stop? What happens when we push everything onto the edge device? What's left of the network and its role?

### What is "the Internet"?

And finally, what defines "the Internet" in all this?

We used to claim that "the Internet" was a common network, a common protocol, and a common address pool. Any connected device could send a IP packet to any other connected device. That was the Internet. If you used addresses from the Internet's address pool, then you were a part of the Internet. This common address pool essentially defined what was the Internet.

These days that's just not the case and as we continue to fracture the network, fracture the protocol framework, fracture the address space, and even fracture the name space, what's left to define "the Internet"? Perhaps all that will be left of "the Internet" as a unifying concept is a somewhat amorphic characterisation of disparate collection of services that share common referential mechanisms.

However, there is one thing I would like to see over the next 50 years that has been a feature of the past 50 years. It's been a wild ride. We've successfully challenged what we understood about the capabilities of this technology time and time again, and along the way performed some amazing technical feats. I would like to see us do no less than that over the coming 50 years!

## 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