Revisiting the Arguments for Edge Computing Research

Varghese, Blesson; De Lara, Eyal; Ding, Aaron Yi; Hong, Cheol Ho; Bonomi, Flavio; Dustdar, Schahram; Harvey, Paul; Hewkin, Peter; Shi, Weisong; More Authors

DOI
10.1109/MIC.2021.3093924

Publication date
2021

Document Version
Accepted author manuscript

Published in
IEEE Internet Computing

Citation (APA)
Varghese, B., De Lara, E., Ding, A. Y., Hong, C. H., Bonomi, F., Dustdar, S., Harvey, P., Hewkin, P., Shi, W., & More Authors (2021). Revisiting the Arguments for Edge Computing Research. *IEEE Internet Computing*, *25*(5), 36-42. https://doi.org/10.1109/MIC.2021.3093924

Revisiting the Arguments for Edge Computing Research

Blesson Varghese[1], **Eyal de Lara**[2], **Aaron Ding**[3], **Cheol-Ho Hong**[4], **Flavio Bonomi**[5], **Schahram Dustdar**[6], **Paul Harvey**[7], **Peter Hewkin**[8], **Weisong Shi**[9], **Mark Thiele**[8], **and Peter Willis**[10]

[1]Queen's University Belfast, UK; [2]University of Toronto, Canada; [3]TU Delft, Netherlands; [4]Chung-Ang University, South Korea; [5]Nebbiolo Technologies, USA; [6]TU Wien, Austria; [7]Rakuten Mobile, Japan; [8]SmartEdge Datacentres Ltd., UK/USA; [9]Wayne State University, USA; [10]British Telecommunications plc, UK

This article argues that low latency, high bandwidth, device proliferation, sustainable digital infrastructure, and data privacy and sovereignty continue to motivate the need for edge computing research even though the initial concepts of edge research were formulated more than a decade ago.

The initial concepts of what is today referred to as edge computing were formulated more than a decade ago (1). Although a nascent research area, it is generally understood that edge computing enables the (pre)processing of data closer to the source outside a centralized and geographically distant data center (2–4). In previous decades, although not articulated in its current form, there were several notions of such geography-aware computing with a premise to bring compute services closer to where data was generated.

'Edge' although generally refers to a location rather than computing using any specific technology, it has started to emerge that it may be more than just the location. Recent advances made in computer processor, 5G and AI technologies and their application in novel domains have necessitated a strong need for geography-aware computing (and much more). These have brought edge computing to the limelight and has inadvertently coupled the notion of the edge as a location with certain technologies. The differentiating lines between technologies that may be required for realizing edge applications have therefore blurred.

An exemplar of edge computing that is commercially used is Content Delivery Networks (CDNs). They are commonly used to deliver digital content (such as web, gaming, AR/VR, videos) from servers to end-users by Internet Service Providers, carriers and network operators. More than half of today's consumer traffic is generated in delivering digital content to users in the Internet using CDNs. Digital content is replicated and stored across many edge servers in different geographic locations, a concept referred to as 'edge caching', which is commercially used for improving application responsiveness and reducing latencies.

When the cloud was rapidly being adopted within the technology landscape, it was argued that extremely centralized compute resources of the cloud would not be suitable for a wide-range of sensor-rich applications that were to emerge in the future. These applications generate data by end-users that is required to be transferred elsewhere for processing (as opposed to delivering content from servers to end-users). Such applications would be latency-critical, bandwidth-intensive, and privacy-craving. A few hyperscalers and comparatively low network speeds observed then mandated the need for more decentralized data centers to be placed and used at the edge. However, it was always recognized that hyperscalers as economies at scale were essential and could not become redundant infrastructure.

Times have now changed - there are plenty of cloud locations scattered across the globe and data can travel through fibre optic communication channels at (near) speed of light. Do the arguments that initially mandated the need for edge computing still hold?

Recent research articles examined cloud reachability across the globe to measure the average round-trip communication latency for an end-user when communicating with the cloud (5, 6). The authors concluded that current clouds in the United States were sufficient for many latency-critical applications and noted that the motivation for realizing edge computing as a mere *'enthusiasm for newer computing paradigms'* (the data used in the above mentioned research and the conclusions drawn will be briefly examined in the next section).

Contrary to the above, we note that cloud and edge computing are not necessarily competing paradigms; rather they are compatriots in delivering computing as a ubiquitous utility by appealing to arguments that will be discussed in this article. In light of the above and a renewed interest in determining whether there is still a need for edge computing as a concept and an avenue of research, this article (re)examines five different arguments, namely (1) Latency, (2) Bandwidth, (3) Proliferation, (4) Sustainability, and (5) Privacy and Sovereignty.

Latency

Reducing the overall latency in processing data at the source or delivering data from servers to end-users has been a key argument in favor of edge computing. These arguments have been supported by predictions of Gartner, for example, anticipating that by 2025, over 50% of enterprise data will be created and processed outside the typical data center[1].

We note that different technology providers consider latency in diverse ways. Therefore, some clarity is required on what should constitute the latency metric. For example, consider an end-user connected via a wired broadband connection - latency should refer to the sum of the times for raising a request from source (browser on device), for transporting the request over the network (including the delays incurred on routers and on different hops), for processing the request on the receiving server, for sending the response back to the source, and for taking an action on the source. The transport time from the source to the server and back only accounts for the round-trip communication latency. Often computational latencies are

[1] https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-impacting-infrastructure-operations-for-2020/

[1]Corresponding e-mail: b.varghese@qub.ac.uk

ignored. When considering a mobile network, the round trip latency between the source and the access network should also be accounted for.

The Federal Communications Commission of the United States (US) carried out a performance measurement study of broadband services in the US. Ten major Internet Service Providers (ISPs) and an additional nine organisations participated in the exercise[2]. The measurement servers were located in thirteen cities across the US with multiple locations in each city. The median round trip communication latencies observed on fibre optic cables were between 10 ms to 27 ms to edge locations.

Broadband connection latencies enable us to quantify what delays will be incurred within an enclosed environment, such as a home or office. Given that a vast number of users rely on mobile devices and that machine-to-machine, vehicle-to-vehicle, machine-to-everything and vehicle-to-everything will need to rely on telecommunication infrastructure, it is worthwhile considering mobile network latencies. 4G, which is the most available global mobile network model has observed communication latencies of over 30 ms. In 2019, Opensignal reported that only 13 countries had a communication latency of between 30-40 ms which excluded North America and many parts of Europe[3]. These reported latencies are access network latencies and do not include the latency for reaching an edge compute location via the mobile network or for actually performing computations. With 5G, although a theoretical 1 ms communication latency is envisioned, early deployments in the US in 2019 had demonstrated nearly a 30 ms communication latency for the access network. In the UK, the 5G deployments in 2020 had a communication latency of at least 20 ms for the access network.

The above communication latencies can indeed support many interactive internet applications that are already in use today. However, they will not be adequate to support (near) real-time computing (sub millisecond), such as those required for rapid responsiveness of autonomous cars or robots that share spaces with humans. For these contexts, the overall latency will need to be guaranteed. Therefore, it would not be sufficient for any latency measuring exercise to merely highlight the average of a distribution of observed communication latencies without considering computation latencies and the type of application. In addition, the the tail-end and outlier latencies in a distribution may be substantially higher than the average latency which also need to be accounted for.

At this point, the dataset employed by the research articles investigating cloud reachability is considered (5, 6). The dataset employed is from RIPE Atlas, an Internet measurement network that provides hardware probes for network measurements (for example, ping) that is publicly available[4]. We note that these measurements reflect only network communication latencies and do not include computation latencies associated with the execution of application code.

We analyzed the dataset and focused on the data for the United States containing 3091 different probe locations. For each location, there are measurements for up to 102 different data centers. Only the closest data center for each probe as determined by the lowest average latency was considered. Since 80% of the locations have less than 64 measurements per data center, we focused on the remaining 650 locations that have at least 100 measurements to their closest data center; the average no. of measurements per probe is 2611.

Figure 1 shows the results of our analysis. Figure 1a shows a box plot of the latency distributions sorted by increasing average latency. For clarity, the plot only includes 1 out of every 7 probes (the plot of the complete dataset shows a similar pattern but is is very hard to read due to clutter). The top and bottom of the box represent 25% and 75% latencies, and the whiskers show the minimum and 99% latency. Measurements outside of this range are shown as individual outlayers. Figure 1b shows the cumulative distribution for the proportion of probes which experience median, 95%, 99%, and 99.9% latency below a given threshold. For example, the figure shows that 25.4% of probe locations experience a median latency to their closest data center of 10 milliseconds or less.

We observed that the majority of locations had a round-trip communication latency of more than 10 milliseconds. Moreover, even probe locations that experience low median latency observe very substantial variations. For example, only 6.7% of the 650 locations were able to reach their closest data center within 10 milliseconds 99.9% of the time. This rose to 18% of the locations when lowering to 95% of the time.

The current communication latencies observed to the nearest cloud locations are undoubtedly an improvement over the average of 80 ms that was observed to cloud locations when edge computing was initially formulated as a concept (7). Overall latencies under 10 milliseconds (let alone sub-milliseconds) cannot be guaranteed today on current public clouds for applications that require performance guarantees. Latency measurement studies are required to better understand edge computing. However, focusing on average latency (5, 6) does not paint a correct or complete picture as it inherently hides significant variations in network latency over time.

The above have led to new industry trends that will potentially lead to the convergence of what is today known as the cloud and edge. For example, cloud providers are embracing edge locations for setting up data centers on the last mile network (for example, Amazon Outpost) together with dedicated hardware, such as the AZ1 neural edge processors for the extreme edge to reduce communication latencies.

However, edge as a location is only one aspect of the latency argument. If only communication latencies had to be considered, then edge compute locations would need to be placed every 60 miles for theoretically achieving a 1 ms round trip communication latency using current fibre optic technologies (based on the speed of light travelling through a medium with a refractive index of 1.5) between two endpoints ignoring latencies in the access network, processing delays on the hops, network congestion or computation latencies on servers. Telecom providers are experimenting with hollow core fibre optics that can transmit data at (near) speed of light to reduce latencies[5]. The invasiveness and substantial increase in costs of infrastructure may not be pragmatic for a global rollout and that by reducing communication latencies alone will not be sufficient for minimizing the overall latency.

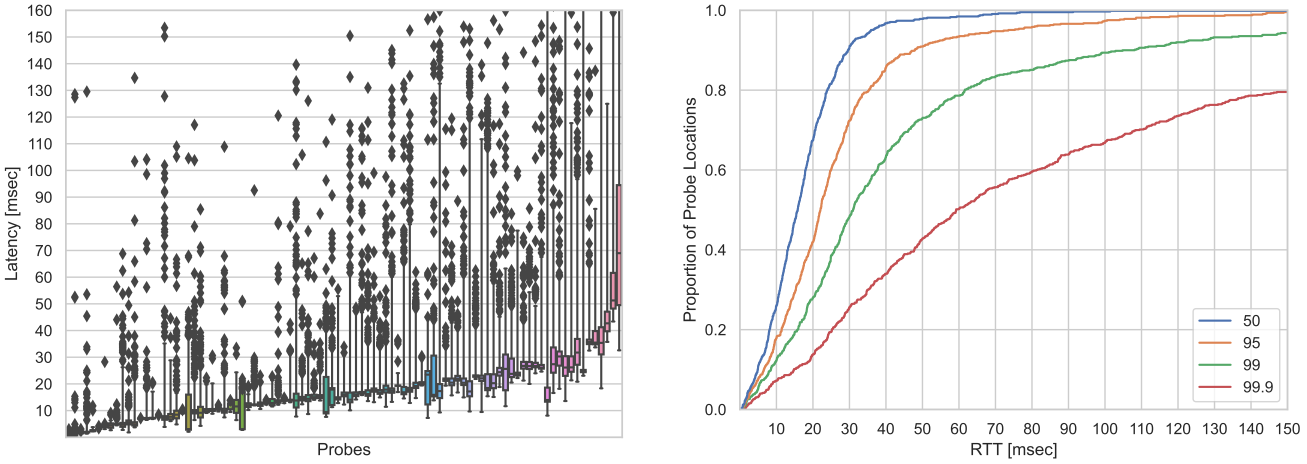
There are a select few locations around the globe by virtue of geographic location or proximity to traditional data cen-

[2] https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-fixed-broadband-tenth-report

[3] https://www.opensignal.com/sites/opensignal-com/files/data/reports/global/data-2019-05/the_state_of_mobile_experience_may_2019_0.pdf

[4] https://doi.org/10.1145/3442381.3449854

[5] https://newsroom.bt.com/bt-kicks-off-trials-of-revolutionary-new-optical-fibre/



(a) Boxplot of probe location latencies shorter in order of increasing average latency. Whiskers show minimum and 99% latency.

(b) CDF of latencies for probe locations.

Fig. 1. Latency to closest cloud datacenter for probe locations in the USA with at least 100 measurements.

ters that can achieve an average communication latency of 10 ms. Nonetheless, delivering low overall latency globally for emerging and futuristic applications still remains a challenge to be surmounted and a vision to be fulfilled. Transformative advancements are still required both on the networking and computing fronts to achieve this. Thus, latency continues to be a first-class argument for edge computing research.

Bandwidth

Although an abundance of processing servers are available on the cloud, the network bandwidth bottleneck of the wide-area network (WAN) to the cloud has been an argument in favor of edge computing (3). It was demonstrated that the network bandwidth on the WAN is restrictive due to the number of traversed hops ranging from 9 to 20 (8). When measured a few years ago, the TCP bandwidth between two *m3.xlarge* instances of Amazon EC2 in the same data center was 900 Mbps (9). However, when the WAN is involved, the bandwidth to the same instance was 30 – 160 Mbps (8), which is far from the peak performance. Furthermore, most cloud providers throttle the bandwidth when the total data transfer reaches a certain threshold. Therefore, a geographically distant cloud is not adequate for emerging applications that require high network throughput.

Emerging applications including AR/VR, remote-controlled factories, and autonomous vehicles employ a wide range of devices and sensors at the edge of the network and increasingly generate and consume a large volume of data. Therefore, a high network bandwidth is required for meeting Quality-of-Service (QoS) objectives. Consider the example of autonomous vehicles. The Automotive Edge Computing Consortium (AECC) estimates that more than 30% of video data produced on the vehicle will need to be offloaded. This is to increase safety thresholds by processing offloaded data with external data for augmenting awareness of the moving vehicle. The volume of data that will need to be offloaded is expected to be between 400 GB to 5 TB per hour. If all the data is sent to the cloud, the response time would be significantly increased owing to the limited bandwidth. Therefore, exploiting the edge that efficiently processes the data near the source is required for such emerging applications.

Many devices and sensors will be connected to the edge using the mobile network. The latest commercial 5G cellular network implements the millimeter wave (mmWave) technology, which theoretically offers bandwidth up to 20 Gbps for download and 10 Gbps for upload utilizing higher frequencies between 24 GHz and 53 GHz (10). A recent measurement study (11) performed field tests of Verizon, Sprint, and T-Mobile on 5G mmWave performance in three major U.S. cities. While the 5G peak download speeds that were observed range from 600 Mbps to 1.7 Gbps, the upload speeds were limited to between 30 and 60 Mbps. Another measurement study of commercial 5G in China (12) reported that the peak download bandwidth was 1.2 Gbps and the peak upload speed was 218 Mbps. Because 5G has just begun commercialization, its performance is still far from reaching the theoretical speed but offers notably higher bandwidth than 4G LTE. 5G is expected to achieve the target speed of 20 Gbps with the development of advanced terminal chips.

The current peak download speed of 5G mmWave is acceptable for many existing applications including video streaming and gaming. For example, high resolution cameras in a stadium can transmit a video stream directly to an edge server without sending the data to the cloud. The edge server then routes the stream to mobile devices in the same venue in order to avoid a latency delay. As the bandwidth required for 8K video streaming is 300 Mbps[6], the current bandwidth of 5G can sufficiently support this application scenario. An emerging real-time streaming system such as volumetric videos, which capture three-dimensional space, demands throughput at least 1.1 Gbps (13). The peak speed of current 5G can satisfy such a requirement, and the advances in 5G will be able to support more high quality volumetric videos in future.

The current upload speed of 5G can meet the bandwidth requirements of non-bandwidth-hungry applications in edge computing. For example, 4K panoramic video telephony does

[6] https://www.huawei.com/~/media/CORPORATE/PDF/whitepaper/Big-Data-Video-Top-Ten-Most-Demanding-Videos-en

not exceed the 5G upload capacity when sending all HD resolution videos up to 5.7K whereas 4G cannot support 5.7K (14). The uploaded video data can be processed at the edge in order to reduce the data volume, which will be transferred to users in different locations. This efficient data processing can provide low latency communication without exploiting the cloud. However, recent deep learning applications employed in IoT devices face challenges when exploiting edge computing. A massive amount of information exchanged between a user device and the edge node presents a network bottleneck in edge training or inference. More communication-efficient training and inference systems are being explored (15).

The current upload and download bandwidths available in 5G and to public clouds can satisfy the requirements of many existing applications. However, bandwidth is still a limiting factor that hinders the emergence of certain safety-critical applications. Therefore, bandwidth continues to be an argument that motivates edge computing research.

Proliferation

It is estimated that by 2025 more than 55 billion devices, sensors and instruments will be connected[7]. This anticipated increase will consequently expose a larger attack surface. One key challenge is cybersecurity - detecting malicious users and containing breaches.

Detecting malicious activity is usually a data-driven approach and using extremely centralized resources to monitor are known to be challenging. Preceding versions of distributed computing paradigms have taught us that centralized monitoring is generally not scalable (16, 17). Therefore, more distributed and hierarchical monitoring strategies are required which can find home on the edge (18). In addition, intrusion detection and prevention systems, such as those used in vehicular ad-hoc networks are latency sensitive and the edge of the network is considered to be an ideal location (19).

The edge appeals to providing more distributed locations for monitoring and data aggregation thereby inherently providing containment zones. Recent years have seen an increasing number of botnet and malware based attacks originating from IoT devices. Edge computing offers the opportunity for localized detection and isolation of such devices (20, 21). Network segmentation for example is one approach that can be adopted at the edge to contain the access of a potentially malicious device beyond the edge.

Many existing edge applications only achieve a functionality improvement by using the edge - they may meet satisfactory performance thresholds even if what is known today as the cloud is available to them. However, looking forward, as edge-native workloads start to emerge, running services on the edge will eventually become necessities for people, factories, cities, and transportation that use them. Thus, even if networks beyond the edge were to fail, the edge can independently operate without any central control, thereby making our people and infrastructure more resilient.

In relation to the device proliferation argument, edge computing is likely to pave way for achieving scalable decentralized management of security, enabling effective containment zones to isolate malicious activities originating from devices, and delivering network independence for more resilience.

Sustainability

Sustainability may be understood in terms of electricity consumption, the amount of electricity to transmit data and the consequent carbon footprint.

The arguments on sustainability in complete favor of edge computing are not sufficiently well articulated and sometimes also send a mixed message. For example, on one hand Nature[8] reported that it is anticipated by 2030 that nearly 21% (other estimates say at least 8%) of the worlds electricity consumption will be driven by increase in networks, requiring nearly 5,000 terawatt hours (TWh) per year and increase in data centers, requiring nearly 3,000 terawatt hours (TWh) per year (22). The estimates presented assumed a year on year increase in electricity requirements for data centers and networks. The exponential growth reported was attributed to expanding telecoms infrastructure and exponentially increasing internet traffic to and from data centers generated by end user devices/sensors and emerging applications.

On the other hand, the IEA reported that the global data centre energy demand has remained largely flat for the last ten years and data transmission networks have become more energy efficient[9]. Data centers and networks indeed consume a large amount of electricity, but whether edge computing with existing technology can substantially shift this trend is not abundantly clear.

Similarly, there have been numerous attempts to estimate the kilowatt hour per gigabyte of data (KWh/GB) transferred over the internet, but has resulted in values ranging across different orders of magnitude (23). All of the above suggests room for more large-scale measurement studies on further articulating the sustainability arguments.

Nonetheless, it is commonly understood that there are costs involved in sending data over the networks. The energy required for transmitting data over the networks is at the least directly proportional to the distance that data needs to travel. With increasing data traffic it is only logical to consider localized data processing to reduce the overall amount of energy required by the networks. It was recently noted that data flowing through the internet is a primary driver for CO_2 emissions; other sources include from the Radio Access Network (RAN) and servers (24). By computing on the edge in a 5G network it was noted that the CO_2 footprint could be reduced by up to 50%.

Sustainability is therefore an important argument to support edge computing research from an electricity consumption and carbon footprint point-of-view, which are both major global concerns. However, further insight from large-scale measurements are required to make a more informed case.

Privacy and Sovereignty

Undoubtedly, data has become the fuel for the digital economy. Social welfare and advancement now relies on protecting critical data. Creating a trusted environment for all stakeholders (for example, public sector organizations, private organizations, governments and individual citizens) is underpinned by data privacy and sovereignty.

Although the cloud is a demonstrable business success in the technology and economic landscape, more recently,

[7] https://www.idc.com/getdoc.jsp?containerId=prAP46737220

[8] https://www.nature.com/articles/d41586-018-06610-y

[9] https://www.iea.org/reports/data-centres-and-data-transmission-networks

the capability growth in the cloud is pressed by the large-scale IoT deployments and data-driven services, such as smart cars[10]. One visible trend is the shift of connected devices from mere data consumers towards data producers. For example, YouTube users generate nearly 100 hours of video content and Instagram users post over 2 million photos every minute (25). This shift raises privacy concerns, specifically pertaining to large-scale machine learning in the cloud over data that is crowd-sourced from individual users that may contain private information (26). As highlighted earlier, large-scale IoT also increases the attack surface and thereby aggravates privacy threats (27).

The edge is generally understood to meet this evident privacy gap by providing the unique capability of enforcing localized privacy control and establishing a trust proxy or firewall (28). The opportunities for edge computing to complement the cloud for addressing the data privacy challenge has been highlighted (25).

As a compute resource offering layer between the data source and distant clouds, By leveraging the resource-rich layer offered between relatively weak devices that generate data and distant clouds, the edge has been demonstrated in the context of distributed machine learning (such as federated learning) to achieve differential privacy for devices while meeting the regulatory and legislative requirements of data sovereignty such as the General Data Protection Regulation (GDPR) (29). This development aligns with the demand for indigenous data sovereignty in Canada, New Zealand, Australia and USA (30). In relation to personal user data, a more secure and trusted way of using them on the user edge has been demonstrated through the 'Data Box' approach (31).

In relation to practical operation, the edge can better utilize local contexts to strike a balance between privacy and usability. Recent studies reveal the synergistic potential of edge, advanced machine learning and privacy-enhancing mechanisms (32–36). Lightweight virtualization (37) has also made it feasible for the edge to quickly adopt novel mechanisms for data privacy and sovereignty (38).

The edge as an enabler for data privacy and sovereignty is an argument that will be further developed as we aim to transform the Internet into a more ethical system. Early research on privacy and sovereignty enhanced by the edge is encouraging. Therefore more collaborative efforts with researchers from law, ethics and public policy, which are from outside the immediate technical envelope of edge computing are required to advance this front.

Conclusions

There are several arguments both technical and non-technical that continue to motivate edge computing research and innovation. The democratization of the future internet is yet another argument in favor of the edge (39). The edge introduces new stakeholders (for example, providers, applications and users), enables the convergence of different technologies that have traditionally operated in silos and takes monopoly away from a select few global players and countries. As a part of this endeavor, the European initiative on the federated data infrastructure for Europe (GAIA-X[11]) and the concept of the Global Data Plane (40) recognize the edge as an essential building block for delivering open, transparent and trustworthy digital infrastructure.

This article argues that the motivation for edge computing research has not diminished since it was first formulated. Ongoing edge research and the wide range of edge-native and edge-accelerated applications that are emerging are indications of the benefits of using the edge. Edge computing as an enabler for advancing new frontiers in space-based systems by reducing communication times and energy is one example among many[12] [13]. While the case for edge computing in private networks and applications is clear, the value in a public rollout will need to be more precisely calculated.

Acknowledgments

The insight and feedback provided by numerous academics and industry experts during consultation when producing this article is acknowledged. The first author is supported by a Royal Society Short Industry Fellowship.

References

1. M Satyanarayanan, P Bahl, R Caceres, N Davies, The Case for VM-Based Cloudlets in Mobile Computing. *IEEE Pervasive Comput.* **8**, 14–23 (2009).
2. B Varghese, N Wang, S Barbhuiya, P Kilpatrick, DS Nikolopoulos, Challenges and Opportunities in Edge Computing in *IEEE International Conference on Smart Cloud.* pp. 20–26 (2016).
3. W Shi, J Cao, Q Zhang, Y Li, L Xu, Edge Computing: Vision and Challenges. *IEEE Internet Things J.* **3**, 637–646 (2016).
4. M Satyanarayanan, The Emergence of Edge Computing. *Computer* **50**, 30–39 (2017).
5. N Mohan, et al., Pruning Edge Research with Latency Shears in *Proceedings of the 19th ACM Workshop on Hot Topics in Networks.* p. 182–189 (2020).
6. L Corneo, et al., Surrounded by the Clouds: a Comprehensive Cloud Reachability Study in *Proceedings of the ACM Web Conference.* (2021).
7. A Li, X Yang, S Kandula, M Zhang, CloudCmp: Comparing Public Cloud Providers in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement.* p. 1–14 (2010).
8. V Persico, A Botta, P Marchetta, A Montieri, A Pescape, On the performance of the wide-area networks interconnecting public-cloud datacenters around the globe. *Comput. Networks* **112**, 67–83 (2017).
9. V Persico, P Marchetta, A Botta, A Pescapé, Measuring network throughput in the cloud: The case of amazon ec2. *Comput. Networks* **93**, 408–422 (2015).
10. A Narayanan, et al., Lumos5g: Mapping and predicting commercial mmwave 5g throughput in *Proceedings of the ACM Internet Measurement Conference.* pp. 176–193 (2020).
11. A Narayanan, et al., A first look at commercial 5g performance on smartphones in *Proceedings of The Web Conference 2020.* pp. 894–905 (2020).
12. T Liu, J Pan, Y Tian, Detect the bottleneck of commercial 5g in china in *2020 IEEE 6th International Conference on Computer and Communications (ICCC).* (IEEE), pp. 941–945 (2020).
13. K Lee, J Yi, Y Lee, S Choi, YM Kim, Groot: a real-time streaming system of high-fidelity volumetric videos in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking.* pp. 1–14 (2020).
14. D Xu, et al., Understanding operational 5g: A first measurement study on its coverage, performance and energy consumption in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication.* pp. 479–494 (2020).
15. Y Shi, K Yang, T Jiang, J Zhang, KB Letaief, Communication-efficient edge ai: Algorithms and systems. *IEEE Commun. Surv. & Tutorials* **22**, 2167–2191 (2020).
16. ML Massie, BN Chun, DE Culler, The Ganglia Distributed Monitoring System: Design, Implementation, and Experience. *Parallel Comput.* **30**, 817–840 (2004).
17. JS Ward, A Barker, . *J. Cloud Comput.* **3** (2014).
18. R Pueyo Centelles, M Selimi, F Freitag, L Navarro, DIMON: Distributed Monitoring System for Decentralized Edge Clouds in Guifi.net in *IEEE International Conference on Service-Oriented Computing and Applications.* pp. 1–8 (2019).
19. M Xiong, et al., Reinforcement Learning Empowered IDPS for Vehicular Networks in Edge Computing. *IEEE Netw.* **34**, 57–63 (2020).
20. J Ni, X Lin, XS Shen, Toward Edge-Assisted Internet of Things: From Security and Efficiency Perspectives. *IEEE Netw.* **33**, 50–57 (2019).
21. A Alwarafy, KA Al-Thelaya, M Abdallah, J Schneider, M Hamdi, A Survey on Security and Privacy Issues in Edge-Computing-Assisted Internet of Things. *IEEE Internet Things J.* **8**, 4004–4022 (2021).
22. ASG Andrae, T Edler, On global electricity usage of communication technology: Trends to 2030. *Challenges* **6**, 117–157 (2015).
23. J Aslan, K Mayers, JG Koomey, C France, Electricity Intensity of Internet Data Transmission: Untangling the Estimates. *J. Ind. Ecol.* **22**, 785–798 (2018).

[10] https://www.tuxera.com/blog/autonomous-cars-300-tb-of-data-per-year/

[11] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

[12] https://www.ibm.com/blogs/industries/ibm-space-tech-cloud-edge-communication-breakthrough/

[13] https://www.hpe.com/us/en/insights/articles/one-giant-leap-for-edge-computing-2102.html

24. B Ramprasad, A da Silva Veith, M Gabel, E de Lara, Sustainable Computing on the Edge: A System Dynamics Perspective in *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications.* p. 64–70 (2021).
25. J Zhang, B Chen, Y Zhao, X Cheng, F Hu, Data security and privacy-preserving in edge computing paradigm: Survey and open issues. *IEEE Access* **6**, 18209–18237 (2018).
26. L Yu, L Liu, C Pu, ME Gursoy, S Truex, Differentially private model publishing for deep learning in *2019 IEEE Symposium on Security and Privacy (SP).* pp. 332–349 (2019).
27. I Hafeez, M Antikainen, AY Ding, S Tarkoma, Iot-keeper: Detecting malicious iot network activity using online traffic analysis at the edge. *IEEE Transactions on Netw. Serv. Manag.* **17**, 45–59 (2020).
28. W Toussaint, AY Ding, Machine learning systems in the iot: Trustworthiness trade-offs for edge intelligence in *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI).* pp. 177–184 (2020).
29. EU, General data protection regulation (gdpr) (2016).
30. SC Rainie, JL Schultz, E Briggs, P Riggs, NL Palmanteer-Holder, Data as a strategic resource: Self-determination, governance, and the data challenge for indigenous nations in the united states. *The Int. Indig. Policy J.* **8** (2017).
31. R Mortier, et al., Personal Data Management with the Databox: What's Inside the Box? in *Proceedings of the ACM Workshop on Cloud-Assisted Networking.* p. 49–54 (2016).
32. S Truex, L Liu, ME Gursoy, W Wei, L Yu, Effects of differential privacy and data skewness on membership inference vulnerability in *Proceedings of the 1st IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications.* pp. 82–91 (2019).
33. S Truex, L Liu, KH Chow, ME Gursoy, W Wei, Ldp-fed: Federated learning with local differential privacy in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, EdgeSys '20. (Association for Computing Machinery, New York, NY, USA), p. 61–66 (2020).
34. ME Gursoy, A Tamersoy, S Truex, W Wei, L Liu, Secure and utility-aware data collection with condensed local differential privacy. *IEEE Transactions on Dependable Secur. Comput.* (2019).
35. L Lockhart, P Harvey, P Imai, P Willis, B Varghese, Scission: Performance-driven and context-aware cloud-edge distribution of deep neural networks in *Proceedings of the 13th IEEE/ACM 13th International Conference on Utility and Cloud Computing.* pp. 257–268 (2020).
36. H Ahn, M Lee, CH Hong, B Varghese, Scissionlite: Accelerating distributed deep neural networks using transfer layer (2021).
37. R Morabito, V Cozzolino, AY Ding, N Beijar, J Ott, Consolidate iot edge computing with lightweight virtualization. *IEEE Netw.* **32**, 102–111 (2018).
38. M Du, K Wang, Y Chen, X Wang, Y Sun, Big data privacy preserving in multi-access edge computing for heterogeneous internet of things. *IEEE Commun. Mag.* **56**, 62–67 (2018).
39. L Peterson, et al., Democratizing the network edge. *SIGCOMM Comput. Commun. Rev.* **49**, 31–36 (2019).
40. N Mor, R Pratt, E Allman, K Lutz, J Kubiatowicz, Global data plane: A federated vision for secure data in edge computing in *39th IEEE International Conference on Distributed Computing Systems*. pp. 1652–1663 (2019).