Edge Computing: digital infrastructure beyond broadband connectivity

William Lehr
MIT

Douglas C. Sicker
U.Colorado
(Denver/Anschutz)

Dipankar Raychaudhuri
Vivek Singh
Rutgers University

TPRC51 (September 2023)[1]

Abstract

Nations around the globe are investing billions in deploying next generation broadband (BB) networks. Those networks make it feasible to expand the reach of computing resources into a wider array of devices, locations, and activities than ever before. The expansion of edge computing resources is both a necessary and unavoidable outcome of our expanded digital connectivity. What is uncertain is how the edge computing infrastructure may evolve: its architecture, what and how services will be provided, and who will own/control the resources necessary to support edge computing. It seems obvious to us that large national-scale service providers of connectivity and digital platform services will play an important role in the expansion of cloud services into edge computing, although how the Internet infrastructure ecosystem may be reshuffled as a result is unclear. What is also unclear is what role end-users may play in the provision of edge-computing. In this paper, we explore the whys and hows of edge computing's emergence, while also making the case for the importance of end-user deployed edge computing as a potentially valuable option for edge computing deployment. We discuss what this may mean for digital infrastructure policy as we move beyond our focus on broadband connectivity.

Keywords: Edge Computing, Mobile, Broadband, Internet, Regulation, Infrastructure, 5G

JEL Codes: O32, O33, L51, L86, L96

Table of Contents

[1] The authors would like to acknowledge support from NSF Grant #2228471, and helpful comments from Volker Stocker.

1. Introduction

Nations around the globe are investing billions in deploying next generation broadband networks, motivated by the view that ubiquitous availability of high-quality digital connectivity is essential infrastructure to remain competitive in the digital economy future.[2] For the broadband investments to deliver the hoped for economic growth and other social benefits, digital connectivity resources need to enable networked access to digital computing and storage resources.

In this paper, we take the long view of the co-evolution of digital computing and communications infrastructures. Continuing technical and economic progress have made it feasible to integrate and distribute computing and storage into an ever-wider array of devices, locations, and activities. From the earliest mainframes to today's Internet of Things (IoT), improvements in computing and telecommunications have made it increasingly technically and economically (i.e., cost-effectively) feasible to unbundle and distribute in space and time the various activities required for digital operations -- or, in short, to support networked computing. The policy justification for viewing broadband as critical infrastructure is because it enables participation in the Internet ecosystem, which in its broadest construction comprises the preeminent global platform for networked computing. For researchers interested in the Internet ecosystems future and the co-evolution of information technologies, industry value chains, and digital policy, the future of digital infrastructure beyond digital connectivity is of special interest.

In Section 2 of this paper, we review the history of the cl networked computing to make the case for why the emergence of Edge Computing is an obvious and necessary next step in the evolution of our critical digital infrastructures. If our investments in broadband are to be productive beyond the demands of current applications, we need edge computing capabilities. We explain what edge computing is, why it is needed, and how it is evolving.

In Section 3, we explore alternative models for how edge computing infrastructures may evolve. We organize our discussion in terms of two distinct economic models or paradigms for deploying edge computing resources: service providers and end-users. This organization of our review is to motivate our case for the value of the end-user deployment option, which we refer to as "pMEC,"

[2] For example, the U.S. has allocated $42 billion toward expanding (mostly) fiber-to-the-home quality broadband services to every location; and Europe has targeted universal service to 1Gbps broadband for everyone by 2030. For the U.S. Broadband, Equity, Access, and Deployment Program (BEAD) see https://broadbandusa.ntia.doc.gov/resources/grant-programs/broadband-equity-access-and-deployment-bead-program; and for European Digital Decade targets, see https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en. For discussion of BEAD program, see Lehr (2023).

which is short for *public*[3] *Mobile Edge Computing* infrastructure.[4] It seems obvious to us that most edge computing services in the future and the associated networking and digital infrastructure needed to support those services will likely be provided by large, national scale (or larger) providers of digital connectivity and platform services (e.g., ISPs, cloud and other content and application service providers). However, we believe that need not and should not be the only option for enabling edge computing resources. Although the end-user deployment model has always existed as part of the digital landscape in the form of private networks and computing, the future promises the potential for new models and new importance for how end-user and service-provider infrastructures and services may interact, with important technical, economic and policy implications. We explain our pMEC model as a potential hybrid development path for end-user deployed infrastructure, and review some of the challenges for pMEC and possible evolutionary paths it may follow.[5]

In Section 4, we explore the implications of expanded edge computing and the pMEC concept for several key communications policy issues. Those include the implications for broadband competition, universal service policy, interconnection, and related digital infrastructure-related policies.[6] In short, we will argue that the need to deploy edge computing resources will necessitate acceptance of more flexible and complex passive and active sharing among value chain participants, which will require a reassessment – and likely reframing – of legacy models of industry structure and regulation and a re-thinking of the role of end-users and edge-community networks.[7] It will also open new avenues to study trust and cooperation in digitally mediated settings. We see the pMEC option as an important test case worthy of further exploration. If successful, it has the potential to introduce an important new vector for the emergence of competitive discipline and for reasserting end-user autonomy. If unsuccessful, a clearer understanding of why edge computing infrastructure is and should be a capability provided by large service providers in a relatively tight oligopoly, ought to help drive consensus toward appropriate regulatory models.[8]

---

[3] One might argue that "public" is not a good description of what we are proposing. In fact, these would be private (end user) provided services and only through some open federation would they become accessible to some or all of the public. For further discussion, see Note 5 *infra*.

[4] For discussions of Mobile Edge Computing see, for example, Abbas et al. (2018), ETSI (2014), Vhora & Gandhi (2020), Khan et al. (2019), and Wang et al. (2023).

[5] We will use the acronym pMEC to refer to this model, which is our version of how privately-owned, edge-computing resources might be combined to form an edge-based community cloud, as an alternative to a service-provider cloud. Calling this a "public" MEC turns out to be potentially confusing for reasons we explain later, but the acronym proves useful because it connects with an on-going research effort we have been engaged in to develop the technical-economic case for a pMEC (see NSF Award #228470 : Collaborative Research: SAI-P: Public Multi-Access Edge Cloud (pMEC) as a Community-based Distributed Computing Infrastructure for Emerging Real-time Applications, awarded August 2022).

[6] For a discussion of how interconnection has evolved in Internet, see Clark, Lehr & Bauer (2016).

[7] For discussion of edge sharing, see Lehr & Stocker (2023).

[8] For example, in the mobile device ecosystem, there are two dominant mobile OS platforms (iOS and Android) and, in the U.S., a handful of high-speed broadband access options in each local market (two fixed

Section 5 concludes by summing up the key themes in this paper and highlighting directions for future research.

2. What is Edge Computing, Why is it Needed and How is it Evolving?

As noted earlier, nations around the globe have concluded that high-quality broadband access networks are critical infrastructure for the digital future.[9] This conclusion does not seem justifiable to us based on the economic value of the applications driving most of the traffic on today's broadband networks. That is because the dominant driver of traffic today is associated with streaming entertainment video traffic, and the mass-market application driving the current need for very low latency, interactive (2-way) network services is gaming and video-conferencing.[10] During the Covid pandemic that started in the spring of 2020, demand for broadband access to enable remote work and education (as well as expanded demand for on-line entertainment) proved surprisingly important, allowing a vast amount of economic activity that might otherwise have been impossible to shift online. Moreover, the suffering of those without adequate broadband access helped raise collective awareness of the economic development need to ensure universal access to high-quality broadband.[11] In spite of these facts, however, ubiquitous access to high quality broadband is not sufficient to meet the need for enhanced digital infrastructure to support the more demanding applications that the ICT industry hopes to enable.

Cloud computing, and increasingly cloud computing resources proximate to the network edge and end-users is needed to support the next generation of immersive, delay-intolerant, highly-interactive applications that are identified as justification for continued investment by industry and

---

providers – a cableco and a telco, and three national MNOs, two of which are subsidiaries of fixed broadband telco providers).

[9] See Note 2 *supra*. The focus of the discussion herein is on the U.S., Europe, and other developed markets. Developing and less developed nations are also pursuing expanded broadband infrastructure, although the goals for what constitutes an acceptable target for universally available digital connectivity is typically much less ambitious than in developed markets, where the 100Mbps and faster broadband that is desired is already widely available to most citizens. In the U.S., a location is defined as unserved if there is no service provider offering service at that location with data rates of 25/3 or better (where that refers to at least 25Mbps downstream and 3Mbps upstream). A location is underserved if there is no availability of 100/20 service.

[10] See Lehr & Sicker, 2017. Although estimates of the share of video traffic have changed somewhat since 2017, the fundamental conclusions provided in that paper still apply. To the extent the principal applications accounting for most of the mass-market broadband traffic are entertainment oriented, the potential for those to drive significant economic growth are limited. The principal effect is to shift revenue toward new applications and platforms. However, even though most of the traffic may be associated with entertainment video, the digital connectivity fabric that mass market-broadband networks are part of is also critical infrastructure for businesses, and businesses have been investing for decades in the ICT infrastructure and services that have facilitated the significant progress in the digitalization of business processes that has already occurred.

[11] See Whalley et al. (2023).

governments to support 5G+ infrastructure.[12] Those applications include Augmented Reality (AR), Virtual Reality (VR), Autonomous Vehicles (AV), Unmanned Aerial Vehicles (UAV), and a host of Artificial Intelligence (AI) and Internet of Things (IoT) applications that depend on real-time, ubiquitous, highly-interactive data streams and are augmented by high-performance computing-supported automation. The digital infrastructure needed to support this future vision is one of pervasive computing: anywhere/always on-demand access to digital connectivity, computing, and storage resources to support any economic or social activity (*Smart-X*) that may benefit from ICT augmentation.[13] Ubiquitous, high-speed, high-reliability broadband connectivity is part of the necessary infrastructure, but it is insufficient in itself to deliver the performance required of more demanding and ambitious applications. To meet aggressive millisecond response times,[14] reduce overall costs, protect data security and privacy, and facilitate edge-user control (or at least choice and participation in the design and provisioning of Next Generation digital services), edge computing infrastructure is also necessary.

In the following sub-sections, we expand on our explanation for why edge computing is a natural and necessary next step development.

2.1. From mainframes to IoT: edge computing as the inexorable outcome of progress

Among the first commonly used computers were mainframes. In the early days of computing, the need for high data throughput and fast response times required the sub-components communicate via a high-speed data bus interface (on the CPU chip and on the computer backplane). Remote users connected to the computing resources via dumb (limited capability) terminals connected via low-speed data lines to the centralized computer resources, which were accessed in a client-server architecture. Fast computations had high-latency and low-interactivity capabilities as data was subject to batch processing. By the 1950s and during the 1960s, IBM had emerged as the dominant mainframe provider with a vertically integrated architecture of customized software and hardware for IBM systems that competed with a group of other vertically-integrated mainframe providers identified by the acronym, BUNCH.[15] The dominant telecommunications provider was AT&T's Bell System, whose principal service was voice telephony via its narrowband, circuit-switched,

[12] See Jiang et al. (2021), Lehr, Queder & Haucap (2021), Lehr (2019, 2022), Lin et al. (2021), and Oughton et al. (2021).

[13] We refer to such applications of ICT augmentation or automation generically as *Smart-X*, where X can be replaced by the activity benefiting from ICT augmentation. For example, Smart Grids, Smart Healthcare, Smart Cities, Smart Supply Chains, etc.

[14] ITU (2015) visions of requirements for 5G specify sub-10msec response times and order of magnitude improvements (relative to 4G performance targets) in reliability, connection density, and other performance dimensions. In-process 6G standards target further improvements to enable networks capable of AI-native support (e.g., see https://www.6gworld.com/exclusives/immersion-ai-and-reliability-next-g-alliance-details-roadmap-to-6g/). Moreover, future computing and networking visions anticipate realization of the "Tactile Internet," enabling real-time control and haptic response capabilities (see Aijaz & Sooriyabandara, 2019; Fettweis, 2014; and Gupta et al., 2019).

[15] BUNCH stood for Burroughs, Sperry's UNIVAC, NCR, Control Data Corporation (CDC), and Honeywell mainframe computers.

analog telephone network.[16] There was no wide-area data communications network to enable distribute computing, and the costs of mainframes and the specialized resources required to make use of them were too high to make it feasible for any but large enterprises to self-provision their own needs, and midsize businesses accessed computing resources provided by time-sharing mainframe computing service providers.

With complementary and re-enforcing innovations across the entire ICT value chain and computing and telecommunications technologies, (exemplified by Moore's Law exponential improvements in semiconductor capacity/cost performance and the digitalization of communications networks, and later softwarization and virtualization technologies), it became feasible to distribute computing resources more widely.[17] Simultaneously, new parallel and distributed computing architectures emerged to allow new algorithms to be applied for solving computational problems. In the 1970s, companies like Digital Equipment Corporation (DEC), Hewlett Packard (HP), and others emerged offering cheaper (than mainframe) minicomputers that expanded the range of customers who could own their own computing resources in-house. Networks of smaller computers could begin to compete with the mainframes of old. By the 1980s, the PC revolution was ready to take off, enabling businesses to provide PCs for every employee and for homes to have PCs. In businesses, networks of PCs were connected via wired LANs (Local Area Networks) to support networked computing first in businesses and later in homes. By the 1990s, mass market adoption of the Internet took off, fueled first by narrowband, dial-up access which began to give way to first generation broadband access by the mid-1990s.

The era of mass market computing services with most people having regular access to computing services at work and at home, and simultaneously, having personal communication devices (cell phones) had arrived by 2000, but it was not until around 2010 that mass market mobile broadband was feasible. It was only with the emergence of smartphones and the rise of 3G/WiFi access connectivity that users could realize the benefits of converged mobile broadband services on a hand-held device.

Notice, however, that at each stage along the way, improvements in computing architectures supporting more distributed hardware and software architectures and improvements in digital connectivity via the rise of broadband networking, computing infrastructure and resources have become more widely distributed, geographically and in terms of the population of users that can own and use those resources.[18] Edge computing has been coming for a long time as what had been yesterday's new digital technology migrates from the network core to make way for the latest

[16] In the U.S., competition policy restricted telecommunications and computer competition and convergence. IBM and AT&T were both bound by consent decrees limiting their ability to enter each other's markets.

[17] Faster components and interfaces at many end-to-end points created slack in time link budgets, opening the way for new designs that distributed components and added interfaces and additional functionality (e.g., protocol conversions) as part of the overall task.

[18] Cheaper, faster, smaller components means that the computing functionality enabled by those components can be placed in more locations at affordable cost. And lower costs mean that small and medium sized businesses could opt to own their ICT infrastructure instead of either relying on time-sharing, or worse, foregoing the benefits afforded by adopting ICT technologies.

generation of technology and to meet the growing need for capacity that follows the growth of core traffic, but with a lag. In hierarchical telecommunication networks, the most powerful, highest-capacity, and most expensive switches and routers are located in the core of the backbone networks. Then, as the next generation of high-performance, high-speed computing and routers become available, the large service providers migrate yesterday's devices toward the edge when they upgrade to the highest capacity core routers. Similarly, in fiber networks, Dense Wave Division Multiplexing (DWDM) was initially deployed to increase the capacity of undersea and other expensive-to-deploy long-haul fiber transport where the need to accommodate the continually increasing core-network traffic. The migration of DWDM to access networks occurred as the costs of DWDM dropped with learning and scale economies and as demand for middle-mile capacity interconnects rose to levels previously only seen in core interconnects a few years earlier.

Today, most consumers have multiple devices with embedded CPUs and a number of these have significant on-board digital computing and storage capabilities and multiple wireless ways to connect to networks and other digital components. Those include smartphones, tablets, smart watches, smart TVs, eReaders, etcetera. Similarly, many (most) businesses have PCs for every employee, as well as shared servers and even minicomputers or mainframes as part of their enterprise ICT infrastructure.

At the same time, the Internet ecosystem has expanded to include multiple layers of platform service providers offering a range of functionality and services that complement or substitute for capabilities that are otherwise accessible via the Internet, offering an array of cloud services to other service providers as well as end-users. Those include Content Delivery Networks (CDNs) that assist in the delivery of content, Data centers, Content and Application service providers (CAPs), as well as the large cloud computing service providers. The largest of these digital platform providers in the Internet ecosystem are referred to via the new acronym, GANFAM, for Google, Amazon, Netflix, Facebook, Apple, and Microsoft. All these players have been deploying a growing array of distributed computing and storage assets (servers, data centers) around the Internet and across the globe.[19]

An early taxonomy of cloud services differentiated between Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SasS) models, highlighting the different and increasing levels of computing functionality that the cloud service providers offered to their business customers.[20] Today, the largest cloud providers (Microsoft Azure, Google, and Amazon AWS) and a growing array of specialty cloud service providers offer an even larger array of cloud computing services, ranging from bare iron (rack space in distributed data centers) to fully-managed, on-demand distributed software.

Indeed, with the shift of many business processes from in-house computing networks toward increased reliance on cloud services, the trajectory of computing and telecommunications appears

[19] See Lehr, Clark et al. (2019a, b) and Stocker et al. (2017).

[20] With IaaS, the cloud service providers offered access to basic computing capacity to augment enterprise computing resources; with PaaS, the cloud service providers offered additional tools to make it easy for customers to develop custom computer applications; and with SaaS, the cloud service providers offered a turn-key solution to allow customers to operate distributed software applications. See Armbrust et al. (2009).

to have come full circle. It started with remote computing resources connected via dumb terminals (thin clients) via smarter and more computing-capable edge PCs (fat clients) connected via dumb pipes toward once again seeking to enable expanded use of thin client (IoT devices and Chromebooks) connected to computing resources in the network. In the beginning, technical limitations foreclosed the option of distributed computing; whereas today, the option to manage computing resources in different ways is an option. Although the rise of IoT and other less-capable edge devices suggests an expanded need for cloud computing; the rise of AVs, robots, and ever-more-powerful personal computing devices (PCs and smartphones) demonstrates that the future is likely to be more of a hybrid with both thin and fat client options providing more flexible choices for where computing and digital storage activity should take place.

All these distributed digital computing and storage resources whether owned by ISPs, cloud service providers, or end-users are capacity that, at least in principle, could be integrated to serve as edge-computing resources. Thus, the emergence of edge computing has to a large extent already occurred from one point of view. But this emergence of edge computing as the outcome of the inexorable march of ICT technical progress and market growth misses a key question motivating our inquiry. Namely, in today's world where there are many ways to provide edge computing resources, what are the economic and policy implications of choosing one strategy or another?

We summarize this computing evolution in the Table 1 below and highlight what this evolution means for edge of the network and related computing

Table 1: Evolution of Computing and Rise of the Edge

Date	Era	Technical	Economic	Industry structure	Edge Computing
1960s and earlier	Mainframes & Client-Server (1960s and earlier)	On-board/local bus for High Performance Computing (HPC) to sustain data throughput and response times (latency) needed for HPC.	High-costs for computers and support costs to use. Specialized MIS departments, computing only for large enterprises and governments.	IBM and BUNCH v. AT&T Bell System. Consent decrees kept Computing & Telecom separate.	Connectivity infrastructure technically and economically (too expensive, insufficient capacity/coverage) to enable edge computing.
1960s-1970s	Distributed Computing	DEC and rise of mini-computers able to take advantage of expanding data communications connectivity.	More businesses can adopt ICTs as costs fall, but high costs for computers and data coms, limited skill to use, and limit	DEC, Xerox, Digitalization of telecoms	Core data networks provide connectivity. Semiconductor advances and CPU architectures make possible for emergence of parallel computing and cheaper boxes for mainframe computing in more venues (smaller enterprises).
1980s-1990s	PC and Cell phones	Internet and WLANs. Networked computers/servers. UNIX/WIN make feasible for distributed apps ecosystem to emerge separate from hardware	Falling costs, expanding use. From IT departments to everyone at work and then mass market. Rise of personal computing and communications.	Broadband, Cell phones everyone, PCs everywhere. Separate comms and computing still.	SMEs and consumers have computing resources and are connected. Parallel processing and rise.
2000s-2025	Connected Internet and emergence of IoT and 5G+	PCs and post-PC devices emerge (tablets, smartphones, smart TVs). Softwarization with rise of Network Function Virtualization (NFV).	Internet: everyone on-line (B2C) and all business on Internet (B2B). Rise of sharing economy (GigEconomy, Matching platforms, etc.)	eCommerce, social media/digital platforms. Rise of Digital Clouds and Platforms (platforms of platforms with Apple, Android, Amazon,, Microsoft, etc. ecosystems). ISPs v. GANGFAM.	Clouds emerge. Connectivity from to 1Gbps access and beyond, including fixed and ubiquitous mobile connectivity. Converged communication networks. Focus on Mobile Broadband.
Future	AI Connected, 6G+	IoT Pervasive Computing. Distributed processing. Softwarization, virtualization means all ICT can run in software and software could be anywhere	On-demand ICT resources anywhere. Who provides and controls?	Much more sharing and blurring of industry boundaries. Rise of AVs, Robots, UAVs, Space-based computing, etc.	CPU commodity hardware can host software, but still need to provide hardware and power and other capabilities and essential resources (e.g., Intellectual property, spectrum, etc.). And, will always have specialized hardware (ASICs such as GPUs, etc.); and software architectures are unclear (Kubernetes, etc.)

2.2. Applications that require Edge Computing

Any application that needs computing resources could be a candidate for edge computing, especially if there is an edge computing alternative that offers better price-cost performance than the available non-edge alternatives. However, there are certain classes of applications that may require computing take place at the network edge. Three key reasons for this include (a) Low latency requirements; (b) Disruption tolerance requirements; and (c) Form/factor requirements. Each of these motivations is discussed in somewhat different form in the following sub-sections.

2.2.1. Low Latency Required

One of the first and most important criteria justifying the need for edge computing resources is an application's need for media intensive, high-speed, low latency interactivity. That may involve applications that need to respond very quickly to changing local environmental information. Classic examples are dynamic control problems like controlling an AV or UAV. Closely related, are situations where the application needs the support of computational resources that exceed the capacity of those that can be maintained on-board (on the device) either for fundamental structure of the problem, configuration, or cost reasons. An example of the first sort of problem are situations where the answer to the decision problem needing a fast response requires accessing data that is not available locally (e.g., requires integration with data from a large data set such as a machine learning AI application). An example of the second is if the size, power, or other device design considerations make it technically infeasible to include the necessary computing resources on-board. Finally, even when it is possible to rely on on-board computing resources, relying on computing resources that are shared may benefit from scale, scope, or other economies that make using cloud services less costly. AR and VR applications of the sort anticipated for the Metaverse,[21] which would allow individuals (and other devices or applications) in the real and/or virtual worlds to interact with each other across the real/virtual interfaces may require tight-interactivity (low latency) bounds.

With the speed of today's networks and cloud service architectures, it is feasible to support interactivity with cloud computing that is not edge computing but still delivers latency performance in the sub-50msec or even sub-20msec range. That is very fast. Of course, for some applications, it is possible to imagine situations where even the speed of light imposes a distance (proximity) constraint on how far a server can be from the communicating device to support the required degree of interactivity.

Although it is possible to hypothesize applications for where the latency between the device and server providing the computational resources needs to be in the 10msec range, those needs are today most likely addressed via re-architecting the problem to obviate the need for such low latencies, potentially by shifting the problem into another domain. For example, in many cases, while real-time information may be best, it is often possible to use predictive strategies to substitute

[21] For discussion of the Metaverse see Note 29 *infra*.

a forecast when real-time responses are not available fast enough, and then correct with a lag when the needed data is available.[22]

Of course, when one provider or solution can offer reduced latency, that will create competitive pressure for others to offer competitive solutions. In the world of securities trading, the opportunity to exploit millisecond differences in transport speeds across fiber optic cables connecting New York and Chicago gave rise to high-frequency arbitrage trading strategies.[23]

Another issue to consider is that although today's modern cloud networks can support low latency access, the 20-50msec performance may be doubled or more if one or more inter-ISP interconnections are needed to use the cloud computing resources. Even when cloud providers maintain rich and distributed networks of distributed servers with lots of servers close to the edge where the accessing customers are located, the traffic from any particular end-user may need to be routed to an interconnection point between the carriers that may not be close to the end-user, requiring one or more round-trips and exposing the user to the risk of Internet congestion that may further add to the realized latency and degrade the quality of performance.

2.2.2. Network disruption intolerant

Another important reason to need edge computing is to minimize the exposure or risk of losing connectivity between the device and the networked resource. Of course, it is quite possible that a distant cloud server or connectivity to a fabric of redundant cloud servers may offer more reliable and robust connectivity for an edge device than an edge computing server that is proximate to the end-user (so not vulnerable to loss of middle-mile or long-haul connectivity). If the last-hop connectivity to the cloud is lost, the edge computing resource is not available if hosted in the cloud. The need to address such situations is an important motivation for self-provisioning or relying on edge cloud services provided by a local provider. Remote factories, mines or other businesses that have ICT-intensive operations or that may be extremely intolerant of any network connectivity outages (e.g., robots, hospitals, UAVs, hospitals, etc.) may be forced to rely on edge computing resources to enable continued safe operations in the event of a loss of broadband connectivity.

Closely related to the above are situations where the application cannot rely on wide-area connectivity to cloud resources because the infrastructure needed does not exist. It may not exist because no one has deployed the requisite connectivity yet (e.g., the location is unserved or underserved), or because of some emergency (in a location where deploying infrastructure does not make sense, or a location where the infrastructure was destroyed). An example here is for public safety and first responders, or for special events (concerts) where existing infrastructure is either not available or too limited to address the anticipated demand.

[22] Such strategies are common in variable rate encoding (e.g., compression algorithms) and location awareness (e.g., predicting future location indoors based on trajectory when last connected to GPS).

[23] Apparently, Spread Networks deployed a fiber network that could shave off 3msec in transit time between Chicago and New York (see "What's the big deal about Michael Lewis and high-frequency trading?" NPR, April 2, 2014, available at https://www.pbs.org/newshour/nation/whats-big-deal-michael-lewis-high-frequency-trading).

In such cases, the best solution to address such on-demand (unanticipated, ad hoc) needs for computing infrastructure may be to deploy mobile computing infrastructure. One way to do this is via portable mobile cell sites or data centers that can be packed in a tractor trailer, driven to the site needing the capacity, and quickly brought online. Several Non-terrestrial Networking platforms using high-altitude drones or balloon platforms demonstrate the potential for such ad hoc deployment strategies.[24]

2.3. Nextgen Edge Computing: MECs instead of general cloud computing

Although for many activities/tasks there are many computing architectures and business strategies (centralized v. distributed, in-house v. outsourced, etc.) for meeting the digital infrastructure needs of end-users, there are a range of design characteristics that are unique to mobile/portable applications. Those include mobility in geospace at both high-speeds (the sort of mobile connectivity support that has been the flagship capability of cellular phones[25]) and nomadic speeds (where users move between access points slowly or remain within the coverage area of a single access point while connected). It also includes portability and cable-free connectivity (which facilitates portability). The need to be portable as well as mobile often imposes size and power constraints. The radio devices must be small enough to be carried or be attachable to the device it provides connectivity to, and cable-free access requires portable power. The need to economize on size and power constraints is a reason for separating the computing logic from the device, but also keeping it relatively close. For example, thin-client devices (those that provide connectivity but only limited on-board ICT resources) are often preferred in many IoT circumstances, or where access to good connectivity and cloud-based ICT services are available. That is why Google Chromebooks and eReaders as substitutes for more powerful personal computers have been able to compete successfully with personal computers (fat-client devices). For consumers, the choice of a thin- or fat-client solution for accessing a range of Internet (cloud) based applications and services is a common decision that balances a range of considerations like cost, weight (convenience), and performance. Indeed, when confronted with so many options for accessing and making use of ICT resources in our daily lives, many end-users have multiple devices (tablets,

[24] For example, the High Altitude Platforms (HAPs, www.hapsalliance.com) alliance promotes a range of communication platforms that can operate in the stratosphere (18-50km above earth surface) that have the advantage of being able to offer wider-coverage than terrestrial access points, being removed from terrestrial disasters, but also being able to be launched and brought back to earth for upgrading and specialized redeployment in new locations easily. Altaeros, an MIT spinout, has offered its "SuperTower" aerostat since 2019 as an aerial platform for a full cellular macro-cell site (or platform for other digital connectivity and computing capabilities). The aerostat is an AI-controlled blimp that is capable of being remotely launched from the deployment truck it is tethered to, and with AI-controlled stabilizers and sensors it can keep itself stable in even high winds but can also be hauled in if needed and then redeployed (see, www.altaeros.com). Other options for such NTN digital platforms include high-altitude UAV drones and LEO satellite constellations that can deliver very high-speed connectivity and low latency to focused locations.

[25] Smartphone services that can support continuously connected mobile broadband when devices are moving at high speeds presented special networking challenges, including wide-area coverage and the ability to rapidly hand-off connections to new access points. Enabling such capabilities adds costs and complexity that can be avoided with RAN technologies such as for WiFi where users are expected to mostly be within range of a single or few access points.

smartphones, smart TVs, portable, and/or desktop PCs, etc.). Each of these devices provide a range of edge-computing and digital connectivity options but different ones may be preferred by users under different circumstances.

What all these edge computing resources have in common is that they are closer to the network edge than are the cloud computing resources that may be located relatively distant from an end-user and the device on which an application and the end-user's user interface (UI) is running. Moreover, when these infrastructures are designed to support mobile broadband networks, they are distinct from cloud applications which could be located anywhere. Some of the additional advantages or reasons why MEC may be preferred to more distant computing resources are because of the latency and DTN benefits discussed earlier. Another potential performance discriminator is the fact that local processing of local data may offer security benefits and reduces the utilization of unnecessary wide-area connectivity resources. From a cost perspective, keeping data processing local for data that is inherently local is analogous to the motivation behind the deployment of edge-caching in Content Delivery Networks (CDNs).[26] Thus, it may be desirable to process visual data locally to reduce the load on long-haul transmission capacity (e.g., by compressing the images). Additionally, processing the data locally reduces the potential risk of unwanted exposure (e.g., privacy violations). For example, visual data may be transformed to protect against surveillance tracking.[27]

3. Models of Edge Computing

As noted already, the need for ubiquitous high-quality broadband connectivity is now generally accepted and part of national digital infrastructure policy globally. For example, the US identifies locations with less than 100/20 Mbps (downstream/upstream) as "unserved" and the European Parliament has set a goal of universal access to 1Gbps service to everyone by 2030. These goals cannot be justified based on current needs for anyone with BB today.[28]

---

[26] See Stocker, Smaragdakis et al. (2017).

[27] The surveillance tracking might even be unintentional, exposing those deploying edge devices to liability for privacy violations. For example, with the proliferation of inexpensive wireless cameras, those may become attractive as general-purpose IoT devices for collecting data like temperature or precipitation information (see Lehr & Sicker, 2017). Indeed, it may be less expensive to use existing cameras or even deploy new ones rather than specialized IoT sensors. However, for such applications a drawback is that they collect too much information in visual data. An inappropriately aimed camera might capture images unintentionally from a homeowner's window, leading to unintentional liability for privacy violations. Processing the video locally to extract only the desired IoT information (rainfall, temperature, etc.) and sending only that upstream could reduce such risks.

[28] In the U.S. today, there are few applications and few users who would notice any significant performance improvements with service speeds of more than 50/50Mbps downstream/upstream, and most users are not significantly impaired in using most of the most common Internet applications with access services offering at least 25Mbps/3Mbps. This raises the question of whether the broadband quality targets may be overly ambitious. However, if one accepts that there is a need for much more interactive and resource intensive applications of the sort noted earlier (for AR/VR, robotic automation, and real-time AI control of complex systems), then much higher speed and symmetric broadband service may be needed.

Consequently, the economic case for ubiquitous access to high-quality broadband needs to rely on much more data-intensive, highly interactive applications. While there is certainly demand for such applications for entertainment and training purposes (e.g., gaming, immersive simulator and training applications, and VR/AR applications like the Metaverse[29] and Digital Twins[30]), high quality digital connectivity would not be enough to enable such applications. They are also computationally intensive, and although the computing resources do not always have to be local (i.e., close to the real-world devices in network distance, which may be measured in geodistance or time), sometimes they do. For reasons noted above, some computational tasks need to be local.

Moreover, from an economic perspective, the applications ought to be employed to realize Smart-X automation tasks rather than be limited to entertainment applications.[31] If limited to entertainment, the potential to lead to significant economic growth or productivity benefits are limited since are an important but relatively small share of total consumption. Better digital entertainment is likely to simply displace poorer digital and non-digital entertainment options. On the other hand, Smart-X applications may address *any* economic or social task or activity that can benefit from digital augmentation or automation.[32] That includes smart-cities, smart-supply chains, smart-healthcare, and smart-business processes. Embedding computing and storage capabilities in this expanding array of locations, devices, and contexts is part of the infrastructure needed to enable ambitious visions of Smart-X. That is what is driving the expansion of IoT.[33]

---

[29] Definitions of what constitutes the Metaverse vary. One vision is that it will allow users to seamlessly interact in and with others in a virtual world. This could be via an avatar to allow a user to act within the digital world, or via various IO devices and actuators to connect virtual world events to the real world. The metaverse represents an evolution and convergence of AR/VR systems (see for example, https://www.nytimes.com/2022/01/18/technology/personaltech/metaverse-gaming-definition.html, visited 5/16/2022). Facebook rebranded itself as "Meta" in October 2021 with the plan "to bring the metaverse to life and help people connect, find communities and grow businesses" with a much more immersive Internet capable of interacting with the physical world in much richer and seamless ways" (see https://about.fb.com/news/2021/10/facebook-company-is-now-meta/).

[30] A Digital Twin is a virtual world representation of a real-world system, replicating the behavior of the real-world system in the digital world. If the system is complex (e.g., a digital model of a factory or supply chain) and the model a good representation (not an over-simplified abstraction) then it can be used to test out scenarios such as the outcome of environmental changes, bug testing for new software updates, the outcome of the failure of sub-components, changes in operating procedures, and a host of other things. Those tests can run faster than the real-world and so can simulate real-world outcomes and then analyze the scenario outcomes to enable better real-world decision-making. That is the hope, but obviously it depends on the fidelity of the Digital Twin to its real-world counterpart. However, if the hope is realized, it could dramatically alter production and operating processes for any business with access to good Digital Twin models.

[31] See Lehr & Sicker (2017).

[32] AI advances challenge the question whether, indeed, there are *any* tasks that may be beyond the potential of ICT automation, and its implications for restructuring how humans interact with the world they inhabit (see Lehr, 2022).

[33] For example, see https://sociable.co/business/future-proofing-for-industry-4-0-the-role-of-ai-blockchain-and-edge-computing/ (visited 5/16/2023).

Today, virtually all businesses have embedded demand for real-time access to ubiquitous digital computing and communication services into all aspects of their operations. Most white-collar workers have both personal computers and smart phones, and regularly access and make use of software applications like email, chat, videoconferencing, word processing, and a host of other applications. This growth in the use of on-demand computing and storage resources and recognition these are already essential infrastructure for modern business is demonstrated by the rise of cloud computing and data centers. The emergence of cloud computing and data centers has not caused end-users to get rid of their personal (fat-client) devices, but with the expansion in connectivity and remote and distributed computing options, demand has increased for cloud services to augment and sometimes substitute for on-board (on-device) computing and storage. Indeed, in some cases, more data intensive cloud applications drive more intensive need for edge computing resources.

For example, the rise of digital cameras, digital music, eBooks, and expansion of websites (blogs, wikis, and other venues for digital media) has dramatically increased the cumulative volume of digital data that individuals and businesses want to store, access, share, and manage (including secure and occasionally delete). In response, a growing number of businesses and end-users are now using cloud storage to back-up the growing compendium of digital artefacts (written, audio, and image data) that is being continuously collected. Networked cloud storage facilitates more convenient individual data accessibility from multiple devices and locations, as well as enabling users to share data on social media and other digital platforms.

Architectures for both cloud and data centers have been changing and evolving (as we will discuss later in this section), in response to and enabling market growth[34] and technological advances. Increasingly, in the same way that data networking infrastructure became essential for business, eventually morphing into today's Internet that is now essential for everyone, so too will access to on-demand computing and storage become critical infrastructure for everyone.

[34] The demand for computing resources (cycles, disk storage, communications in MBps) has grown significantly in aggregate and on a per capita basis. More individuals are using more digital applications, and their usage is more data-intensive across almost all demographics. This is analogous to the growth of computer uses from days of mainframes (when only large enterprises could use) to today when nearly everyone has a smartphone, and a growing number of computing chips are being embedded in our environment. Corporate computing has gone from VPNs to public clouds to hybrid clouds for large and increasingly even smaller enterprises (SME adoption growth). This growth has been fueled by modularization of technology and drop in prices for services and equipment. There is cycle where marginal (cost-conscious) users can first only access as a service (e.g., time-sharing on mainframes), then equipment becomes inexpensive enough that users can self-provision (e.g., PBXs and PCs), and then services become so inexpensive and competitive that (some) users shift back to services (e.g., clouds and thin client architectures). That also happened in telecoms with CENTREX being displaced by PBXs which were augmented by SDN to challenge newly intelligent NFV-enabled networks. What is different is the choice of whether to self-provision or rely on a service. When the capability can be offered in many ways, choice becomes an option and control of choice increases as a key system concern.

From a technical perspective, the provisioning of computing infrastructure is also evolving with the emergence of newer, more distributed architectures.[35] Where previously (as discussed earlier), the only viable architectures were centralized, it is now possible to have distributed architectures wherein resources are shared across networks that may be owned and managed by many different, independent entities.[36] Those newer architectures include distributed, federated edge clouds. Figure 1 compares these new architectures for providing computing resources with the more traditional centralized cloud architectures.

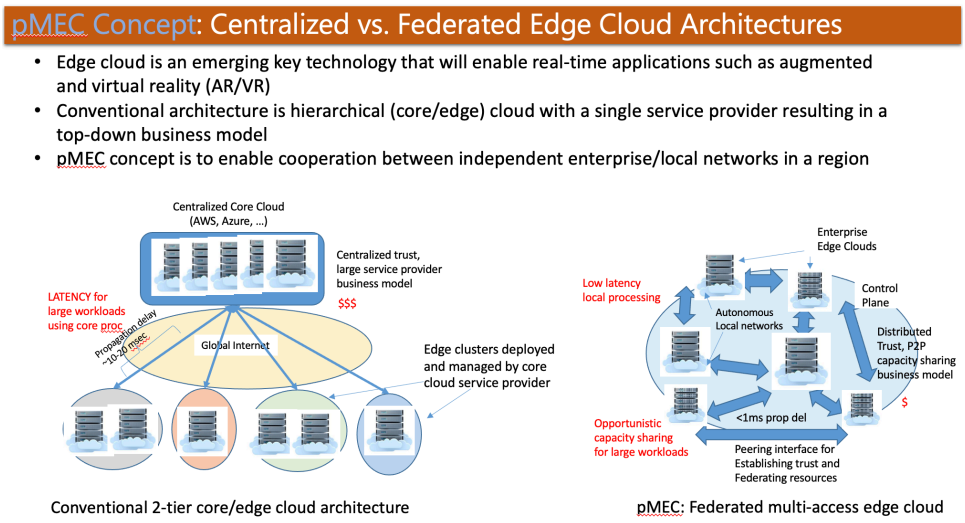


Figure 1

As depicted and described in the figure, an edge model allows for real time services and supports novel thinner edge clients. Moving from a centralized to a distributed and federated model allows for a shared service provider model, where new models of cooperative services could evolve.

3.1. Service Provider Models of Edge Computing

As described above, there are two obvious service-provider business models for meeting future needs for edge computing services:[37] (1) Cloud, Data Center, and Content Distribution Providers,

[35] For a discussion of research into a future Internet architecture research, see Sollins & Lehr (2020), and for a discussion of mobility-centric architecture, see Venkataramani et al. (2014),

[36] End-to-end traffic on the Internet is typically multi-hop traffic, although with the changing topology of the Internet, the number of inter-ISP hops has changed (e.g., rise of hyper-giants, see Clark et al., 2016, or Stocker et al., 2017).

[37] Other service provider models are also possible, and examples of those already exist. In focusing here on the examples of edge-providers or ISP broadband access service providers, we focus on the two types of players we anticipate will play the most obvious prominent roles. For example, niche service providers of integrated network and application services (e.g., financial service providers like banks that already operate ATM and other widely accessible ICT networking infrastructure) or other resource providers like antenna companies (e.g., Crown Castle, American Tower) are examples of other national scale, service providers that may become providers of edge computing services. Indeed, if the pMEC concept we explore here is successful, then we anticipate that it will give rise to service provider business models as entrepreneurs

which are sometimes referred to as Edge Providers, to differentiate them from the traditional providers of telecommunications infrastructure and services; and (2) Internet Service Providers of broadband access services, or the traditional providers of telecommunication network services.

The recent emergence of high-speed wired and wireless access at Gbps speeds (e.g., 5G, WiFi 6) is once again tipping the balance away from computing at the client and more towards computing inside the network. This time, the client-server or client-cloud model is evolving into a three-tier model with the introduction of a close-in computing layer known as the "Multi-access Edge Cloud" (MEC). This middle tier of computing infrastructure in the network is driven both by technical and economic forces. On the technical side, the fact that cloud computing latency at centralized data centers is lower bounded by the propagation delay (~20-50 msec) through long distance fiber, has motivated early deployments of a new edge cloud infrastructure designed to achieve sub 10 msec latency as needed to support real-time AI-enhanced applications such as augmented reality (AR), virtual reality (VR) and industrial control.

The economic forces include the falling prices for edge-servers – which is a result of continuing innovation, hardware modularization and commodification, and lifecycle hardware pricing trends. As already noted, the continuation of Moore's law technical progress has resulted in falling prices for digital computing and storage hardware, allowing digital CPU and storage to become available in modular, commodity hardware in a wide array of physical configurations and capacities. Also, lifecycle hardware pricing means that equipment prices fall for older-generation equipment, as yesterday's top-of-the-line equipment is supplanted by the latest versions. This makes it feasible for smaller data centers to install hardware and software capabilities that were previously uneconomic except for hyperscale data centers; and for the hyperscalers to migrate those older servers into their expanding edge-networks to make room for the latest, high-performance servers in their flagship hyperscale core data centers. In this way, the cloud computing equipment economics results in computing resources becoming more widely distributed throughout the cloud service provider networks.

3.1.1. Edge Provider's expansion into edge computing

The largest Cloud/Data/Content Providers include such companies as Google, Amazon, Microsoft, Meta, Cloudflare, Akamai, and IBM. These global providers of cloud services are perhaps the most obvious providers of edge computing services since they already provide extensive cloud-based application support and have been expanding their portfolios of offerings for years. As digital connectivity has expanded, computing costs have fallen, and the need for increased computing resources by all types and sizes of enterprises has grown, the market for cloud services has grown apace. For example, Gartner estimated that the worldwide public cloud services market grew over 20% per year since 2020.[38] The largest enterprises were the first to adopt cloud services, but as the

adopt the model for deployment in newer markets and seek to realize the scale, scope, and learning economies that growth in geographic and market size portend.

[38] Raza, M. and C. Kidd (2021), "The Cloud in 2022: Growth, Trends, Market Share & Outlook," BMC Blog, October 21, 2021, available at https://www.bmc.com/blogs/cloud-growth-trends/. The Infrastructure-as-a-Service (IaaS) market grew to $90.9B in 2021, and over 80% of the revenue was contributed by the

market has grown, adoption has spread further into all size categories, with small and medium sized enterprises (SMEs) seeing the greatest expansion in accounts.

The Cloud/Data/Content providers benefit from the ability to realize significant scale economies when their data facilities are centralized.[39] One disadvantage that centralized locations pose, however, is that they are not geographically close to their enterprise customers. Of course, this begs the question of how close is close enough and based on some initial back-of-the-envelope work, we have found that with the expansion in cloud service provider networks (with multiple hyper-scale data centers located strategically across their markets, and with expanding arrays of smaller data centers in smaller, secondary locations), the distance to customers has been shrinking. With distances to customer locations of up to 50 miles, these cloud providers can still support many applications requiring 10 msec (or faster) round trip latency. Given such numbers, it is reasonable to assume that today's cloud providers could reach most of the population without significant further expansion of the physical topology of their server networks.[40]

3.1.2. ISP Broadband access provider expansion into edge computing

The second obvious class of players that we expect to see offering edge computing services are the fixed and mobile Internet Service Providers (ISPs). They are well positioned with digital connectivity infrastructure close to where businesses and consumers connect to wider-area networks. These ISPs include providers of both fixed and mobile broadband services, which increasingly is the same company as ISPs derived from legacy cable networks service providers like Comcast, Charter, and Cox add mobile services; and cellular-based Mobile Network Operators (MNOs) like Verizon, AT&T, and T-Mobile expand into offering fixed wireless broadband.[41] These network-based operators are expanding their broadband and other service offerings to expand their product lines and integrating into application and service provisioning to avoid being

top 5 global service providers: Amazon, Microsoft, Alibaba, Google, and Huawei (in descending order of market shares) (see "Gartner says worldwide IaaS Public Cloud Services Market Grew 41.4% in 2021," Gartner Press Release, June 2, 2022, available at https://www.gartner.com/en/newsroom/press-releases/2022-06-02-gartner-says-worldwide-iaas-public-cloud-services-market-grew-41-percent-in-2021).

[39] The scale/scope economies arise with respect to many cost elements, including power, real-estate, maintenance personnel, and other cost components.

[40] Of course, this does not address the specialized needs of businesses that located in remote locations (e.g., mines).

[41] The legacy fixed telephony operators were among the first to offer cellular mobile telephony services based on the family of cellular technologies that is currently deployed based on the 3GPP 4G LTE and 5G family of standards. More recently with the growth of WiFi coverage, legacy cable television operators have expanded into mobile telephony, purchasing wholesale coverage from the cellular network operators where WiFi coverage or capacity is inadequate. In addition, a host of other technologies including Non-terrestrial Network (NTN) providers are offering an expanding array of satellite-based broadband connectivity platforms such as Starlink and Amazon's Project Kuiper (LEO-based networks) seek to compete with MEO/GEO-based platforms from providers like Hughes and Viasat and an expanding array of local, smaller fixed wired and wireless ISPs (e.g., WISPA).

relegated to being mere bit pipes by edge-provider. With the expansion of CDNs,[42] the last-mile ISPs are playing a larger role in hosting edge servers and with the transition to software architectures, the capabilities of last-mile ISPs to offer an expanded array of computing services has expanded.[43] Moreover, the last-mile ISPs and legacy cloud/data center/content delivery service providers can collaborate to offer an expanding array of hybrid service-provider edge-computing services.

Finally, the next generation of MNO cellular technologies anticipated for 5G+ networks call for enabling expanded computing platforms in the base stations of 5G+ cell sites. Additionally, to take advantage of higher-frequency spectrum at 10GHz and above (into millimeter wave frequencies),[44] to facilitate spectral reuse, and to expand wireless capacity, 5G+ networks call for the deployment of many more, smaller cell sites.[45] Indeed, the acronym MEC was first used to describe the Mobile Edge Computing resources that are expected to be part of the next-generation for 5G+ mobile networks. Some mobile providers are currently in the process of building out 5G capable networks and have plans for deploying MEC platforms closer to end-users with the goal of competing on latency and performance to capture a share of the huge and growing market for cloud computing services.

3.1.3. Mainstream Service Provider Models will still fall short

While we anticipate that these traditional large service providers will play an important and likely dominant role in meeting the future need for edge-computing resources, relying solely on such providers will likely leave many end-users and communities underserved.[46] In the most challenging cases where the economic value to be captured by service providers is inadequate to support profitable service provider operations because the costs are too high (e.g., in low density rural or other high-network-cost environments) or per-subscriber revenue potential is too low (e.g., in poor or otherwise economically disadvantaged communities), public subsidies and infrastructure investment may be needed to ensure that those citizens and communities have

[42] See Stocker, Smaragadakis et al. (2017).

[43] Network Function Virtualization (NFV) is the term used to describe the wholesale transformation of the back-end software-control architecture of legacy telecommunications networks toward a layered software architecture which enables operators to better dynamically control and virtualize their underlying network infrastructure. NFV often goes together with Software Defined Networking (SDN), which is complementary of movement of functionality from what used to be implemented in dedicated hardware into software. For further discussion, see for example, Han et al. (2015).

[44] For a discussion of spectrum management challenges, including for use of higher-frequency spectrum, see Bhattarai et al. (2016), Singh, Sicker & Lehr (2019) and Lehr (2016).

[45] See Lehr, Queder, & Haucap (2021), Lehr & Oliver (2014).

[46] The determination of what constitutes being "underserved" or the appropriate timing for addressing those needs raises contentious debates. With limited resources and capacity constraints at multiple levels, it is not generally feasible to address all gaps simultaneously, and so some queuing of efforts to address gaps is unavoidable. What the pMEC option provides, in part, is an opportunity for end-users to step in with their own efforts to meet gaps when other efforts are deemed inadequate by those end-users.

affordable access to the digital connectivity and computing resources they need to participate in the digital economy.

Additionally, many communities that are served may be served by only a single provider. In such situations, the services offered may not be affordable or may poorly match the needs of subscribers, leaving those communities and subscribers underserved. Indeed, a goal of broadband policy is to enable competition and consumer choice which depends on their being multiple broadband and computing services being available in the market from multiple service providers. Thus, even after the basic goal of delivering at least a minimal level of digital connectivity and computing resources accessibility to all citizens is met, there will be a need to expand competitive options and the quality of services to keep pace with evolving market and technology needs. Finally, there may be communities and users (in some part of communities) that have demand for capabilities that they view as being inadequately met by existing service provider options. They may wish for capabilities that exceed the standard that is deemed as an adequate threshold, disqualifying those users from claims for public infrastructure support. Although such users/communities may not be eligible for universal service programs, they are still among the population of under-served demand if the only solutions available depend on the offerings from the large, for-profit, service providers.

To avoid these failures of relying solely on service provider business models to address the needs of unserved and underserved communities, we propose the concept of an open participatory public multi-access edge cloud (pMEC) and then identify key enabling technologies, business models, regulatory policies, and human factors necessary to enable cost-effective and sustainable shared infrastructure services to the entire community. This approach to edge clouds draws inspiration from the Internet, which grew organically during the 1980's and 1990's after release of fully distributed inter-domain routing protocols and associated service level agreements (SLAs) that encouraged independent network domains to peer and contribute resources to the global network.

3.2. An Alternative to the Service Provider MEC Model

We now ask the question – is there an alternative to the service provider model of MEC and in what ways is pMEC different from what the carriers and the service providers are proposing? We argue that the pMEC concept provides an alternative model for enabling edge computing cloud capabilities, while still supporting emerging real-time mobile applications with tight latency constraints. It may also provide other additional benefits relative to a service provider model, such as lower total cost, improved privacy, potentially greater resilience, and end-user autonomy (or control).[47]

[47] End-users may wish to own and control their computing infrastructure for individual reasons, even when to do so incurs additional costs or offers reduced performance, at least in the short run. There are many strategic motivations for why end-users may choose to play a direct role in ownership, provisioning, and management of their digital infrastructure. At one extreme, end-users choose to self-provision (buy and operate their own computers, develop custom applications, etc.); while at the other, they may choose to outsource those activities to service providers. Likewise, groups of end-users may choose to collaborate to self-provision for a portion of their computing needs; and even when end-users opt to outsource a portion of their IT service needs from a service provider, many enterprises opt to outsource from multiple service providers, even when bundled offerings may offer discounts and other benefits. There are often bargaining

The difference between pMEC and what other service providers are offering depends on the context being considered. From an economic perspective, the value of pMEC is determined relative to its next best alternative.[48] In cases where the local needs for edge-computing are perceived to be large, no service provider solutions are available, and end-users are sufficiently motivated, a pMEC may offer the only alternative. The pMEC's value proposition in such cases is measured by the value of meeting that need (e.g., value of reduced latency that only achievable today with a pMEC). In other cases where service provider alternatives exist, the pMEC value proposition depends on how its cost/quality compares incrementally to the available alternatives. Those valuations may vary by end-user context. For example, even if the pMEC alternative offers reduced reliability or other desirable features/capabilities, the pMEC may still offer an attractive option. For example, it may be less expensive (e.g., because it takes advantage of already existing excess local computing resources),[49] offer benefits in local control/autonomy,[50] or provide robustness benefits in the event of a loss of wide-area connectivity. Moreover, adoption of the pMEC model need not foreclose adoption of service provider options also. It is reasonable to anticipate that if the pMEC option is sufficiently attractive for some end-users that those end-users will meet their ICT needs for edge computing resources via a hybrid model mixing locally owned and service-provider resources.

---

and resiliency benefits from acquiring services from multiple service providers, rather than sole sourcing from a single provider.

[48] The adoption of the pMEC ought to be core stable. That is, participation in a pMEC ought to be individually rational for each participant. Different participants may have different net valuations for pMEC participation, and those may be contingent on the identities of the other participants in the pMEC. Thus, there may be a critical mass of participants needed to form a pMEC and that may vary by context. There may be a natural distribution in each context of early and later pMEC adopters. Moreover, if a pMEC grows sufficiently large, it may be desirable to create a custom-service provider to meet the pMEC user groups needs or the potential market offered by the pMEC may induce service providers to change their offerings to render the economic case for the pMEC no longer viable. Thus, pMECs may require a degree of activation effort to launch and may cease to be viable (or desirable) once they reach a level of success. That does not render the pMEC concept unimportant, since the potential for a pMEC to emerge provides a source of competitive discipline – or contestability – that can discipline the behavior of service providers and induce them to better match their offerings to end-user needs.

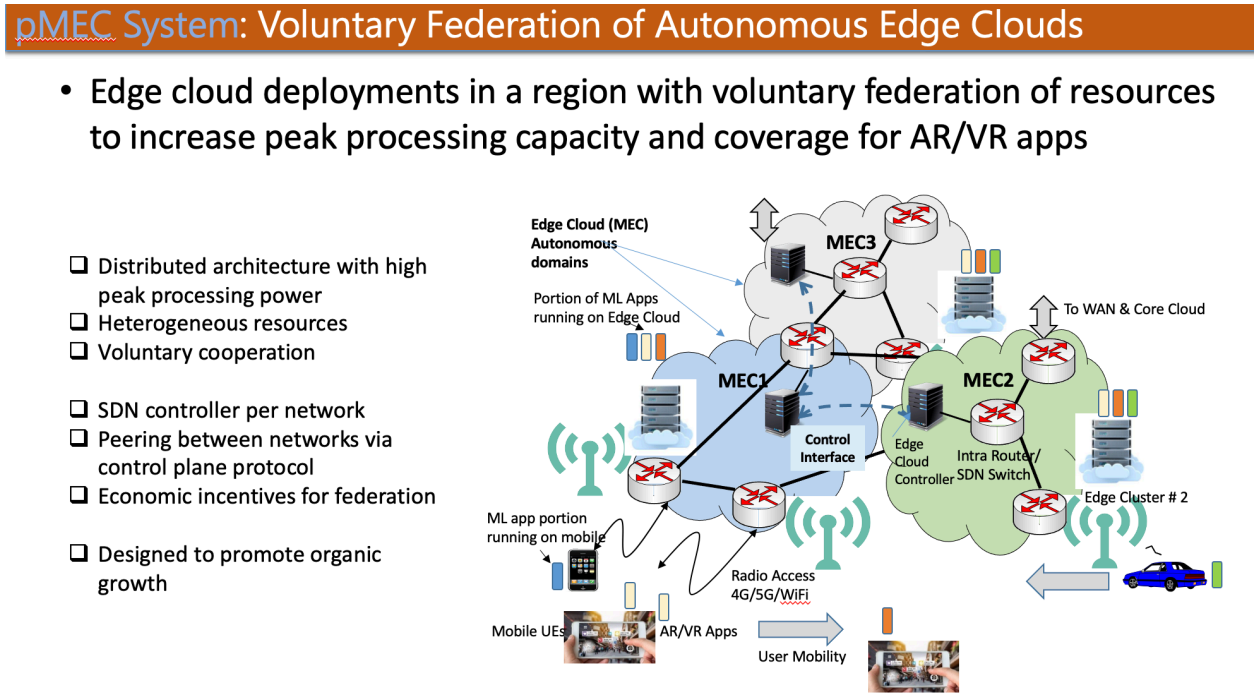
[49] Comparing the cost of participating in a pMEC versus purchasing services from a service provider involves a make-vs.-buy decision, and such analyses are inherently complex. Although making use of excess local computing resources may avoid having to pay usage fees to a service provider, there are end-user costs of taking advantage of shared local computing resources that need to be factored into the decision. Even if the overall market trend suggests that more end-users are opting to outsource their ICT needs to large service providers, that still leaves substantial room for end-user alternatives (if feasible) to prove cost-effective, especially if those can be designed to confront adopters with low adoption and use costs. That is part of the technical and business design challenge confronting the pMEC concept.

[50] As computing resources become more mission-critical for more businesses, businesses may perceive benefits in retaining strategic control over critical business resources and functionality, even if that incurs higher costs. Retaining such control may enable more rapid strategic responses in product design and offer better control over business confidential information (e.g., trade secrets).

From the preceding, it should be clear that although we view the pMEC concept as a promising focus for research, the challenges it confronts to offer a viable commercializable economic solution will confront significant technical and non-technical challenges. However, as should also be clear, much has changed to bring the concept closer to realization.

From a technical perspective, the proposed pMEC approach is based on the concept of edge cloud as a ubiquitous local infrastructure that self-organizes into a high-performance computing system via peer cooperation among multiple autonomous systems. The pMEC architecture outlined in the figure above comprises a collaborating group of autonomous edge cloud domains consisting of a mobile core network along with collocated edge compute clusters as shown. Mobile end users (such as pedestrians or vehicles) access the network wirelessly using existing or emerging radio standards (including 4G, 5G or WiFi) to access applications such as AR for urban navigation or V2X (vehicle to everything)/autonomous driving, with most applications typically involving some form of machine learning (ML) and human interactivity. As shown, the user devices run a portion of the application locally and migrate the more compute intensive portion to edge cloud servers distributed across the mobile services network. The architecture assumes the existence of an edge cloud domain controller (ECDC) for routing/assignment of tasks to multiple edge servers to meet application latency demands while balancing the traffic load. The proposed architecture is illustrated in Figure 2.

Figure 2



In the above architecture, the edge clusters use a suitable distributed cloud computing platform, such as Apache Storm or Kubernetes, as the base layer. This software platform can be used to

distribute tasks to multiple edge servers for parallel processing and subsequent collection of results (i.e., map-reduce) and deliver the high-performance computing resources at lower end-user cost needed to meet the technical demands of highly interactive, low latency or local outage intolerant applications while offering the potential for performance improvements over more centralized computing infrastructures. While some may argue that a service provider will always be able to deliver a higher degree of availability, reliability, and resilience, this depends more on how each MEC instance is deployed and managed. What can be said with more confidence is that the pMEC model will keep data local and out of the hands and eyes of the service providers who may not be trusted not to continuously look for ways to monetize user data.

4. Implications of Edge Computing

Here we briefly consider the implications of expanded edge computing and the pMEC concept for several key communications policy issues. These include the implications on broadband architectures, broadband competition, universal service policy, interconnection, and related digital infrastructure-related policies.

The concept of a pMEC in its simplest form is really nothing more than the concept that a group of end-users in a geographically contiguous area (a town, a neighborhood, an industrial park) having distributed computing resources that are mismatched in terms of local supply and demand that are capable of being shared over a local network. Generically, as individual user needs for local computing resources have expanded and become more resource intensive and bursty (i.e., peak to average usage has expanded), most users have excess (under-utilized) local computing resources at least some of the time. Most PCs and servers do not use 100% of their capacity 100% of the time. Indeed, one way to increase utilization is to time-shift tasks to reduce intertemporal peaks (e.g., shift non-time-critical tasks such as file maintenance and software updates to overnight); however, users may reasonably fail to provision sufficient capacity to meet all of their peak needs, requiring them to fail in meeting those needs, reduce the performance expected by those needs (e.g., due to congestion), or be forced to acquire additional resources elsewhere. A key benefit of cloud services is their ability to scale dynamically, allowing end-users to meet their peak demand resource requirements without relying on excess self-provisioned capacity. However, since local user peak needs are unlikely to be perfectly correlated in time at both short and longer-time scales, there is significant potential for sharing local excess capacity to meet local peak demand needs.[51]

[51] At longer, investment time scales, local users are likely to have excess capacity since computing capacity is added in fixed, lumpy increments. That is rational since expanding capacity incurs a fixed (and often sunk, one-time) cost elements that may be optimized if excess capacity to provide for future growth is installed. Other users, who are anticipating but have not yet invested in an upgrade, may be systematically more likely to have inadequate self-provisioned capacity and so present as viable customers for the excess capacity available from local users who have recently upgraded. Moreover, due to the uncertainty of demand and vagaries of market share shifts and aggregate shocks or outages which may impact the self-provisioning supply/demand capacity matching for individual users differently, there is likely to be locally under-utilized and resource-seeking local users even over the short-run. Finally, the pMEC model may help stimulate the emergence of more active two-sided resource sharing among end-users and among service

Obviously, the ability to share computing resources depends on their being adequate digital connectivity. In the U.S. and most of the most developed markets broadband digital infrastructure is already extensive, and significant public investment programs and national policy goals to ensure ubiquitous availability of broadband services with data rates of 100Mbps or greater suggests that the digital connectivity infrastructure needed will be widely available. However, that does not mean that the local digital connectivity needed to ensure low-latency performance will be readily available. First, users even in the same local market with high-quality broadband services provided by different service providers may find their traffic has to be routed to distant locations to be interconnected. Such interconnection failures can result in single-provider roundtrip latency performance in the low 10s of milliseconds to be multiplied several times (once other interconnection overheads are added), resulting in excessive latencies such as those justifying the original interest in edge computing. Second, there are a wide array of wireless technologies and options for enabling digital connectivity that may not be supported by large service providers (e.g., free-space optics or millimeter wave private point-to-point links) but which may be economical for end-users in particular local contexts. Thus, the fabric of local digital connectivity, when coupled to the fabric of large-scale service provider digital connectivity may enable a much wider array of network sharing options than would be feasible with only one or the other fabric alone.

If such a rich digital connectivity fabric exists, then a wide array of more demanding applications in need of abundant computing resources may be feasible. This is the hope of the Metaverse visions. Such a fabric could support a user with a thin-client (lacking on-board computing resources) utilizing a virtual reality app or a computationally intensive gaming app. The pMEC resources could support this locally or complement the support of the application that may also utilizes service-provider cloud resources. The pMEC resources could be shared among local resident users that may also contribute to making the resources available, or when local users roam abroad, may allow those roaming users to access the pMEC resources of affiliated pMECs under a federated model. There might even be multiple federated pMECs in a local area, differentiated by their core user models and the resources they make available to their members (as opposed to members by virtue of their federation membership). This is akin to the EDUROAM model that has emerged to support the sharing of broadband access services among EDUROAM member networks, mostly academic institutions. Under the EDUROAM federation model, faculty and students with authenticated access rights to their home institutions network can easily access the networks of other academic members of the EDUROAM federation when they are traveling. A similar arrangement exists among cellular service providers wherein cellular users may be granted roaming rights on other service provider networks, potentially with a lower level of access privileges and at higher prices than for the subscribers of the hosting network. Today, those cellular

providers. For example, today, most service providers of cloud services stand ready to meet peak customer demands with resources that they can provide dynamically, but potentially with surge pricing. The rise of pMECs, like the rise of ride-sharing platforms like Uber or solar-energy-connected renewable power grids, may enable the two-sided matching of excess capacity among user edge networks and service provider networks to more cost-effectively and resiliently meet peak demand needs and provision for baseload utilization (e.g., by shifting which computing resources are utilized to meet off-peak demands also).This assumes that the large service provider market is one-sided (that is, large service providers are willing to sell capacity but potentially subject to surge pricing -- like Uber – but are unwilling to purchase excess capacity from end-users today).

providers are adding additional services that may also enable roaming (e.g., from voice calls to data roaming, where the latter is often subject to speed and MB caps) to provide a richer selection of service provider offerings. The EDUROAM network offers an attractive alternative for universities to collectively self-provision for their capacity. The pMEC concept may offer a way to extend the EDUROAM model to other digital resources in high-performance computing that are increasingly under-utilized and mismatched in their allocation.[52]

In short, we see the need to deploy edge computing resources will necessitate acceptance of more flexible and complex passive and active sharing among value chain participants, which will require a reassessment – and likely reframing – of legacy models of industry structure and regulation and a re-thinking of the role of end-users and edge-community networks. We therefore see the pMEC option as an important test case worthy of further exploration. If successful, it has the potential to introduce an important new vector for the emergence of competitive discipline and for reasserting end-user autonomy. If unsuccessful, a clearer understanding of why edge computing infrastructure is and should be a capability provided by large service providers in a relatively tight oligopoly, ought to help drive consensus toward appropriate regulatory models.

In the near future, mobile broadband access will not just focus on how big of a "bit pipe" or the reliability of that pipe, but will also focus on latency, and as latency becomes more important, it will be necessary to consider how and where computing should occur in the network. This inevitably means how do we craft policies to ensure that the public has access to mobile edge computing in a manner that serves their needs. As the network, data and application architectures evolve, edge computing will be another essential part of universal service and drive new shared models of investment and competition. It will also mean that interconnection models will need to consider not simply the "bit pipe" of data but also how to fairly deploy, control and access applications at the edge of the network.

Additionally, to the extent the "bit pipe" needs to be shared and is a natural monopoly resource, then it is a potential bottleneck facility that may require regulation, especially if it is provisioned as a public utility with public funds.[53] Moreover, today's potential bottleneck facility might be used as a vehicle for creating a new bottleneck resource that may leverage market power into new markets or further extend the market power afforded by today's bottleneck resource.[54]

---

[52] Many institutions lack significant HPC resources, while others with significant HPC find those resources under-utilized. Sharing those resources nationally is one motivation for Internet2 and mirrors the economic driver behind cloud computing. Enabling the local sharing of those resources across on-campus communities and for visitors is what the pMEC model would facilitate.

[53] See Lehr & Sicker (2018, 2019) on regulating a bottleneck facility. If public funds are used, then there is a public interest in ensuring that those resources are used to benefit the public, and not just private interests.

[54] Historically, telephone service was regarded as a natural monopoly, but over time, the components required to deliver telephone service and the network infrastructures and market services those components have enabled have significantly redefined industry value chains. We have gone from the end-to-end monopoly AT&T Bell System to today's much more complicated and decentralized industry value chain, but one in which control over legacy core elements have continued to provide a basis for significant market power. For example, control over last-mile infrastructure, key resources (like access to RF spectrum or rights of way), or key technologies (like software essential patents) may give rise to new bottleneck resource concerns as digital technologies, markets, and industry value chains continue to evolve. Examples of

One policy response is to work to enable actual or potential competition for the bottleneck resource (i.e., make the market for the bottleneck resource contestable).

This might be accomplished by enabling sufficient intermodal competition. Intermodal competition is competition via alternative technologies or business models and may arise in multiple ways. For example, legacy cable television and telephone companies offer duopoly intermodal competition in many markets already. Over time, the technologies and capabilities of their network infrastructures have grown closer as each have evolved toward being general-purpose, IP-based broadband platforms. Intermodal competition also arises when Direct Broadcast Satellite and over-the-air TV services compete for video and when mobile services compete with fixed services. One regulatory approach to enabling intermodal competition is to enable and promote more flexible Network Sharing Agreements (NSAs).[55] NSAs may focus on active, passive, or mixed network components or services and may be implemented in many ways.[56]

Whether existing or potential intermodal competition is sufficiently robust to eliminate market failures which may arise when bottleneck resources are not shared adequately (either foreclosed to some users or available only at monopoly prices) is an enduring regulatory debate. Those who argue that existing competition is adequate (inadequate) often use such arguments to justify further deregulation (re-regulation). And, both sides continue to debate what regulatory interventions might best assist in addressing the problem. A key focus and motivation for our pMEC work[57] is to expand or at least better understand options for a new source of intermodal competition.

---

resources or services that have attracted such concerns include control over on-line search, databases of consumer information, and cloud computing resources. Some of these have been postulated as winner-take-all services or markets, which if true, could result in some of these becoming bottleneck resources. Our goal here is not to comment on which components of our communications infrastructure might require regulatory oversight, but only to highlight that these questions remain worthy of continued investigation today and into the future.

[55] See Lehr & Stocker, 2023.

[56] For example, fiber-optic transmission facilities can be decomposed into elements and shared as shared access to conduit (to allow multiple deployments of fiber cables), as dark fiber (to allow leasing of strands), as lit fiber (to allow leasing of light waves), as "bit-pipes" (Layer 2 or Layer 3 wholesale network services), etcetera. Higher-level NFV functionality might also be shared and options for doing so are expanding because of the progression of distributed software control in modern broadband networks.

[57] Intermodal competition is via alternative technologies or business models and may arise in multiple ways. For example, legacy cable television and telephone companies offer duopoly intermodal competition in many markets already. Over time, the technologies and capabilities of the network infrastructures have grown closer as each have evolved toward being general-purpose, IP-based broadband platforms with significant high-capacity wired infrastructure (including significant fiber optic facilities). Intermodal competition also arises when Direct Broadcast Satellite and over-the-air TV services compete for video and mobile services compete with fixed services. Whether such intermodal competition is sufficiently robust to eliminate market failures which may arise when bottleneck resources are not shared adequately (either foreclosed to some users or available only at monopoly prices) is an enduring regulatory debate, as are debates over what regulatory interventions might best assist in addressing the problem. A key focus and motivation for our pMEC work

Furthermore, exploration and further development of the pMEC concept could provide a "living lab" for social science experiments. Such "living labs" can provide opportunities to study social science phenomena in real world naturalistic settings.[58] For example, past research has used phone logs, wearables, and digital activity logs to study trust, cooperation, social capital etc. in real-world settings.[59] Specifically, the pMEC architecture would provide opportunities for studying the key challenges and opportunities for enabling expanded end-user choices in how critical communications and computing local digital infrastructure are provided and managed, including shared.[60] The pMEC "living lab" could enable active on-going research on such topics as resource allocation, incentive design, security and privacy, and social impact.

For instance, an important question the pMEC concept raises is the need to better understand the needs, preferences, and behavior of potential pMEC users and stakeholders. Different user groups and communities may have different expectations and motivations for participating in pMEC, as well as different pain points and challenges in their current Internet and edge computing usage. Therefore, we hypothesize that there is a need to co-create user-centric and context-aware pMEC solutions that meet the diverse needs of different user segments. For example, rural residents, low-income households, students, and small businesses are expected to benefit from pMEC in different ways, such as improving their access to information, education, health care, and economic opportunities.[61] Participatory design methods have been useful in similar applications (e.g., designing COVID apps[62]) and can be used to elicit feedback and provide input from such demographic groups on the pMEC architecture and approach. This can help to ensure that the pMEC system is designed with the users in mind and that it provides value and benefits for them. Exploration of such user-centric engagement in the design and deployment of local digital infrastructure should contribute to promoting end-user engagement, as well as both the feeling and reality of end-user autonomy and choice.

Another important research goal for the living lab will be to better understand cooperation and trust in emerging participatory computing scenarios. Cooperation and trust are essential for the

[58] See Aharony, Pan et al. (2011) and Shmueli, Singh et al. (2014).

[59] See Singh & Agarwal (2016), Bati & Singh (2018), and Singh & Ghosh (2017).

[60] The rise of eBay, AirBnB, Uber, and a host of other examples of platform models for the "sharing economy" suggests significant expanded interests in using digital technologies to enable resource and asset sharing of an expanding array of assets. The pMEC extends that model to edge-computing resources.

[61] Moreover, the needs of they different types of users will vary a lot by the local context where those needs arise. For example, is it in a remote rural area or urban locale? During the Covid pandemic, even when broadband connectivity was widely available, the edge-circumstances of remote users varied significantly in terms of the physical spaces where remote workers or students were connecting to cloud services (from crowded tenement apartments in lower-income urban areas to vacation homes for those lucky enough to have such access) and the devices and other complementary resources needed to have a productive on-line experience. The conditions and quality of local infrastructure and plans for expanding those will also factor into users incentives and receptivity to the pMEC concept. For example, a pMEC may offer an attractive interim or bridge solution to meet needs before anticipated "better" alternatives become available. If the pMEC concept is to fit these myriad circumstances, it will have to be low cost to adopt (which also implies being low cost to switch to other alternatives).

[62] See Park, Ahmed et al. (2022).

success of pMEC, but the insights obtained can also be useful in broader range of scenarios that require trust and cooperation (e.g., open-source software development, multi-player online games, online commerce etc.[63]). Joint study of cooperation and trust are important because they are complex and dynamic phenomena that depend on multiple factors and contexts that vary across local contexts that are difficult to observe. For instance, how do users decide whether to share their resources or not? How do they evaluate the trustworthiness and reliability of other users and the system? How do they cope with potential risks and uncertainties? How do they balance their individual interests and collective welfare? These are some of the questions that can be explored in the living lab using experimental methods. These questions are not only pertinent to the design of the pMEC architecture but could also be important for the theories in the fields of trust and cooperation.[64] By manipulating different variables that could affect user cooperation and trust, such as incentive mechanisms, reputation systems, social norms, and information cues, researchers can measure the outcomes of user cooperation and trust, such as task performance, resource utilization, user satisfaction, and social welfare. The findings can be useful for identifying the best practices and design principles for fostering cooperation and trust in pMEC settings and yield design guidelines for broader social processes and activities in digitally mediated settings.

In addition to the benefits for promoting efficient and equitable access to the next generation of critical digital infrastructure – beyond broadband connectivity – further development of the pMEC concept is expected to provide insights applicable to the broader challenge of how best to regulate and manage the human-digital interface in the emerging Artificial Intelligence-powered future that is rapidly approaching.

5. Conclusions

We now sum up the key themes in this paper and highlighting directions for future research.

The initial key point we make is that edge computing is part of a long running technical and economic evolution of our global computing and digital communications networks. Today, that is what we refer to as the Internet and our fixed and mobile broadband networks. We see value in exploring the ways that edge computing might evolve in the US and other industrialized nations and the consequence of the evolution on the structure of the network and related digital infrastructure policies. Most notably, we see an exploration of edge computing as helping to inform the future of broadband policy. We argue that if our investments in broadband are to be productive beyond the demands of current applications, we need edge computing capabilities.

A further point illustrated by the story above is how technology and markets have interacted and helped give rise to the situation we find ourselves in today. In the early days, the limits of technology-imposed constraints on the viable architectures and services that could be offered. Those constraints limited the interfaces and degree to which control could be decentralized and distributed while still enabling end-to-end services. The early limitations, in effect, dictated thin-client models for distributed networked computing. Over time, as markets grew and technology at

[63] See Singh, Jain & Kankanhalli (2009).

[64] See Horton, Rand, and Zeckhauser (2011).

all levels advanced, more distributed architectures with more capable fat end-clients emerged, expanding the economic opportunities and technical options for more distributed and decentralized networking and computing architectures and industry structures to emerge. Convergence further advanced these trends as legacy silo-service/network providers moved toward general-purpose IP platforms and intermodal competition. Today, technical and market limitations pose less restrictive constraints on how networked computing might evolve, leaving us will many possible models. In this world, a key issue to better understand is how those different models may alter the economic control, efficiency, and equity implications of alternative models. Highlighting that question is a key motivation for our effort to differentiate between service-provider and end-user provided models for providing future critical digital infrastructure.

Moreover, however this infrastructure emerges, there will be a greater need to share critical resources and infrastructure elements and services. Therefore, we also argue that the need to deploy edge computing resources will necessitate acceptance of more flexible and complex passive and active sharing among value chain participants, which will require a reframing of legacy models of industry structure, particularly the roles of end-users. The future of digital infrastructure cannot and should not wait solely on the decisions of large, national-scale service providers whether those be the traditional telecommunications providers of the digital connectivity infrastructure (i.e., the ISPs and last-mile broadband providers of fixed and mobile networks) or the edge providers of digital platforms, content, or application services. Certainly, both the ISPs and the large digital platform service providers can be expected to play a significant and even dominant role in meeting our need for edge computing services. However, we believe that those service providers need not and should not be the only option for enabling edge computing resources. Enabling end-user deployed solutions to emerge has the potential to deliver significant economic and technical benefits, accelerating the realization of edge-computing capabilities for all citizens and communities and offering the potential for valuable competitive discipline and end-user autonomy (choice) over how digital services are delivered.

Although the end-user deployment model has always existed as part of the digital landscape in the form of private networks and computing, the future promises the potential for new models and new importance for how end-user and service-provider infrastructures and services may interact, with important technical, economic and policy implications. We explain our pMEC model as a potential hybrid development path for end-user deployed infrastructure, and review some of the challenges for pMEC and possible evolutionary paths it may follow.

Next, we argue that the technical evolution to include end-user edge computing (i.e., pMEC) is technically and architectural straightforward and could have numerous advantages in terms of timely deployment, control, security, and privacy.

In future work, we plan to build an open-source model of pMEC and test the various technical, economic, and user issues that we highlighted in this paper. We see that a deployable system at the very least will provide a platform for a living laboratory to allow end users to explore and innovate with new ways of exploiting edge computing. We see this living-lab opportunity as offering a valuable research resource to research a significant set of use and adoption questions that have surfaced as we have developed this pMEC concept. Because we do not think any single path to future critical edge computing infrastructure is optimal, and the collection of paths will

need to co-evolve, the pMEC concept offers a laboratory for helping to build consensus on the appropriate roles of end-users in enabling and managing critical digital infrastructure.

6. References

1. Abbas, N., Y. Zhang, A. Taherkordi and T. Skeie (2018), "Mobile Edge Computing: A Survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450-465, Feb. 2018, doi: 10.1109/JIOT.2017.2750180.

2. Aharony, N., Pan, W., Ip, C., Khayal, I., & Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and mobile computing*, *7*(6), 643-659.

3. Aijaz, A. and M. Sooriyabandara (2019), "The Tactile Internet for Industries: A Review," in Proceedings of the IEEE, vol. 107, no. 2, pp. 414-435, Feb. 2019, doi: 10.1109/JPROC.2018.2878265.

4. Armbrust, M., A. Fox, R. Griffith, A. D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, and I. Stoica. (2009) "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley

5. Bati, G. F., & Singh, V. K. (2018, April). "Trust Us" Mobile Phone Use Patterns Can Predict Individual Trust Propensity. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-14). Chicago

6. Bhattarai, S., J. Park, B. Gao, K. Bian, and W. Lehr (2016), "An Overview of Dynamic Spectrum Sharing: Ongoing Initiatives, Challenges, and a Roadmap for Future Research," *IEEE Transactions on Cognitive Communications and Networking*, 2(2), 110-28

7. Clark, D., W. Lehr, and S. Bauer (2016), "Interconnection in the Internet: peering, interoperability, and content delivery," Chapter 16 in J. Bauer and M. Latzer (eds) Handbook on the Economics of the Internet, Edward Elgar: Northampton MA, 2016

8. ETSI (2014), "Mobile-edge computing—Introductory technical white paper," White Paper, Mobile-Edge Computing (MEC) Industry Initiative, European Telecommunications Standards Institute (ETSI), September 2014, available at https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge_computing_-_introductory_technical_white_paper_v1%2018-09-14.pdf.

9. Fettweis, G.P., 2014. The tactile internet: Applications and challenges. *IEEE vehicular technology magazine*, *9*(1), pp.64-70.

10. Gupta, R., Tanwar, S., Tyagi, S., & Kumar, N. (2019). Tactile internet and its applications in 5G era: A comprehensive review. International Journal of Communication Systems, 32(14), e3981. doi:https://doi.org/10.1002/dac.3981

11. Han, B., Gopalakrishnan, V., Ji, L. and Lee, S., 2015. Network function virtualization: Challenges and opportunities for innovations. IEEE communications magazine, 53(2), pp.90-97.

12. Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, *14*, 399-425

13. ITU (2015), "IMT Vision: Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," International Telecommunication Union (ITU), Recommendation ITU-R M.2083-0, available at https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf.

14. Jiang, W., B. Han, M. A. Habibi and H. D. Schotten (2021), "The Road Towards 6G: a Comprehensive Survey," *IEEE Open Journal of the Communications Society*, 2, 334-66

15. Khan, W., E. Ahmed, H. Saqib, I. Yaqoob and A. Ahmed (2019), "Edge Computing: A Survey," *Future Generation Computer Systems*, 97, 219-35.

16. Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J. and Jahanian, F., 2010. Internet traffic and content consolidation. *77th Internet Engineering Task Force*.

17. Lehr, W. and M. Oliver (2014), "Small cells and the mobile broadband ecosystem," Euro ITS2014, Brussels, June 2014, available at http://econpapers.repec.org/paper/zbwitse14/101406.htm

18. Lehr, W. and D. Sicker (2017), "Would you like your Internet with or without Video," *Journal of Law, Technology & Policy*, vol 2017 (issue 1 Spring), available at http://illinoisjltp.com/journal/wp-content/uploads/2017/05/Lehr.pdf.

19. Lehr, W. and D. Sicker (2018), "Communications Act 2021," *Journal of High Technology Law*, Volume 18, Number 2 (2018) 270-330, available at https://cpb-us-e1.wpmucdn.com/sites.suffolk.edu/dist/5/1153/files/2018/05/Communications-Act-2021-1fkmij5.pdf

20. Lehr, W. and D. Sicker (2019), "Telecom déjà vu: a model for sharing in the broadband Internet," *Information & Communications Technology Law*, 28:3, 275-310, DOI: 10.1080/13600834.2019.1653546

21. Lehr, W. (2020), "Economics of Spectrum Sharing, Valuation and Secondary Markets," Chapter 18 in C. Papadias, T. Ratnarajah, and D. Slock (eds), *Spectrum Sharing*, John Wiley & Sons: New York, 2020

22. Lehr, W. (2022), "5G and AI Convergence, and the Challenge of Regulating Smart Contracts," in Europe's Future Connected: Policies and Challenges for 5G and 6G Networks, edited by E. Bohlin and F. Cappelletti, European Liberal Forum (ELF), pages

72-80, available at https://liberalforum.eu/publication/europes-future-connected-policies-and-challenges-for-5g-and-6g-networks/

23. Lehr, W., D. Clark, S. Bauer, A. Berger, and P. Richter (2019a), "Whither the Public Internet?" *Journal of Information Policy* 9 (2019): 1-42. doi:10.5325/jinfopoli.9.2019.0001.

24. Lehr, W., D. Clark, S. Bauer, and K. Claffy (2019b), "Regulation When Platforms Are Layered," TPRC47:Research Conference on Communications, Information and Internet Policy, Washington, DC, September 2019, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3427499.

25. Lehr, W., F. Queder, and J. Haucap (2021), "5G: A new future for Mobile Network Operators, or not?," *Telecommunications Policy*, 45(3), 102086. doi:https://doi.org/10.1016/j.telpol.2020.102086

26. Lehr, W. (2019), "5G and the Future of Broadband," in G. Kneips & V. Stocker (eds.), The Future of the Internet - Innovation, Integration, and Sustainability, Baden-Baden: Nomos, 2019

27. Lehr, W. (2023), "Getting to the Broadband Future Efficiently with BEAD funding," white paper prepared with support from WISPA, January 2023, available at https://www.wispa.org/docs/Lehr_White_Paper_Final.pdf

28. Lehr W. and V. Stocker (2023) "Next-generation networks: Necessity of edge sharing," Frontiers Computer Science, 5:1099582, available at https://doi.org/10.3389/fcomp.2023.1099582

29. Lin, X., S. Rommer, S. Euler, E. A. Yavuz and R. S. Karlsson (2021) "5G from Space: An Overview of 3gpp Non-Terrestrial Networks," IEEE Communications Standards Magazine, 5(4), 147-53.

30. Oughton, E., W. Lehr, K. Katsaros, I. Selinis, D. Bubley, and J. Kusuma (2021), "Revisiting Wireless Internet Connectivity: 5G vs. Wi-Fi 6," *Telecommunications Policy*, 45 (2021) 102127, available at https://authors.elsevier.com/sd/article/S0308-5961(21)00032-X.

31. Park, J., Ahmed, E., Asif, H., Vaidya, J., & Singh, V. (2022, February). Privacy attitudes and COVID symptom tracking apps: Understanding active boundary management by users. In *International Conference on Information* (pp. 332-346). Cham: Springer International Publishing.

32. Ren, J., D. Zhang, S. He, Y. Zhang, and T. Li (2019), "A survey on end-edge-cloud orchestrated network computing paradigms: Transparent computing, mobile edge computing, fog computing, and cloudlet," *ACM Computing Surveys (CSUR)*, 52(6), pp.1-36.

33. Shmueli, E., Singh, V. K., Lepri, B., & Pentland, A. (2014). Sensing, understanding, and shaping social behavior. *IEEE Transactions on Computational Social Systems*, *1*(1), 22-34.

34. Singh, V. K., Jain, R., & Kankanhalli, M. S. (2009, October). Motivating contributors in social media networks. In *Proceedings of the first SIGMM workshop on Social media* (pp. 11-18).

35. Singh, V. K., & Agarwal, R. R. (2016, September). Cooperative phoneotypes: exploring phone-based behavioral markers of cooperation. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 646-657).

36. Singh, V. K., & Ghosh, I. (2017). Inferring individual social capital automatically via phone logs. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1-12.

37. Singh, R., D. Sicker, and W. Lehr (2019), "Beyond 5G: The Role of THz Spectrum," TPRC47: Research Conference on Communications, Information and Internet Policy, Washington DC, September 2019.

38. Sollins, K. and W. Lehr (2020), "Exploring the Intersection of Technology and Policy in the Future Internet Architecture Effort," TPRC48: The 48th Research Conference on Communication, Information and Internet Policy, December 2020, available at https://ssrn.com/abstract=3748638.

39. Stocker, V., G. Smaragdakis, W. Lehr, and S. Bauer (2017), "The Growing Complexity of Content Delivery Networks: Challenges and Implications for the Internet Ecosystem," *Telecommunications Policy*, Vol 41 (10) 1003-1016, https://doi.org/10.1016/j.telpol.2017.02.004

40. Venkataramani, A., J.F. Kurose, D. Raychaudhuri, K. Nagaraja, M. Mao, and S. Banerjee (2014), "MobilityFirst: a mobility-centric and trustworthy internet architecture," *ACM SIGCOMM Computer Communication Review*, 44(3), pp.74-80.

41. Vhora, F. and J. Gandhi (2020), "A comprehensive survey on mobile edge computing: challenges, tools, applications," IEEE 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 49-55).

42. Wang, X., J. Li, Z. Ning, Q. Song, L. Guo, S. Guo and M.S. Obaidat (2023), "Wireless powered mobile edge computing networks: A survey," *ACM Computing Surveys*, forthcoming 2023.

43. Whalley, J., V. Stocker, and W. Lehr (eds) (2023), Beyond the Pandemic? Exploring the Impact of Covid-19 on Telecommunications and the Internet, Emerald Publishing: United Kingdom, May 2023, available at https://www.emerald.com/insight/publication/doi/10.1108/9781802620498