# Pruning Edge Research with Latency Shears

Nitinder Mohan
Technical University of Munich
mohan@in.tum.de

Lorenzo Corneo
Uppsala Universitet
lorenzo.corneo@it.uu.se

Aleksandr Zavodovski
University of Helsinki
aleksandr.zavodovski@helsinki.fi

Suzan Bayhan
University of Twente
s.bayhan@utwente.nl

Walter Wong
University of Helsinki
walter.wong@helsinki.fi

Jussi Kangasharju
University of Helsinki
jussi.kangasharju@helsinki.fi

## ABSTRACT

Edge computing has gained attention from both academia and industry by pursuing two significant challenges: 1) moving latency critical services closer to the users, 2) saving network bandwidth by aggregating large flows before sending them to the cloud. While the rationale appeared sound at its inception almost a decade ago, several current trends are impacting it. Clouds have spread geographically reducing end-user latency, mobile phones' computing capabilities are improving, and network bandwidth at the core keeps increasing. In this paper, we scrutinize edge computing, examining its outlook and future in the context of these trends. We perform extensive client-to-cloud measurements using RIPE Atlas, and show that latency reduction as motivation for edge is not as persuasive as once believed; for most applications the cloud is already "close enough" for majority of the world's population. This implies that edge computing may only be applicable for certain application niches, as opposed to a general-purpose solution.

## CCS CONCEPTS

• **Networks** → **Network measurement**; **Cloud computing**.

## KEYWORDS

Edge computing; Cloud computing; Internet measurements; Cloud reachability

**ACM Reference Format:**
Nitinder Mohan, Lorenzo Corneo, Aleksandr Zavodovski, Suzan Bayhan, Walter Wong, and Jussi Kangasharju. 2020. Pruning Edge Research with Latency Shears. In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks (HotNets '20), November 4–6, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3422604.3425943

## 1 INTRODUCTION

Edge computing has emerged as a new, compellingly sounding solution for improving and enabling many network applications. One selling point of edge is improving latency by moving services closer to end-users and pre-processing at the "edge" to save the network (and cloud) from being overwhelmed by unforeseen amounts of data [17]. This enables sophisticated applications, e.g., augmented and virtual reality [23, 42], robotic control [14, 21, 45], smart homes/cities [4], etc. Many concrete scenarios have been developed [11, 28] and industrial standardization initiatives, e.g., multi-access edge computing are actively promoted by telcos [24, 67].

However, since edge's inception around a decade ago, several current trends have emerged which *may* impact its utility. *First*, cloud infrastructure is spreading geographically by installing new datacenters. For example, since 2010, Amazon has expanded its cloud network from 3 to 16 countries. Furthermore, cloud providers are establishing (and incorporating) specialized facilities to tackle edge needs, e.g., CloudFront [2]. *Second*, modern smartphones come equipped with considerable processing power, including specialized chipsets, enabling them to process complex tasks, such as AR. On the other hand, last-mile access being the latency bottleneck over (cellular) wireless remains true [12, 31]. While new technologies, such as 5G, show promise, initial tests report their performance to be deficient in practice [49, 71]. *Third*, offloading [19, 58, 59], and cyber foraging [8, 30], have become a reality with services like Siri or Cortana. Products with tight latency constraints, e.g., cloud-based gaming, are already on the market [3, 29, 44], implying improved cloud access latencies. Recent study from Facebook reveals that most users can reach their services in the cloud within 40 ms [60].

We believe these trends necessiate *pruning* popular assumptions driving edge computing research and *identifying* more promising future directions. We achieve this by examining the latency for connecting to cloud globally. Our contributions are as follows.

**(1)** We conduct large-scale measurements over RIPE Atlas analyzing user reachability to datacenters owned and operated by seven cloud providers. Our measurements targeted 101 datacenters in 21 countries and lasted several months. Only [36] has conducted a similar study to ours but it is limited to single cloud provider; the most recent multi-cloud measurement is a decade old [40].

**(2)** We take a critical look at edge computing and its future potential, by analyzing latency and bandwidth thresholds of several applications reputedly enabled by edge computing. Extrapolating our measurement results, we find that, contrary to popular belief, the effectiveness of edge computing is limited to a few applications, such as traffic monitoring, gaming, etc., which, incidentally, are not the primary drivers of edge hype. Other applications can be either supported by current cloud infrastructure (smart home, wearables) or *will* require onboard processing for optimal operation.

Edge computing is still in flux. Some [27] see edge taking over from the cloud; others see a combination [55]. Peterson et al. [52] see edge and the democratization it offers as a cure for Internet ossification. Some argue for wide-spread in-network computation [57],



*HotNets '20, November 4–6, 2020, Virtual Event, USA*

ACM ISBN 978-1-4503-8145-1/20/11.
https://doi.org/10.1145/3422604.3425943

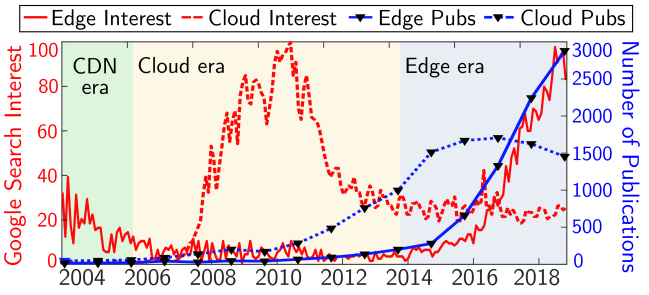



**Figure 1: The popularity (in red) and publications (in blue) of keywords "edge computing" (in solid line) and "cloud computing" (in dashed line) in Google web searches and Google scholar respectively.**

blurring the borders of cloud and edge [32, 75]. Our focus, in this paper, is on a *general-purpose edge deployed by telcos/ISPs* for a wide range of applications [47]. While we show this path to have substantial hurdles, there may be better-suited scenarios for edge. We return to these at the end of the paper.

## 2 A RETROSPECTIVE ON EDGE

Figure 1 captures the zeitgeist of "edge computing" over the past fifteen years. It compares the frequency of Google web searches[1] and scientific publications[2] for "edge computing" and "cloud computing" from 2004 to 2019. Resultingly, three eras can be distinguished: **content delivery networks (CDN)**, **cloud**, and **edge**.

The term *edge* emerged when ***CDNs*** started to deploy edge servers near their clients [20]. They acted as caches of content, speeding up content delivery and reducing bandwidth usage. At the same time, centralized, large-scale datacenter deployments emerged, heralding the ***Cloud era***. Cloud was a success as the type and volume of application's resources could be elastically adjusted to meet the current demand on-the-fly. Application developers could also take advantage of a flexible *"pay-as-you-go"* model for resource utilization.

*Cloudlets* [59] in 2009 started the ***Edge era*** and similar concepts, such as *fog computing* [9]. Back then, the cloud was limited to a few datacenters and unable to address the stringent latency and data transport requirements of new use cases, such as the Internet-of-Things (IoT). Therefore, the research community, including industry, jumped at the opportunity to decouple network latency from the computation time, and devised "edge computing". Many edge architectures have been proposed [46, 48], including exploiting last-mile access points [13], crowdsourcing [26], and using IoT sensors [53]. We will next take a systematic look at various applications driving the hype on edge computing and analyze their requirements.

## 3 DRIVERS OF THE EDGE HYPE

We capture applications used to motivate edge computing in Figure 2. The y-axis is the required latency scale, ranging from a few milliseconds (ms) to an hour (hr). The x-axis shows the amount of

[1]Results obtained from https://trends.google.com/.
[2]Data was collected by a custom web crawler for Google Scholar, based on an open source implementation [38].

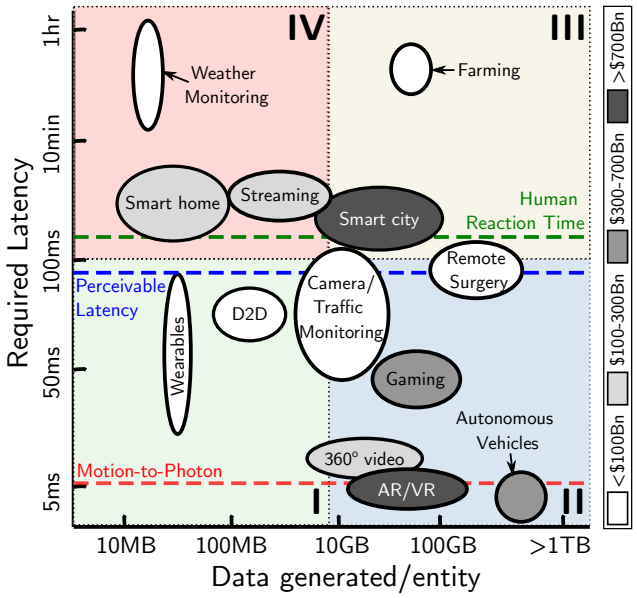
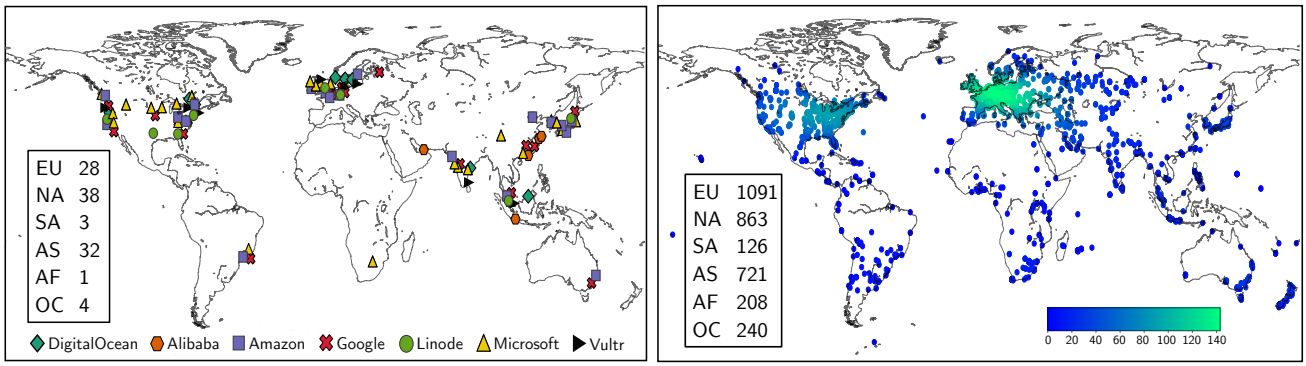



**Figure 2: Driving edge applications represented as ellipses. The orientation and width signify strictness in bandwidth and latency requirements. Color denotes the expected market share by the year 2025.**

data an entity of each application generates, e.g., a camera, which naturally correlates with the network bandwidth requirements. We estimate application requirements by relying on theoretical analysis and preliminary implementations from previously published results [7, 37, 42, 54, 64]. Each application is represented as an ellipse to overcompensate for any estimation errors. The form and orientation of the ellipse represent the application's *strictness* towards latency/bandwidth constraints. The ellipse's color denotes application's expected market share by 2025 in US dollars (data is from [63]). Majority applications in Figure 2 are human-centric – taking inputs and providing feedback to users, e.g. gaming. The QoE of such applications is governed by strict latency thresholds as human senses require, which we also draw in the figure.

*(1) Motion-to-Photon (MTP)* is the delay between user input and its effect to be reflected on a display screen. MTP is guided by the human vestibular system, which requires sensory inputs and interactions to be in complete sync; failure of which results in motion sickness and dizziness. Maintaining latency below MTP, i.e., $\lesssim$ 20 ms, is critical for immersive applications, such as AR/VR, 360° streaming, etc. [43]. Of this, $\approx$ 13 ms is taken up by the display technology due to refresh rate, pixel switching, etc. which leaves a budget of $\approx$ 7 ms for computing and rendering (including RTT to server) [16]. Studies by NASA concludes that certain HUD systems may require the compute part of MTP to be as low as 2.5 ms [7].

*(2) Perceivable Latency (PL)* is the threshold when the delay between user input and visual feedback becomes large enough to be detected by the human eye [54]. PL threshold plays a vital role in the QoE of applications where the user interaction with the system is fully




**(a) Distribution of cloud regions from seven major cloud providers.** **(b) Distribution of 3200+ RIPE Atlas probes.**

**Figure 3: Cloud regions with compute DCs in (a) represent our targets, and probes in (b) are the vantage points for our study.**

or semi passive, e.g., video streaming (stuttering), gaming (input lags), etc. It is roughly estimated to be 100 ms.

*(3) Human Reaction Time (HRT)* is the delay between the presentation of a stimulus and the associated motor response by a human. While HRT is highly dependent on the individual (and can be improved by training), its value is reported to be ≈ 250 ms [73]. Applications that require active human engagement, e.g. remote surgery, teleoperated vehicles etc., must operate within HRT.

Considering the similarities in operational thresholds, we group the emerging applications by quadrants.

**Quadrant I** - *Low Latency & Low Bandwidth*: The bottom-left (green) quadrant represents applications that produce only a small volume of data but impose strict latency constraints for optimal operation. Typical examples include wearables, health monitoring, and other individual-focused applications. The core aim of applications in *Q1* is to perform "naturally", i.e., to operate within the PL threshold. Hence, they can benefit from the low latency computation promised by the edge.

**Quadrant II** - *Low Latency & High Bandwidth*: The blue quadrant at bottom-right encompasses applications that generate large data volumes *and* impose strict latency constraints, e.g. autonomous vehicles, AR/VR, cloud gaming, etc. Edge computing is considered to be the key enabler for applications in *Q2*, as additional latencies to compute at traditional cloud and bandwidth strain on backhaul networks to transport generated data may break the "immersiveness" of end-users [37]. As most applications in this quadrant are expected to garner large market shares, these are popularly heralded as the driving force behind edge computing.

**Quadrant III** - *High Latency & High Bandwidth*: The top-right quadrant (yellow) is composed of applications that generate large volumes of data but with "somewhat" relaxed timing constraints. Take, for example, a smart city that implies automatic updates on buses timetables, smart parking meters, and overall maintenance with control mechanisms. The demand *Q3* applications place on edge computing is usually limited to data aggregation and preprocessing to reduce network bandwidth load.

**Quadrant IV** - *High Latency & Low Bandwidth*: The final quadrant in red, located top-left, comprises of applications that neither generate data of large volumes nor require strict latency for operation, e.g. smart homes, weather monitoring, etc. While such applications can leverage the existence of edge computing, they do not offer compelling reasons for deploying edge servers.

## 4 AT THE EDGE OF THE CLOUD

One key driver of edge computing is its claimed ability to provide services at lower latencies than the cloud. While this claim was valid at the emergence of edge computing (circa 2009) due to sparse cloud deployment [40] and higher latencies in the core network than at the last-mile [39]; the world (and cloud) has changed in the past decade. For instance, Amazon's cloud has increased from 3 to 22 datacenter locations [1], and wide area latencies to Google's CDN have decreased from 100 ms to 10-25 ms [61]. On the other hand, edge computing is still in its infancy with no (wide-area) deployment to date, so the claims seem more speculative than real.

We re-evaluate the latency-centered claims of edge computing via extensive global wide-area measurements to datacenters of major cloud providers. We aim to understand if the cloud access latency is still too high in satiating the requirements of emerging applications, or are the clouds already "close enough".

### 4.1 Measurement Methodology

**End-Points.** We chose 101 cloud regions with compute datacenters (e.g. ec2) from seven cloud providers, Amazon, Google, Microsoft Azure, Digital Ocean, Linode, Alibaba, and Vultr, as end-points, shown in Figure 3a, and established a VM in every selected location. The chosen operators are widely used and provide global coverage with distinct network infrastructure. Some, e.g. Amazon, Google etc. have installed *private*, large bandwidth, low latency network backbones with wide-scale ISP peering, while others, e.g. Linode, largely rely on the *public* Internet for connectivity.

**Vantage Points.** We used 3200+ RIPE Atlas probes [62] distributed in 166 countries as vantage points for our measurements (shown in Figure 3b). RIPE Atlas is a global Internet measurement platform that is widely used for reachability, connectivity, and performance studies. Atlas probes are installed in varying network environments

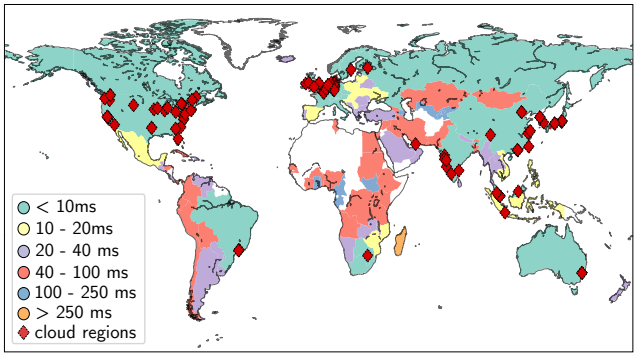



**Figure 4: Minimum latency to nearest datacenter globally.**

(core, access, or home), allowing us to analyze the user reachability to the cloud globally. We filter out all the probes that are clearly installed in privileged locations (e.g., datacenters, cloud network) from our measurements using their user-defined tags [6].

**Experiment.** We measured end-to-end latencies between users (Atlas probes) and cloud datacenters within the same continent via `ping` every three hours. For probes in continents with low datacenter density, e.g., Africa and South America, we also measured latencies to datacenters in adjacent continents, i.e., Europe and North America. Our measurements are ongoing since *September 2019*, and the results in this work are drawn from *nine months* of data collection. Overall, our dataset includes 3.2 million datapoints spanning several GBs and is available for public use [18].

## 4.2 Proximity to the Cloud

*What is the least latency with which countries can access the nearest datacenter?* The question allows us to analyze the spread of cloud across the globe in terms of latency. We extract the minimum `ping` latency observed by the best-performing probe for every country to any cloud datacenter. Figure 4 shows the map of latency distribution per country. The results show that 32 countries can access the cloud with RTTs less than 10 ms, and another 21 countries with RTTs between 10 to 20 ms (MTP threshold). Our findings become more intuitive upon correlating geographical latencies to locations of targeted datacenters (red diamonds) in Figure 4. Countries with cloud access latency less than 10 ms typically have one or more local datacenters, and those with latencies less than 20 ms either share borders or have direct fiber connectivity [68] to the country housing a datacenter. In fact, all *but* 16 countries (mostly in Africa) can access the cloud within PL threshold bounds (100 ms).

While the above shows only the best probe in every country, Figure 5 plots the CDF of the minimum latency observed from *every* probe in our dataset to *any* datacenter, grouped by continents. The result includes probes without a stable Internet connection, or with wireless connectivity. Despite this, the results support the findings in Figure 4. Around 80% probes in Europe and North America – ≈50% of our total probes – can access a cloud datacenter within MTP (20 ms). Probes in Oceania follow similar performance pattern as almost all of them can access the cloud within 50 ms RTT. Surprisingly, despite the low availability of cloud regions and substandard network deployment, ≈75% probes in Africa and Latin America achieve less than 100 ms cloud access latency and meet PL thresholds. While the results in this section are "optimistic" – in that they show the *minimum* latency – they also indicate that the cloud potentially can support latency requirements for applications driving edge computing.

## 4.3 Where is the Delay?

**Insufficient Infrastructure Deployment.** Our results above focused on best-case scenarios to illustrate the *potential* reach of the cloud. We now turn to our entire latency dataset to shed light on the *reality* of the cloud. Figure 6 shows the latency distribution of all measurements grouped by continent. Probes in North America, Europe, and Oceania exhibit excellent cloud reachability, with *more than 75%* of the probes achieving RTT below the PL threshold. The top 25% probes in NA and EU can even support MTP threshold required by edge-compelling applications, e.g. AR/VR, autonomous vehicles, etc. The reason for this exceptional performance (also hinted in §4.2) is the concentrated efforts of cloud providers to deploy datacenters in these continents. Note that the long tail of latency distribution for EU is largely missing from NA. On deeper analysis, we found that the primary contributors to the tail are probes in eastern EU and countries without local or neighboring datacenters, in line with our assessment from Figure 4.

We now turn our focus on the remaining continents, i.e., Latin America, Asia, and Africa. Cloud reachability from within these continents is quite poor, and only a fraction of probes can satisfy the PL threshold. Probes in Asia show much diverse latencies primarily due to scattered datacenter deployment favoring certain countries, like China and India. Unsurprisingly, the worst performance is in Africa as it is severely under-served, both in cloud presence (only one operating region) and network infrastructure [15].

**Nature of last-mile access.** Many studies analyzing the performance of wireless access have been conducted in the past. Be it WiFi in home networks [66] or LTE in public spaces [50], the consensus of last-mile being the bottleneck is well established. Reasons for lack of wireless performance can be many, from packet drops due to contention, to network bufferbloating because of handovers [35]. As most applications in Figure 2 rely on wireless, we also analyze its impact as access medium to the cloud.

We leverage RIPE Atlas user-provided tags [6] to filter probes which indicate the type of access link, e.g. *ethernet, broadband* for wired and *lte, wifi, wlan* for probes connected to network through wireless links. We further filter probes deployed in similar regions in both sets and verify that their baseline latency is in line with their country's average. Figure 7 compares the latencies observed by both sets throughout our measurement period. We find that probes tagged with wireless keywords perform consistently worse than their wired counterparts – taking ≈2.5× longer to access the nearest cloud region. Our result is in line with previous studies showing that users can experience 10-40 ms of added latency while using wireless as last-mile [65, 66]. While these results might improve in future with solutions such as 5G promising much shorter wireless latencies, however, the technology is still in its nascent stages and these promises are waiting to be delivered [71].

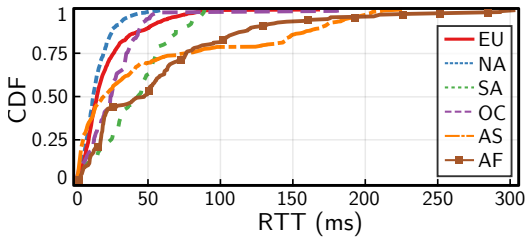



**Figure 5: CDF of minimum RTT of all probes to nearest datacenter by continent.**

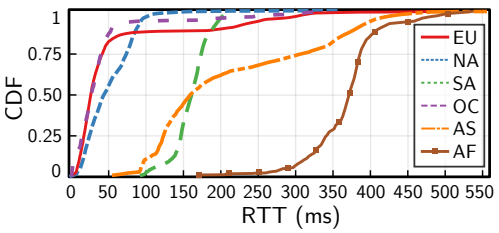



**Figure 6: CDF of all `ping` measurements from all probes to their closest datacenter.**

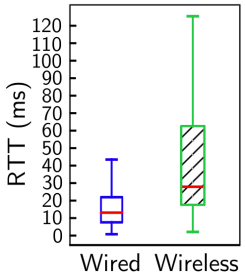



**Figure 7: Wired vs. wireless access RTT.**

## 5 DISCUSSION

***Revisiting Edge Applications.*** We reconsider Figure 2 with our updated knowledge on real-world latencies from our measurements. As per §4.3 and previous measurement studies, current wireless technologies do not support access link latencies below 10 ms. While new wireless standards *promise* to improve the situation, e.g., 10 Gbps speeds with WiFi-6 [22] and 1 ms latency with 5G [34], the reality may differ from claims. For example, at its inception in 2011, LTE promised access latencies below 10 ms while the user is near the base station [25]. However, recent measurements show that the standard commonly experiences delays lasting several seconds due to queue build-ups [35] and handovers [72]. Recent investigations report performance of preliminary deployment of 5G in the real-world to be sub-optimal [49, 71]. Likewise, Hadzic et al. [31] and Cartas et al. [12] find that latency gains for accessing edge server colocated with an LTE basestation is minimal compared to accessing a datacenter located ≈ 1000 km away. While the "true" gains of 5G are yet to be seen, considering supporting strict MTP thresholds, even with edge servers located at basestations, seems uncertain. From §4.2, we can conclude that for most of Europe and North America (and majority of the world in best case), cloud latencies are low enough to support applications operating under perceivable latency. On the contrary, due to lack of network infrastructure, some countries (in Asia & Africa) see cloud access latencies of 150–200 ms, making perceivable latency unachievable but HRT-based applications feasible by the cloud.

Figure 8 recaps the edge applications and their network requirements but now adds latency (red) and bandwidth (blue) "reality" boundaries, as shaded regions (based on the results in §4). The lower bound on latency is ≈10 ms, i.e., the current state of wireless access latency. The upper bound is the human reaction time – as this is supported by the cloud almost globally. For bandwidth, edge is most useful for applications generating enough data to congest the network. Specifically, benefits from the edge are greatest close to the users, and decrease with increasing distance. Contrarily, it is also well-established that the primary bandwidth bottleneck is usually the last-mile [66]. While last-mile bandwidth congestion also depends on contention and competition, based on previous studies [35], we estimate 1GB/entity data generation to be a fitting threshold for edge's bandwidth aggregation gains.

The overlap is the "*feasibility zone*" (FZ) of edge computing. Applications in this zone, e.g., traffic camera monitoring, cloud gaming,

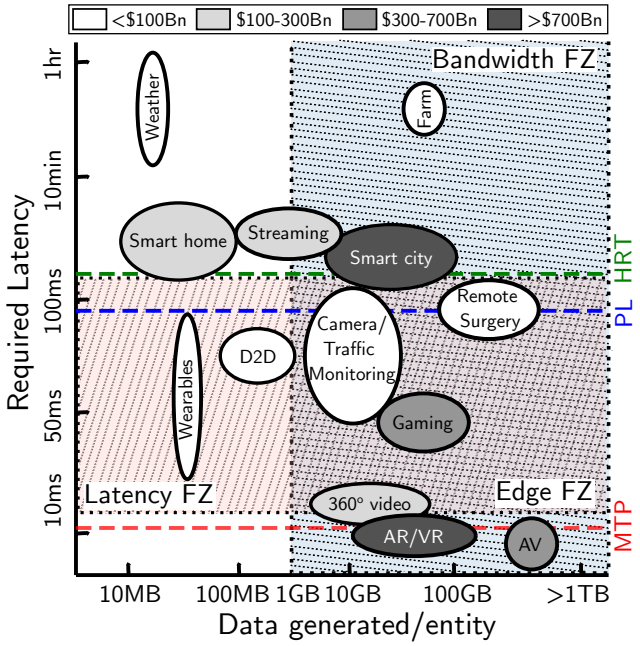



**Figure 8: Edge applications but with edge computing feasibility zones (FZ). The red shaded area represents potential latency gains, and the blue shaded region is the bandwidth gain zone for utilizing edge.**

etc., clearly benefit from a wide deployment of edge as they impose both latency and bandwidth constraints. Surprisingly, the primary drivers of edge computing research (the ones with the most hype) *do not* fall in this zone. For some, it is due to low bandwidth requirements (e.g., wearables), and for others, it is either too stringent (e.g., autonomous vehicles) or too relaxed (e.g., smart cities) latency constraints. Interestingly, the predicted market share of applications within the edge FZ pales compared to those for which edge does not provide much benefit. Further, many applications in the edge FZ can be supported by a wider deployment of cloud/network infrastructure, especially in Asia, Latin America, and Africa.

***Other Considerations.*** We recognize that our critique above may not fully encompass the utility of edge computing due to possible limitations in our methodology, emerging application diversity or

other complications that we may not have considered. For example, while we place applications in edge FZ based on their latency and bandwidth requirements, factors – like privacy – may push other neighboring applications into the FZ envelope. We identify several such factors that may impact our findings and require further research investigation.

**Network vs. application latency.** It may be argued that our viewpoint does not include additional processing delays imposed by applications as we derive network latency from `ping`. However, a recent study from Facebook report similar results as ours and shows that clients rarely observe latencies above 40 ms while accessing their services hosted in the cloud globally [60]. Furthermore, we plan to extend our measurements to include TCP-based probing techniques [41] that may better reflect behavior of application traffic inbound cloud networks.

**Computing power**: Our discussion in this paper did not consider the differences in computation power between the cloud and edge servers. The more pervasive deployment edge needs, the lower the likely processing capabilities of individual edge servers become. It is thus quite possible that despite extensive edge deployment, faster processing and availability of specialized hardware (like GPUs) offered by the cloud may far exceed the network latency gains from deploying applications at the edge [12].

**Economies of scale:** One decisive advantage cloud computing is economies of scale, which edge is unlikely to meet. For cloud, aggregating a large number of servers in a single location achieves substantial savings on building, maintaining, and securing the infrastructure. For edge, marked gains in latency are possible only via a wide and expensive deployment. While last-mile ISPs are best placed to exploit this, recently, cloud providers have begun to utilize the ISP edge [10, 70] and CDN infrastructure [2], further bringing cloud closer to users. However, as the cloud footprint expands to support lower latency requirements of emerging applications, the cloud infrastructure cost will also increase [33].

## 6 FUTURE RESEARCH DIRECTIONS

With our global measurements, we showed that one of the compelling motivators behind the edge – *reduction in latency* – has lost much of its importance since the inception of the field almost a decade ago. Through the course of this study, we found that latency is only one piece in the puzzle and other considerations may serve as more convincing drivers for edge. To conclude, we outline some more promising directions for future research in edge computing.

**Plausible deployments.** General-purpose edge yields little benefits in well-connected areas, but in developing regions, gains are more significant, making edge deployment more compelling. Efforts, therefore, should instead focus on those regions for deployment. Noghabi et al. [51] lists some application-specific deployments where edge may offer benefits even in developed regions, e.g. handling video feeds from traffic cameras. Such deployments are typically purpose-built to support an organization's workload and may emerge as the preferred solution in lieu of generic telco-hosted edge. However, the race towards deploying an edge infrastructure can be viewed as tussle between ISPs and cloud providers, both competing for bigger share in "compute" market, which may sway plausible deployments favoring one network type over the other. In this case, research related to tradeoffs in placement and utilization of processing capacity may yield interesting insights.

**Privacy** has been brought up as an advantage of edge, as it obviates the need to send (sensitive) data to a central cloud. Encryption alone may not be sufficient to hide all details [5, 56]. As concerns for monitoring [69] and data collection mount, processing local data locally and not sending it to the cloud oligopoly may become more attractive. We see the potential for edge computing to address these concerns, especially for (i) applications with a geographically limited scope of interest and (ii) deployments offered by multiple local providers [74] which are safeguarded by rules of the land.

**Trust and security.** Cloud operators invest heavily in infrastructure security, but how would the situation translate for a widespread deployment of edge in remote locations? Could the same be assumed of the (possibly myriad of) edge operators? What guarantees would an application provider have in these cases? In the cloud the terms-of-operation agreement is between service and a cloud provider and, in case of problems, litigation can be used. Translating this to an edge with multiple participating (somewhat transparent) entities requires much additional work.

## 7 CONCLUSION

In this paper, we investigated the rationale behind edge computing in light of recent trends in cloud computing. Still much favored by research and industry, we showed that original motivations for edge computing are weak in today's Internet. We performed an extensive measurement study lasting several months, using probes from RIPE Atlas platform in 166 countries, to measure the proximity towards deployment of modern cloud providers. We found that in well-connected areas, like Europe or North America, the cloud is able to satisfy almost all application requirements that have been envisioned for edge. The remaining ones may continue to remain infeasible for immediately foreseeable future, as they depend on the last-mile wireless access latency. While new technologies, like 5G, promise to improve the last-mile connectivity, related studies measuring its performance over initial deployment show sub-optimal results and full-fledged roll-out of 5G will take several years to complete. While there may be other, non-technical drivers for edge computing, our results clearly showed that from a performance point of view, the potential benefits of edge computing remain small. Only in less-connected areas, such as Africa, Latin America, or parts of Asia have we discovered larger benefits from edge computing. *In conclusion*, we believe that the research in edge computing should shy away from latency-centric views and instead focus its efforts in problem areas that are unresolved and can truly benefit from the attention.

## ACKNOWLEDGEMENT

We would like to acknowledge RIPE Atlas team for providing us access to their platform and supporting our measurements with increased quota limits. We also appreciate the valuable feedback and insights from Prof. Dr. Jörg Ott for shaping this paper. This work was supported by the Academy of Finland in the BCDC (314167), AIDA (317086), WMD (313477) projects and from the Swedish Foundation for Strategic Research with grant number GMT-14-0032 (Future Factories in the Cloud).

## REFERENCES


[1] Amazon. AWS Global Infrastructure. https://aws.amazon.com/about-aws/global-infrastructure/.

[2] Amazon. CloudFront. https://aws.amazon.com/cloudfront/, 2020.

[3] Amazon. Project Luna, 2020. https://www.amazon.com/luna/landing-page.

[4] G. Ananthanarayanan, P. Bahl, P. Bodík, K. Chintalapudi, M. Philipose, L. Ravindranath, and S. Sinha. Real-time video analytics: The killer app for edge computing. *computer*, 50(10):58–67, 2017.

[5] N. Apthorpe, D. Reisman, and N. Feamster. A smart home is no castle: Privacy vulnerabilities of encrypted iot traffic. *CoRR*, 2017.

[6] R. Atlas. Probe tags. "https://atlas.ripe.net/docs/probe-tags/", 2020.

[7] R. E. Bailey, J. J. Arthur, and S. P. Williams. Latency requirements for head-worn display s/evs applications. In *Enhanced and Synthetic Vision 2004*. International Society for Optics and Photonics, 2004.

[8] R. Balan, J. Flinn, M. Satyanarayanan, S. Sinnamohideen, and H.-I. Yang. The case for cyber foraging. In *Proceedings of the 10th workshop on ACM SIGOPS European workshop*, pages 87–92. ACM, 2002.

[9] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu. Fog computing: A platform for internet of things and analytics. In *Big data and internet of things: A roadmap for smart environments*, pages 169–186. Springer, 2014.

[10] T. Böttger, F. Cuadrado, G. Tyson, I. Castro, and S. Uhlig. Open connect everywhere: A glimpse at the internet ecosystem through the lens of the NetFlix CDN. *ACM SIGCOMM Computer Communication Review*, 48(1):28–34, 2018.

[11] Y. Cao, S. Chen, P. Hou, and D. Brown. Fast: A fog computing assisted distributed analytics system to monitor fall for stroke mitigation. In *2015 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pages 2–11. IEEE, 2015.

[12] A. Cartas, M. Kocour, A. Raman, I. Leontiadis, J. Luque, N. Sastry, J. Nuñez-Martinez, D. Perino, and C. Segura. A reality check on inference at mobile networks edge. In *Proceedings of the 2nd International Workshop on Edge Systems, Analytics and Networking*. ACM, 2019.

[13] C. Chang, S. N. Srirama, and R. Buyya. Indie fog: An efficient fog-computing infrastructure for the internet of things. *Computer*, 50(9):92–98, 2017.

[14] Y. Chen, Q. Feng, and W. Shi. An industrial robot system based on edge computing: An early experience. In *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*, Boston, MA, 2018. USENIX Association.

[15] M. Chetty, S. Sundaresan, S. Muckaden, N. Feamster, and E. Calandro. Measuring broadband performance in south africa. In *Proceedings of the 4th Annual Symposium on Computing for Development*, ACM DEV-4 '13, New York, NY, USA, 2013. Association for Computing Machinery.

[16] S.-W. Choi, S. Lee, M.-W. Seo, and S.-J. Kang. Time sequential motion-to-photon latency measurement system for virtual reality head-mounted displays. *Electronics*, 7(9):171, 2018.

[17] CISCO. Cisco visual networking index: Forecast and methodology, 2016-2021 (whitepaper). 2017.

[18] L. Corneo, N. Mohan, A. Zavodovski, S. Bayhan, W. Wong, and J. Kangasharju. Pruning Edge Research with Latency Shears: Dataset. "https://mediatum.ub.tum.de/1574595", 2020.

[19] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl. Maui: Making smartphones last longer with code offload. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, pages 49–62, New York, NY, USA, 2010. ACM.

[20] A. Davis, J. Parikh, and W. E. Weihl. Edgecomputing: Extending enterprise applications to the edge of the internet. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers &Amp; Posters*, WWW Alt. '04, pages 180–187, New York, NY, USA, 2004. ACM.

[21] S. Dey and A. Mukherjee. Robotic slam: a review from fog computing and mobile edge computing perspective. pages 153–158, 11 2016.

[22] Digital Trends. How Wi-Fi 6 will transform connectivity in your home, at the office, and beyond. "https://www.digitaltrends.com/computing/wi-fi-6-transform-pc-cisco/", 2019.

[23] M. Erol-Kantarci and S. Sukhmani. Caching and computing at the edge for mobile augmented reality and virtual reality (ar/vr) in 5g. In *Ad Hoc Networks*, pages 169–177. Springer, 2018.

[24] ETSI. Multi-access Edge Computing (MEC). https://www.etsi.org/technologies/multi-access-edge-computing, 2019.

[25] ETSI Technical Report. Feasibility study for Further Advancements for E-UTRA. "https://www.etsi.org/deliver/etsi_tr/136900_136999/136912/09.03.00_60/tr_136912v090300p.pdf", 2011.

[26] P. Garcia Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, and E. Riviere. Edge-centric computing: Vision and challenges. *ACM SIGCOMM Computer Communication Review*, 45(5):37–42, 2015.

[27] Gartner Maverick Research. The Edge Will Eat the Cloud. https://blogs.gartner.com/thomas_bittman/2017/03/06/the-edge-will-eat-the-cloud/, 2017.

[28] T. N. Gia, M. Jiang, A.-M. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen. Fog computing in healthcare internet of things: A case study on ecg feature extraction. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 356–363. IEEE, 2015.

[29] Google. Stadia. https://stadia.dev/, 2019.

[30] K. Ha, P. Pillai, W. Richter, Y. Abe, and M. Satyanarayanan. Just-in-time provisioning for cyber foraging. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 153–166. ACM, 2013.

[31] I. Hadžić, Y. Abe, and H. C. Woithe. Edge computing in the epc: A reality check. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, SEC '17, pages 13:1–13:10, New York, NY, USA, 2017. ACM.

[32] D. Haja, M. Szabo, M. Szalay, A. Nagy, A. Kern, L. Toka, and B. Sonkoly. How to orchestrate a distributed openstack. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 293–298. IEEE, 2018.

[33] S. Hasan, S. Gorinsky, C. Dovrolis, and R. K. Sitaraman. Trade-offs in optimizing the cache deployments of cdns. In *IEEE INFOCOM 2014-IEEE conference on computer communications*, pages 460–468. IEEE, 2014.

[34] International Telecommunication Union. Minimum requirements related to technical performance for IMT-2020 radio interface(s). https://www.itu.int/pub/R-REP-M.2410-2017, 2017.

[35] H. Jiang, Y. Wang, K. Lee, and I. Rhee. Tackling bufferbloat in 3g/4g networks. In *Proceedings of the 2012 Internet Measurement Conference*. ACM, 2012.

[36] Y. Jin, S. Renganathan, G. Ananthanarayanan, J. Jiang, V. N. Padmanabhan, M. Schroder, M. Calder, and A. Krishnamurthy. Zooming in on wide-area latencies to a global cloud provider. In *Proceedings of the ACM Special Interest Group on Data Communication*, SIGCOMM '19, page 104–116, New York, NY, USA, 2019. Association for Computing Machinery.

[37] T. Kämäräinen, M. Siekkinen, A. Ylä-Jääski, W. Zhang, and P. Hui. A measurement study on achieving imperceptible latency in mobile cloud gaming. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 88–99, New York, NY, USA, 2017. ACM.

[38] C. Kreibich. Parser for Google Scholar. https://github.com/ckreibich/scholar.py, 2017.

[39] R. Krishnan, H. V. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao. Moving beyond end-to-end path information to optimize cdn performance. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 190–201, 2009.

[40] A. Li, X. Yang, S. Kandula, and M. Zhang. Cloudcmp: Comparing public cloud providers. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 1–14, New York, NY, USA, 2010. Association for Computing Machinery.

[41] Linux Manpage. ITCP Traceroute. "https://linux.die.net/man/1/tcptraceroute", 2020.

[42] S. Mangiante, G. Klas, A. Navon, Z. GuanHua, J. Ran, and M. D. Silva. Vr is on the edge: How to deliver 360 videos in mobile networks. In *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, pages 30–35. ACM, 2017.

[43] K. Mania, B. D. Adelstein, S. R. Ellis, and M. I. Hill. Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity. In *Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, pages 39–47. ACM, 2004.

[44] Microsoft. Xbox Project xCloud, 2019. https://www.techradar.com/news/project-xcloud-everything-we-know-about-microsofts-cloud-streaming-service.

[45] N. Mohamed, J. Al-Jaroodi, and I. Jawhar. Utilizing fog computing for multi-robot systems. In *2018 Second IEEE International Conference on Robotic Computing (IRC)*, pages 102–105, Jan 2018.

[46] N. Mohan and J. Kangasharju. Edge-fog cloud: A distributed cloud for internet of things computations. In *2016 Cloudification of the Internet of Things (CIoT)*, pages 1–6, Nov 2016.

[47] N. Mohan, A. Zavodovski, P. Zhou, and J. Kangasharju. Anveshak: Placing edge servers in the wild. In *Proceedings of the 2018 Workshop on Mobile Edge Communications*, pages 7–12, 2018.

[48] S. H. Mortazavi, M. Salehe, C. S. Gomes, C. Phillips, and E. de Lara. Cloudpath: A multi-tier cloud computing framework. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, page 20. ACM, 2017.

[49] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, and Z.-L. Zhang. A first look at commercial 5g performance on smartphones. In *Proceedings of The Web Conference 2020*, WWW '20, page 894–905, New York, NY, USA, 2020. Association for Computing Machinery.

[50] B. Nguyen, A. Banerjee, V. Gopalakrishnan, S. Kasera, S. Lee, A. Shaikh, and J. Van der Merwe. Towards understanding tcp performance on lte/epc mobile networks. In *Proceedings of the 4th workshop on All things cellular: operations, applications, & challenges*. ACM, 2014.

[51] S. A. Noghabi, L. Cox, S. Agarwal, and G. Ananthanarayanan. The emerging landscape of edge computing. *GetMobile: Mobile Comp. and Comm.*, 23(4):11–20, May 2020.

[52] L. Peterson, T. Anderson, S. Katti, N. McKeown, G. Parulkar, J. Rexford, M. Satyanarayanan, O. Sunay, and A. Vahdat. Democratizing the network edge. *SIGCOMM*

*Comput. Commun. Rev.*, 49(2):31–36, May 2019.

[53] J. S. Preden, K. Tammemäe, A. Jantsch, M. Leier, A. Riid, and E. Calis. The benefits of self-awareness and attention in fog and mist computing. *Computer*, 48(7):37–45, 2015.

[54] K. Raaen, R. Eg, and C. Griwodz. Can gamers detect cloud delay? In *2014 13th Annual Workshop on Network and Systems Support for Games*, pages 1–3. IEEE, 2014.

[55] S. Rambo and E. Sperling. Racing To The Edge. "https://semiengineering.com/racing-to-the-edge/".

[56] J. Ren, D. J. Dubois, D. Choffnes, A. M. Mandalari, R. Kolcun, and H. Haddadi. Information exposure from consumer iot devices: A multidimensional, network-informed measurement approach. In *Proceedings of the Internet Measurement Conference*, IMC '19, pages 267–279, New York, NY, USA, 2019. Association for Computing Machinery.

[57] A. Sapio, I. Abdelaziz, A. Aldilaijan, M. Canini, and P. Kalnis. In-network computation is a dumb idea whose time has come. In *Proceedings of the 16th ACM Workshop on Hot Topics in Networks*, HotNets-XVI, pages 150–156, New York, NY, USA, 2017. ACM.

[58] M. Satyanarayanan. A brief history of cloud offload: A personal journey from odyssey through cyber foraging to cloudlets. *GetMobile: Mobile Computing and Communications*, 18(4):19–23, 2015.

[59] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies. The case for vm-based cloudlets in mobile computing. *IEEE pervasive Computing*, (4):14–23, 2009.

[60] B. Schlinker, I. Cunha, Y.-C. Chiu, S. Sundaresan, and E. Katz-Bassett. Internet performance from facebook's edge. In *Proceedings of the Internet Measurement Conference*, IMC '19, page 179–194, New York, NY, USA, 2019. Association for Computing Machinery.

[61] R. Singh, A. Dunna, and P. Gill. Characterizing the deployment and performance of multi-cdns. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, pages 168–174, New York, NY, USA, 2018. Association for Computing Machinery.

[62] R. Staff. RIPE Atlas: A global internet measurement network. *Internet Protocol Journal*, 18(3), 2015.

[63] Statista. The statistics portal for market data. "https://www.statista.com/", 2019.

[64] L. Sun, F. Duanmu, Y. Liu, Y. Wang, Y. Ye, H. Shi, and D. Dai. Multi-path multi-tier 360-degree video streaming in 5g networks. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, pages 162–173, New York, NY, USA, 2018. ACM.

[65] S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, and A. Pescapè. Measuring home broadband performance. *Commun. ACM*, 55(11):100–109, Nov. 2012.

[66] S. Sundaresan, N. Feamster, and R. Teixeira. Home network or access link? locating last-mile downstream throughput bottlenecks. In *International Conference on Passive and Active Network Measurement*, pages 111–123. Springer, 2016.

[67] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella. On multi-access edge computing: A survey of the emerging 5g network edge cloud architecture and orchestration. *IEEE Communications Surveys & Tutorials*, 19(3):1657–1681, 2017.

[68] TeleGeography. Submarine Cable Map. "https://www.submarinecablemap.com/", 2019.

[69] The New York Times. Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras. "https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html", 2019.

[70] M. Trevisan, D. Giordano, I. Drago, M. Mellia, and M. Munafo. Five years at the edge: Watching internet from the isp network. In *Proceedings of the 14th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '18, pages 1–12, New York, NY, USA, 2018. ACM.

[71] Verizon. Customers in Chicago and Minneapolis are first in the world to get 5G-enabled smartphones connected to a 5G network. https://www.verizon.com/about/news/customers-chicago-and-minneapolis-are-first-world-get-5g-enabled-smartphones-connected-5g.

[72] J. Wang, Y. Zheng, Y. Ni, C. Xu, F. Qian, W. Li, W. Jiang, Y. Cheng, Z. Cheng, Y. Li, et al. An active-passive measurement study of tcp performance over lte on high-speed rails. *International Conference on Mobile Computing and Networking (MobiCom)*, 2019.

[73] D. L. Woods, J. M. Wyma, E. W. Yund, T. J. Herron, and B. Reed. Factors influencing the latency of simple reaction time. *Frontiers in human neuroscience*, 9:131, 2015.

[74] A. Zavodovski, S. Bayhan, N. Mohan, P. Zhou, W. Wong, and J. Kangasharju. Decloud: truthful decentralized double auction for edge clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 2157–2167. IEEE, 2019.

[75] A. Zavodovski, N. Mohan, S. Bayhan, W. Wong, and J. Kangasharju. ICON: Intelligent Container Overlays. In *Proceedings of the 17th ACM Workshop on Hot Topics in Networks*, HotNets '18, pages 15–21, New York, NY, USA, 2018. ACM.