The Once And Future Internet

James McCauley, Arvind Krishnamurthy, Tejas Narechania, Aurojit Panda, Scott Shenker

The Internet is the centerpiece of the world's communication infrastructure, and it touches almost all aspects of our lives. Thus, if there is any technology that should be designed in a socially conscious manner, it is the Internet. This short paper discusses what the Internet was, what it is now, and what it could become, all from the perspective of (quoting from this workshop's CfP) "potential ethical concerns arising from system and architecture design choices." Given the limitations on length, this paper focuses on identifying a few crucial, but often overlooked, social concerns related to the architectural design choices for the Internet, and leaves the details of our proposed solution to longer descriptions available elsewhere [1].

The Early Internet (up to roughly 2010): When people talk about the brilliance of the Internet's design, they are typically referring to its technical architecture. The elegantly simple layered approach has enabled the Internet to withstand radical changes in numerical scale, geographic scope, and communication speed, all while supporting an ever-increasing range of applications.

While the technical design was the result of much care and debate, the Internet's economic architecture was more of an evolutionary accident. A conceptual (but not historically precise) description of this evolution contains two steps. First there was the recognition that the Internet could not be managed as a single entity, but had to be broken into separate Autonomous Systems or domains. This required both a technical design for how to route packets between those domains (what we now know as BGP) and a set of economic agreements between these domains for how to pay for services rendered, which were as follows: (i) users paid their access ISP, (ii) a class of transit ISPs arose that would carry traffic between ISPs for a fee, and (iii) the financial arrangements between two connected ISPs was either customer-provider (one would pay the other for the connection) or settlement-free peering (where neither party paid). The routing algorithm BGP is only be guaranteed to be stable if the economic arrangements were "valley-free" [2] but, quite fortuitously, that appears to be the typical case.

Sometime after these arrangements were in place, the need for "network neutrality" became apparent because some carriers were blocking or deprioritizing various classes of traffic (*e.g.,* VPN and VOIP). The imposition of network neutrality allowed these and other applications to flourish without further interference. While network neutrality regulations are no longer in effect in the US, they remain in place in the EU.

These economics arrangements gave the early Internet three crucial properties:

(1) *Interconnection*: The global delivery provided by the Internet resulted from the interconnection of many ISPs, most having very limited geographic scope. The ability to interconnect made it easier to enter the ISP market because ISPs did not have to be a global carrier to provide customers access to global delivery. This in turn meant it was easier to expand Internet coverage, because ISPs could deploy in underserved areas and then interconnect with a close-by transit ISP without having to build a global infrastructure.

(2) *Any-to-any delivery*: ISPs were willing to provide service to all paying customers, and route traffic to/from any of them. The resulting interconnected infrastructure provided connectivity between any pair of customers.

(3) *Neutrality*: The widespread-but-not-perfect adherence to network neutrality in many parts of the world allowed a wide variety of applications to flourish without fear of restrictions from ISPs based on their competitive offerings. To be clear, governmental interference in some regions was (and indeed remains) a major problem, but it did not involve networks stifling users who were offering services that competed with those of the carriers.

These three properties – interconnection, any-to-any delivery, and neutrality – were crucial in making the early Internet a platform on which the emerging cloud and content providers could flourish. The telephony infrastructure also had these three properties, enabling it to support the deployment of the Internet itself in a previous generation. Why are these three properties so crucial? As we discuss in more detail later, interconnection lowers the barriers for entering the market for providing service, increasing both coverage and competition. The latter two properties make the infrastructure a good platform for ongoing innovation in user-to-user services.

The Current Internet: Two recent developments make today's Internet quite different from what came before it.

In-network enhancements: The early Internet offered basic packet delivery between two endpoints, along with caching from CDNs. Now there is a growing number of in-network services – such as DDoS protection and zero-trust-network-access – that enhance the Internet's original service model with additional functionality. These functions are typically implemented in compute clusters positioned at the network edge.

Large private user-facing networks: Several hyperscaled cloud and content providers have built large private user-facing networks that reach many access ISPs, so traffic to/from these locations never has to use a transit ISP.

The infrastructures providing these enhancements (call them ESPs) are not interconnected, in the sense that they do not collaborate in order to jointly provide the required service to the union of their customers (which ISPs do). Thus, to achieve global coverage with an enhancement, a content provider must hire an ESP whose infrastructure is global, or (more painfully) stitch together coverage by hiring several ESPs whose union is global.

The private networks only carry traffic to/from the hyperscaler's backend datacenters, and they are not required to be neutral. In addition, they also implement several enhancements (such as load balancing and caching).

As a result of these private networks and the caching offered at the network edge, a large fraction of Internet traffic never passes through a transit ISP. Instead, traffic is either served by a nearby cache or goes directly between the client's access ISP and a hyperscaler's private network [3]. Thus, the dominant delivery paradigm of the current Internet is radically different from that of the early Internet: in this dominant paradigm, there is no role for transit networks, and the service involves enhancements that go beyond simple packet delivery.

This new paradigm offers significantly better service to clients, which is why the hyperscalers and ESPs have built out these large-scale infrastructures. While this might seem like a socially responsible development, it represents a dramatic turn away from the three crucial properties of: (i) interconnection (the ESPs do not interconnect), (ii) any-to-any delivery (private networks only carry traffic to/from the hyperscaler), and (iii) neutrality (the private networks need not be neutral).

To be clear, some traffic is still served by the traditional Internet model (using transit ISPs with no enhancements), but this is not the common case today. As a result, we are facing a future where our central communications infrastructure will have:

An increased digital divide: These private networks and enhancement infrastructures are focused on the most lucrative areas (primarily North America and Europe). Other areas will be more distant from these private networks and ESPs, and won't derive as much benefit.

Less competition: Deploying these large-scale infrastructures is expensive, and few companies can compete at this scale. For example, the leading CDN owns 70% of the market [4].

An inability to nurture what comes next: While the telephony infrastructure and the early Internet each allowed the creation of their successor, the current Internet – with its dominant paradigm not having any of the key properties of interconnection, any-to-any delivery, and neutrality – is not likely to do so. This is the most worrisome aspect of today's Internet; it may not provide the platform on which we can build what should come next.

The Future Internet: From a socially responsible perspective, what would we want from the future Internet? We think the answer is less about its performance or specific functions, and more about whether it can support interconnection, any-to-any delivery, and neutrality, but without eliminating its ability to support edge-based enhancements. This would retain the better performance of the current Internet, while restoring the properties that were crucial in the early Internet.

To this end, we are working on a design called the "InterEdge." The InterEdge leaves the current IP-level Internet untouched, but provides a way to interconnect the various edge-based enhancements, much like the original Internet connected various incompatible networks. More detailed description of, and arguments for, the InterEdge can be found in [1] and (hopefully) forthcoming technical publications. But here we describe the main conceptual challenge in creating the InterEdge.

Interconnection in the early Internet relied on (i) a technical standard (IP), (ii) a routing algorithm (BGP), and (iii) a set of financial arrangements (peering). For each enhancement that we want to interconnect – such as caching, or pub/sub delivery, or a mixnet – we will need the same components: a technical standard (which we envision will be an open-source implementation, not a written standard), a way of routing traffic between the enhancement providers so that each enhancement is executed at the appropriate edges, and a mechanism for exchanging payments among the enhancement providers. These are the technical challenges the InterEdge design addresses, which should be augmented by neutrality regulations.

References

[1] Marjory Blumenthal, Ramesh Govindan, Ethan Katz-Bassett, Arvind Krishnamurthy, James McCauley, Nick Merrill, Tejas Narechania, Aurojit Panda, and Scott Shenker. Can we save the public internet? *SIGCOMM Comput. Commun. Rev.*, 53(3):18–22, 2024.

[2] Lixin Gao and Jennifer Rexford. Stable internet routing without global coordination. *IEEE/ACM Trans. Netw.*, 9(6):681–692, 2001.

[3] Geoff Huston. The death of transit? Web, Oct 2016. https://blog.apnic.net/2016/10/28/the-death-of-transit/.

[4] Nick Merrill and Tejas Narechania. Inside the internet. *Duke Law Journal*, 73(35), 2023.