August 2022
Geoff Huston

Bigger, Faster, Better (and Cheaper!)

Let's take a second to look back some 50 years to the world of 1972, and the technology and telecommunications environment at that time. The world of 1972 was one populated by a relatively small collection of massive (and eye-wateringly expensive) mainframe computers that were tended by a set of computer operators working around the clock and directed by specialised programmers, trained in the obscure symbol set used by the job control systems on these computers. In the average household of that time, the most complex item of consumer technology was the television set. And it was an analogue device. Our clocks still ran on clockwork. Yet, changes were underway. The enthralling technology achievements of manned space flight had captured not only the imagination of an entire generation but given us a glimpse into the power and utility of technology. Collectively we became obsessed with technology.

In computing, Moore's Law[1] has been truly prodigious over these 50 years. Mainframe computers became more capable, faster, and progressively cheaper. At the same time, we were building computers that were not necessarily faster, nor more capable, but were smaller and cheaper. These *minicomputers* were progressively refined in size, cost, and ease of use to the point where they became a personal consumer product. At the same time computer networks were also changing. They were originally used to extend the reach of the computer by using remotely located peripheral devices. First there were card readers and printers, then terminals. However, this was an unstable situation and as the investment in the network and peripherals increased relative to the cost of the mainframe computer it was no longer an option to throw away everything each time the mainframe computer supplier was changed. With the introduction of minicomputers into the mix then it was no longer all about the mainframe computer. The network was used to interconnect a collection of computers and peripherals. We wanted open standards to drive these networks and they interaction with peripherals so that the network assumed a more central role in the computing environment.

In the 1990's the momentum of this market for computers as a consumer product had an impact on the architecture of the technology landscape. We were making the distinction between the mainframe server and the constellation of personal computer clients 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 c*lients* 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 a paradigm that was a lot closer to 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. Just as computers were now consumer devices, computer communications were now entering the realm of the public utility service space, challenging the incumbency of the telephone network. While the telephone world would've like to treat

[1] Moore, Gorden E, "Cramming more components onto integrated circuits", intel.com. Electronics Magazine, 1965. https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf, retrieved April 1, 2020.

this as just another application that was integrated into the existing telephone environment in the same manner as the fax systems were absorbed into the telephone space in the 1970's and 1980's, the computer world had an entirely different service model in mind.

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 of increasing intensity of use did not exist and the rise of "flat rate" access tariffs for network services exacerbated this 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 acute 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[2].

Perhaps the most fundamental change by the late 2000's was the emergence of content distribution networks. Rather than bringing back all the clients to a single service delivery point[3] 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 centres. In effect 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.

---

[2] Ferguson, Paul, and Huston, Geoff, "Quality of Service: Delivering Qos on the Internet and in Corporate Networks", John Wiley & Sons, February 1998.

[3] At that time Microsoft was trying to service all online updates to their Windows product from their server farm located in Seattle, Washington, which was a significant challenge in both computing and communications terms.

The last 50 years has also seen a profound evolution in networking infrastructure. We've taken the packet-focussed network model used by Ethernet local networks and pushed it into high-speed long-distance infrastructure. We haven't constructed additional SDH-managed circuit capacity for decades, and these days the packet switches of the Internet connect directly to the transmission fabric. Yet in all these transitions we are still operating these packets using the Internet Protocol.

Why and how has this happened? 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 onto 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 Internet-based technologies across successive generations of communications technologies but increased its value in line with every expansion of the Internet's use and users.

This now allows us 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 it's not clear that this has been the case for a 50-year technology prediction. The case can be made that much of today's technology 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 integrated single chip 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 and design 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 within our current world 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 communications service 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 across the seafloor of the Pacific Ocean will be constructed with 12 fibre pairs, each with a design capacity of 12Tb/s. That's an aggregate cable capacity of 144Tb/s. Google's Dunant cable system delivers an aggregate capacity of 250Tbs across the Atlantic, which will be complemented by the 352Tbps Grace Hopper cable system. We are throwing everything we can at this to build ever-larger capacity transmission systems with photonic amplifiers, wavelength multiplexing and incorporating phase/amplitude/polarisation modulation as well as pushing digital signal processing to extreme levels to extract significant improvements in cable capacity.

Moore's Law may have been prodigious, but frankly the numbers behind 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 achieved this consumption 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.

When we consider "bigger" it's not just human use of the network that's a critical consideration. This packet network is a computer network, and the usage realms include the emerging world of the so-called *Internet of Things*. When we look at this world, we have two questions which appear to be unanswerable, at least with any precision. How many "things" are using the Internet today? How many will be using the Internet tomorrow?

There are various estimates as to the device population of the Internet today[4]. There is some consensus around a figure of between 20 and 50 billion devices, but these rely on various estimates rather than more robust analytical measurements. Production volumes for microprocessors run into billions of units per year, so the expectations of growth in the sector are all extremely uncertain but generally incredibly high. Five-year growth projections in this market segment start at around a total of 50B devices and just get higher and higher.

Behind this is the observation that in growing bigger the Internet is no longer tracking the population of humans and the level of human use. The growth of the Internet is no longer bounded by human population growth, nor the number of hours in the day when humans are awake. We're changing this network to serve a collection of computer devices whose use is based on a model of abundance. Abundant processing capacity, abundant storage, and abundant network capacity. We really don't understand what "bigger" truly means in terms of demands. The best we can do is what we've been doing over the past couple of decades: deploy capital, expertise, and resources as fast as these inputs can be assembled. We still seem to be in the phase of trying to keep up with demand, and however big we build this network, the use model has proved more than capable of saturating it.

Faster

At the same time as we are building bigger networks, both in terms of the number of connected devices and 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 claimed peak speed of 10Gb/s[5]. Now this may be a "downhill, wind at your back, no-one else around" extreme measurement, but it belies a reasonable consumer expectation that these mobile networks can now deliver 100's of megabits per second to connected devices. In the wired world DSL technology, and the more generic form of guided digital signal propagation over a legacy telco twisted copper pair conductor, is largely irrelevant these days and continued use of legacy copper infrastructure access technology is waning. We are rewiring our wired environment with fibre optics, and here the language we use to describe a unit of capacity of these wired services are moving away from megabits to gigabits per second.

But speed is not just the speed of the transmission system but the speed of the transmission 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 both low latency and high capacity, and the only way we can achieve this is to reduce the distance of 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 conversation is faster.

[4] https://techjury.net/blog/how-many-iot-devices-are-there/#gref

[5] https://www.tomsguide.com/features/5g-vs-4g

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.

The elements of a faster network are:

- Increasing bandwidth capacity in "last mile" access networks.
- Pushing all forms of content and servicer delivery into highly replicated content distribution networks.
- Increasing the density of content distribution points to place the servers and service closer to the user.
- Pre-provision content so that the entire service transaction occurs over the last mile access network.
- Engineer the application environment to be more responsive.
- Improve the protocol performance of transport protocols.

What we are trying to do is remove the long-haul transit element from network transactions. Also, by anticipating demands and pre-provisioning content in content data centre delivery points, we can eliminate the inevitable capacity choke points associated with distance. In networking terms "closer" is essential for "faster". It's not all that "faster" needs, but without close proximity between sender and receiver "faster" is simply not possible.

Better

This is a more abstract quality, but if "better" means "more trustworthy" and "verifiable authenticity" then it appears that we are finally making headway in this most challenging task.

The use of HTTPS, or encrypted and authenticated content sessions, is close to ubiquitous in today's web service environment. We've been working on sealing up the last open peephole in the TLS protocol by using encrypted Server Name Indication in the TLS Client Hello message in TLS 1.3[6]. We are even taking this a step further with the approaches proposed in Oblivious DNS[7] and Oblivious HTTP[8], where we can isolate any other party, even the service operator, from knowledge of the combination of the identity of the client and the transaction being performed. This would imply that nobody other than the client has a priori knowledge of this coupling of identity and transaction.

The content, application and platform sectors have all taken up selected aspects of the privacy and authenticity agenda with enthusiasm, and the question of the extent to which networks are implicitly trustable or not really does not matter anymore. If the network has no ability to obtain privileged information in the first place, then the question as to whether the network can be trusted with this information is no longer relevant. This question of trust includes the payload, the transaction metadata, such as DNS queries, and even the control parameters of the transport protocol. In today's networks we

---

[6] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018, <https://www.rfc-editor.org/info/rfc8446>.

[7] Schmitt, Paul, et al. "Oblivious DNS: Practical privacy for DNS queries." Proceedings on Privacy Enhancing Technologies 2019.2 (2019): 228-244. (https://odns.cs.princeton.edu/pdf/pets.pdf)

[8] Thomson, Martin, "Oblivious HTTP" work in progress, Internet draft, February 2022. https://www.ietf.org/archive/id/draft-thomson-http-oblivious-01.html

deliver a "better" outcome to users and the services they choose to use by taking the stance that all network infrastructure is regarded as untrustable!

It is likely 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 now occurs 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 and in application design.

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 by 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.

At the same time, we've shrunk the network, so that increasingly service transactions are local. As we've already observed, the rise of the CDN model has changed the Internet. By pre-provisioning content close to every edge, the subsequent on-demand transaction from server to client occurs over a small distance. Not only are smaller distances faster for service transactions, but smaller distances are cheaper to build and operate. Smaller distances consumer less power and have super signal to noise characteristics. This increase in transmission efficiency also implies lower cost.

It goes further than just the reduction in cost, however. Some of these services are funded indirectly and to the consumer they operate without any visible cost imposed on the user. 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 possible because Google has assembled a rich profile of its users and sells this profile information to advertisers through their management of advertising campaigns. If an individual user were to attempt top sell their individual profile to advertisers, the exercise would fail. When this is aggregated into a large collection of profiles, this collection represents a very valuable asset.

It can be argued that much of the Internet's 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 a mass-market commodity service that is available to all. It's not just available at an affordable rate. In many cases it's affordable as in free of any charges at all.

Bigger, Faster, Better and Cheaper

It was often said that in the communications industry it was impossible to meet all these objectives at once. Somehow the Internet's 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 not a baroquely ornamented "New IP"[9] network, or anything remotely close. These factors represent the complete antithesis of these conventional attributes of a so-called 'smater' 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.

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

Longer Term Trends

Where is all this going? It seems that 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 replicated 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. The desire here to position a service ever closer to the client ends in the question of why should we 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 two further fundamental questions 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?

More critically, the entire concept of virtual circuits, packets and common networks was a recognition that shared communications infrastructure was more efficient than each application and each client having dedicated access to their own infrastructure. The distinction between circuits and packets was about how the common resource was shared, but neither fundamentally questioned whether sharing was needed (or not).

[9] Internet Society, "Huawei's "New IP" Proposal – Frequently Asked Questions", February 2022. (web page)

However, in many parts of today's network infrastructure shared infrastructure is looking somewhat old-fashioned. In the submarine cable industry, the largest of the content enterprises is dispensing with a shared infrastructure model and installing fully-owned cables. Data centres are another case in point, where the very largest of the content distribution enterprises operates fully dedicated infrastructure.

It's reasonable to ask where this is heading. Does sharing still matter? Will it matter? Or is sharing a response to particular circumstances but not others? The answer to this question is by no means clear, but the original networking axiom, that networks are a way of sharing a common transmission resource is a lot less obvious now than it was some 50 years ago.

What of the Internet?

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 an 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