REPORT FOR APNIC AND LACNIC

STUDY ON THE INTERNET'S TECHNICAL SUCCESS FACTORS

Michael Kende, Amund Kvalbein, Julia Allford and David Abecassis

DECEMBER 2021

Disclaimer and acknowledgements

This report was commissioned and sponsored by APNIC and LACNIC, the regional Internet address registries for the Asia–Pacific and Latin America and Caribbean regions respectively, and written by Analysys Mason, a global consultancy specialising in telecoms, media and technology.

Analysys Mason is grateful for all inputs during the research and preparation of this report, including detailed discussions with individuals including Vint Cerf, Alissa Cooper, Steve Crocker, Paul Gampe, Geoff Huston, Olaf Kolkman, Danish Lakhani, Xing Li, Yoshinobu Matsuzaki, Ricardo Person, Onno Purbo, Alvaro Retana, Andrei Robachevsy, Steve Song, Michael Stanton and Dirk Wolter. The quotes presented throughout this report are attributed to individuals with their permission, and when they are anonymised they stem from these discussions.

The analysis contained in this document is the sole responsibility of Analysys Mason and does not necessarily reflect the views of APNIC, LACNIC or other contributors to the research.

Copyright © 2021. Analysys Mason has produced the information contained herein for APNIC and LACNIC ('the Clients'). The ownership, use and disclosure of this information are subject to the Commercial Terms contained in the contract between Analysys Mason and the Clients.

Analysys Mason Limited
North West Wing, Bush House
Aldwych
London WC2B 4PJ
UK
Tel: +44 (0)20 7395 9000

enquiries@analysysmason.com

www.analysysmason.com

Registered in England and
Wales No. 5177472

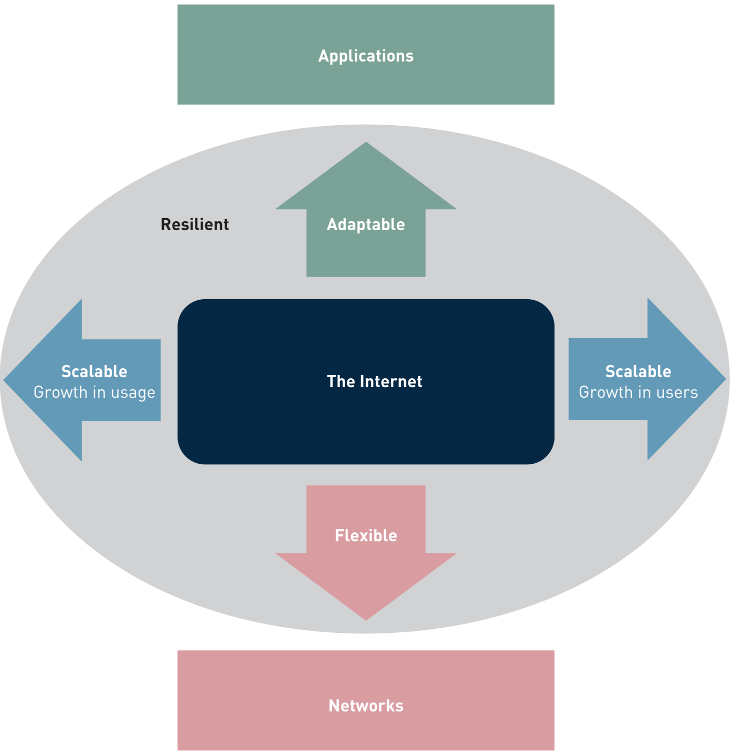
Contents

Executive summary

The global success of the Internet is often taken for granted today, based on almost any measurement – the number of users and the breadth of their geographical scope, and the level of usage and the depth of reliance on the Internet. This success has been built on protocols, standards, operational practices and communities developed over the past 50 years. Some of the foundations for this success were built into the Internet architecture from the beginning, while others have emerged as the Internet has evolved.

However, the success of the Internet should not be taken as given. Throughout the history of the Internet until the present day, questions have been raised about the future viability of the Internet. These questions often arise from an idealistic view of the Internet in the face of assorted threats or challenges and have led to proposed alternative network architectures and approaches. These may be motivated by genuine concerns about the suitability of the Internet or a particular protocol, by commercial, social or socioeconomic considerations, geopolitics, or a combination of these motivations. This report deliberately avoids assessing these motivations and is instead grounded in an objective analysis of the Internet as it exists today, based on highlighting technical success factors and taking a realistic approach to addressing challenges.

In this report, we argue that the technical success of the Internet is manifested through four **dimensions of success.** The Internet has successfully *scaled* to the increased demand from new users and usage, it has been *flexible* to new underlying network technologies, it has adapted to new applications, and the whole has been *resilient* to shocks and changes. We provide evidence of how the Internet is scalable, flexible, adaptable and resilient, and discuss the technical properties of the Internet that underlie these four dimensions of success. The four dimensions of success are illustrated in the figure below.



The early development of the Internet was based on *openness, simplicity and decentralisation*. The adoption of these three **guiding ideals** resulted from conscious decisions taken by the early Internet developers, and they have shaped the technical, organisational and operational development of the Internet. These ideals were not general practice at the time, and they represented a change from the design of the existing dominant network of the day, the telephone network.

Three well-known **design principles** that sprung out of the guiding ideals have been central to the Internet's development and are useful concepts to understand the dimensions of success: these are the *layering, network-of-networks* and *end-to-end principles*. We explain how the variety of underlying network technologies that provide Internet access and carry traffic, the protocols that allow communication within and across the networks that constitute the Internet, the interconnection model that allows individual networks to form business relationships with other networks, the openness to new applications, and operational practices prevalent in service provider networks, have developed in light of these three design principles.

The main contribution of this report is to highlight the technical success of the Internet, by exploring the **four dimensions of success** we introduced above and relating them to the guiding ideals and the design principles. We show how key protocols, design principles, and operational practices have enabled the Internet to scale to its current size (and without obvious future constraint), how it has been flexible in the use of underlying network technologies, how the Internet has been able to adapt to the requirements of new applications, and how the Internet has shown resilience to attacks and sudden changes.

We also highlight that while we believe the design principles have been important to this success, they are not universal rules that are always abided by in practice. In fact, we argue that one key to the Internet's success has been the ability to bend, or even break, its own technical principles when needed, without compromising the dimensions of success. The intrinsic resilience of the Internet has allowed its continual operation even in the face of these unanticipated compromises.

The Internet continually scales, adapts to new applications, and remains flexible to new network technologies. Towards the end of this report, we discuss potential challenges to the Internet model. These can come from perceived technical limitations that may prevent the Internet from offering (for instance) guaranteed security and performance, or they can come from the interests of governments or large Internet companies that are transforming the Internet, but not in ways that impact the dimensions of success. While there are proposals for fundamental changes to the Internet, we conclude that a continuation of the constant evolution of the Internet can address the potential challenges, while continuing to enjoy success across the identified dimensions.

1 Foreword

Over the years we have been hearing critiques of the Internet by users, vendors, standard bodies and governments: Is it currently fit for purpose? Is the Internet secure? Can the Internet continue to evolve, or will it need to renew? These critiques have intensified recently, as cybersecurity threats have become more serious and the geopolitics of the Internet more divided. Some stakeholders have sought to answer these questions by justifying alternative proposals for new protocols and networking standards that compromise those factors that we consider important for the Internet to continue to be successful. Even the original designers of the Internet have asked these questions themselves and have expressed frustration about how much the Internet of today deviates from its early ideals.

This is the context in which APNIC and LACNIC decided to partner in a project to study the Internet's technical success factors, based on a strong belief that an assessment of such factors should be objective and not idealistic. After an open call for proposals, and a difficult evaluation of 14 qualifying contestants, we commissioned Analysys Mason to produce this study to help us refresh the narrative in which we explain the key technical factors that have contributed to the growth and evolution of the Internet in the past 50 years.

The study by Analysys Mason provides an innovative framework based on four "dimensions of success" which are identified: 1) scalability supporting the growth of the Internet; 2) flexibility in network technologies; 3) adaptability to new applications; 4) resilience in the face of shocks and changes. By describing the actual state of the Internet, according to its technical implementation in different geographies, economies and societies, the study explains the evolution of Internet standards and protocols, as well as its architecture-design and system structures, in relation to these four dimensions of success.

We like this framework because it describes the Internet as it is and not as it should be. We looked for a study about the Internet's actual implementation by choice of industry players, and other actors, with all efforts to disentangle the geopolitics of the Internet from an objective assessment of its success. We believe Analysys Mason have delivered in this regard, explaining the Internet's success using technical measures rather than subjective arguments or opinions regarding the Internet's functioning. The study also offers a prospective outlook of the technical factors that have proven successful over the years, and risks affecting their continuity or threats to their stability.

As Regional Internet Registries, APNIC and LACNIC serve a diverse community of network operators - those who literally build and run the Internet in our respective regions. The success of the Internet belongs to them, (and those in the rest of the world of course); and while undeniable, it is a success that we should not take for granted. We feel that by shedding light on it, we can better understand the complex reasons for that success.

We hope our members and community find value in this study. As we slowly begin to reconnect face-to-face, this study should equip us as well with an arsenal of fresh arguments about why the Internet has been technically successful and how to evolve it without compromising its scalability, flexibility, adaptability, or resilience.

Paul Wilson, Director General, APNIC and
Oscar Robles, CEO, LACNIC

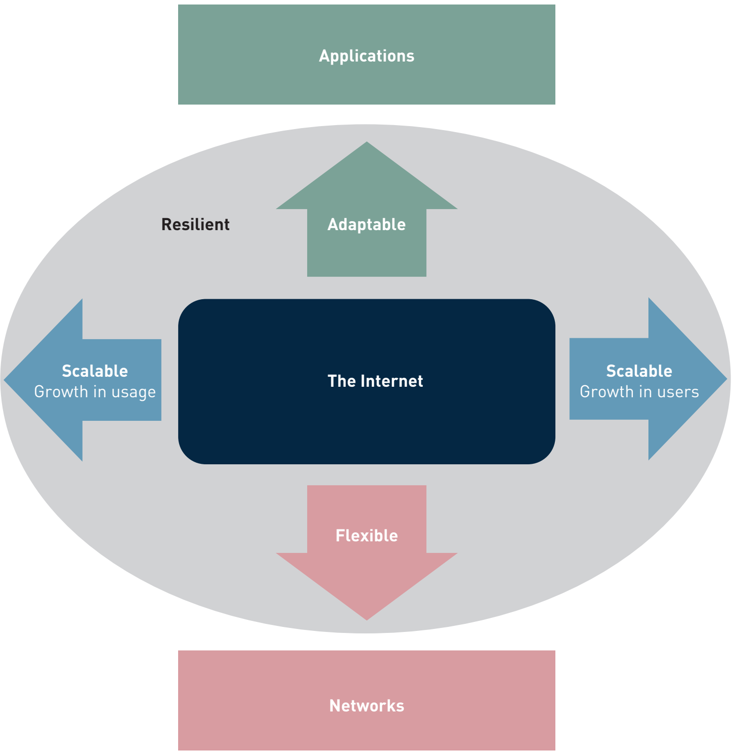
2 Introduction

From its inception among academics and researchers in the USA over 50 years ago, the Internet has grown to a global network approaching 4 billion users across all geographies, and has become deeply integrated in the functioning of modern society and economies. While work needs to be done to increase the availability and affordability of the Internet in order to close the digital divide, there is no disputing that the Internet has been and continues to be one of the most successful infrastructure systems of any kind ever developed. In this report, we argue that the success of the Internet can be described and understood through four dimensions of success:

- The Internet is scalable in its technical architecture and operational and business models, which has enabled it to grow quickly and with few imposed constraints in terms of the number of users and the usage per user, while average Internet speeds keep rising.
- The Internet is flexible to different types of underlying networks ranging from high-speed optical networks to ad-hoc wireless networks, each of them suited to different user requirements, geographies and socioeconomic characteristics of countries, regions and people.
- The Internet is adaptable in that it keeps supporting new applications that are continually emerging, including services historically provided by dedicated networks (converged communications and broadcast services) as well as newly digitised and networked services such as online banking, remote health and ride sharing.
- The Internet has been resilient to a range of shocks, including very recently the dramatic increase in usage and changes in traffic types and usage patterns resulting from the Covid 19 pandemic; it has also proven broadly resilient to attacks and challenges thrown at its underlying design principles.

The four dimensions of success are illustrated in Figure 2.1.

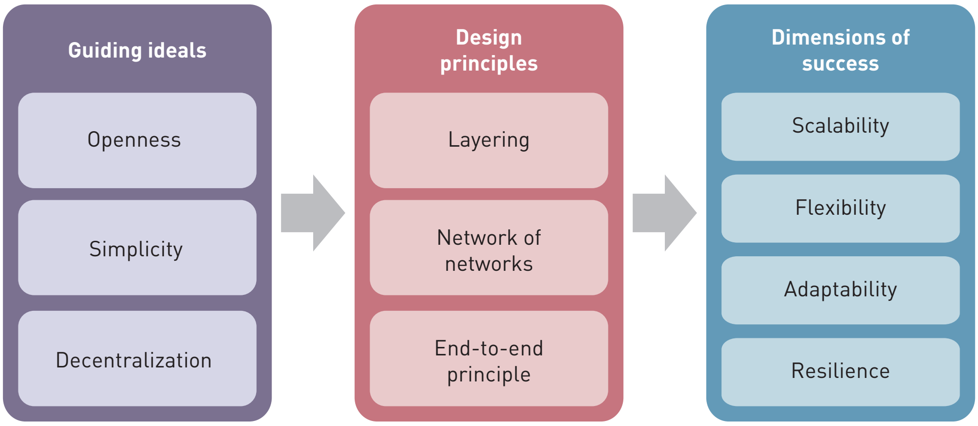
FIGURE 2.1: THE FOUR DIMENSIONS OF SUCCESS OF THE INTERNET [SOURCE: ANALYSYS MASON, 2021]



These dimensions of success explain how the Internet has been able to grow from its roots using dial-up access over traditional copper telephone lines, to being used by terminals to access mainframe computers, to simple text-based uses such as email and file-sharing, and into the modern, high-speed Internet used for multimedia, real-time services by a wide variety of devices. The dimensions of success are discussed in more detail in Section 3 to Section 6 of this report.

The goal of this study is to explore and describe the **four dimensions of success,** and to show how they are based on three fundamental **guiding ideals** that led the development of the Internet and resulted in three **design principles** that have been embedded in the Internet from the start. These guiding ideals and design principles are illustrated in Figure 2.2.

FIGURE 2.2: THE THREE FUNDAMENTAL GUIDING IDEALS AND THE THREE DESIGN PRINCIPLES [SOURCE: ANALYSYS MASON, 2021]



2.1 Three guiding ideals shaped the Internet from the beginning

While many ideas and technical constraints have contributed to the development of the Internet, this report highlights three guiding ideals that we argue have shaped both the technical and organisational development of the Internet, and therefore have been central to its success. The Internet was developed and operated with openness, it adopted simple solutions, and was decentralised with no owner. These guiding ideals led to a number of important design choices and operational practices, which were highlighted in a number of interviews during the course of the project. They are ideals in that while they may not be always achieved or adhered to in practice, they do provide critical guidance for the design and overall development of the Internet.

Credit for these guiding ideals, and the success of the Internet that resulted, belongs to the pioneers who developed and were guided by these ideals, inspiring countless others who have worked hard, and often without acknowledgement, to carry forward the ideals and develop the Internet as we know it today.

> *If you want me to say what the successful factors of the Internet were, I would say first that there were great people there – pioneers. Without the great people there is no great success for the Internet.*
>
> Professor Xing Li, Electronic Engineering Department, Tsinghua University

In the following subsections, we discuss each guiding ideal in turn.

2.1.1 Openness

> *The Internet protocols were given away deliberately, the World Wide Web protocols were given away deliberately. And the idea behind that was just simply to enable people to explore various ways to use these capabilities. It was not proprietary, we [Vint Cerf and Bob Kahn] deliberately thought our way through this and said let’s just make this open. I honestly believe that the openness is key to the success of the Internet.*
>
> Vint Cerf, Internet Pioneer

The Internet is open in a number of ways: the development of many new standards is open, allowing anyone to contribute, while making choices in a meritocratic way; the results of development – the architecture, standards, protocols and code – are open to be used by all, typically without the need to pay any royalties; and the Internet is open to networks arranging their own internetworking with other networks to exchange traffic, enabling new networks to emerge without centralised control or permission. This openness has been carried through time, famously for instance with Tim Berners-Lee ensuring that the World Wide Web protocols were made open to be adopted and adapted by others. The result is that Internet protocols are continually evolving, and they can be freely adopted by others to use and to adapt in turn.

2.1.2 Simplicity

> *In the beginning, the use of the Internet was not clear. There were only some very general applications like file transfer. When you have simple problems, you end up with simple solutions. And simple solutions are often very elegant solutions that, because they lack complexity, can end up scaling and being extensible. The protocols didn’t have too much complexity to start, which means they are general and flexible instead of specialized protocols with limited use.*
>
> Alvaro Retana, VP Technology Strategy, Future Networks, Futurewei Technologies

The Internet is essentially built from a number of mostly simple protocols that each perform a limited and simple task. The protocols are modular building blocks, stacked or sitting side by side in layers. These blocks can be selected and assembled in different ways to solve more complex tasks. The Internet founders did not know everything the Internet would be used for, instead developing a general-purpose technology. The result is that one protocol can be updated or even replaced without changing the other ones at different layers, applications can work without changing the protocols, and networks can be developed independently. This has had a powerful impact on the development of the Internet, as innovation and growth can take place at low cost by being made independently of the other building blocks. The result is affordable networks with economies of scale, and services with broad network effects.

The simplicity and fungibility of the Internet building blocks have turned out to be a very successful recipe compared to other competing technologies. Often, the Internet has not offered the most optimal solution for specific application needs. The telephone network, for example, offers a more optimised approach for voice calls, with less protocol overhead and better service guarantees. Similarly, broadcast networks are unrivalled in efficiently distributing linear TV channels efficiently to a large audience. Despite these advantages, the general-purpose Internet is taking over as the dominant distribution channel for both voice calls and video distribution, as shown below in Section 5, thanks to the economies of scale enabled by the ability of the Internet to carry any type of traffic for any service.

2.1.3 Decentralisation

> *The Internet protocols were given away deliberately, the World Wide Web protocols were given away deliberately. And the idea behind that was just simply to enable people to explore various ways to use these capabilities. It was not proprietary, we [Vint Cerf and Bob Kahn] deliberately thought our way through this and said let's just make this open. I honestly believe that the openness is key to the success of the Internet.*
>
> Vint Cerf, Internet Pioneer

The Internet is decentralised in several ways. Most fundamentally, at an organisational level there is no central authority that owns, operates or controls the Internet as a whole. Rather, these roles are distributed among various organisations, network providers, businesses, developers, and users, with players often acting in more than one role. The result is that there is minimal centralised administrative overhead, and no restrictions imposed by any central authority.

This decentralised nature of the Internet is also reflected at a technical level, featuring distributed protocols and autonomous networks with a large degree of freedom in how they implement their services. This makes it very simple for a new network to become a part of the Internet. There is no central acceptance process or complex co-ordination, all that is needed is a connection to at least one other network that agrees to transport traffic to and from the rest of the Internet. The routing protocol that is used to transport traffic between networks is also built in this decentralised way, automatically allowing new networks to be visible without any manual intervention required, as discussed in Section 3.2.3.

In the box below we provide a practical example of how these guiding ideals have impacted the development of organisational and governance practices on the Internet. We will use such boxes throughout the report to discuss specific concepts or trends that are relevant for the success of the Internet.

The impact of the guiding ideals in practice

The guiding ideals have a significant organisational impact which has contributed to the success of the Internet. In place of a central developer, owner or operator of the Internet, a form of open, multi-stakeholder governance has emerged, in which different stakeholders play different roles based on the forum, nature and location of the issue. At the same time, the practices of companies are also impacted by these general ideals. While governance is very broad, we focus here on the technical aspects that contribute to success.

Standards development: Several standard development organisations (SDOs) have arisen from or adapted to developing relevant standards, including the Internet Engineering Task Force (IETF) for the Internet protocols; the World Wide Web Consortium (W3C) for web standards; and the Institute of Electrical and Electronics Engineers (IEEE) for network standards. These organisations have adopted a set of open principles for identifying where solutions are needed, have developed standards that are market driven and chosen based on merit, and have managed the balancing act of ensuring the continuity of key protocols, while enabling the Internet to evolve to meet new opportunities and challenges.[1] Many of these standards are available royalty free, and they all form the basis for a global Internet built on thousands of independent networks, millions of applications and services, and billions of devices.

[1] See https://open-stand.org for a set of open standards principles developed by IETF, W3C and IEEE along with the Internet Society (ISOC) and the Internet Architecture Board (IAB).

> *Internet standards are meritocracies in that they are adopted based on their own functionality and utility as opposed to anything else, which is one of the success factors of the Internet.*
>
> Paul Gampe, CTO at PCCW

Unique identifiers and protocols: Several essential identifiers, including domain names, autonomous system (AS) numbers and IP addresses, as well as related protocol identifiers, must be managed across the Internet to avoid duplication and confusion. The Internet Corporation for Assigned Names and Numbers (ICANN) assigns these common resources and co-ordinates the Domain Name System (DNS), using a multi-stakeholder model. The DNS plays a central role, by translating human-readable domain names (such as info.cern.ch) to a routable IP address. It was introduced in response to the need for a scalable way of distributing the names of hosts connected to the Internet, and it has scaled to support the hundreds of millions of domain names registered today.

ICANN is also responsible for the Internet Assigned Numbers Authority (IANA) functions, by which protocol parameters are administered, through its affiliate organisation, PTI (Public Technical Identifiers). The five regional Internet registries (RIRs) manage the allocation and registration of Internet number resources within their region in line with regional growth, demands and policies.[2] Given the vast scale of the Internet today, the overhead costs of these functions are very low, and balance the need for central co-ordination and point of contact with regional differences.

Operator collaboration: Since the earliest days of the Internet, the ability to interconnect networks has been fundamental to the Internet, and the operators of the networks play an important role in developing and implementing a set of common operational practices. This has required collaboration among operators, including between those who compete with one another, and those on opposite sides of the world with no direct connections. One clear and unavoidable area of co-operation is for interconnection to exchange traffic, which is often done using peering agreements, as discussed further below in Section 4.2.2. More broadly, network operator groups (NOGs), such as NANOG in North America have emerged to develop communities of operators for learning, developing relationships, and addressing challenges.

> *The fact that you can have this collaboration of private-sector entities across borders that makes it all happen without needing to get permission from a regulator in every country speaks to the Internet's success.*
>
> Dr Alissa Cooper, VP/CTO at Cisco

Commercial role: Companies other than operators play a wide variety of roles in the development and advancement of the Internet. Vendors help to develop standards by participating in the relevant SDOs and implement the standards into hardware that is used to operate and use the Internet. Software companies, which include the globally successful online platforms, develop applications and services using existing standards and help to develop new ones. Depending on their size and role, these Internet companies work together for operational and security resolution. More generally, companies in many industries are significant users of the Internet, benefiting from low costs from the economies of scale of Internet products and services, along with the network effects of being able to transact online.

[2] The two sponsors of this report, APNIC and LACNIC, are the RIRs for Asia–Pacific, and Latin America and the Caribbean, respectively.

Government role: In many countries, the government role in promoting the development of the domestic Internet was indirect, primarily through deregulating the telecoms market and promoting competition, enabling existing operators to become ISPs and new ISPs to enter the market, while making relatively little direct investment in funding its development. The Internet has now proven successful in a variety of political and socioeconomic contexts, delivering economic growth and development and enabling efficient delivery of government services, and jobs and income for citizens. As a result, many governments have increased their role – both domestically and at the international level – to help bridge the digital divide in unserved or underserved regions to ensure everyone benefits from the Internet.

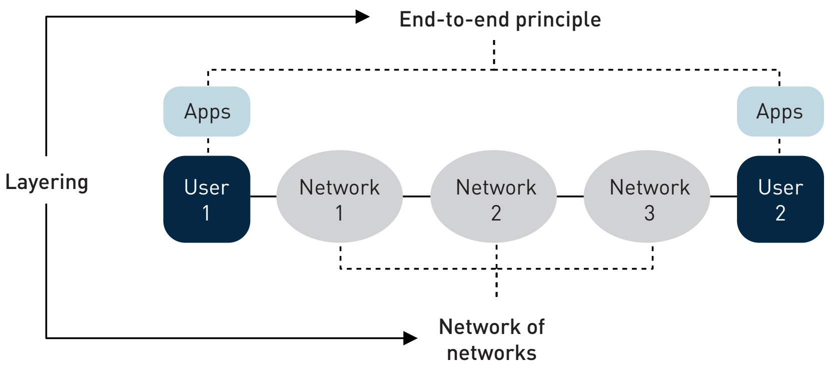
At the same time, governments initially took a largely *laissez faire* approach to regulating the Internet. Over time they began to appreciate the impact that the Internet would have on their economies and societies, and that the hands-off approach may not be sustainable. At the first multilateral meeting addressing the Internet, the World Summit on the Information Society (WSIS) in Geneva in 2003, the topic of Internet governance was raised and debated. The result was the definition of the multi-stakeholder approach to Internet governance, in which governments, the private sector and civil society all have critical ongoing roles in governing the Internet.[3]

2.2 Three design principles underpin the Internet's success

While the Internet continues to develop, with new applications, protocols and networks updating or replacing older ones, the Internet is generally characterised by three design principles, which we argue have been central to the success of the Internet:[4]

- **Layering principle.** Under the layering principle, applications are separated from the underlying networks, allowing evolution to occur in some parts of the Internet without affecting others. The Internet Protocol (IP) is the central, stable building block separating the layers. This principle enables the following two principles.
- **Network-of-networks principle.** Each network can be operated independently of the other networks as long as they all use the common Internet protocols to route traffic. This allows existing networks to be connected to the Internet while new networks are deployed, each adapted to its technology and environment.
- **End-to-end principle.** Under this principle, the intelligence sits in end devices at the edge of the network, rather than in routers in the core of networks. As a result, applications can be developed and installed in the wide and growing variety of Internet-enabled devices, without making changes in all the networks.

FIGURE 2.3: THREE DESIGN PRINCIPLES CENTRAL TO THE SUCCESS OF THE INTERNET [SOURCE: ANALYSYS MASON, 2021]



[3] The Working Group on Internet Governance, set up during the WSIS, provided the following definition:

Internet governance is the development and application by Governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet.

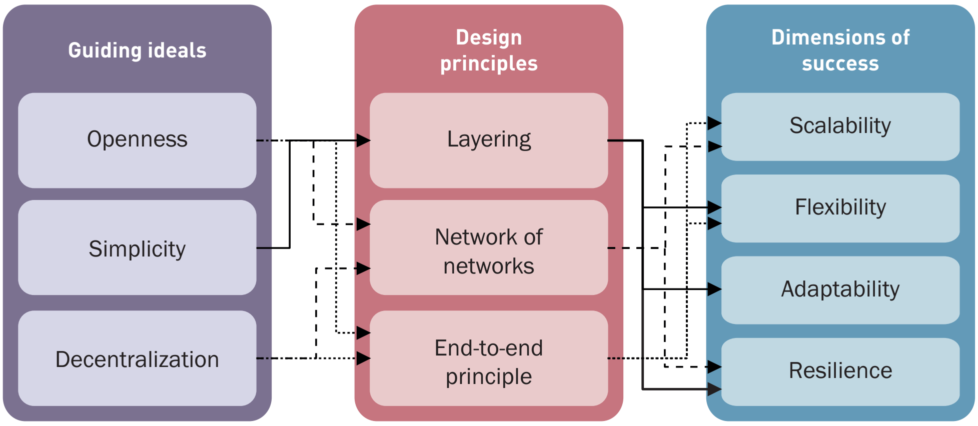
See Report of the Working Group on Internet Governance, June 2005.

[4] These design principles have been presented in various forms in different contexts. For a related discussion, see for instance Clark, D., The Design Philosophy of the DARPA Internet Protocols, in Proc. SIGCOMM '88, Computer Communication Review, 18(4), August 1988.

As shown in Figure 2.4 below, the three design principles of **layering, end to end** and **network of networks** are built on the fundamental guiding ideals discussed above, and in turn lead to the different dimensions of the success of the Internet. The **openness** of the Internet allows anyone to develop new apps and anyone to access them, and to develop new networks and connect them to others. The **simplicity** of the network is expressed through separating the layers so that what happens at one layer can be independent of the other layers. And finally, the **decentralisation** distributes intelligence and function to the edges, such that no network or entity controls the Internet.

The three design principles in turn have contributed to the different dimensions of the success of the Internet. They support the **scalability** of the Internet, as new networks can continue to emerge independently, while relying on the intelligence in the end systems to operate applications. The separation of networks from applications gives **flexibility** by allowing networks to develop and grow independently. The same separation allows applications to develop independently from the underlying network technologies, and thus makes the Internet **adaptable**. The technical principles also contribute to the **resilience** of the Internet: if a particular network goes offline, the Internet can route around that network to deliver traffic, without requiring applications running on another layer of the Internet to even be aware of the change.

FIGURE 2.4: EXAMPLE RELATIONSHIPS DEVELOPING THE SUCCESS OF THE INTERNET [SOURCE: ANALYSYS MASON, 2021]



The design principles have shaped the Internet that we experience today and have played a central role in its technical success. They are, however, neither set in stone nor inviolable. There are, in fact, many examples of Internet technologies and practices that explicitly violate one or more principles. Later in this report (Section 7.1), we examine examples of such violations, how the Internet has coped with them and discuss whether the principles are still relevant for the Internet of today and tomorrow.

In the following subsections, we discuss each design principle in turn.

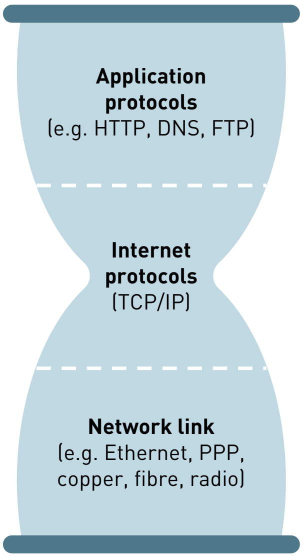
2.2.1 Layering principle

> *The architecture of the host protocol is layered. We learned very quickly that layering is your friend, that you can change all the stuff in the middle of a layer without affecting anybody else as long as the interfaces stay the same. And so that stability was already injected into the thinking for the ARPANET protocol suite, and that carried over into the Internet as well.*
>
> Vint Cerf, Internet Pioneer

Layering is a central principle in the Internet architecture. Internet protocols are layered on top of each other, and data is passed from a protocol operating at one layer to a protocol operating at the next, through standardised interfaces (i.e. procedure calls and packet formats). A classic depiction of the layering approach is to use an hourglass, as shown in Figure 2.5 below. Application protocols are depicted at the top of the hourglass, and the network link protocols that control the physical layer are at the bottom, separated by a limited number of protocols, notably the Internet Protocol (IP) and the Transmission Control Protocol (TCP) at the 'waist'. The result is sometimes referred to as allowing 'everything over IP' – all applications to run on any Internet-enabled network, and 'IP over everything' – the Internet to run on any network.

The narrow waist is also illustrative of the simplicity of the 'Internet' layer, which consists of relatively few distinct protocols, compared with a plethora of underlying infrastructures over which it operates, and of overlying applications which it supports. This layer has been stable, while the layers above and below have continued to change and grow continually over time.

FIGURE 2.5: THE HOURGLASS DEPICTION OF THE INTERNET PROTOCOL SUITE[5] [SOURCE: ANALYSYS MASON, 2021]


Application protocols
(e.g. HTTP, DNS, FTP)
Internet protocols
(TCP/IP)
Network link
(e.g. Ethernet, PPP, copper, fibre, radio)

This separation allows lower-layer protocols to operate without regard to how the higher-layer protocols perform their tasks, and vice versa. This again gives much flexibility. Protocols can be developed to perform small, simple tasks, while the overall communications solution is realised by combining several protocols into a working whole. This way of breaking tasks down into simpler components that are solved by different protocols has enabled innovation and contributed to several dimensions of Internet success. It has allowed innovation at the application layer to take place mostly independently from developments at the network or physical layers.

Layering has always been implemented in a pragmatic way in the Internet. While a reference architecture like ISO's Open System Interconnect (OSI) model describes a clean and well-defined seven-layer protocol architecture, the Internet represents a more ad-hoc approach where protocols can be added on top of each other (sometimes duplicating the same functionality, like encryption) in a less regimented way. This lack of stringency has arguably given the flexibility needed to enable new and innovative solutions.

The Internet Protocol (IP) is the most central building block in the suite of Internet protocols. IP is a simple protocol, which defines a packet format and an address space. The main purpose of IP is to convey the address of the source and destination of an IP packet, so that it can be delivered from any source network on the Internet to any destination network. While lower-layer protocols can operate on single links or within a single physical network, and higher-layer protocols often run only in the end hosts (clients or servers), the Internet Protocol must run both in end systems and in every network router along the path that connects these.

[5] Adapted from https://www.researchgate.net/figure/The-hourglass-architectural-model-of-the-Internet-Protocol_fig2_251419252; see also https://www.iab.org/wp-content/IAB-uploads/2010/11/hourglass-london-ietf.pdf.

IP has remained essentially unchanged over four decades. IP version 6 (IPv6) was introduced over 20 years ago in a move that expanded the IP address space dramatically, but this change has not yet been fully adopted and the previous version, IPv4, is still used extensively. This is discussed in more detail in the box below.

IPv6

By the early 1990s, it became clear that the Internet would grow so large that the number of available IP addresses would become a problem. The 32-bit IPv4 address space makes it possible to address just over 4 billion end points, although the practical implementation of the protocols gives a lower number in practice. The last IPv4 address block was assigned from IANA to the regional RIRs in 2011, and almost all RIRs have now depleted their IPv4 resource pool. The solution to this was the introduction of IPv6, a new protocol version with a 128-bit address space; enough to assign 100 addresses to every atom on the surface of the Earth.

Adopting IPv6 has, however, been more difficult than anticipated. One reason for this is the lack of a short-term incentive. Transitioning to IPv6 involves some cost in terms of upgrading equipment or configurations and training, but without everyone else adopting it, the benefit is less clear. Recently, there has been a notable increase in the use of IPv6, but currently there are still many end devices and access routers that do not support IPv6.

The difficulties in replacing IPv4 are not entirely unexpected. IP is the most central protocol in the Internet, and must be supported both in network elements and in end devices. Communication over IPv6 requires that all devices in the end-to-end path support this protocol, or alternatively use potentially resource-constrained gateways that translate between IP versions. In some ways, the slow adoption of IPv6 is also an attestation to the flexibility of the Internet protocol suite. This 'flip side of flexibility' has made it possible to deploy network address translation (NAT), split IPv4 address blocks into smaller units and to trade IPv4 addresses. All these efforts may be economically and operationally rational, but in sum they contribute to slowing down the adoption of IPv6.

Figure 2.6 and Figure 2.7 show the percentage of websites using IPv6 and the percentage of IPv6 traffic.

FIGURE 2.6: PERCENTAGE OF WEBSITES USING IPV6

[SOURCE: W3TECHS, 2021]

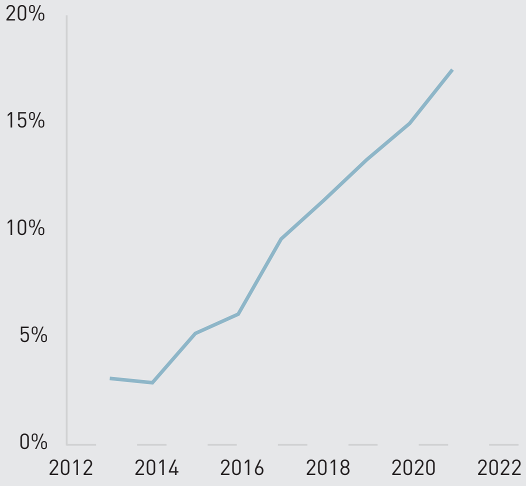
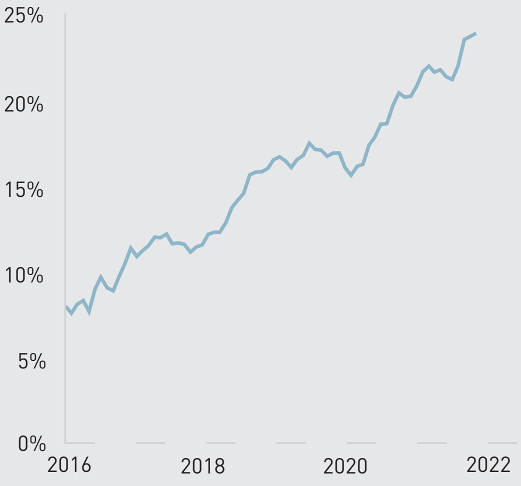


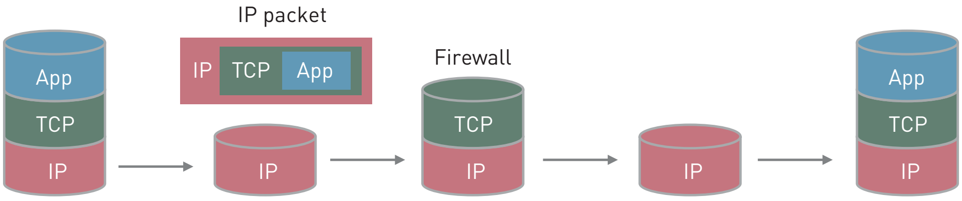
FIGURE 2.7: PERCENTAGE OF TRAFFIC THAT IS IPV6

[SOURCE: AKAMAI, 2021]



In its pure form, layering implies that a function or protocol operating at a given layer in the protocol suite should perform its duties based only on information contained within the protocol itself, without relying on information contained in a higher-layer protocol. This ideal rule is, however, frequently broken, leading to so-called layering violations. Examples of layering violations include NATs, firewalls, load balancers and other traffic management devices that manipulate Internet packets in various ways based on their content. Such devices, often called middleboxes, operate at the network layer of the layered protocol suite since they operate on IP packets and are not concerned with end-to-end communication with a remote peer. Yet they normally rely on higher-layer information like port numbers or various application identifiers to perform their task, as illustrated in Figure 2.8.[6]

FIGURE 2.8: A FIREWALL IS A MIDDLEBOX THAT ACTS ON HIGHER-LAYER PROTOCOL INFORMATION IN THE NETWORK TO FILTER OUT UNWANTED TRAFFIC [SOURCE: ANALYSYS MASON, 2021]



2.2.2 Network-of-networks principle

> *When the ARPANET was put together, it was a network of heterogeneous computers connected with a homogeneous transport mechanism and there was a single point of control. It was clear immediately that you can only scale that up to a certain degree, that that would become a bottleneck right away, in multiple ways. Operationally, from a management point of view, from an economic point of view, and unquestionably from a political point of view. And so, there would be multiple networks and so the only question is, how are they going to interoperate and you could err too far in each direction, you could try to make them all one network and as I said that will fail, or you could try to ignore them, basically, and have them just sprout up independently, and then people would cross connect them in some fashion and you would have an ungainly mess of gateways, and to a certain extent, there was some of that – you had cross connections between ARPANET, BitNet, UUNET and so forth and they were messy.*
>
> *So an extraordinary challenge to figure out, how do you get a useful and effective interconnection of these multiple networks that had the positive aspects of separation and distribution – separate economics, separate motivation, separate new technologies coming in and so forth, and at the same time, a tight enough co-ordination so that from a usability point of view it functions pretty much as if it's one integrated system. I don't know how much more there is to say about that, I mean you see what the results are, and you know it's easy to go and find particular problems. But I think any list of the problems has to be against the backdrop of how extraordinarily successful it's been.*
>
> Steve Crocker, Internet Pioneer

[6] We discuss the role of layering, layering violations and the reinvention of the Internet protocol suite due to recent developments like end-to-end encryption and new approaches to flow control in Sections 5.2.2 and 6.2.3 of the report.

It was clear from the outset that what later became the Internet would serve to interconnect many existing, established networks; it would be built from these networks, not replace them. One of the original motivations was military command and control, which implied putting network nodes on ships and in planes. Another challenge was to connect to Europe using satellite. The Internet therefore had to work over wired, wireless and satellite networks. To cope with the vastly different characteristics of different underlying networks in terms of latency, variations and error rates, the early developers of the Internet decided to split the end-to-end paths into separate underlying networks. These would connect to each other through Interface Message Processors (IMPs), which were the routers of the day. This gave birth to the notion of the Internet as a network of networks.

The network-of-networks principle has been key to the development of the Internet. The choice to separate the complexity of how traffic is transported within a given network from the problem of how these networks should connect to each other has shaped both technical and organisational aspects of the Internet. The notion of networks as autonomous systems (ASes) that largely decide both their internal organisation and their rules for bilateral connections with other ASes has allowed for innovation and fostered the development of a rich ecosystem of networks.

As a result of this principle, the problem of moving a piece of content across the Internet becomes a simple question of selecting which neighbouring network can be used to get the data closer to the destination. The internal workings of each network are mostly hidden from the protocol that makes these decisions. Each AS independently decides how to move traffic internally, and independently selects to which neighbouring AS to expose its routes. Traffic is routed between networks using an inter-domain routing protocol. Each AS has its own number for identification, and traffic is routed based on IP addresses identifying the source and destination within the network. Further details on the role of inter-domain routing and traffic exchange are found in Sections 3.2.3 and 4.2.2.

The network-of-networks principle works in concert with *packet switching*, which is another important characteristic of the Internet. In the traditional telephone network, an end-to-end path must be signalled, and resources reserved between two end systems before communication can start, a concept known as circuit switching. On the Internet, in contrast, traffic is segmented into packets, which are individually forwarded towards the destination by routers along the path. Packet switching reduces the amount of signalling and avoids per-flow state in the network, thus contributing to simplicity. It is also an efficient way to share network capacity under rapidly varying loads, which are typical for Internet traffic.

The Internet routing and addressing plan does not contain a notion of countries or national borders, instead making the AS the key building block of the Internet, with minimal barriers to connecting them together (reflecting the openness of the Internet).

RIRs assign IP addresses directly to ASes, and not to countries. The non-discriminatory, open, transparent and hassle-free distribution of IP addresses has supported the success of the Internet, since it has made it very easy to create new networks and connect them to the Internet, normally without any national or otherwise centralised qualification process.

2.2.3 End-to-end principle

> “
>
> *End to end is derived from the layered protocol architecture, basically, but it also meant that important to the success of the Internet is the ignorance of the Internet Protocol layer. By this, I simply mean that it didn't know how things were being carried or what was carried, like a postcard doesn't know how it's carried or what is written on it. This meant that new applications requiring new interpretations of the payload of the Internet packets did not require any change to the underlying network, just re-interpretation of payload by the serving computers at the edge of the Internet.*
>
> Vint Cerf, Internet Pioneer

The end-to-end principle states that the network should be kept simple with limited functionality, while more complex functions should normally be performed in the end systems.[7] This implies that the architecture should preserve the ability for end systems to communicate directly, without intermediation. In the context of the open network-of-networks principle and corresponding decentralisation of the Internet as discussed above, there is no central network to install or orchestrate functionality for an end-to-end path across networks, and it is difficult to co-ordinate across networks. Combined with the physical limits on bandwidth and processing capacity in routers, this has contributed to limited functionality being retained in the networks. Thus, complex functions, such as guaranteed packet delivery, congestion control or encryption, can only be achieved with information that is available at the end systems.

Keeping the network simple and pushing complexity to the edges has had a significant positive impact on the Internet's ability to support new applications and use cases (this is explored in Section 5). New applications can be introduced in the end systems without requiring new software or other changes in the network, instead relying on intelligence in the end points and the user devices themselves. The relative simplicity of the network has arguably also played a positive role in the development of sustainable business models. By keeping complexity in applications in end systems, including devices, the investments needed to build networks are reduced. Setting up a network and connecting it to the Internet is relatively simple, which has contributed to a large diversity in the network provider landscape (this is explored in Section 4).

The end-to-end principle has been challenged in several ways through the history of the Internet. The prevalence of middleboxes has to some extent reduced the relevance of the end-to-end principle. Middleboxes in this context are primarily NATs (used to allow several end hosts to share the same public IP address and thereby ration scarce IPv4 addresses) and firewalls (used to protect networks from unwanted traffic and attacks). In addition to NATs, caching devices and content delivery networks (CDNs) are also prominent examples of technologies that move intelligence from the end points and into the network.

Middleboxes challenge the end-to-end principle by putting application- or flow-specific state in the network. This can in some cases limit the flexibility of end systems by requiring certain protocols to be used, or by restricting which end points can reach a network. Protocol headers in IP packets have traditionally been transmitted in an unencrypted text format, which has allowed middleboxes to access information about higher-layer protocols or even applications in the network in order to treat the relevant traffic in a particular way. Recently, however, we have seen a trend where most Internet traffic is encrypted in an end-to-end manner. This limits the possibility of middleboxes to classify traffic and may in this way contribute to strengthening the relevance of the end-to-end principle. As noted above in Section 2.2.1, some middleboxes also violate the layering principle – we examine the resilience to such challenges in Section 7.

[7] The case for the end-to-end principle is made in the seminal paper, End-to-End Arguments in System Design (1984) by Saltzer, J. H., Reed, D. P. and D. D. Clark.

Best efforts and quality of service

> *The Internet is simple, and the simplicity has value in its own right. And the consequence of that simplicity is that it's not perfect in the sense that you have these sorts of degrees of uncertainty as to whether a packet is going to be delivered and a degree of uncertainty as to when it's going to be delivered. But the simplicity is important, and the efficiency of it is enormously important. It costs you a huge amount to squeeze out that last little bit of uncertainty.*
>
> Steve Crocker, Internet Pioneer

The Internet, in general, offers a best-effort service without guarantees on important performance metrics such as bandwidth or delays. This is a design feature of packet switching, which was adopted early in the development of the Internet as an efficient means to multiplex several data streams with variable load onto a telecoms link. This contrasts with other networks such as the traditional telephone network, which in its original form used circuit-switching to reserve capacity for each flow to guarantee a certain quality of service (QoS). Several standards have been developed to enable relative or absolute service guarantees in IP networks, and they are often used internally in individual networks. We have not, however, seen widespread adoption of QoS protocols across the Internet.

There are several reasons for this, which can be tracked back to the three design principles discussed earlier. End-to-end QoS can only be guaranteed if each network along the path agrees to provide such guarantees. This would, however, involve placing more complex functionality and more state in the networks, going against the end-to-end principle. Also, the distributed control in the network of networks means that each network makes independent routing decisions, which makes it hard to enforce a given path with guaranteed performance.

Furthermore, we will show below that the technical and commercial arrangements for exchanging traffic between networks have not embodied any service guarantees (Section 4.2.2). At the same time, service quality has been increasing in spite of drastic increases in demand. This is mainly a result of increased investment in capacity, some of which is being provided by the content and application providers, and not traditional operators (Section 3.2.2). At the same time, a new business model of distributing content and applications to the edges of the networks, closer to the end users, has helped lower the latency of delivery, albeit with no guarantees (Section 5.2.3).

The Internet community seems to have found ways to work with best efforts. However, the lack of stricter service guarantees continues to be one of the most frequent critiques of the Internet, and a main motivation for other alternative architectures, as discussed in Section 7.

In the remainder of the report, we examine:

- the scalability of the Internet (Section 3)
- the Internet's flexibility in networks (Section 4)
- the Internet's adaptability to new applications (Section 5)
- the resilience of the Internet in the face of shocks and changes (Section 6)
- the prospects for the Internet's further success (Section 7).

3 Success dimension: Scalability supporting the growth of the Internet

> *The Internet has grown by orders of magnitude over the last 60 years, from an Internet that was used by a small group of researchers to one that is used worldwide for every purpose imaginable, and it still works today – a major success. The first thing that highlights the Internet's success would be scale. The fact that underpinning the Internet today is a series of protocols which have demonstrated their ability to scale to provide seamless connectivity to all corners of the world. And the other attribute that I think is critical with that scalability is the fact that it has been multi-vendor and multi geographic. Also the fact that we've been able to have nonlinear scalability in the amount of traffic that we can move across the Internet.*
>
> Paul Gampe, CTO at PCCW

3.1 Observations

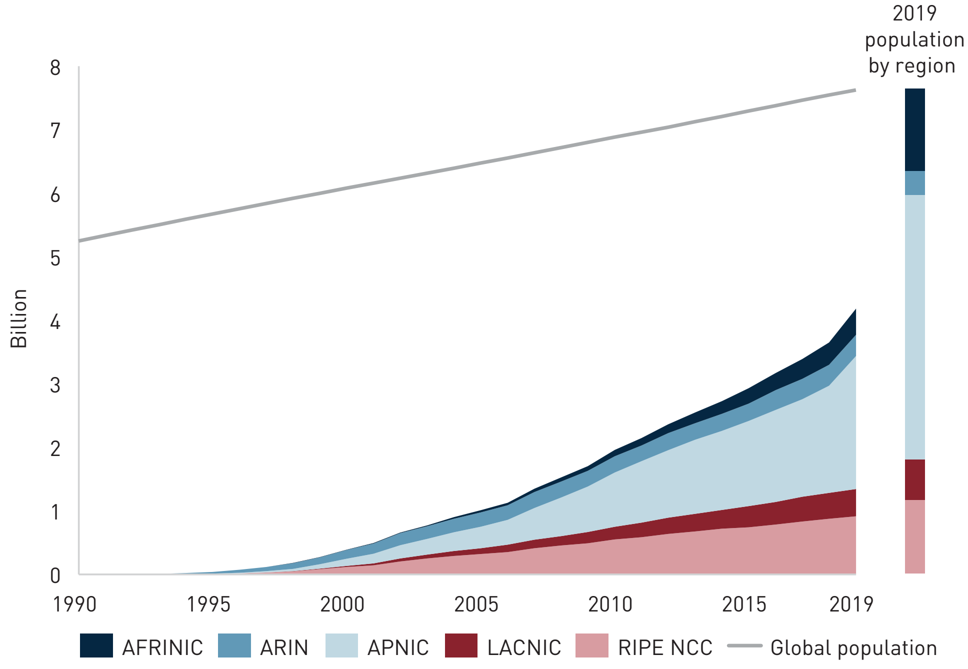
The successful growth of the Internet has occurred at multiple levels: not only has the number of Internet users grown rapidly, but each user's traffic volume has also grown, all while connection speeds have become faster. The technical factors behind the Internet's scalability that enabled this success are explored in Section 3.2.

Figure 3.1 illustrates the rapid growth of the Internet's user base from an estimated 2.6 million in 1990 to 3.9 billion (more than half of the global population) only 30 years later. This growth is ongoing. Even in regions where per capita Internet adoption is nearing or at 100%, there is still exponential growth of the network in terms the number of devices connected. According to Cisco, in 2003 there were 0.08 connected devices per person worldwide, 1.84 in 2010, 2.2 in 2015 and 2.4 in 2018.[8] Obviously these numbers are still small, if we consider the potential number of personal and household devices, not to mention industrial machines, that will become Internet connected in future.

Most current and future growth, however, will occur in Africa and Asia–Pacific, where only 31% and 51% of the respective populations used the Internet in 2019. As investment continues and barriers to adoption are addressed, penetration will increase in these regions and the Internet will continue to grow and become more representative of the geographical distribution of the global population (see the right-hand side of Figure 3.1).

FIGURE 3.1: INTERNET USERS BY RIR REGION,[9] OVERLAID WITH GLOBAL POPULATION

[SOURCE: ANALYSYS MASON, ITU, WORLD BANK, 2021]

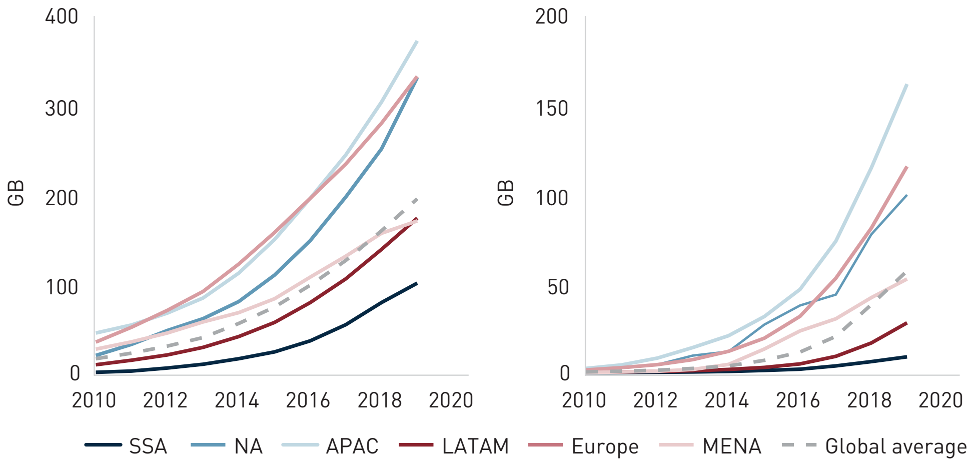


[8] https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf, https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2015_Year_in_Review.pdf and Cisco Annual Internet Report – Cisco Annual Internet Report (2018–2023) White Paper – Cisco

[9] RIR regional data built up from ITU country data. AFRINIC is the African Network Information Centre; ARIN is the American Registry for Internet Numbers; and RIPE NCC is the Réseaux IP Européens Network Coordination Centre.

Not only is the number of users growing, which increases the total amount of traffic generated, but the traffic per user is also growing, indicating a remarkable ability to scale. Figure 3.2 shows how the amount of traffic generated per user has increased over time, from a global average of 19.0GB of fixed data per annum in 2010 to 198.3GB in 2019 (a factor of 10), and from 0.6GB to 57.1GB (a factor of 100) for mobile data in the same time period.

FIGURE 3.2: FIXED (LEFT) AND MOBILE (RIGHT) DATA TRAFFIC PER INTERNET USER BY REGION[10] [SOURCE: ANALYSYS MASON, 2021]



The increase in individual traffic is the result of two key factors. The first is increased usage – for example, daily time spent online in the USA increased 170% in 12 years, from 2.9 hours a day in 2008[11] to 7.8 hours a day in 2020.[12] The second factor is more data-intensive usage – for example, video constituted 70% of global consumer traffic by 2017.[13] Traffic per user will continue to grow as time spent online increases. Traffic will also grow as usage becomes increasingly data-intensive, due to rising demand for higher-quality video, a surge in the number of devices per user connected to the Internet, and decreases in data pricing.

As a result of the growth in the number of users and individual traffic usage, total Internet traffic has increased exponentially over time. Combined traffic exceeded 2841EB (exabytes) in 2019 – 30 times the data traffic volume of 95EB in 2010 (Figure 3.3).[14]

[10] Data is regional based on regions as per Analysys Mason DataHub (https://www.analysysmason.com/what-we-do/practices/research/datahub/). The data is not built up into RIR regions. MENA covers two RIR regions, so an assumption would have to be made in order to split it between the two.

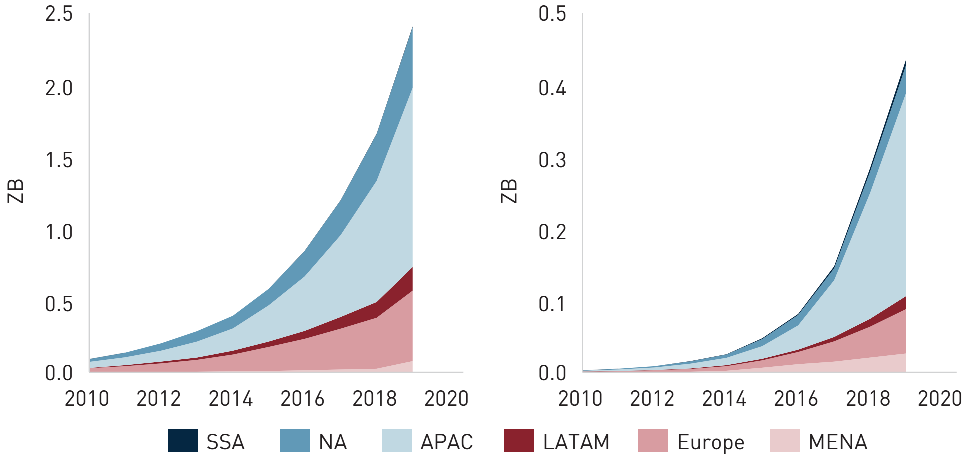
[11] https://www.bondcap.com/report/it08/

[12] https://www.emarketer.com/content/us-time-spent-with-media-2021-update

[13] https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2015_Year_in_Review.pdf and https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2017_Year_in_Review.pdf

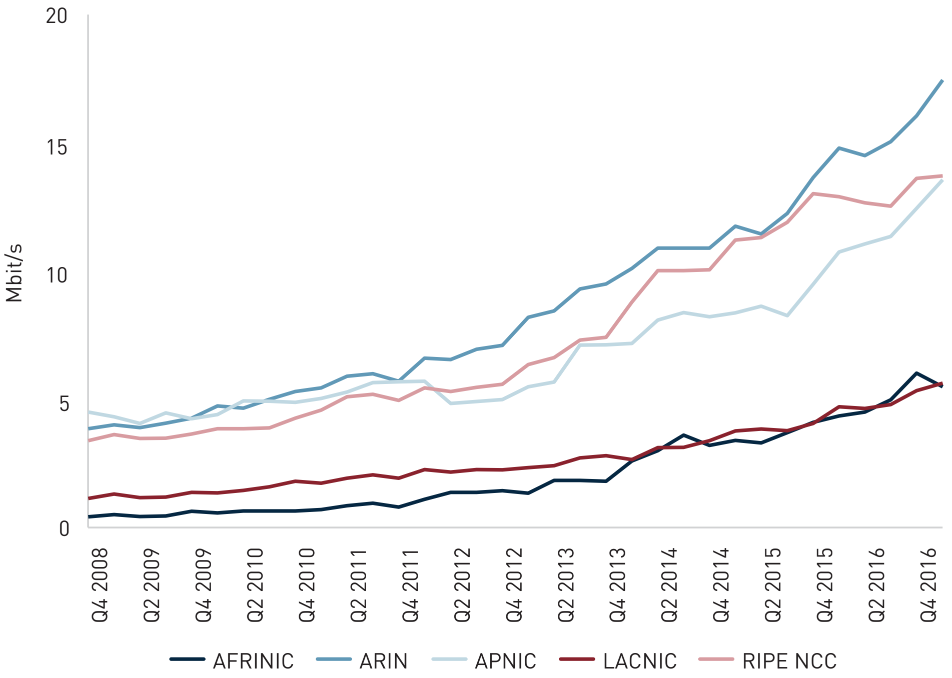
[14] An exabyte is 1018 bytes. A zettabyte (represented by the abbreviation 'ZB' in Figure 3.3), is 1000 EBs.

FIGURE 3.3: FIXED (LEFT) AND MOBILE (RIGHT) DATA TRAFFIC BY REGION[15] [SOURCE: ANALYSYS MASON, 2021]



The Internet has had to scale to keep up with this incredible demand, both in terms of available capacity and network architecture.[16] Since demand continues to grow, the Internet will have to continue to scale. Remarkably, despite more users, and more traffic per user, connections speeds are increasing over time, as shown in Figure 3.4. For example, LACNIC region countries' average connection speed saw a 200% increase over eight years, and AFRINIC region countries saw a 1200% increase. The Internet has thus scaled such that it has not only met growing demand, but has improved user experience.

FIGURE 3.4: AVERAGE CONNECTION SPEED BY RIR REGION[17] [SOURCE: ANALYSYS MASON, AKAMAI, 2021]



[15] Data is regional based on regions as per Analysys Mason DataHub (https://www.analysysmason.com/what-we-do/practices/research/datahub/). The data is not built up into RIR regions. MENA covers two RIR regions, so an assumption would have to be made in order to split it between the two.

[16] As explained further in Section 3.2.

[17] RIR regional data built up from Akamai's country data

The next subsection explores the technical factors that allowed the Internet to scale successfully.

3.2 Explanation

The technical and commercial fabric of the Internet has been able to scale well with increased demand. The underlying design principles have enabled growth in the geographical reach of the Internet, while also scaling so that speeds are improving over time (Section 3.2.1). This has allowed the introduction of new networks that give increased capacity (Section 3.2.2), while the interdomain routing system has scaled well to the increased number of networks (Section 3.2.3).

3.2.1 The scalability of the Internet is enabled by the underlying design principles

Critical to the scalability of the Internet has been the layering principle, which allows networks and applications to be modified or replaced independently of each other.[18] This demonstrates the benefits of designing the Internet around simplicity, such that changes at one layer do not impact the others layers.

As a result of the network-of-networks principle, and openness, anyone can invest in networks and capacity, as discussed in Section 3.2.2. Another central factor in the success of the Internet has been the scalability of the system that routes traffic between the networks, as discussed in Section 3.2.3.[19]

At the same time, the end-to-end principle allows applications and services to grow independently, to keep up with demand, based on the investments of the content providers and using the intelligence in the devices rather than the network.[20]

3.2.2 Supply of network capacity has kept pace with growth in traffic

To keep up with the growing demand described above, the number of networks, and their capacity, has had to increase dramatically.

The demand for Internet connectivity has driven a steady increase in the number of ASes. In 1986, there were only 80 ASes connected to the Internet. By 2000, this had increased to around 10 000, and today the number of ASes that are visible in the global routing table has passed 90 000. This growth reflects the growth in users and geographical footprint of the Internet, but it is also an attestation to the openness and simplicity of acquiring the numbering resources (IP addresses and AS numbers) that are needed to set up a new network and connect it to the Internet.

The increased user base and traffic have also been matched by increased capacity in the global transport networks that move traffic between cities and continents. Figure 3.5 shows how subsea transport capacity has grown over the last decade via various regional routes. The growth in the total potential subsea capacity in this period has been significant, from 240Tbit/s in 2010 to 3177Tbit/s in 2020, an average compound annual growth rate of 38%. By 2020, about one-third of this fibre capacity was being used, leaving significant spare capacity that can be set in production relatively quickly if there are rapid increases in demand (as happened during the initial phase of the Covid 19 pandemic).

The growth in subsea cable capacity is driven to a large extent by the capacity needs of large Internet companies that offer content and applications, and who are now themselves investing in supply. TeleGeography reports that Internet companies' share of undersea capacity rose to 66% of total capacity by 2020, up from less than 10% in 2012. Much of these capacity needs are currently being met by the Internet companies investing directly in cables – USD8 billion in investment has been announced over the next three years.[21] Investments in long-distance subsea cables are normally joint ventures between several companies, and the investments made by the large Internet companies have therefore also helped increase capacity and route diversity for traditional transit providers.

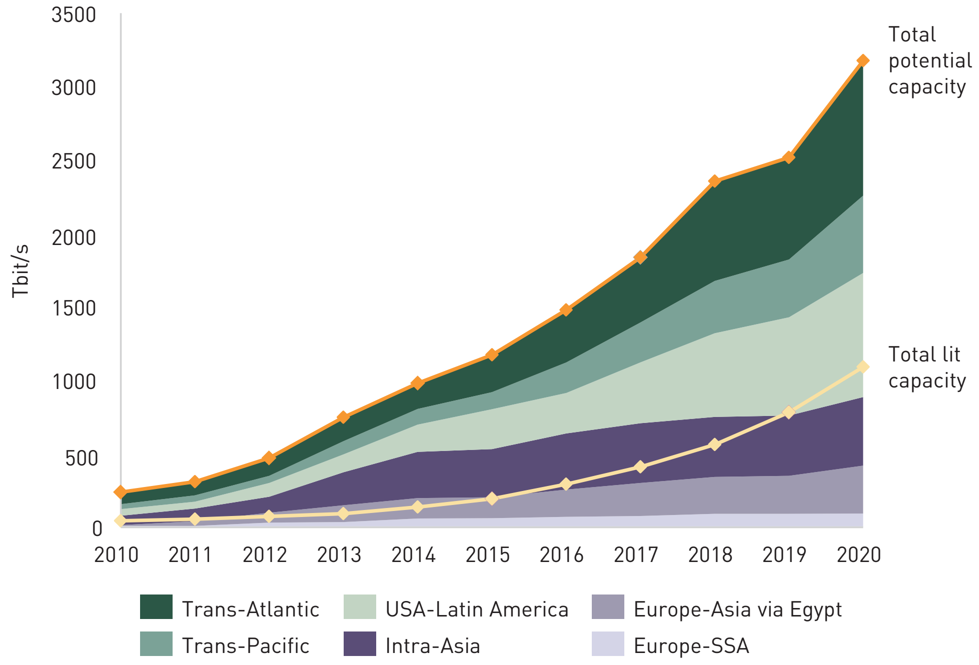
[18] As illustrated in the depiction of the Internet protocol suite as an hourglass shape in Figure 2.5.

[19] The flexibility of the Internet towards new network technologies is addressed further in Section 4.

[20] The adaptability of the Internet toward new applications is addressed further in Section 5.

[21] https://blog.telegeography.com/telecom-headlines-june-2021

FIGURE 3.5: POTENTIAL SUBSEA CAPACITY BY ROUTE, OVERLAID WITH TOTAL LIT CAPACITY [SOURCE: TELEGEOGRAPHY, 2021]



The strong increase in international transport capacity goes hand in hand with reduced prices, which in turn helps drive demand. Figure 3.6 and Figure 3.7 show the evolution of weighted median global IP transit prices per Mbit/s from 2017 to 2020 in different regions. The figures illustrate how prices are falling rapidly, with over 50% reduction in prices over this period in several cities. They also show, however, that there are still large regional differences based on usage levels. Prices are significantly higher in Africa, Latin America and Asia than in Europe and North America (noting the different vertical scales on Figure 3.6 and Figure 3.7).

FIGURE 3.6: SELECTION OF CITIES IN EUROPE AND NORTH AMERICA: WEIGHTED MEDIAN GLOBAL IP TRANSIT PRICES PER MBIT/S, FOR 10GE CAPACITY LINKS [SOURCE: TELEGEOGRAPHY, 2021]

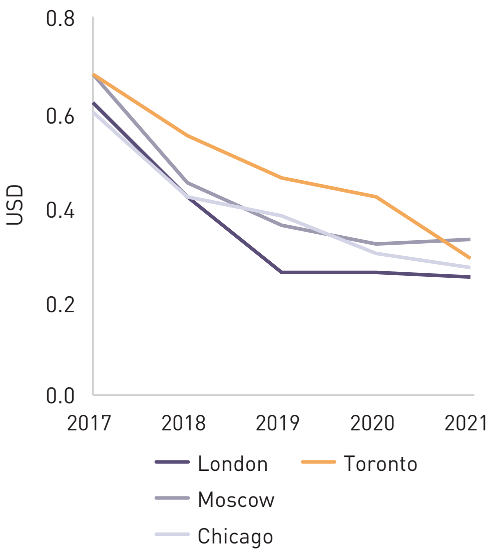
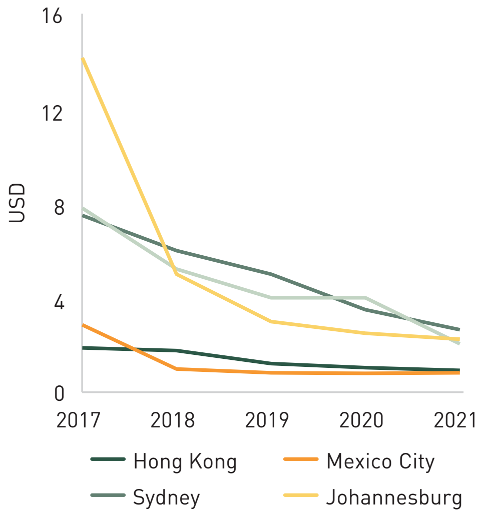


FIGURE 3.7: SELECTION OF CITIES IN ASIA, OCEANIA, AFRICA AND SOUTH AMERICA: WEIGHTED MEDIAN GLOBAL IP TRANSIT PRICES PER MBIT/S, FOR 10GE CAPACITY LINKS [SOURCE: TELEGEOGRAPHY, 2021]



The rapid increase in international capacity and falling prices are important to help understand how the Internet, based on best efforts rather than service guarantees, has been able to support an ever-wider range of applications, often with strict requirements on bandwidth and latency.[22] The availability of abundant capacity at affordable prices has helped make it possible to compensate for the lack of service guarantees through increasing capacity. Additional investments taken in end-user networks across the RIR regions in the past 15 years to expand capacity and help deliver traffic efficiently are shown in Figure 3.8 (for fixed fibre connections) and Figure 3.9 (for active 4G mobile connections).

FIGURE 3.8: FIBRE-TO-THE-PREMISES/BUILDINGS CONNECTIONS BY RIR REGION[23]
[SOURCE: ANALYSYS MASON DATAHUB, 2021]

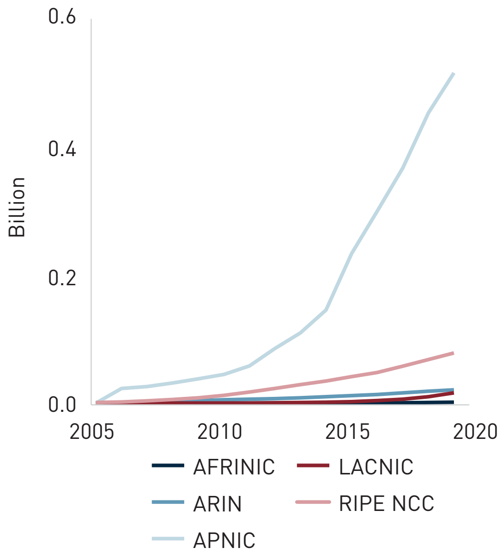
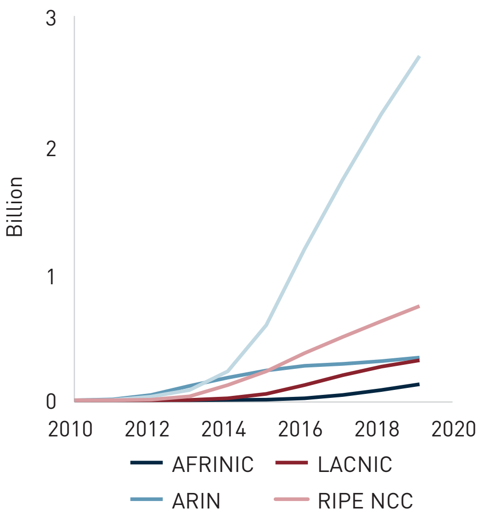


FIGURE 3.9: ACTIVE 4G CONNECTIONS BY RIR REGION[24]
[SOURCE: ANALYSYS MASON DATAHUB, 2021]



3.2.3 The interdomain routing system has scaled well

The key to making the network-of-networks principle work in practice is that the networks must be able to interconnect and exchange traffic with one another. This entails agreements between networks on how and with whom to exchange traffic, and a technical protocol for implementing these agreements and routing traffic between networks. The main forms of interconnection agreements are *peering* and *transit*, and we cover these in Section 4.2.2. In this section, we cover the underlying technical protocol and show how it has scaled with the growth of the number of networks and corresponding exponential growth in the number of potential routes for traffic.

Interdomain routing is the process of determining how traffic should be forwarded from one network (also known as a domain, or routing domain) to the next to ultimately reach its destination. The de-facto standard for interdomain routing is the Border Gateway Protocol (BGP), which is one of the most important Internet protocols. It was first standardised in 1989 as the successor of the earlier Exterior Gateway Protocol (EGP) and is currently in its fourth version (BGP-4).

BGP is a simple algorithm. A router sitting at the border of one AS makes a connection to a router sitting at the border of a neighbouring AS with which it has an agreement to exchange traffic. The two routers then exchange information about the routes (in the form of a

[22] As discussed in the box in Section 2.2.3

[23] RIR regional data built up from Analysys Mason Research country data

[24] RIR regional data built up from Analysys Mason Research country data

sequence of ASes, or paths) that they can make available to the other party. The paths announced are determined by the interconnection arrangement between the neighbouring ASes.[25] BGP has been central to the success of the Internet, in spite of its known flaws and perceived shortcomings.[26] It allows each AS to independently make agreements with other ASes to exchange traffic, and over which routes.

> *I like BGP because it's simple. You can control your outgoing traffic flow by putting some preferences on your BGP policy, but you cannot control your incoming traffic by BGP. And this limitation is actually a factor of the success of the Internet. You do not have full control yourself, so you need to co-ordinate with your peers. This c-oordination is an important factor for the Internet.*
>
> Yoshinobu Matsuzaki, Senior Engineer,
> Internet Initiative Japan

As the Internet has globalised and grown, the potential number of networks through which traffic could travel to reach its destination has multiplied. This could have made end-to-end interconnection slow to a crawl.
At the same time, the increased number of networks could overwhelm BGP with more and more announcements of new and changing routes. However, a set of simple and well-known routing configurations and operational practices have emerged that has made it simple for networks to connect to each other and allowed interdomain routing to scale well.[27] These are discussed in turn in the following subsections.

The average path length stays virtually constant as the Internet grows

Figure 3.10 shows how the average minimum path length between two discrete networks in the Internet has evolved over the last 20 years. The values shown in the figure are averaged over four different monitoring points across the Internet. As can be seen, the path lengths have been remarkably stable, almost constant at around 4, for both IPv4 and IPv6 over this period (left-hand axis). At the same time, the number of ASes has increased from around 10 000 in 2000 to more than 90 000 (right-hand axis).

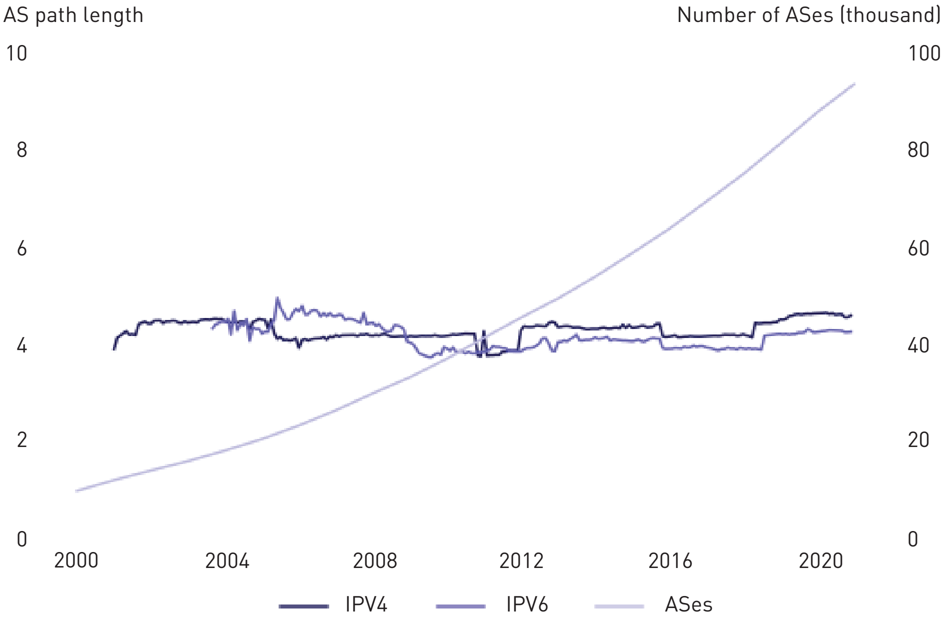
The forming of neighbouring connections between ASes defines the Internet topology. The stable path length is related to the small-world characteristic of the Internet topology, popularised by the famous theory of 'six degrees of separation', in this case demonstrating that there are typically four degrees of network separation between any two end users on the Internet. This shows that the system is flexible enough for new routes to easily be adopted that limit the growth in the path length.

[25] A **provider** AS (which sells transit to connect its customer AS to the rest of the Internet) will typically announce a path to all known destination networks, while a **peer** AS will only announce paths to its own destination addresses and those of its customers (see Section 4.2.2 for more details).

[26] Among the often-mentioned critiques of BGP is its vulnerability to prefix hijacking and how its flexible configuration options also increase the risk of misconfigurations.

[27] These practices include the emergence of open peering policies and Internet exchange points (IXPs), as discussed in Section 4.2.2, which enable multiple networks in a region or country to directly interconnect with one another, reducing the path length accordingly.

FIGURE 3.10: AVERAGE AS PATH LENGTH[28] AND NUMBER OF ASES [SOURCE: ANALYSYS MASON, NRO, RIPE RIS[29], 2021]



The rate of routing updates received by routers scales well as the Internet grows

BGP does not exchange messages unless needed, and a router only sends updates to its peers when there is a change in the route to a given prefix. Still, the scalability of BGP routing in terms of the update rate (known as churn) was, for a while, a concern in the networking community. There are several factors that could potentially drive the BGP churn rate to unmanageable levels. As the number of ASes keeps growing along with the number of AS-level interconnections, the number of potential paths that can be announced by a router also keeps growing. At the same time, there have been worries that the use of more selective announcements of certain routes for traffic engineering purposes was increasing. The fear was that the combination of these factors would lead to a situation where the BGP update rate would become unmanageable, so that routers would be overwhelmed by updates and constantly be recalculating their routing tables.

Figure 3.11 shows how the number of BGP updates received per day has evolved over the last 20 years for the IPv4 Internet.[30] The data shown is collected from a number of different routing monitors, which are routers that collect BGP updates in centrally located networks. The graph shows the churn rate for four large networks (using the left-hand vertical axis) compared with the growth in the routing table size (using the right-hand vertical axis)

[28] Calculated using data from RRC1 (London), RRC4 (Geneva), RRC5 (Vienna) and RRC6 (Tokyo)

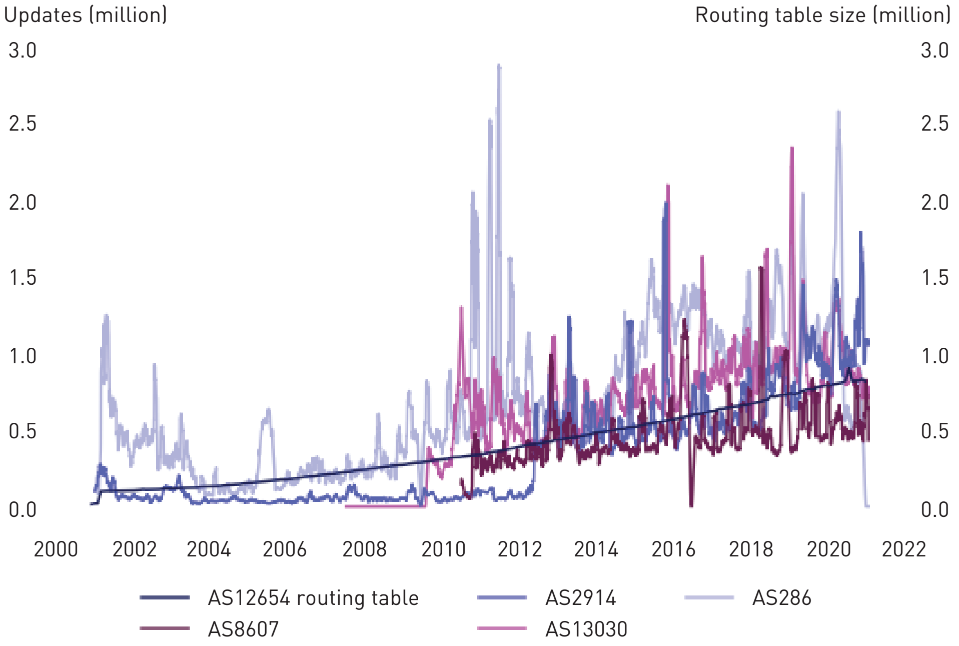
[29] The RIPE Routing Information Service (RIS) is a RIPE NCC service. With the help of network operators all over the world, RIS employs a globally distributed set of Remote Route Collectors (RRCs), typically located at IXPs, to collect and store Internet routing data. Volunteers peer with the RRCs using the BGP protocol and RIS stores the update and withdraws messages. RIS data can be accessed via:

- RIPEstat, the 'one-stop shop' for all available information about Internet number resources. RIPEstat uses individual widgets to display routing and other information
- RIS Live, a real-time BGP streaming API allowing server-side filtering of BGP messages by prefix or autonomous system
- RIS Raw Data, available for each route collector, with state dumps and batches of updates made available periodically
- RISwhois, which searches the latest RIS data for details of an IP address using a plaintext 'whois'-style interface. It is useful when querying RIS data using scripts.

The website can be found at https://www.ripe.net/ris

[30] The corresponding plot for IPv6 is similar.]

FIGURE 3.11: 30 DAY ROLLING AVERAGE OF DAILY BGP IPV4 UPDATE ACTIVITY FOR FOUR ASES, OVERLAID WITH THE IPV4 ROUTING TABLE SIZE[31] [SOURCE: ANALYSYS MASON, RIPE RIS, 2021]



There are several important observations to be made from the figure. First, there is significant variation in the update rate. The number of routing updates can be several times higher over a day, week or month, depending on the topology changes. Second, the update rate looks different depending on where it is measured. The figure shows significant differences between the four monitors, which are caused by local differences in connectivity around the monitored networks. However, for the purpose of our discussion, the most important observation from this figure is that the long-term trend in churn rates does not seem to exceed the growth in the routing table size. In other words, the average number of updates per prefix seems to be relatively stable over time. This is a critical element of the scalability of the Internet given the network-of-networks principle. If the frequency at which a prefix needs to be updated had instead increased as the Internet grows, due to a higher number of possible routes to each prefix, we could have ended up in a situation where the number of routing updates became overwhelming. Ultimately, this could have led to routing inconsistencies and the collapse of the routing system.

In addition to the results shown here, there are also other indications that the routing system scales well. In addition to the *frequency* of updates to routes, the time it takes for the change to propagate across the Internet, also known as the *convergence time* of BGP, is an important metric. It has been shown that this convergence time has not increased over the last ten years, despite the strong growth in the number of networks and in the number of available paths in the same period. This is also important for the stability of the Internet; if the time it takes to propagate a routing update through the Internet was growing as the Internet expands, we could have approached a situation where destinations were unreachable due to constant routing updates.

The properties of the Internet topology that have kept it scalable were not designed or planned by its creators, but have emerged as the Internet has grown, shaped by the Internet protocols and by the thousands of independent choices made by individual network operators. This emergent behaviour results from the flexibility of the Internet and has kept the Internet scalable, with improving performance, in the face of significant growth in networks and users.

[31] The four ASes used in this figure are as follows: AS8607 is TIMICO – Digital Space Group Limited; AS2914 is NTT Communications; AS13030 is Init7; and AS286 is GTT. The numbers are collected from monitoring sessions with AS12654 which is LINX – London Internet Exchange.

[32] Garcia-Martinez, A. and Bagnulo, M. (2019). Measuring BGP Route Propagation Times. IEEE Communications Letters, 23(12), pp. 2432–2436.

3.3 Implications

The Internet has scaled remarkably well over sustained growth in the number of users and their usage, based on the underlying design principles. New networks can easily continue to emerge, while the capacity of existing ones keeps growing, and measures of traffic exchange between the networks has remained relatively stable, while the number of interconnections keeps growing.

This is encouraging, as just over 50% of the world is online and many services can still be moved online around the world. The Internet will be required to continue to scale for more users more devices per user, and more usage per user. Addressing the challenges associated with these requirements will be helped by the flexibility of the Internet to new network technologies and its adaptability to new applications, as discussed in the following sections.

4 Success dimension: Flexibility in network technologies

> *Wajan" is wok in Indonesian, as in the dish used for cooking. Combining that with the Indonesian for parabolic, you get "WajanBolic", which you can find in many online shops. It's becoming commercial. They are selling actual woks to be used for WiFi networks.*
>
> Onno Purbo, lecturer at IBI Darmajaya

The network-of-networks principle allows networks to be developed and run independently and with different network technologies. From the beginning, the Internet's flexible protocol suite has supported a wide range of underlying physical network technologies. As the Internet has grown and matured, new underlying network technologies are added, from terabit optical networks to high-speed mobile and low-power Internet of Things (IoT) networks, alongside the likes of low-tech 'WajanBolic' or so-called WokFi networks in Indonesia.[33] Furthermore, these networks are all able to interconnect directly or indirectly to exchange traffic.

4.1 Observations

IP is a very flexible protocol, designed to be able to run over any underlying network.[34] As such, the Internet has been run over most existing physical communication networks, including traditional copper telephone networks, coaxial cable-TV networks, cellular mobile networks, various wireless radio networks, and satellite networks. The general nature of the Internet protocols has allowed the use of these networks to carry IP traffic, with increasing efficiency. Gradually, carrying IP traffic has taken over as the main purpose of most of these networks, as well as new ones purpose built for IP traffic. Today, based on its adaptability, IP dominates as the preferred technology to carry most types of traffic, including voice, video and general data transfers, as covered in Section 5.

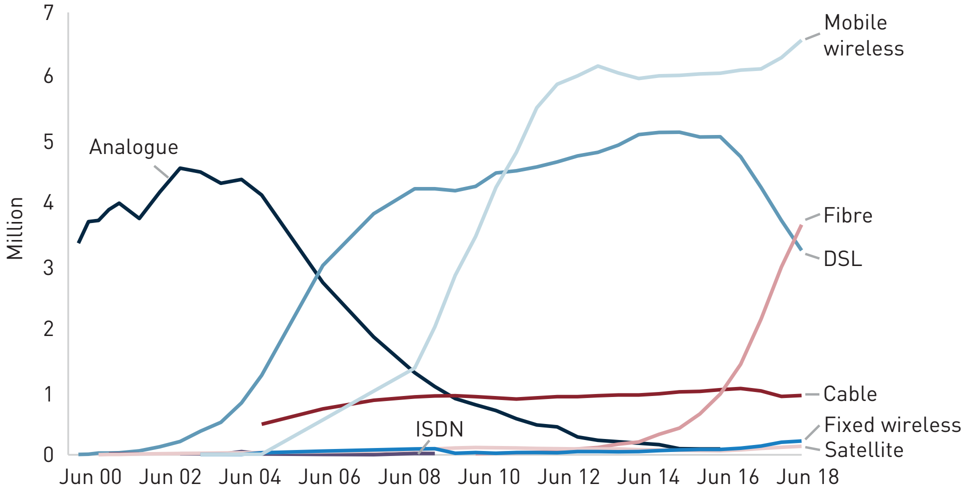
Figure 4.1 illustrates how the preferred Internet access technology has changed over time in Australia.[35] While the numbers in this figure are for a specific country, the general pattern of development that moves through a number of different access technologies with increasing bandwidth and quality is common to all countries. In the early days of mass-market adoption, the use of the existing fixed telephone networks using copper lines was predominant. This started first with Internet dial-up connections using analogue modems, which were then replaced with digital ISDN technology, and later with digital subscriber line (DSL) broadband. The latter was 'always on', without the need to establish a new connection when a subscriber wanted to access the Internet. The use of cable-TV networks for Internet access with DOCSIS technology allowed a boost in access speeds. The development of 3G, followed by 4G and now 5G, meant that mobile networks became a broadband access technology that is the main form of Internet access for many users. Today, fibre access networks have become the dominant fixed-line access technology in many markets. Given their capacity, fibre networks look poised to remain the dominant wire-based access technology for the foreseeable future.

[33] https://en.wikipedia.org/wiki/WokFi.

[34] This is famously illustrated in the experimental RFC1149, A Standard for the Transmission of IP Datagrams on Avian Carriers, available at https://datatracker.ietf.org/doc/html/rfc1149. The use of this standard unfortunately flies below the radar in Figure 4.1.

[35] Australia is one of the few countries that gathers detailed historical data that include analogue, ISDN and satellite.

FIGURE 4.1: NUMBER OF INTERNET SUBSCRIPTIONS BY ACCESS CONNECTION, AUSTRALIA
[SOURCE: AUSTRALIAN BUREAU OF STATISTICS, 2021]



Interestingly, countries with less developed copper-based telephone networks have often led the transition to mobile access to the Internet (as they did in case of voice telephony in the 1990s), and in some cases fibre. In Europe, for example, countries with the most developed fixed telephone networks, such as Germany, France and the UK, have been among the slowest in building out fibre access networks. In developing countries in Africa and parts of Asia with little existing fixed networks, mobile networks have quickly become the most important networks for accessing the Internet. The role of mobile networks for accessing the Internet is discussed more in the box below.

[8] https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf, https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2015_Year_in_Review.pdf and Cisco Annual Internet Report – Cisco Annual Internet Report (2018–2023) White Paper – Cisco

[9] RIR regional data built up from ITU country data. AFRINIC is the African Network Information Centre; ARIN is the American Registry for Internet Numbers; and RIPE NCC is the Réseaux IP Européens Network Coordination Centre.

The rise of Internet access over mobile networks and devices

The first cellular mobile networks were launched in the 1970s and 1980s and were analogue networks that could only transmit voice calls. The second generation (2G) mobile networks launched in the 1990s took the first small steps towards mobile data by allowing text messages and narrowband data transfers capable of tens of kilobits per second. The data capabilities of mobile networks were further improved with the advent of 3G, which was rolled out in the early 2000s. It was not until the introduction of 4G from around 2010 that the use of mobile networks for broadband Internet access really gained momentum. The introduction of 4G coincided with the widespread adoption of smartphones and other mobile devices, which revolutionised how people access and use the Internet.

FIGURE 4.2: FIXED (LEFT) AND MOBILE (RIGHT) BROADBAND SUBSCRIPTIONS BY RIR REGION[36]
[SOURCE: ANALYSYS MASON, ITU, 2021]

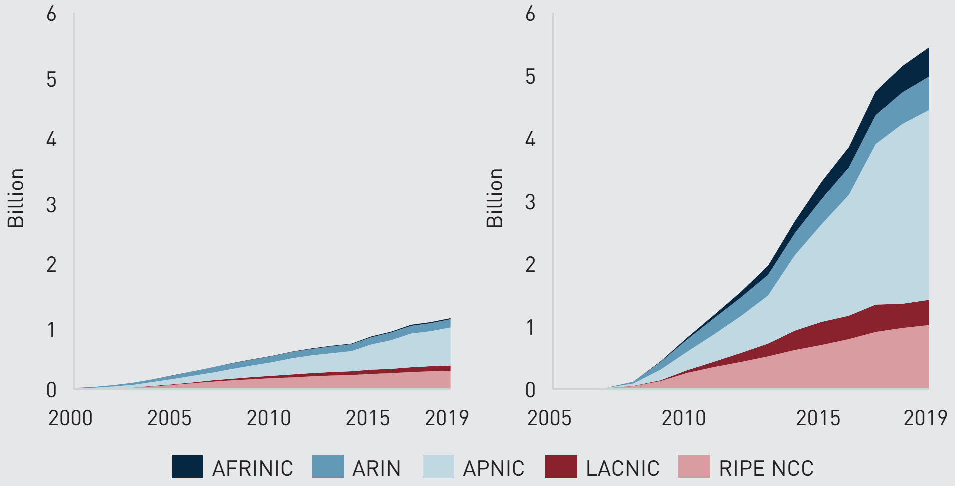


Figure 4.2 shows how the number of fixed and mobile broadband subscriptions have evolved over the last two decades. The graphs illustrate the enormous popularity of mobile broadband, which had almost five times as many subscriptions as fixed broadband in 2019. Smartphones have individualised Internet access and allowed a wide range of new Internet-based services tied to mobility. Mobile networks are now arguably the most important technology used to access the Internet. It is the primary access form for hundreds of millions of people, and in practice the only access form in some regions. In many countries, the rise of high-speed mobile broadband has allowed people to leapfrog several steps of technological evolution. However, as noted in Figure 3.3, total fixed data exceeds mobile data in all regions except Sub-Saharan Africa, reflecting the more data-intensive use of fixed networks.

[36] RIR regional data built up from ITU country data

4.2 Explanation

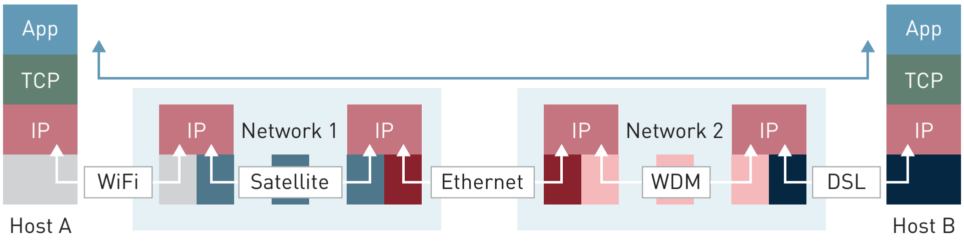
The flexibility of the Internet to accommodate new network technologies starts from the layering principle, which separates the end-to-end routing of IP traffic from the underlying network technologies. Also, the network-of-networks principle states that each network is independent and can use its own technology as long as it runs IP. These networks must interconnect to create the Internet, and this is done with BGP as the technology, and using arrangements negotiated between the networks to exchange traffic, including peering and transit. Finally, as the number of users in various regions has grown, the number of networks can grow correspondingly, although there are differences in the markets and consumer usage patterns across and within regions.

4.2.1 The layering and network-of-networks principles drive flexibility

The layering and network-of-networks principles are central for the flexibility in allowing different network technologies to be used for individual networks, as long as they run IP. As discussed in Section 2.2.1, in the hourglass depiction of the layered protocol suite, IP is at the waist of the suite and separates the applications from the underlying networks. As a result, an end-to-end path between two end hosts can run through networks that use different underlying technologies, as shown in Figure 4.3 below. The layering principle allows an application to communicate end to end over a transport protocol (TCP in this example), without any knowledge of the networks that actually carry the traffic.

Below IP, each network can use separate physical networks with their corresponding protocols. In this example, the end-to-end path goes over WiFi, satellite, Ethernet, wavelength division multiplexing (WDM) and DSL connections. The network-of-networks principle allows each network to independently choose which underlying technology is used internally, as long as it delivers IP traffic at the interfaces towards its neighbours.

FIGURE 4.3: ILLUSTRATION OF THE FLEXIBILITY TO CARRY THE SAME PAYLOAD INDEPENDENTLY FROM THE UNDERLYING NETWORK [SOURCE: ANALYSYS MASON, 2021]



Using IP as a layer that separates applications and end-to-end protocols from the underlying networks is not without cost. IP adds an overhead that must be carried in every data packet, and each IP packet must be processed in all IP routers along the path. Coupled with the best-effort nature of the Internet, this has historically given root to scepticism in traditional telecoms networks about the use of IP and the Internet to transport applications such as voice. Over time, however, the simplicity and flexibility of IP, and the lower prices that came from economies of scale, have been shown to outweigh these initial performance drawbacks.

4.2.2 The interconnection model is flexible

In addition to the flexibility in allowing content to be carried in an identical form over different networks, there is also flexibility in networks because they can operate independently from one another, with different technologies, users, devices and applications. However, to create a network of networks, by definition each individual network must interconnect with other networks – and in fact with every other network on the Internet, directly or indirectly.

We showed in Section 4.2.1 how different networks can physically connect – here we show how they interconnect to exchange traffic. This has a technical aspect and a commercial aspect. As noted above, BGP is used to technically exchange traffic between networks, using best efforts. When private ISPs began to emerge in the 1990s and the US government fully commercialised the Internet, a small number of commercial models governed interconnection between networks.[37]

Interconnection could have been regulated, as was the norm between telecoms operators at the time. A rate would be set for the operator originating the traffic to pay to the operator to transit or terminate the traffic. Regulating Internet interconnection would have resulted in trying to impose settlements between ISPs.[38] National telecoms regulators, beginning with the US Federal Communications Commission, chose to leave Internet interconnection decisions in private hands.[39] Settlements between providers could have become the default even in the absence of regulation, to recoup the costs of delivering traffic on behalf of other providers. However, the ISPs at the time developed another path, without settlements, which reflected the ethos of co-operation that pervaded the growing Internet.

Two main forms of interconnection that have emerged are known as peering and transit. To this day, these arrangements are commercially negotiated between the providers.

- In a *peering* arrangement, two providers agree to exchange their own traffic with each other. In general, if two networks send each other roughly the same amount of traffic (that is, they are peers), any settlements would cancel out, so peering is typically settlement free, as a pragmatic way to avoid the need to measure and bill for traffic. As networks have grown and evolved over time, peering has remained a constant, typically without even a formal arrangement.[40] Where there are differences in the amount of traffic exchanged, peering can sometimes include settlements – the constant is that the two parties would continue only to exchange traffic originating on their network and terminating on the other network.

- A peer will not allow traffic to transit its network to reach other networks, hence many peering agreements are needed to obtain access to the entire Internet, and only the largest networks are able to achieve this. As an alternative, a transit arrangement allows a smaller network to buy access to the entire Internet from a larger network, which delivers this traffic to and from its peers and any other *transit* arrangements it may have.

In order to facilitate traffic exchange, and make it efficient, Internet exchange points (IXPs) have emerged around the world.

[37] https://www.fcc.gov/reports-research/working-papers/digital-handshake-connecting-Internet-backbones

[38] One challenge would have been to determine who 'originated' the Internet traffic. This is easy with a telephone call, however with the Internet it is more difficult. For instance, is the originator the website that sends a page to a user, or the user who asked for it in the first place?

[39] There were a number of overlapping reasons for this. First, the emergence of the commercial Internet took place during a time when countries were beginning to liberalise their telecoms sectors, with a view to introduce competition that would reduce or eliminate the need for regulation. Second, having established regulations allowing regulated access to telecoms infrastructure, there were few entry barriers for ISPs to enter the market and compete. And finally, the Internet was nascent and a niche phenomenon at the time, only growing into the Internet as we know it today in part because of the decisions not to regulate it at the time. For a discussion of the situation in the US, see https://transition.fcc.gov/Bureaus/OPP/working_papers/oppwp31.pdf.

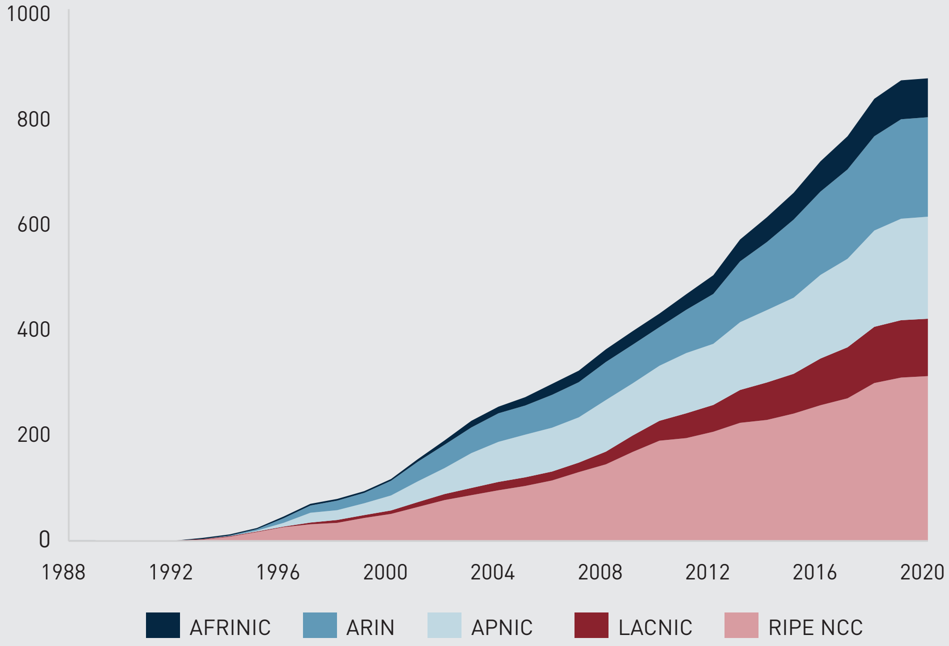
[40] Packet Clearing House, 2016 Survey of Internet Carrier Interconnection Agreements, accessible at https://berec.europa.eu/eng/document_register/subject_matter/berec/others/6574-2016-survey-of-Internet-carrier-interconnection-agreements

IXPs

As noted, peering is a bilateral arrangement between two providers. As the number of ISPs increased, first within the USA, it became costly to use a separate link to connect directly to every other ISP to exchange traffic. IXPs emerged, allowing providers to physically connect to a shared switch with one link, allowing them to peer and exchange traffic with any or all of the other connected providers. As the Internet began to grow in other regions, beginning in Europe, it was common at first to use international transit to the US to exchange even domestic traffic. IXPs soon began to emerge in Europe to localise traffic exchange, which lowered the cost and latency of using international capacity.

As the Internet globalised over time, this pattern continued. Local ISPs required international links to access the global Internet and would use those links to exchange traffic with other providers from their own country. This phenomenon is known as 'tromboning' because the traffic would go out and back, tracing the shape of the musical instrument. Through the work of the domestic and regional Internet community, along with the Internet Society and other international organisations, IXPs began to emerge in more and more countries, resulting in significant savings for ISPs and benefits for users.[41] Figure 4.4 below shows how the number of IXPs has grown, matching the growth across the regions.

FIGURE 4.4: NUMBER OF IXPS OVER TIME, GROUPED BY RIR REGION[42] [SOURCE: ANALYSYS MASON, WORLD BANK, 2021]



This growth highlights that there are very low barriers to entry for IXPs. They are generally run as non-profits, on behalf of the members of the IXP who are using it for peering. The main barrier in most countries is lack of awareness of and unwillingness to develop and join an IXP. The growth of content and content providers increases the benefits of IXP, as CDNs put caches in countries and deliver the content through the IXP.[43]

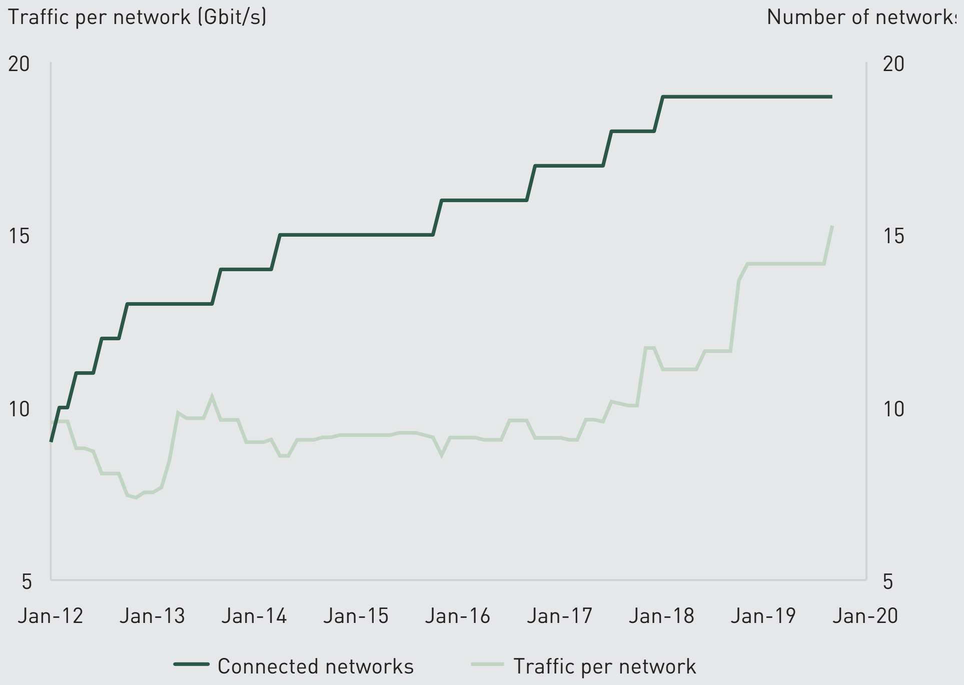
In addition to saving costs and lowering latency, the IXPs have a generative impact on traffic levels. Figure 4.5 below shows the growth in the number of networks connected to the IXP in Vietnam, VNIX, and the amount of traffic exchanged per network. While one would expect the total traffic to increase as a result of an increased number of networks exchanging traffic, the average traffic per network also increases. At VNIX, as the number of networks doubles over the time period, the traffic per network steadily increases, showing how more content is delivered through the IXP as it grows, and the resulting increase in usage.

[41] For more information, see https://www.internetsociety.org/issues/ixps/

[42] RIR regional data built up from World Bank country data

[43] As discussed in Section 5.2.3

FIGURE 4.5: NUMBER OF MEMBERS, AND TRAFFIC PER MEMBER FOR VNIX [SOURCE: ANALYSYS MASON, VNNIC, 2021]



The agreed interconnection in the form of peering or transit relationships is technically implemented in the BGP interdomain routing protocol. The business relationship between two neighbours is translated into a set of routing policy rules in BGP, which determines which routes are announced to a neighbour and consequently which traffic is accepted. A network will potentially receive several alternative routes to a given destination, from different neighbours, and it is free to select whichever route it prefers. A network would usually prefer routes through peers over routes through transit providers (since there is no payment involved in the former) and would prefer shorter routes over longer routes. As noted in the box above, by concentrating peers together, an IXP can improve interconnection outcomes.

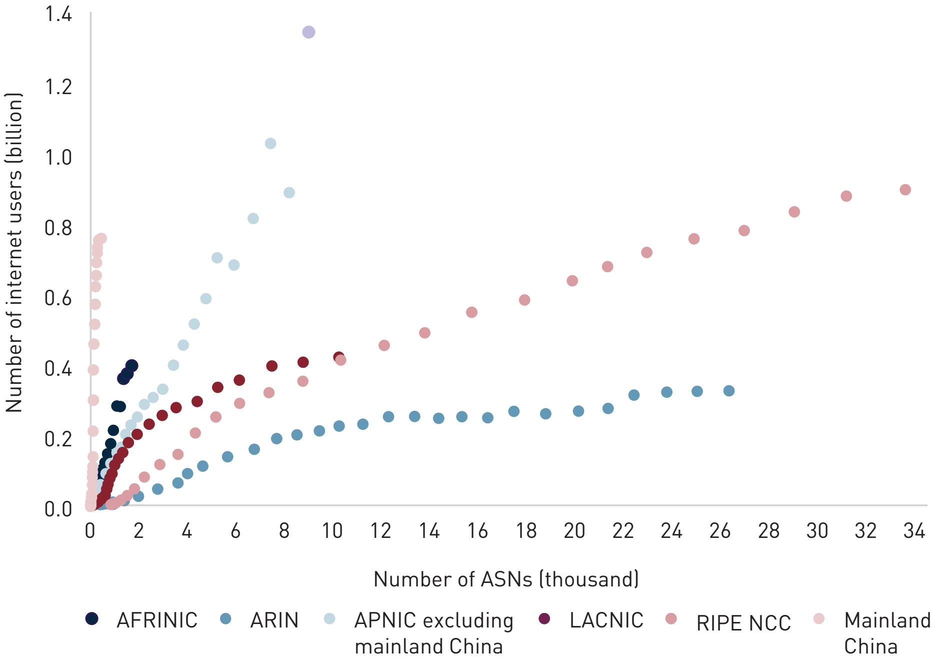
BGP is a simple and flexible routing protocol, which is well suited to implement the common interconnection arrangements between networks. BGP also has limitations that have been significant for the way interconnection works. For example, it is hard to use BGP for so-called inbound traffic engineering, where a network expresses over which (of several) routes it prefers to receive traffic. Several techniques exist (including the use of DNS and the announcement of more specific IP prefixes), but often this has to be implemented through agreements between neighbouring networks. This example highlights the technical collaboration that is often needed between neighbouring networks to secure a smooth operation. This collaboration at a technical level, even between competitors, plays an important role in the operations of the Internet.

The interconnection model is flexible, and based on collaboration and co-operation between operators, even those who compete. It has proven difficult to establish interconnection models that account for different service classes or even service guarantees across networks, as discussed above. A peering or transit agreement can specify that a certain capacity should be available in the neighbouring network, but it has proven hard to establish a system where such guarantees can be extended across several networks to form end-to-end guarantees. We address the implications of this further in Section 7.

4.2.3 Regional differences drive the growth of network numbers

ISP networks connected to the Internet require an AS number to exchange traffic using BGP with peering or transit. However, other organisations with their own networks, including universities, government agencies and enterprises, can also apply to their RIR for AS numbers, for instance as a precursor to be able to join an IXP to exchange traffic. Figure 4.6 shows there is a positive correlation between the number of users over time and the number of ASes assigned. An increased number of users leads to organisations having a more pronounced online presence, and vice versa.

FIGURE 4.6: NUMBER OF INTERNET USERS VERSUS ASES PER RIR REGION[44] (MAINLAND CHINA IS SHOWN SEPARATELY), 1990–2019 [SOURCE: ANALYSYS MASON, ITU, NRO, 2021]



However, the path of growth is not uniform across the regions represented by the RIRs. In particular, we can see that North America has the highest number of ASes per users, followed by Europe, Latin America, Asia–Pacific, and then Africa. This could be a factor relating to the maturity of markets, as growth in users outpaces organisations' adoption; it could also be a reflection of the Internet ecosystem of a region. For instance, Africa has relatively few IXPs, where an AS number is required to peer, and thus there is less incentive for organisations to obtain an AS number.

We note that there can be significant differences within a region. For instance, Mainland China has significantly fewer ASes per user than the rest of Asia–Pacific, while Asia–Pacific has less than other regions, even without including Mainland China. This highlights the flexibility of the Internet, and that adoption and growth of the Internet is possible with fewer AS numbers compared with other regions.

4.3 Implications

From a relatively uniform starting point, where the Internet was primarily accessed over legacy copper telephone networks using dial-up modems, the Internet has shown flexibility to accommodate a large and growing variety of networks, ranging from the almost uniform coverage of fibre-to-the-home (FTTH) broadband in Singapore to WokFi in neighbouring Indonesia. These networks can interconnect with one another and exchange traffic directly or indirectly using arrangements that have been developed by networks themselves such as peering and transit, at IXPs owned and operated by their members for traffic exchange. This flexibility allows new technologies and networks to emerge to meet the needs of their users, while being able to interconnect and exchange traffic with the rest of the Internet.

[44] RIR data on Internet users built up from ITU country data

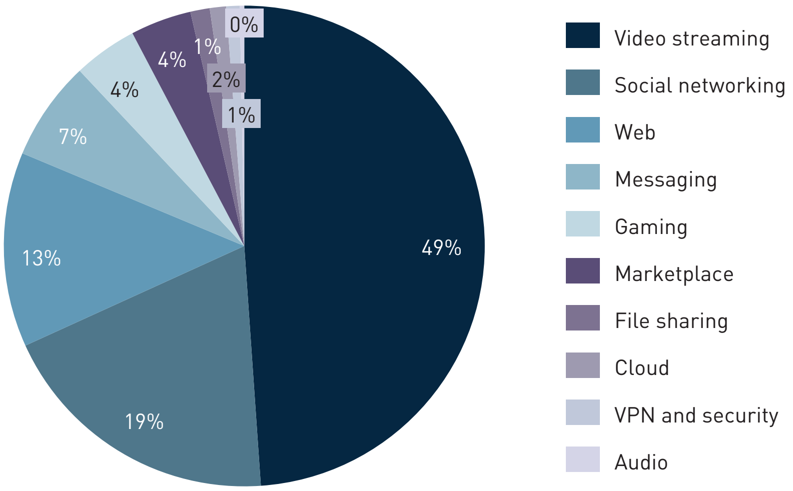
5 Success dimension: Adaptability to new applications

5.1 Observations

As the early Internet grew, it transformed. Users began messaging each other and email was born. The Internet was no longer just a means of accessing computing power and sending files, but became a way of communicating and socialising. The evolution of TCP/IP and increasingly affordable devices led to Internet use within academic and technical communities. As it grew, two application platforms emerged over several decades, and helped drive the uptake of the Internet. The first was the World Wide Web, which emerged in the early 1990s, and the second was mobile app stores, which became popular in the late 2000s. Both platforms were made available owing to the openness of the Internet.

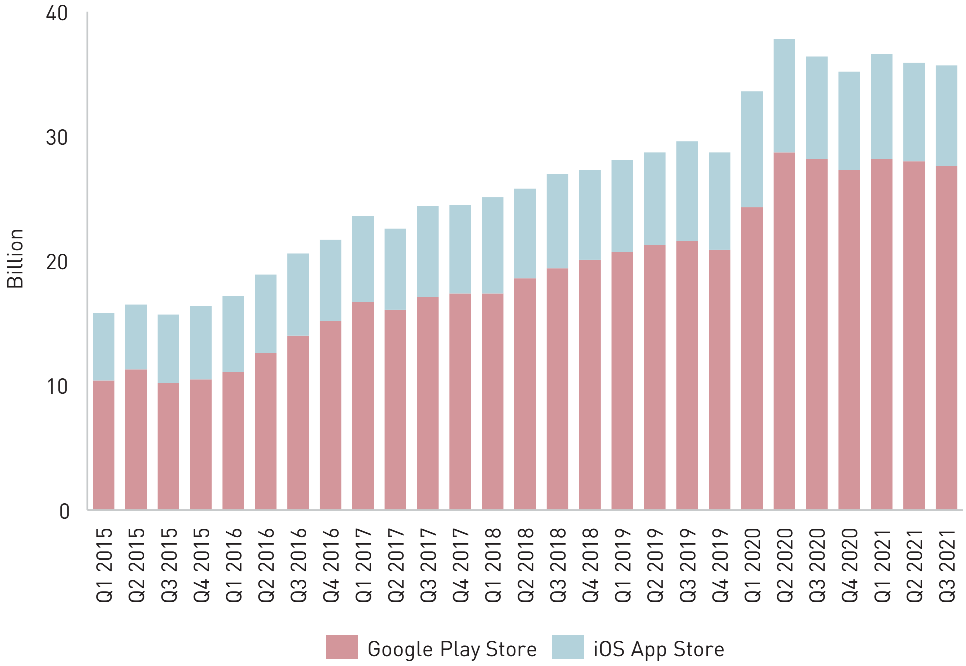
It was only after Tim Berners-Lee proposed a 'web of information' at CERN in 1989, developed HTTP and created the World Wide Web as an open platform that the Internet became easily accessible to non-technical audiences for the first time, through web browsers, starting with Mosaic in 1993 and then the Netscape Navigator browser in 1994. The World Wide Web enabled a wide range of new uses and applications of the Internet, from multimedia websites to social media, as shown in Figure 5.1.

FIGURE 5.1: TOP TEN CATEGORIES OF APPLICATION BY GLOBAL TRAFFIC SHARE, 2021 [SOURCE: SANDVINE, 2021]]



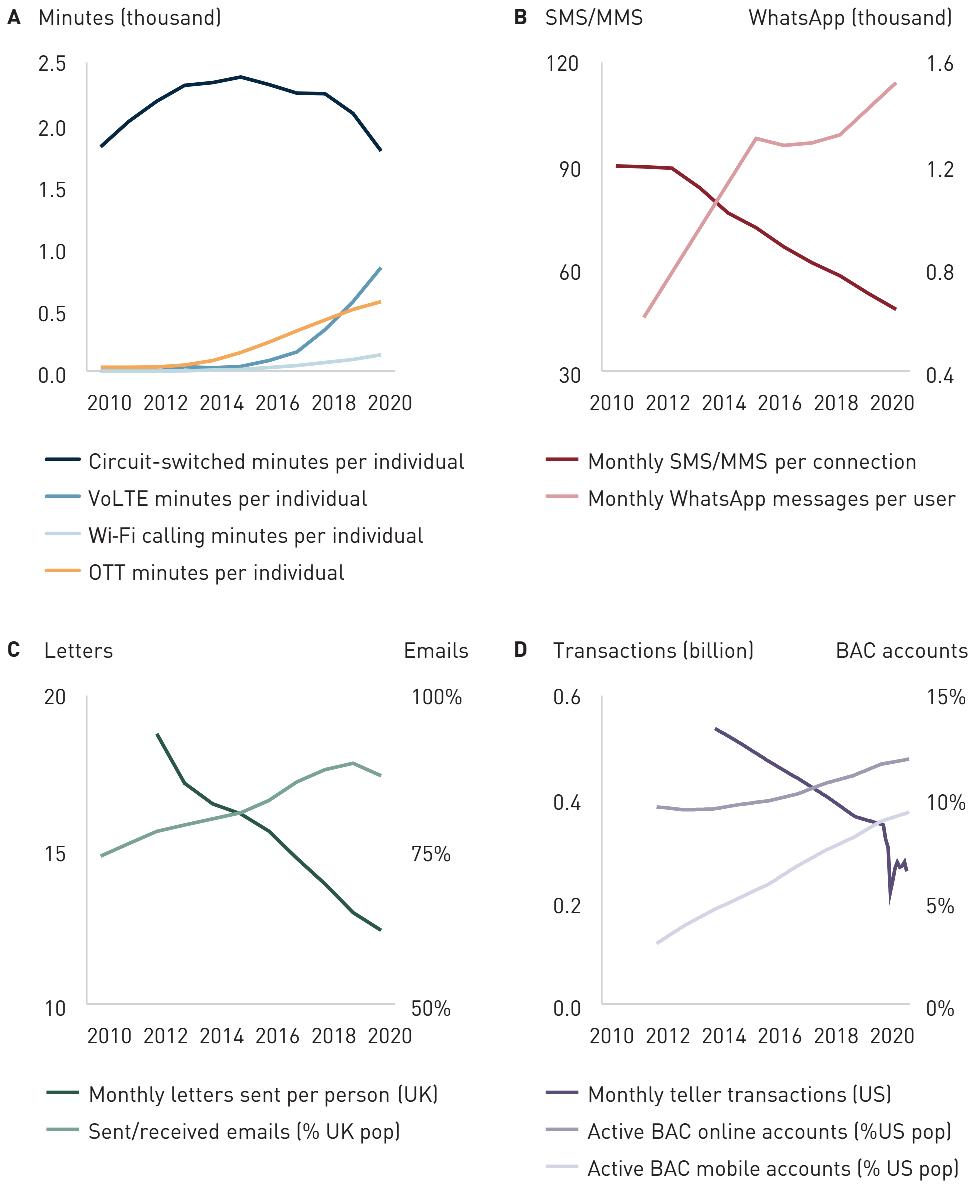
The Apple iPhone introduced smartphones to broad audiences, but the opening of the Apple App Store to third-party apps in 2008 had a fundamental impact, driving the adoption and usage of Internet access over mobile networks. Other companies developed their own smartphones and app stores, notably including Google's Android operating system and the Google Play store, and apps became a leading way for many to access the Internet. The apps effectively act as an entry to a website, while incorporating attributes of smartphones and mobility, such as easy access to cameras and location awareness. Figure 5.2 shows how popular apps have become – there have been more than 35 billion downloads in the Google Play and Apple App stores in the first quarter of 2021 alone. The sharp increase in downloads in 2020 corresponds with the lockdowns during the Covid-19 pandemic and increased use of the Internet, as described further in Section 6.1.

FIGURE 5.2: QUARTERLY NUMBER OF APP DOWNLOADS GLOBALLY [SOURCE: SENSOR TOWER, 2021]



Over time, the Internet has become the primary vehicle for delivering many services that existed long before it came into existence. From voice calls to banking, reading the news to watching movies, grocery shopping to schooling, booking vacation lodgings to getting a ride, services have converged on the Internet, and it has adapted accordingly. Figure 5.3 illustrates how the Internet is taking over as a delivery platform for voice calls, messaging, mail and payments.

FIGURE 5.3: CONVERGENCE OF VOICE (A), TEXT (B), LETTERS (C) AND BANKING (D) SERVICES
[SOURCE: ANALYSYS MASON, WHATSAPP, EUROSTAT, OFCOM, NOVANTAS, BANK OF AMERICA, WORLD BANK, 2021]



5.2 Explanation

The Internet has been able to adapt to support a large, diverse, ever-growing collection of applications. As the requirements of applications have evolved, the end-to-end protocols at the transport layer have also progressed accordingly. The last few years have seen significant changes in the higher-layer Internet protocols responsible for flow control and security. Interestingly, several of these changes are driven by content providers, as they start to become involved in how their content is delivered.

5.2.1 Layering and the end-to-end principle are central to support a wide range of applications

> *The invention of new services and fundamentally new activities on the Internet was built in from the very beginning, when we defined the original architecture of multiple layers. I remember saying these layers are provided for a convenience in a sense, rather than as an imposition and you're welcome to add layers to impose intermediate layers or to ignore them and go all the way down to essentially the bottom, which in the case of the current Internet that would mean going down to the IP layer.*
>
> Steve Crocker, Internet Pioneer

The general nature of the most central Internet protocols has proven to be successful in supporting a wide range of diverse applications. IP offers a general best-effort service and leaves it to the applications and protocols in the end systems to implement more advanced functionality such as security, congestion control or other forms of co-ordination. This model has created a common network where applications can access other applications and services and use them in developing innovative new offerings.

The openness of the Internet means that anyone can innovate and provide new applications, to be made available to anyone else to adopt. These new applications do not require any changes in the network to function. Rather, the intelligence to drive the applications is embedded in the end-user devices, which are steadily increasing in variety and power. The openness of the standards process means that protocols can be adapted or new ones developed in order to optimise entire new categories of applications, as we explore below.

As a result, and in the context of the end-to-end principle, applications and protocols in the end systems can treat the Internet as a non-discriminatory entity that will move traffic regardless of its content. The increased use of various types of middleboxes, which filter traffic based on type, has been perceived as a threat to this principle. While this is a concern, it is also clear that the Internet continues to show the ability to adapt and to innovate to meet the changing requirements of applications, as discussed below.

> *I think we have hit the apex of the strength of intermediary network functionality and now we're going back down because of the rise of end-to-end encryption. So we're in the process of figuring out how to do all the things we used to do with the network intermediaries intercepting clear text, in light of the fact that that's not possible anymore. So I actually think the end-to-end principle is highly relevant today and more relevant than it was 10 years ago, because if you're building on top of QUIC or HTTP as many applications are, and even with Web RTC, you have to know how to do everything that you need to do from the end points, because there's no other choice.*
>
> Dr Alissa Cooper, VP/CTO at Cisco

5.2.2 Protocols are changing to support new applications

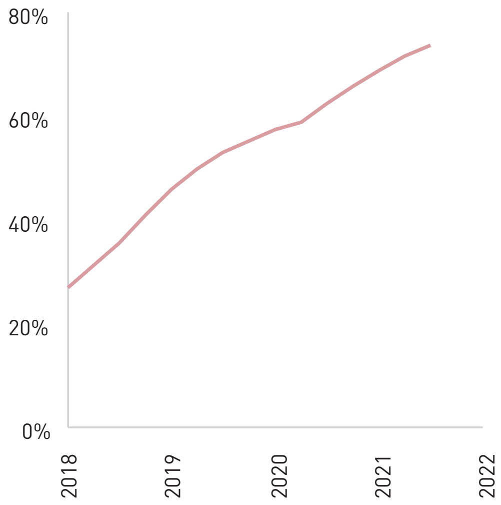
> *Layer separation I think has been a huge contributor to the development of the Internet, simply because it allows physical infrastructure to evolve while the protocols stay the same or content to evolve while the underlying transport stays the same. But at the same time, we've seen the limitations of that, particularly in the recent publication of the QUIC protocol which reveals the limitations of the use of TCP/IP and its need to be aware of the HTTP transport layer above it. To some degree, every rule has a kind of qualification that there are limits to it.*
>
> Steve Song, founder of Village Telco

The end-to-end protocols that control the flow of traffic between two communicating end hosts are called transport protocols. Together with application-layer protocols such as HTTP, they are used by applications to set up connections and move data across the Internet. Traditionally, the dominant transport protocol has been the TCP, which accounted for about 90% of all Internet traffic in the period from 2000 to 2010.[45] TCP offers a reliable packet transport, meaning that it ensures that all packets sent will arrive at the destination, and they will arrive in the same order. In addition, TCP performs congestion control, by adjusting the sending rate of data to the available capacity of the end-to-end path. TCP has been in constant evolution since its introduction, with several variants developed to make more efficient use of bandwidth or reduce delays.

TCP is well suited for moving asynchronous content such as files, web pages, videos or software across the network. Interactive content, however, such as voice calls or video conferencing, have different requirements. For interactive applications, in-order arrival and lossless transport is often less important than timely delivery. Delaying the arrival to ensure the same order or waiting for a lost packet can result in delays which make a conversation difficult to follow. Instead, such applications typically rely on the User Datagram Protocol (UDP), which is a simpler transport protocol without congestion control mechanisms or guarantees with respect to lossless and in-order delivery.

While the protocol mix at the transport layer was relatively stable in the 1990s and 2000s (save modest changes and upgrades in TCP), the last few years have seen more significant changes in this area. These changes are driven by the widespread adoption of Transport Layer Security (TLS), combined with the introduction of QUIC as an alternative to TCP. TLS is a security protocol that uses cryptography to secure communication between end points. End-to-end encryption using TLS has become increasingly popular, in particular for HTTP web traffic, which is known as HTTPS when it is encrypted. As shown in Figure 5.4, the fraction of websites using HTTPS has increased to almost 80% in a few years.

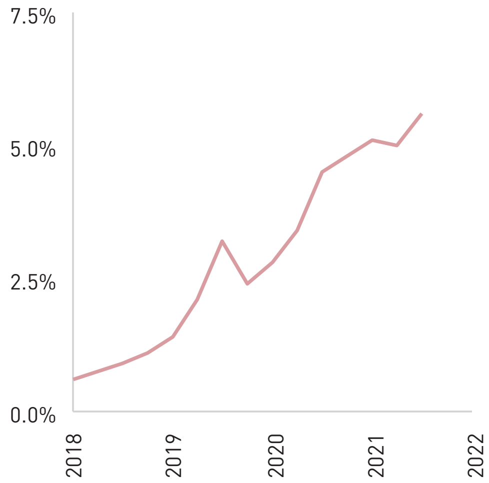
FIGURE 5.4: PERCENTAGE OF WEBSITES USING HTTPS [SOURCE: W3TECHS, 2021]



More recently, Google developed QUIC to reduce the delay in reaching its services; QUIC was first supported in its Chrome web browser. It was later standardised in the IETF as RFC9000, and is now supported by most major browsers, several popular apps and many important content providers. As shown in Figure 5.5, the fraction of websites that use QUIC is still modest, but quickly rising. QUIC is designed to work closely with HTTP. One of its main aims is to reduce latency for traffic that uses TLS for end-to-end encryption, and in particular for HTTPS traffic. This is achieved by making the exchange of cryptographic keys part of the initial connection set-up.

[45] FLee, D., Carpenter, B. E. and Brownlee, N., Media Streaming observations: Trends in UDP to TCP ratio, International Journal on Advances in Systems and Measurements, 3(3&4), 2010, pp. 147–162.

FIGURE 5.5: PERCENTAGE OF WEBSITES USING QUIC
[SOURCE: W3TECHS, 2021]]



The introduction of QUIC demonstrates how protocols have changed in response to altered requirements from applications, in this case the need for encryption and reduced delays. It shows how protocols are not strictly organised in layers as per the layering principle, but rather in a more composable set of building blocks. QUIC, while performing much of the same tasks as TCP, operates on top of UDP, which is another transport layer protocol. Also, QUIC is not implemented as part of the operating system in a host, but as part of the application, again demonstrating the flexibility of the Internet.

The changes in the transport layer with the growing use of HTTPS and QUIC are also interesting in light of the end-to-end principle. NATs and other network middleboxes sometimes filter out traffic from unknown or modified protocols, reducing the flexibility to make changes to transport protocols, a phenomenon known as protocol ossification. Protocol ossification, it is feared, can limit the Internet's ability to adapt. QUIC circumvents middlebox filtering by using UDP and encrypting traffic within this protocol. End-to-end encryption may in this way contribute to making the end-to-end principle more relevant and limit the impact of middleboxes.

5.2.3 Internet companies' business models to deliver content are evolving

The changes in applications over the years and the growing geographical scope of the Internet have led to new business models and investments on the part of Internet companies providing content and applications to help deliver the content.[46] Initially, when most content was text based and communications were asynchronous, content and applications could be hosted in one location, and transit arrangements could be used to access and deliver content, even as the Internet began to grow internationally. However, content (particularly video) has grown in volume, while real-time applications have emerged. The cost of delivering high-bandwidth content across long distances became significant, and the latency limited the demand for and usefulness of real-time applications. This required new approaches to delivering content and applications.

Broadly speaking, content can be divided into static and dynamic content, which impacts how they are distributed. Static content, such as videos, does not change over time or by user and can be stored in caches closer to the end users, and CDNs have emerged to deploy caches and distribute content.[47] Initially the CDNs were independent players distributing the content of other companies; later the largest content providers began to develop CDNs to deliver their own content. Dynamic content, such as social media and video conferencing, cannot be hosted in a cache. However, the largest providers have begun to establish points of presence (PoPs) around the world to deliver dynamic content efficiently and also to fill their caches with static content. The result is a better user experience with lower latency, and lower cost for ISPs who do not have to access content using international transit capacity.

These caches and PoPs are known as 'edge nodes' in the context of CDNs. The edge nodes are served by content and applications stored in data centres in the core of the network, mainly in the USA and Europe and, increasingly, Asia. In order to distribute content among data centres and from there to the edge nodes, a number of content providers have become the biggest

[46] For an overview of Internet companies' infrastructure investments, see https://www.analysysmason.com/consulting-redirect/reports/online-service-providers-Internet-infrastructure-dec2018

[47] For an overview of the benefits of caching, see https://www.analysysmason.com/consulting-redirect/reports/benefits-of-caching-may20

investors in submarine cables connecting continents and coastal countries within continents, as noted in Section 3.2.1.[48] The result is that an increasing amount of content is delivered over 'captive' networks owned and operated by the content providers themselves. This has led Geoff Huston, Chief Scientist at APNIC, to discuss the 'death of transit'.[49] We take up the implications of this in Section 7 where we discuss future challenges to the success of the Internet.

5.3 Implications

Over the years, the Internet has adapted to new types of applications that would have not been imaginable in the early days of text-based services. This includes the emergence of new platforms including the World Wide Web and then mobile apps. However, it is important to note that the Web is not the Internet, it uses the Internet to facilitate access to applications and content, and that mobile Internet is still the Internet, with new features based on smart devices and mobility.

Based on the layering principle, the protocols for transporting traffic over these networks have evolved and new ones have emerged independently of the networks over which they operate. These transport protocols are adapting to changes in applications or have enabled the widespread emergence of new applications.

The change of applications also changed the role of Internet companies across the layers of the Internet. First, as noted above, Google helped to develop QUIC (and then standardised it) to meet the requirements of its services at the transport layer. At the network link layer, several Internet companies have begun to invest in submarine cables and edge nodes, in order to deliver their own content and applications instead of using transit services. In Section 7, we examine whether the growth of application providers and their change in role could fundamentally alter the Internet.

[48] https://datacenterfrontier.com/more-than-8-billion-in-subsea-cable-investment-in-the-pipeline/

[49] Huston, G,. "The death of transit", 28 October 2016, at https://blog.apnic.net/2016/10/28/the-death-of-transit/

6 Success dimension: Resilience in the face of shocks and changes

6.1 Observations

The Internet has sustained significant growth, thanks to its flexibility and adaptability. The Internet has also proven to be resilient over time: it has continued to operate and offer a sufficient service level in the face of noteworthy internal changes and external challenges. The very growth of the Internet in terms of the number of users and their usage, and the emergence of new networks and applications is an indirect confirmation of resilience. While there are technical challenges facing the future success of the Internet, as described in Section 7, individuals, enterprises, governments and others continue to move important social and economic activities online, taking the availability and reliability of the Internet almost entirely for granted.

Over the years there have been many concerns expressed, with more or less validity or justification, about the possible collapse of the Internet, be that due to uncontrolled congestion, the collapse of the interdomain routing system or security threats. By 1990, the collapse of the routing system was being predicted, along with exhaustion of IPv4 space under the 'classful' architecture used at the time. These were very real threats and precipitated the development of classless inter-domain routing (CIDR) and the RIR system.

In 1995, Robert Metcalfe, the inventor of Ethernet and founder of 3Com, published a column predicting that the Internet would collapse in the face of challenges including congestion. In 1996, he conceded he had been wrong, and publicly followed through on his promise to eat his column. While there are regular events where significant services are affected, or where capacity is reduced in certain geographical regions, the Internet as a whole has been able to handle a range of challenges, some of which are highlighted in the following box.

Internet threats

In 1988, in an early inadvertent illustration of threats to the Internet, Robert Morris developed a program to assess the size of the Internet. The program exploited vulnerabilities in several Unix programs, and weak or missing passwords, to spread from system to system. It did not cause damage to files but it was designed to self-replicate 14% of the time, even on systems where it was already present, which caused computers to slow down to the point of becoming unusable – the first viral denial of service (DoS) attack. The Morris Worm was never intended to be a threat, and yet it ended up disabling as many as 10% of computers connected to the Internet at the time. Due to the novel nature of the attack it took some institutions as much as a week before they were able to get systems back online, however the Internet bounced back, the importance of cyber security was understood and work began on counteracting such threats.

Since then, cyber attacks have become malicious and more sophisticated. In the 1990s, viruses evolved to be cross-platform and polymorphic, toolkits were created that enabled easy mass creation of viruses for unskilled programmers, and email became the preferred vector of attack. Mass-email viruses are embedded in email attachments, or within the email itself, and once executed they infect the computer, and the virus sends copies of itself to email addresses from the user's address book. One such virus was the Love Bug virus in 2000 (so called because the subject line of the email was 'ILOVEYOU'), which was designed to steal passwords. Within 10 days, 45 million infections were reported, and users including the Pentagon and the CIA had to shut down their email systems for several hours to respond to the incident.

[50] At the annual conference where he had originally made his prediction, Metcalfe ripped up his column, put it in a blender with water, and ate the result; see https://1995blog.com/2015/12/03/prediction-of-the-year-1995-Internet-will-soon-go-spectacularly-supernova/

Mass-email viruses are still common today, but so too are more sophisticated viruses that include multiple, novel infection vectors, active attacks against anti-virus programs and dangerous payloads including ransomware, spyware and trojan horses. Moreover, attacks have evolved to receive real-time command and control. This has been harnessed to create armies of botnets which can be used to mine for cryptocurrency, steal information, orchestrate distributed denial of service (DDoS) attacks and can even be rented out to other hackers. Although DDoS attacks target a specific resource or network, DDoS attacks against the root servers used to map domain names to IP addresses have the potential to disrupt global DNS operation, affecting a multitude of networks. However, the caching and redundancy features of DNS make it resilient. There have only been a few notable successful attacks, and their impact was nearly imperceptible to end users.

With the rise of IoT and the rapidly increasing number of devices connected to the Internet, botnet armies can now number in the millions. In 2016, Mirai malware began infecting devices using 61 common default username/password combinations, which created an army of IoT botnets, including a large number of CCTV cameras and routers. The creators used their army to orchestrate DDoS attacks on Minecraft servers, and then released the code. A month later, someone else used it in three DDoS attacks against an infrastructure company, Dyn, using tens of millions of devices, and reaching an attack strength of 1.2Tbit/s, bringing down websites including Amazon, Netflix, PayPal, Spotify and Twitter. Measures have been taken to dismantle botnets, and new devices now often prompt users to change default passwords. However, many existing botnets are effectively undetectable, while hackers are always finding new ways to gain access to devices.

Not all events are man-made, deliberate or malicious. Natural disasters such as typhoons and earthquakes have also resulted in broken submarine cables and reduced Internet access, and in the past even sharks posed a threat to the Internet by chewing through cables.

Configuration complexity and errors can also cause network problems. The latest high-profile example of this occurred on 4 October 2021, when Facebook's servers were unavailable globally for almost six hours. The outage was reportedly caused by a network failure that caused Facebook's DNS servers to withdraw their addresses from the BGP routing table.[51] A lot of the tools that are needed to fix the problem depend on the DNS itself, which contributed to prolonging the recovery time.

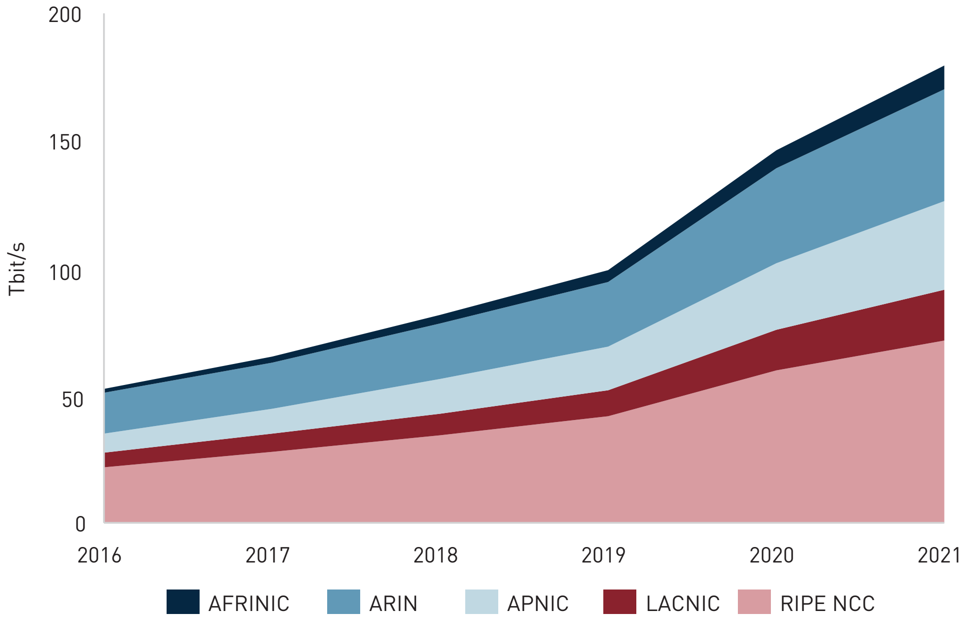
It is unavoidable that hackers will continue their attacks, and human error and accidents will continue to disrupt operations and infrastructure. Nonetheless, the resilience of the Internet drives changes that will help to avoid and mitigate these threats, while the Internet as a whole remains resilient to all the changes and threats.

> *If you think about the things that happened during the pandemic, the ability to scale up the infrastructure on incredibly short notice and grow networks in the span of a month the same amount that they usually grow in a year or two, and be able to support every new need that people had, whether that's video conferencing or sharing virus contagion data. That really speaks to the power of the generality of the network. If you needed to do that with the phone network, you would not be able to do, because it just doesn't support the kinds of things that people want to be able to do.*
>
> Dr Alissa Cooper, VP/CTO at Cisco

[51] https://engineering.fb.com/2021/10/05/networking-traffic/outage-details/

The impact of the Covid-19 pandemic serves as a good illustration of the resilience of the Internet. When the first reports of the virus began to emerge in January 2020, no one could have predicted its profound and far-reaching impact, including the role the Internet would take in response. With more than half of the global population in lockdown by April 2020,[52] Internet usage skyrocketed. For example, a European ISP reported a 50% increase in traffic in April 2020 compared with April 2019, double the usual year-on-year increase.[53] The same trend could be seen globally: Figure 6.1 shows how the average international traffic increased by 47% in 2020, reaching 146Tbit/s, a clear jump from the growth rate in the previous years.[54] The reasons for the increase in usage are clear and ongoing. Remote working and schooling, social video calls, more free time to spend online, all contributed to the rise in traffic. After the shock caused by the pandemic, the growth rate from 2020 to 2021 seems to have returned to a more normal level.

FIGURE 6.1: AVERAGE INTERNATIONAL TRAFFIC BY RIR REGION[55] [SOURCE: ANALYSYS MASON, TELEGEOGRAPHY, 2021]



The types of applications used also changed in line with the changes in activities due to lockdowns and social distancing. Use of business, medical, health and fitness, education and gaming applications increased substantially over pre-Covid 19 levels, while use of weather, sports, navigation and travel applications dropped significantly, as illustrated in Figure 6.2.

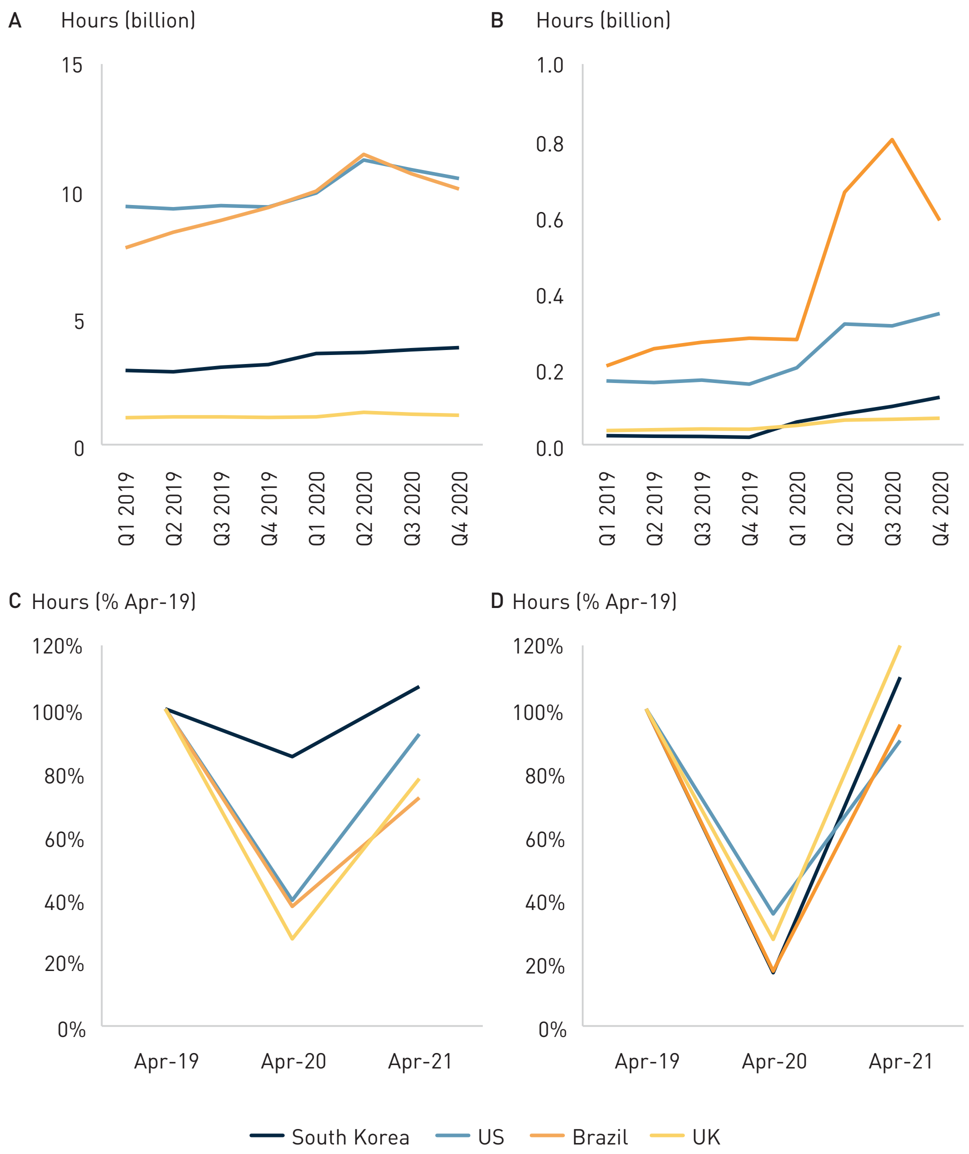
[52] https://www.euronews.com/2020/04/02/coronavirus-in-europe-spain-s-death-toll-hits-10-000-after-record-950-new-deaths-in-24-hou

[53] Feldman, A. et al., A year in Lockdown: How the Waves of COVID-19 Impact Internet Traffic, Communications of the ACM, 64(7), July 2021.

[54] Please note that the data is as of mid-year, and thus the increase shown in 2020 is largely from the beginning of the lockdowns in March 2020 through the reporting period at the end of June 2020, thus concentrated in just over three months.

[55] RIR regional data built up from TeleGeography country data

FIGURE 6.2: QUARTERLY HOURS SPENT IN VIDEO STREAMING APPS (A) AND BUSINESS APPS (B), AND MONTHLY HOURS SPENT IN TRAVEL AND NAVIGATION APPS (C) AND SPORT APPS (D) RELATIVE TO PRE-COVID 19 LEVELS
[SOURCE: APP ANNIE, 2021]



Various measures were taken to help maintain performance while supply was increased to meet demand. Netflix developed a method to reduce the traffic per video stream by 25% while only marginally impacting video quality. This was deployed in Europe in March 2020, and then for ISPs in other areas under 'shelter-in-place' orders. Netflix then worked with ISPs to increase capacity and by the next month had added four times the normal capacity.[56] Similarly, Amazon Prime Video reduced streaming bitrates to help telecoms services handle the increased demand.[57] Operators also played a role in response to Covid 19: 77% of ISPs accelerated domestic capacity upgrades, 73% accelerated international capacity upgrades, 52% increased IP transit purchases, 74% increased peering capacity and 51% increased caching capacity.[58] The above examples illustrate how various players in the Internet ecosystem were able to take action to meet the challenge imposed by the pandemic. These actions were mostly decided and implemented locally and

[56] https://about.netflix.com/en/news/reducing-netflix-traffic-where-its-needed

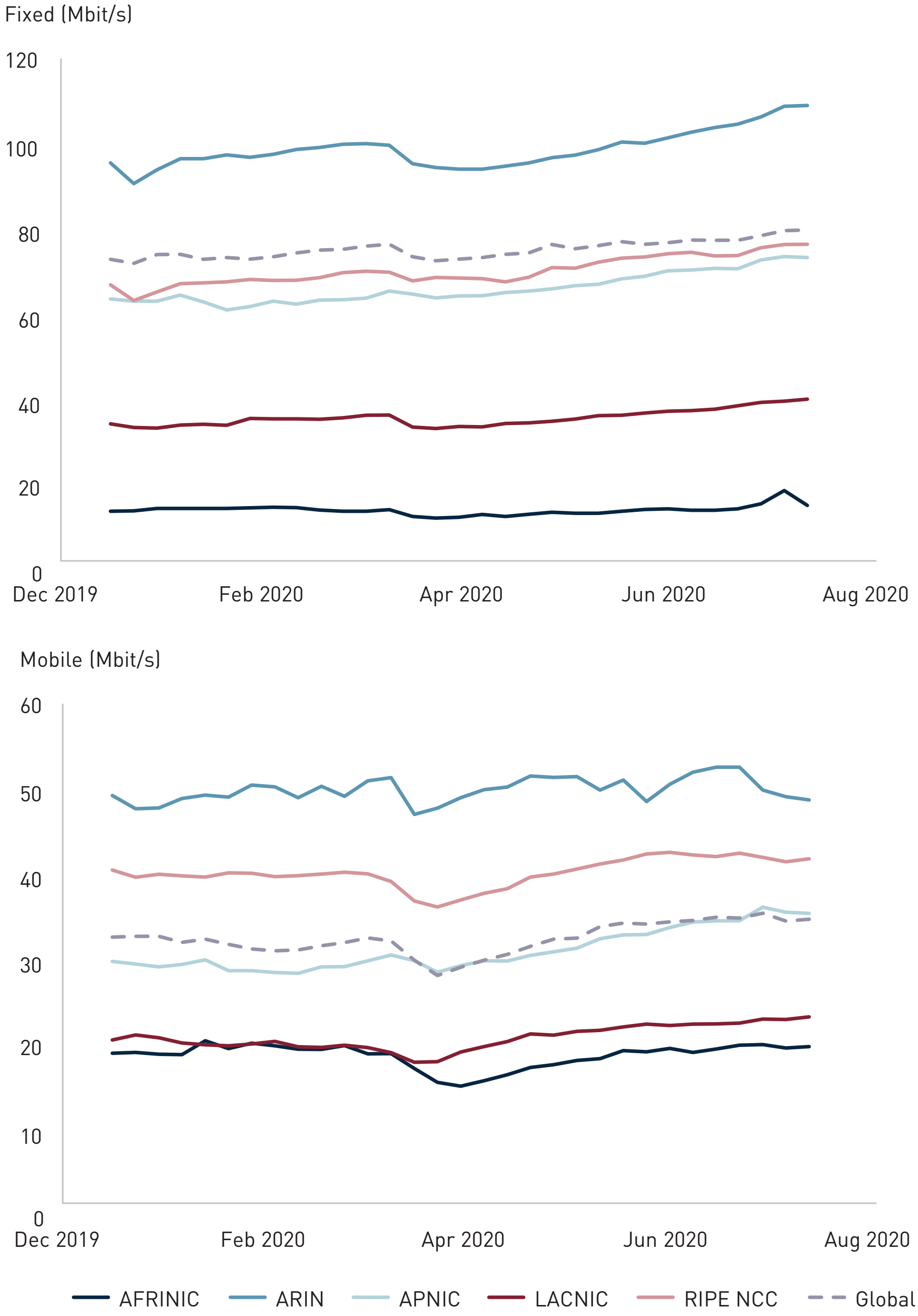
[57] https://techcrunch.com/2020/03/20/amazon-follows-netflixs-lead-reducing-streaming-quality-in-europe

[58] https://www.telegeography.com/products/global-Internet-geography/analysis/capacity-and-traffic-trends/index.html

independently, without requiring much co-ordination. This shows the Internet's flexibility, stemming from the network-of-networks and layering principles.

Despite all the increased traffic and demand placed on it during the initial pandemic lockdown period, the Internet was resilient. Figure 6.3 shows that overall access speeds only dropped marginally as lockdowns began in each region, and within two months were at, or above, pre-lockdown levels. This also demonstrates how increases in capacity can serve to maintain download speeds, rather than implementing any service guarantees.

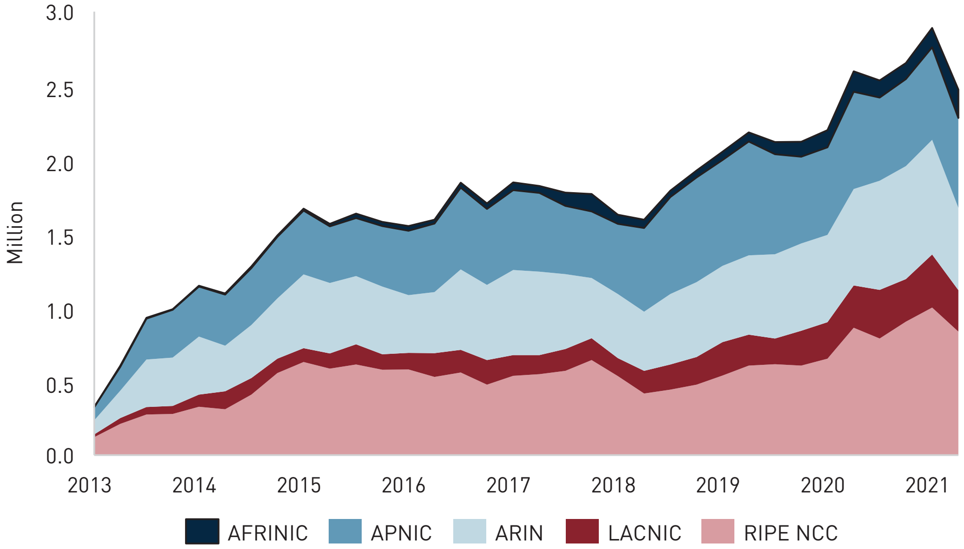
FIGURE 6.3: DOWNLOAD SPEEDS FOR FIXED AND MOBILE BY RIR REGION[59] [SOURCE: ANALYSYS MASON, OOKLA, 2021]



[59] Based on analysis by Analysys Mason of Speedtest Intelligence® data for average mobile and fixed download speeds in select countries, December 2019 - July 2020. https://www.speedtest.net/insights/blog/tracking-covid-19-impact-global-internet-performance/#/ Ookla trademarks and data used under license and reprinted with permission.

Besides the extraordinary strains that have been placed on the Internet during the Covid-19 pandemic, various types of attacks and malicious behaviour are unfortunately a part of the daily operations of the Internet. As mentioned earlier in this section, one of the most common forms of attack on Internet infrastructure is DDoS attacks. A target server is flooded with a massive number of requests from thousands, or millions, of hosts simultaneously, often originating from a botnet. Figure 6.4 shows how the number of DDoS attacks keep increasing over time across the various RIR regions. Simultaneously, the intensity of attacks also increases due to the use of larger and larger botnets. While important services have gained some protection from DDoS attacks through replication and other techniques such as anycast, this form of attack is an ongoing challenge against Internet resilience, one we return to in Section 7.

FIGURE 6.4: EVOLUTION IN THE NUMBER OF DDOS ATTACKS PER QUARTER BY RIR REGION[60] [SOURCE: NETSCOUT, 2021]]



6.2 Explanations

The resilience of the Internet stems both from fundamental technical properties such as the distributed nature of Internet routing protocols, and from operational practices and methods that have been developed over time by network operators. Resilience also encompasses some measures that violate the design principles in response to challenges, without fundamentally altering the nature of the Internet.

6.2.1 The resilience of the Internet is based on simplicity and decentralised operations

The Internet has proven resilient to both strong growth and significant changes in traffic patterns and application requirements. The simplicity of its core protocols, the resilient topology of the Internet, and the decentralisation of the network of networks have all played important roles to achieve this resilience. Within the network of networks, the entire responsibility for maintaining each individual network operation is distributed across individual, fully autonomous operating entities. This distributed responsibility fosters resilience in several ways: through diversity in equipment, diversity in operational practices, and topological diversity, in planning and decision making at all levels.

By way of example, Internet companies are increasingly investing in their own dedicated infrastructure to ensure delivery of content and applications and increase resilience, rather than relying on 'traditional' peering via other networks to carry their traffic.

Individual networks can fail due to cable cuts or other challenges, and robustness against these failures

[60] RIR regional data built up from NetScout regional data

depends on the physical redundancy of the network topology and the design of the Internet routing protocols. The Internet topology has evolved in such a way that there is no single central point of failure that can disrupt the whole Internet. Both engineering practices and physical limitations have led to a network where the centrality of a single router or network is limited. The Internet routing protocols used, both within a network and between networks, will eventually find a new path and reroute traffic after a failure as long as a physical connection exists. They do this in a distributed manner, without relying on a central entity in the network to direct traffic. The distributed nature of the Internet routing protocols is important for resilience.

At the same time, individual Internet companies can also experience failures in their applications or distribution networks. As a result of the layering principle, and the end-to-end nature of application provision, a failure in an application will normally have a very limited impact on the underlying networks or other uses of those networks.

> “
>
> *My sense is that there is a lot of autonomous and independent operation of the interconnecting networks on the Internet and that contributes greatly to its resiliency, and the ability to have multiple choices for how you're going to route traffic, to be able to route around problem areas and the network and the ability to make independent operational decisions within networks based on different jurisdictional and physical and capital constraints in different places of the world allows people to interconnect on their own terms. This idea that you can run your own network and get interconnected to the rest of everybody else as long as you speak the standardized protocols and pay what you owe. I think that is still very much true and is definitely part of the reason for the resiliency and the scalability.*
>
> Dr Alissa Cooper, VP/CTO Cisco

6.2.2 Operational practices and collaboration contribute to resilience

In addition to the distributed nature of the Internet topology and routing protocols, the operational practices in the Internet are also important for resilience. The Internet as a network of networks gives operational resilience in the localisation of problems, where the solutions are in the hands of those affected (as noted in Section 6.1, regarding the response to the Covid-19 pandemic).

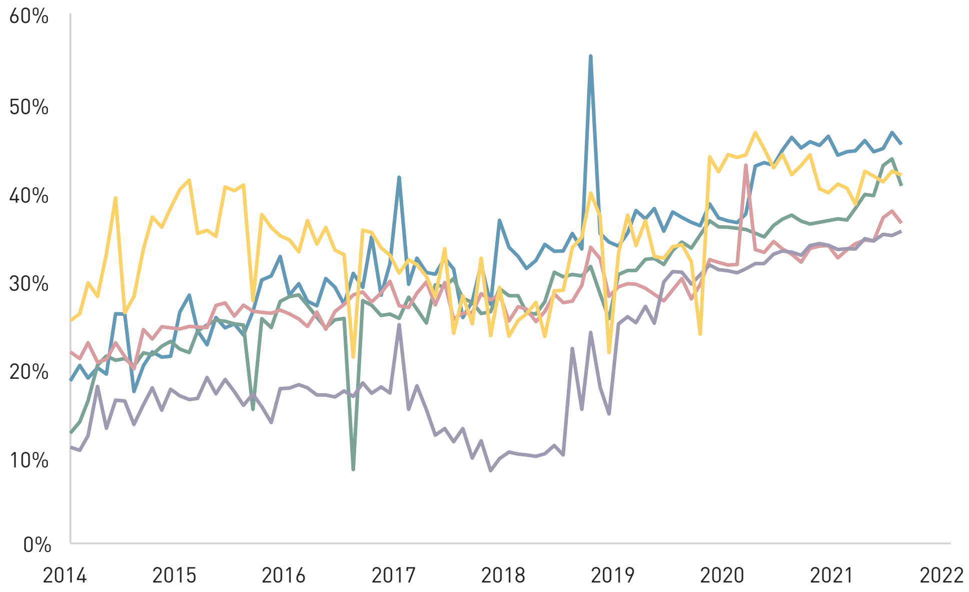
The openness and collaborative spirit that has been a part of the Internet from the beginning, embodied in the guiding ideals identified in this work, has carried through to the network operations community. There are a number of national and regional fora where operators meet and share their insights and experiences with technical and operational issues. This exchange of experience and practices has played an important role in building trust and increasing the quality of network operations. Given the Internet is the network of networks, each network is to some degree dependent on the sound operational practice of its neighbours. The stability of one network can be directly impacted by bad configurations, security breaches or other mistakes in a neighbouring network. Over time, the operational community has developed a long list of best practice recommendations, which are adhered to by most networks. Examples of these that have been adopted in Requests for Comment (RFCs) by the IETF include recommendations on BGP operations (RFC7454), traffic filtering (RFC7126) and many others.

One example of operational practices that have been developed to increase the resilience of the Internet is the efforts to secure the DNS. The DNS is a central Internet building block that has shown remarkable scalability, flexibility and resilience over time. The core application of DNS is to translate human-readable identifiers into routable IP-addresses. It performs this task through a hierarchical system of name servers, where the whole name system of the Internet ultimately depends on a set of 13 root servers to function.

Playing such a central role in the Internet, DNS is a popular target for various attacks. There have been several attempts to take down the roots of the DNS through DDoS attacks. Famously, an attack on a DNS provider in October 2016 caused major Internet services to become unavailable for a large number of users in the Eastern USA. This attack used millions of hosts, including Internet-connected devices that had been infected by malicious code. The DNS root servers have also been the target of several large DDoS attacks, several of which have made headline news.[61] They have, however, proven resilient in the face of such attacks, thanks to massive replication and the use of anycast for the root servers. This resilience builds on the distributed nature of both the DNS itself, and the routing system that carries the traffic.

DNS is also vulnerable to a form of attack known as 'spoofing', where an attacker pretending to be a DNS server sends false responses to address queries with the aim of redirecting traffic to a different host. In response to this threat, the IETF developed the Domain Name System Security Extensions (DNSSEC) protocol, which provides encrypted server authentication and data integrity. As shown in Figure 6.5, however, adoption of DNSSEC has been relatively slow, partly due to challenges in encryption key handling and the non-trivial overhead that comes with this protocol. Vulnerabilities in DNS remain a challenge, and the IETF is now working on several solutions, including DNSSEC and DNS over HTTPS (DoH).

FIGURE 6.5: USE OF DNSSEC VALIDATION (DNS VALIDATES + PARTIAL VALIDATES) BY REGION[62] [SOURCE: APNIC, 2021]]



In addition to DNS, the global routing system based on BGP is a central building block for the Internet. As is the case with DNS, significant effort is being made to secure the interdomain routing system. These efforts include the development of route origin validation (ROV) to verify the legitimate origin of a route, and BGPsec to verify advertised routes. Both depend on the use of the Resource Public Key Infrastructure (RPKI), where the RIRs take on a role in attesting the legitimate owner of an IP address.

[61] In 2007, a DDoS attack on the DNS root servers prompted the US Department of Defense to warn that it might respond to such attacks by physically bombing their origin.

[62] Data is regional based on regions as per APNIC labs (https://stats.labs.apnic.net/dnssec/XA?o=cXFw1v1p1x0l1). The data is not built up into RIR regions. The Americas covers two RIR regions, so an assumption would have to be made in order to split it between the two, and some of Asia would need to be combined with Europe to create RIPE.

6.2.3 The Internet has been resilient to violations of the design principles

> “
>
> *Tolerance of breaking the rules is part of the success of the Internet.*
>
> Yoshinobu Matsuzaki, Senior Engineer,
> Internet Initiative Japan

This report has highlighted several examples where the Internet design principles have been bent or broken:

- The end-to-end principle is broken by NAT boxes that use information from transport layer protocols or applications to manipulate the content of IP packets in the middle of the network.
- The layering principle is broken by practices where DNS, an application layer protocol, is relied upon by the routing system operating at a lower layer.
- Since the early days of the Internet, proxies have been used to trick TCP to work over geostationary satellite connections, which would otherwise have unacceptable performance due to high delays. Proxies and caches were the precursors of today's CDNs, all of which rely on a variety of systematic and complex compromises to an 'idealised' model of the Internet.

In all of these examples, the violation of central design principles has been used as a tool to achieve engineering goals and increase performance. Design principles are thus not absolute rules, but rather norms for how the Internet should generally work, or ideals that are always subject to pragmatic optimisations. The ability to bend or break these principles has arguably been central to the Internet's success, since it gives the flexibility to solve specific problems without the need for a central permission. At the same time, other attempts to address known issues, such as the introduction of IPv6 to increase the IP address space, or DNSSEC to address attacks, have not been fully implemented, leaving challenges that must still be addressed. This is a topic we turn to in the final section of this report.

6.3 Implications

The best proof of Internet resilience is how an increasing number of users are relying on the Internet for an increasing number of services, including sensitive ones such as banking and healthcare.

The resilience of the Internet has been particularly tested recently, during the Covid-19 lockdowns. It has proved itself to be resilient to the challenge, allowing users to increase their reliance on the Internet for work, study, government services and entertainment in the most difficult of circumstances.

At the same time, the Internet is continually impacted by network failures and malicious attacks. The Internet was designed to route around network failures, and it has proven capable of routing around broader problems, even if some of the underlying design principles are challenged. In the next section, we discuss whether this can continue in the face of existing and potentially new challenges.

7 Prospects for further success

In this section, we look at some of the technical challenges facing the Internet today and how these can shape the future development of the Internet. Overall, we show that the Internet can continue to evolve to address these issues as it has been for many years, and a fundamental change in the Internet architecture – even if it could be successful – does not seem warranted. We also briefly discuss other potential changes and challenges resulting from the actions of companies or countries that may pose a challenge to the traditional way of developing and operating the Internet, as embodied in the design principles, and how these developments could impact the four dimensions of success.

7.1 Technical challenges to the success of the Internet

As described in this report, the Internet's success is evident through its scalability, flexibility, adaptability and resilience. The Internet has so far proven able to route around challenges and perceived shortcomings, in order to sustain continued growth in the number of users, traffic and applications. There are, however, a number of challenges to the continued adoption and use of the Internet. These challenges can relate to a perceived lack of functionality, quality or security. Often, technical solutions to these challenges exist, but their adoption has so far been limited. Here we name some of the most well-known challenges that might be important for the future technical development of the Internet.

Security of the Internet infrastructure

As described in Section 6, individual networks and the Internet as a whole are routinely challenged by various forms of attack focused on the technical infrastructure of the Internet.[63] These can be in the form of DDoS attacks, route hijacking or other threats. On a large scale, the Internet has proven itself resilient in the face of these attacks. Still, the damage inflicted on individual networks and their users can be significant.

Several protocols and frameworks have been developed to make the Internet infrastructure more secure.[64] These include DNSSEC and DoH or DNS over TLS (DoT) in order to secure the DNS, and route origin validation to secure the interdomain routing system, as discussed above in Section 6.2.2.

While both DNSSEC and the BGP security extensions are important steps towards securing the Internet infrastructure, significant efforts will still be needed before these protocols are widely deployed and used. Nonetheless, continued efforts in terms of new and updated standards and operational practices are underway to address security issues at the technical level. Without undermining their impact on the Internet, none of the existing security problems, so far, have threatened the design principles or dimensions of success, and continued development and adoption of solutions will strengthen the Internet.

Quality of service to support emerging applications

As discussed in this report, a long-standing question in the context of the Internet has been that of best-effort traffic delivery across domains, versus a service with some form of guaranteed quality. The Internet's best-effort service model is simple and, as argued in this report, this simplicity has been important for the success of the Internet. However, concerns are sometimes raised that the best-effort model will not be sufficient to support the needs of emerging interdomain applications such as augmented/virtual reality or interactive gaming. The lack of service guarantees is sometimes used as a motivation for potential alternative network architectures.

We note that the lack of service guarantees is mainly an issue across network domains. Within a single network, it is often possible to engineer solutions that fulfil the necessary service guarantees. As a result, an Internet with strict end-to-end service-level guarantees might involve a deviation from the network-of-networks principle (where individual networks independently decide their technical parameters and service levels), and may in turn impact the flexibility of the Internet toward a variety of network technologies.

[63] IThis is in addition to other forms of attacks focused on the content, services and users of the Internet, including phishing and identity theft, piracy and so forth, which are not the focus of this study.

[64] While the security of the Internet infrastructure itself is important at a systemic level, most Internet users are more directly concerned with security issues concerning their own data and the services they use. Challenges such as malicious software, phishing attacks and vulnerabilities in applications are a continuing threat to end users. The countermeasures to these challenges are often a mix of end-user solutions such as multi-factor authentication or stronger passwords, and network solutions such as firewalls or other forms of content filtering.

Delays in adopting new protocols

The decentralised management of the Internet, based on the network of networks, means that the decision to adopt new protocols or solutions is taken by individual network operators. Many operators will only adopt new solutions if there is a clear local benefit. This will sometimes lead to difficulties in deploying new protocols for improved security or functionality on a wide scale. The prime example of this is the slow adoption of IPv6, as discussed in Section 2.2.1, but also the slow uptake of DNSSEC (see Figure 6.5). These deployment delays, while significant, do not threaten the dimensions of success. The Internet has adjusted to the slow adoption of IPv6 and continues to push this adoption, along with other new protocols such as DNSSEC.

7.1.2 The Internet continues to evolve to meet technical challenges

Over the last few years, the Internet has gone through significant changes, including the introduction of new network technologies, new and changed protocols, and changes in the roles of Internet companies.

The Internet has historically been successful even if, or perhaps because, it does not inherently offer optimised performance or integrated security solutions. As argued in this report, the simplicity, openness and decentralised nature of the Internet, manifested in the design principles, has driven the success of the Internet through some or all of the identified dimensions. Today, the network effects on the Internet are overwhelming, adding significant value to services that are run over the Internet.

An evolutionary path for the further development of the Internet involves continued improvements in the form of extended infrastructure, increased adoption of existing protocols and development of new protocols and faster and more capable network technologies. There are ongoing initiatives and developments to address the known technical limitations on the Internet, including increased security. These efforts are mainly the result of a market-driven development process, where network operators, equipment vendors, content providers and other relevant stakeholders design new solutions in response to customer demands. Likewise, the success and adoption rate of new and improved technologies are also determined by the market.

The Internet as we know it today is fundamentally one network, with the capability to provide universal reachability. The common address space and the packet format defined by the IP protocol allows communication between all Internet-connected end systems. There are limitations, but they are not embedded in the core Internet protocols or in the decentralised governance structure of the Internet. While the limitations or their solutions may bend or break design principles, they do not threaten the dimensions of success.

7.1.3 A fundamental change to core Internet protocols is unlikely

While there is no disputing the success of the Internet, it is also clear that some potential design goals are hard to reach with the current architecture. In addition to security and quality guarantees (discussed above), other limitations such as difficulties in handling mobility and the complexity of network management have been put forward as problems that may require a radical departure from the current Internet design.[65]

There have been several proposals and efforts towards radical departures from the current Internet. Most of these have come from academia. Among the most significant initiatives are *active networking*,[66] and later the *named data networking* architecture.[67] While elements of these initiatives have influenced later technologies such as software-defined networking, the radical ideas at the core of these proposals have yet to see widespread deployment.

Widespread adoption of a new Internet architecture that radically departs from the current Internet is challenging for several reasons. First, the core Internet protocols are deeply embedded in a global installed base of equipment. Any new technology that is not backwards compatible with this installed base has a large (and likely unsurmountable) disadvantage. Second, the implementation of current Internet protocols has been tested and refined over many years, and a new architecture that promises to provide intrinsic security and performance guarantees will take

[65] http://ccr.sigcomm.org/online/files/p59-feldmannA.pdf

[66] Tennenhouse et.al., A Survey of Active Network Research, IEEE Communications Magazine, 35(1), pp. 80–86, January 1997.

[67] https://named-data.net/

years to develop and may fall short given the complexity, while current Internet protocols will continue to be developed to address the targeted issues. Third, the network effects in the Internet are significant, and it will take time before an alternative architecture can build a competitive ecosystem of equipment, applications and network operators. The challenges involved in making fundamental changes to the Internet are well documented through the long-lasting efforts to implement IPv6.

Further, widespread adoption of any new Internet architecture would put at risk the success of the Internet. We have argued in this report that the guiding ideals and design principles are central to explain the success of the Internet. It is far from certain that a different Internet architecture that deviates from these principles, for example through a tighter coupling between application and network or a move of intelligence from end systems to the network, would be as successful. A novel Internet architecture could also limit the possibilities of networks to select their own technical implementation, at odds with the network-of-networks principle. These violations of the design principles would in turn put at least several of the dimensions of success at risk. For instance, a new architecture may not be as adaptable to new applications, which would have to couple with the network, while at the same time not being as flexible to new network technologies.

A number of technical issues have been identified and do not pose a fundamental threat to the dimensions of success or the underlying design principles. The Internet can continue to evolve to address these issues as it has been doing for many years. A fundamental change in the Internet architecture does not seem warranted to address these issues, and may in fact be counterproductive: new technology and standards brings new risks and attack surfaces that take time to be understood, tested and defended against, and there is no guarantee that radical change would be adopted by the market in place of existing equipment running Internet protocols.

7.2 Possible developments for the Internet of tomorrow

The Internet has been growing and evolving since its inception and has, as described in this report, been hugely successful. Here, we describe and discuss two important developments that are likely to shape the development of the Internet in the years to come. The first is related to the rise of large global companies that constitute a significant part of the Internet economy. The second is related to the role of governments as the Internet plays an ever more important role in society.

7.2.1 Economics: the importance of large Internet companies

The past decade has seen strong growth for companies that offer content and applications over the Internet. These include social media companies, video streaming companies, CDNs and cloud companies that offer applications or infrastructure as a service over the Internet. Some of these companies have been very successful and are among the most valuable companies in the world. We focus here on the potential impact of these companies on the technical issues raised in this report.

While these companies are not traditional infrastructure providers, the quality of their services, and to some extent their cost structures, are influenced by the need to deliver the services through existing network infrastructure. Several of these companies therefore invest heavily in building their own network infrastructure (as discussed in Section 6.2.1) to interconnect their data centres and to push their content closer to the end users. A significant fraction of global IP traffic now consists of data that is moved between the data centres and edge networks of large Internet companies. This data is mostly moved inside their own private networks, without the use of transit network providers. These networks are growing in geographical footprint, as the large Internet companies establish presence at IXPs and other central Internet locations in order to improve the quality of their customers' user experience.

The evolution of large Internet companies could have consequences for the further technical development of the Internet. A smaller proportion of traffic will be carried in the traditional way, through a series of networks interconnected through peering or paid transit arrangements. Over time, we could see the

Internet transform into a more centralised system with a few global private networks carrying most of the content and services.[68]

In this scenario, what remains outside these private networks are primarily ISP networks that move traffic to and from end users, and the user experience would be shaped by how close a user sits to the private network of the relevant Internet company. Such a development would increase our collective dependence on a few major players, with possible implications for the dimensions of success of the Internet:

- Within their own network, large Internet companies could choose to use any protocol or technology they see fit. Although we have seen no evidence of such a development so far, there are concerns that this could happen and take resources away from the development of open Internet standards and protocols. On the other hand, such a development would be consistent with the network-of-networks principle, and the content would nevertheless have to be delivered using Internet protocols over the access networks.

- Increased centralisation could blur the distinction between network and applications, as expressed in the layering principle. An Internet where content and services become more tightly integrated with the network they are delivered over can limit universal access to services. The content a user is served may be determined by the location or type of access network she is connected to, however again the access networks would be used to deliver the content to the users.

- Finally, failures or other problems in these large networks could have severe and far-reaching consequences. However, the same is true for large backbone networks today, and the Internet companies can ensure resilience using third-party networks as needed.

We note that so far major Internet companies have played a constructive role in the multi-stakeholder model of Internet governance. Much of their infrastructure investments are in partnerships with traditional telephone carriers and ISPs, and even without a formal partnership, all of their content is delivered over third-party access networks, which ensures adherence to common protocols. Further, rather than taking resources away from traditional development and standardisation efforts, Internet companies have contributed to the open standardisation processes. QUIC, which was discussed in Section 5.2.2, is one significant example of a novel protocol that was developed by an Internet company and later standardised by the IETF as an open standard, which is seeing increasing levels of adoption.

7.2.2 Governance: balancing control and openness

As discussed in Section 2.2, the Internet is international in nature, and does not include a concept of national borders in its core technical protocols. Furthermore, in many cases, services that were previously produced within the national borders or jurisdiction of a nation state have now been moved to the Internet, and might be provided from outside the jurisdiction.

Given the importance of Internet-based services and applications and their social and economic impact, governments in many parts of the world are looking for ways to control or regulate these services for a variety of reasons, including to enforce national laws. In some cases, a government's attempts to control Internet-based services can be at odds with the technical implementation of the Internet. Examples where such conflicts may arise include:

- Lawful interception of communication. Traditionally, many governments have required that telecoms providers collaborate with authorities to enable the interception of calls or other forms of monitoring of communication. This can be more complicated with Internet-based services, both due to end-to-end encryption and questions around jurisdiction. Demands for an encryption 'backdoor' for governments could impact the security of the content, and the development of encryption, but not any design principles.

- Data localisation. An increasing amount of sensitive data is stored on systems connected to the Internet, and some governments have started to impose requirements that certain types of data are only stored and processed within certain jurisdictions.

[68] https://blog.apnic.net/2016/10/28/the-death-of-transit/

One reason for this is to address the jurisdictional issues with interception addressed in the previous bullet. In any case, while this can have a commercial impact on companies providing the services and on data centres, and impact the technical architecture of their own networks (such as the location of data centres and how data is routed between them and to end users), it does not impact the design principles of the Internet per se.

- Limiting access to illegal content. The separation of networks from content and applications makes it hard to control and limit access to illegal or harmful content. The Internet, with openness as one of its guiding ideals, offers limited technical support for this type of content filtering. Regulating the content providers, rather than filtering the content at the technical level, may be the result, but this would take place at the application level, rather than at any technical level.

- In addition, there is a possibility that countries could begin to impose their own technical standards on companies operating in their jurisdiction. This could be as a way to develop their own industry, or to make it easier to control traffic flows, by developing a gateway between domestic and international networks. This is possible as a result of the network-of-networks principle but would put at risk the end-to-end principle for international traffic, and also potentially violate layering, if the new standards enable more control over applications.

It is inevitable that governments seek to impose laws on the Internet, as more and more social and economic activity shifts to the Internet, and that some governments seek to increase their control over it. These efforts can take several forms, from stricter regulations on networks and Internet companies, to altering the way that the Internet is governed. A development where governments gain more control over the development of the Internet may involve a risk of a more fragmented system, without the common address space and global reachability we have today. However, to date the impact is predominantly with regards to content and applications, and their corresponding providers and business models, rather than any technical fragmentation that puts at risk the design principles and dimensions of success.

7.3 Conclusion

The current Internet is built on the fundamental guiding ideals of openness, simplicity and decentralisation. These guiding ideals are expressed in the three design principles *of layering, network of networks and end-to-end*. The design principles are not absolute rules and we have provided several examples of how they are sometimes bent or broken to achieve certain goals. However, the ability to accommodate such violations of the design principles highlights that these are principles and not absolute design rules, and they have not threatened the success of the Internet.

The success of the Internet has turned it into the world's most important system for connecting people and sharing information. The design principles of the Internet make it simple for anyone to connect a system to the Internet and use it to distribute any type of content. Any proposed radical changes to the Internet should be evaluated against their ability to maintain its dimensions of success. The Internet continues to develop, as described in this report, while maintaining its success. New technical protocols replace or supplement existing ones, new network technologies are developed, and the network continues to adapt to new applications. Such developments are natural and necessary for the Internet to stay relevant. As the Internet continues to evolve, we believe it is important to recognise and maintain the guiding ideals and design principles that have contributed to the scalability, flexibility, adaptability and resilience that represent the dimensions of success for the Internet today.

Stay connected

You can stay connected by following Analysys Mason via Twitter, LinkedIn and YouTube.

in linkedin.com/company/analysys-mason

@AnalysysMason

youtube.com/AnalysysMason

analysysmason.podbean.com