Latency Comparison of Cloud Datacenters and Edge Servers

Batyr Charyyev
School of Systems and Enterprises
Stevens Institute of Technology
bcharyye@stevens.edu

Engin Arslan
Computer Science and Engineering
University of Nevada, Reno
earslan@unr.edu

Mehmet Hadi Gunes
School of Systems and Enterprises
Stevens Institute of Technology
mgunes@stevens.edu

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9322406

***Abstract*—Edge computing has become a recent approach to bring computing resources closer to the end-user. While offline processing and aggregate data reside in the cloud, edge computing is promoted for latency-critical and bandwidth-hungry tasks. In this direction, it is crucial to quantify the expected latency reduction when edge servers are preferred over cloud locations. In this paper, we performed an extensive measurement to assess the latency characteristics of end-users with respect to the edge servers and cloud data centers. We also evaluated the impact of capacity limitations of edge servers on the latency under various user workloads. We measured latency from 8,456 end-users to 6,341 Akamai edge servers and 69 cloud locations. Measurements of latencies show that while 58% of end-users can reach a nearby edge server in less than 10 ms, only 29% of end-users obtain a similar latency from a nearby cloud location. Additionally, we observe that the latency distribution of end-users to edge servers follows a power-law distribution, which emphasizes the need for non-uniform server deployment and load balancing by an edge provider.**

***Index Terms*—Edge computing, Fog computing, Cloud computing, Latency measurement.**

I. Introduction

Growing real-time and cognitive computing applications in daily life moves computing from the central clouds to the network edge [22]. Together with the proliferation of low-power mobile and Internet of Things (IoT) devices, this led to an increase in the demand for computing platforms that can offer low-latency and high-speed communication. While cloud platforms offer dynamically scaled computing power and fit well for compute-intensive jobs, they fall short to meet QoS requirements of delay-sensitive workloads due to high communication time. This is further exacerbated by the fact that inter-AS communications exhibit highly dynamic path selection behavior [9], causing fluctuations in network delay for long haul connections. Therefore, while cloud computing is effective in providing computing architectures at scale, it is unable to meet stringent delay constraints of real-time applications.

Edge computing has recently gained interest by bringing computing resources closer to the end-user to provide low-latency and high-bandwidth communication [12]. The shift toward the edge is propelled both by technological constraints of the centralized data centers and the last mile network capabilities, as well as personal considerations of privacy [22]. As various computing and sensing devices are developed, the variety of edge devices keep growing. The use cases of the edge services include: (i) real-time applications such as connected health, disability aids, and augmented reality, (ii) cognitive computing tasks such as intelligent personal assistants and machine learning for model training/inference, (iii) smart homes to support peak demands from home devices, (iv) video analytic/monitoring for prompt decisions at the edge (e.g., crime detection or prevention, human-device interaction), and (v) autonomous vehicles to support intermittent data access or large data transfers.

Although cloud computing offers high-capacity and reliable services, communication overhead with the cloud may deteriorate the user experience by increasing latency and power consumption. On the other hand, edge servers may lack the capacity to satisfy stringent resource requirements of applications. Hence, complementing edge devices with on-demand cloud services emerged as a potential solution to take advantage of both and improve the quality of experience. While offline processing and aggregate data could reside in the cloud, latency-critical and bandwidth-hungry tasks can be processed at the edge. Researchers found that opportunistically offloading computation tasks to edge servers improve mobile device battery lifetime compared to the cloud and minimize execution time compared to the mobile device [10]. Another study showed that processing mobile applications with the support of the edge platforms speedups applications up to 20 times while reducing energy consumption by 5% [3]. Similarly, performing face recognition at the fog could reduce the response time by 81% [18].

In this paper, we emulate an edge-cloud integrated service scenario using over 8.5k RIPE Atlas nodes to comprehensively assess latency variations for users when they direct their requests to edge servers and cloud data centers. To the best of our knowledge, this work is the first large-scale measurement to quantify the latency benefits when widely deployed edge servers are used instead of cloud data centers.

The major contributions of this paper are:

- We perform large-scale latency measurements from 8,456 end-users to 6,341 edge servers and 69 cloud locations.
- We run a detailed analysis of the edge computing latency for end-users with and without cloud support.
- We share the latency measurements at `github.com/netlab-stevens/cloud-edge-latency`.

978-1-7281-8298-8/20/$31.00 ©2020 IEEE

In the rest, Section II describes the details of the measurement study. Section III compares the latency of edge servers and cloud locations. Section IV analyzes the impact of limited edge servers on supporting end-user demands. Section V presents related work, and Section VI concludes the paper.

II. Measurement Setup

In this section, we describe our experimental setup to measure network latency for edge servers and the cloud providers. We utilized Ripe Atlas nodes around the globe to represent *end-users*, Akamai servers as *edge servers*, and major compute cloud providers as *cloud locations*. We measured the round trip time (RTT) from 8,456 vantage points to 6,341 edge servers and 69 cloud locations. In the rest of the paper, we use *vantage point* and *end-user* interchangeably. Each latency measurement contains five pings from a vantage point to an edge server and each cloud location. The median RTT of five measurements is used as the representative latency for the measurement. Note that, even though Ripe Atlas has more than 10,000 nodes we observed that only portion (i.e., 8,456) of these nodes are continuously active. Thus, we were able to continuously measure latency to all edge and cloud servers from 8,456 vantage points during our measurement campaign. After all measurements, we removed ones where the destination is unreachable or vantage point is not available for measurements to all destinations (i.e., an edge server and all cloud locations). Table I tabulates the number of end-users (i.e., RIPE Atlas nodes), edge servers (i.e., Akamai servers), and cloud locations in each continent. We observe that end-users, edge servers, and clouds are concentrated in America and Europe, with the least being in Africa.

A. Edge Measurements

To model edge computing deployment, we adopted Akamai servers in our measurements. Although Akamai was deployed as a content distribution network (CDN), it has re-focused as an edge computing platform [14]. Akamai currently has a quarter of a million edge servers deployed around the world. Hence, large-scale edge server deployment use cases can be emulated based on Akamai servers [11]. We selected three popular Akamai customer web sites (i.e., Apple, Microsoft, and Yahoo) to discover the closest Akamai server to end-users. Note that not all Akamai servers cache a particular web site, so we observed only 6,341 of the Akamai servers from our vantage points in the measurements.

TABLE I: End-users, edge servers, and cloud locations in each continent

	America	Europe	Asia	Oceania	Africa	Total
Users	1,394	5,419	1,193	238	212	8,456
Edge	2,218	2,210	1,338	331	244	6,341
Clouds	29	20	14	6	0	69
Amazon AWS	6	4	4	1	0	15
Google GCP	6	5	4	1	0	16
IBM Cloud	9	6	5	2	0	22
Oracle Cloud	5	4	0	1	0	10
Rackspace	3	1	1	1	0	6

Since the location and IP of Akamai servers are not public, we rely on Akamai's own mechanism to find the closest server for each request initiated from an end-user. Akamai's mechanism works as follows: let's assume a user wants to access the Microsoft Office website and initiates a request to `office.microsoft.com`, which is served by Akamai servers. The request first tries to resolve an IP address for the `office.microsoft.com` and queries its local DNS server (LDNS), which contacts its name server. The name server returns a canonical name, `officecdn.microsoft.com.edgesuite.net`. Then LDNS performs name translations on the `officecdn.microsoft.com.edgesuite.net` to obtain the IP address of a nearby edge server that contains a cached copy of the `office.microsoft.com`. Note that, `edgesuite.net` is a domain owned by Akamai. For a detailed explanation of the Akamai redirection mechanism please see [16].

In our measurement, to obtain the IP address of the closest edge servers (i.e., Akamai server) to end user we initiated a request to three Akamai customers (i.e., Apple, Microsoft, and Yahoo) from each end user. We used canonical names of customers `appldnld.apple.com.edgesuite.net` for Apple, `officecdn.microsoft.com.edgesuite.net` for Microsoft, and `a943.x.a.yimg.com` for Yahoo, to ensure that an IP address returned from DNS name translation belongs to the Akamai, and not to some centralized Akamai customer's (i.e., Apple, Microsoft, and Yahoo) server. Su et al. [2] observed that median redirection time for Akamai is below 100 seconds. Thus, we used 120 seconds as the time interval between two measurements to avoid DNS caching when making name translation on canonical names. For each of the three Akamai customers, we repeated the experiment two times to obtain IP addresses of the servers close to each end-user.

We discovered 6,341 edge servers for 8,456 end-users after mapping the closest edge servers to each end-user. Table I presents the distribution of edge servers (i.e., Akamai servers) per continent. Note that as we only observed a fraction of Akamai servers, the latency results for edge servers is an upper bound measurement for the Akamai as there could be closer servers that we could not observe. While we can precisely measure latency to each cloud location, we only reach a fraction of about 240,000 Akamai servers that are located in 17,000 AS networks. Table II presents the percentage of end-users reaching to a particular number of edge servers. As we performed measurements to three different websites hosted by Akamai for two measurements, each end-user can see up to six different servers. While most end-users discovered 4 to 6 servers, some users only observed 1 or 2 distinct IPs.

TABLE II: End-users observing a number of edge servers

Observed servers	1	2	3	4	5	6
End-user %	1.3	2.9	16.5	23.9	31.3	24.1

B. Cloud Measurements

We selected five popular compute cloud service providers (i.e., Google, Amazon AWS, Rackspace, Oracle, and IBM) to perform cloud measurements. Among popular providers, we excluded Microsoft Azure. Azure load balancer drops ICMP packets, and hence we could not obtain RTT measurements. Table I tabulates the number of data centers in each continent. Note that, at the time of these experiments, none of the measured cloud providers had a datacenter in Africa. Common locations of the datacenters include Seoul, Singapore, Tokyo and Hong-Kong for Asia; Frankfurt, London, Paris and Amsterdam for Europe; Sao Paola, California, North Virginia and Texas for America; and Sydney for Oceania.

We performed measurements from each end-user (i.e., RIPE Atlas vantage point) to a particular cloud location by deploying an instance in each data center of the cloud provider. We measured the latency from each end-user to all cloud instances. Then, we picked the minimum of all instances as the latency between the user and the cloud provider. Note that cloud providers may have multiple zones inside a location. For example, region us-central1 in Google Cloud has zones us-central1-a, us-central1-b, us-central1-c, and us-central1-f. Since the latency between zones in the same region is usually negligibly small, we measured the latency of only one zone for regions with multiple zones.

III. Cloud-Edge Latency Comparison

In this section, we compare the latency of cloud locations and edge servers to each end-user. We use labels C1-C5 instead of the actual cloud provider names to keep the focus on the edge-cloud comparison and its implications rather than ranking cloud providers. We analyze the latency of edge servers compared to all locations of a cloud provider.

Observation 1: Cloud providers prefer similar locations to deploy data centers. As a result, end-users experience negligible latency enhancement by using multiple cloud providers.

Figure 1 shows the latency comparison of a user to the edge and individual cloud providers. While 55% of end-users can find a nearby edge server with less than 10 ms latency, this value reaches 82% with 20 ms latency. However, for individual cloud providers, the user coverage ratio varies between 3%-21% and 22%-52% for 10 ms and 20 ms latency ranges, respectively. We also combined all cloud providers as if they were a single provider. Even in this case, the user coverage ratio improves slightly to 27% (for 10 ms latency) and 62% (for 20 ms latency). We believe this is because data centers of different cloud providers are in nearby locations such as Sao Paola, California, and Texas in America.

Interestingly there are some end-users with unexpected latency to edge and cloud servers. For instance, we observed that 12 end-users had a latency of more than 250 ms to the closest edge server. Similarly, for cloud providers, there are 6 to 14 end users with a latency larger than 250 ms. They are unexpected latencies because network latency on the equatorial circumference is around 200 ms [7]. These cases might be due to Ripe Atlas nodes connecting through a VPN or a middlebox, or network misconfiguration.

Observation 2: Compared to the cloud, edge servers offer lower latency to 92% end-users.

We also measured the latency difference between edge servers and cloud providers for all end-users. In Figure 2, the x-axis indicate the range of latency difference between end-user to cloud and edge. Negative values indicate that a cloud provider location is closer to the end-users than edge servers. For example, the value of (-100,-10) refers to end-users for which the latency of the closest edge server is 10 to 100 ms greater than the latency of the closest cloud provider location. Note that the buckets are increasing in size, and the center bucket indicates latency difference less than 1ms. We observe that edge servers provide smaller latency to 92% to 97% of end-users compared to different cloud providers. For significant majority of end-users, edge servers are closer than cloud providers with a latency difference of 10 to 100 ms.

Latency characteristics might differ among regions such as Europe and Asia. Thus, it is important to analyze the regional difference in observed latency. Figure 3 further shows user coverage comparison in each continent between edge servers and all cloud locations combined. In Africa, the difference in user coverage is considerably large because there is no cloud datacenter in Africa. Similar behavior is observed for Oceania as there are up to two datacenters for any cloud provider. In Americas and Europe, cloud has the best coverage compared to the edge servers.

Observation 3: Regional latency analysis reveals that cloud

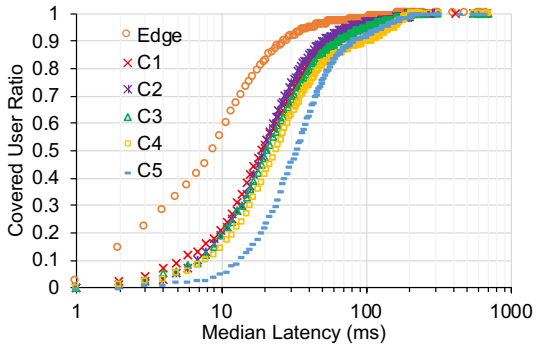

Fig. 1: User latency coverage (CDF)

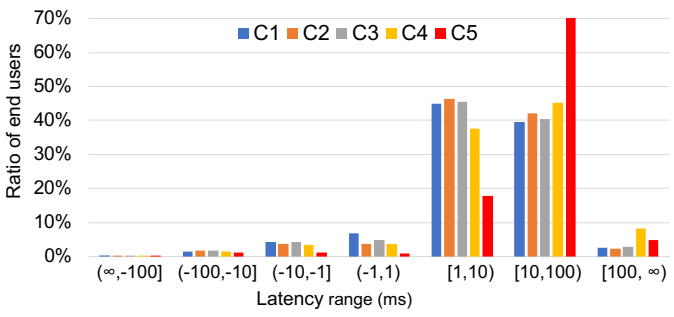

Fig. 2: Latency difference of end user to cloud and edge (negative values correspond to lower cloud latency)

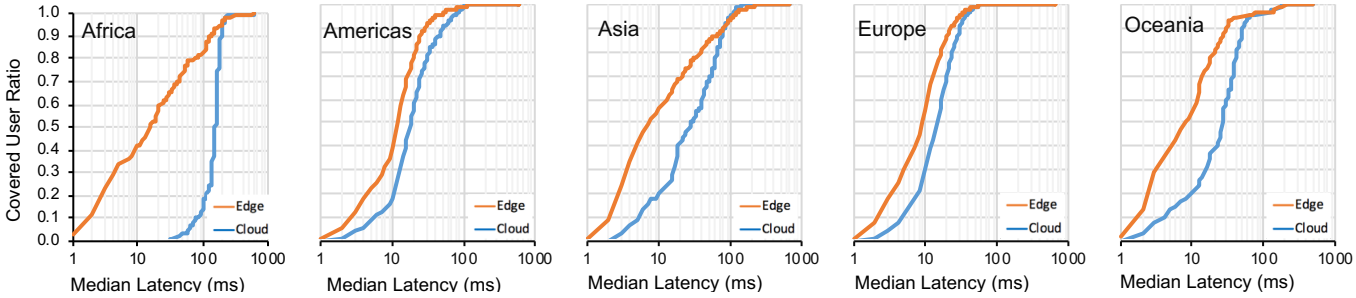

Fig. 3: User coverage by continent for edge servers and combined cloud providers

locations can match the performance of edge servers where data centers are abundant such as in West Europe.

Figure 4 compares the latency difference of cloud provider C1, the best provider in terms of user latency, and edge servers for each individual end-user. The x-axis shows the end-users partitioned by continent and ranked by the latency difference. The y-axis shows the latency difference in a log-scale. Note that y=1 ms indicates a latency difference of less than 1 ms in either direction due to the log-scale of the figure. Results for other clouds had similar distributions with C5 having the highest latency values. We observe that latency between end-users to edge servers is considerably lower than latency between end-users to cloud locations. In all continents, however, there exists a small number of end-users for which the latency to the C1 cloud is much smaller than to the closest edge server. This difference might be due to the lack of edge servers (i.e., Akamai servers) close to the end-users. We also noticed that some of our end-users are actually hosted by the C1 cloud, thus they yield a considerably smaller latency to the cloud in those measurements. In Europe, a non-negligible portion of end-users observes a small latency difference between edge servers and cloud locations.

When we analyze Europe in detail we observe that the average latency difference of edge servers and the combined cloud is around 10 ms except for Western Europe where the difference is even smaller. Note that many of the cloud locations are located in Western Europe such as Frankfurt, London, Paris, and Amsterdam and hence end-users in Western Europe observe a smaller latency difference between the edge and the cloud. Lack of coverage of cloud providers in some continents amplifies the need for edge servers in those locations to improve the quality of service for end-users.

Overall, we observe that edge servers can provide better latency to a significant majority of end-users than the cloud providers. While the cloud can provide comparable latency in certain regions (e.g., West Europe), more than 95% of end-users are better served by the edge servers, in some cases by order of magnitude. These findings confirm that edge servers are well-suited for latency-critical applications.

IV. Impact of Limited Resources

In this section, we analyze the impact of limited resources of edge server on observed latency.

Observation 4: Latency-based distribution of end-users to edge servers follows power-law, requiring non-uniform server deployment to avoid hot spots and increase the quality of experience for end-users.

Cloud datacenters encapsulate thousands of servers deployed in one location [13]. This gives the imagination of the unlimited resource capacity in cloud datacenters. Although edge servers offer significant latency improvement over the cloud, capacity limitations of edge servers may prevent user requests to be completed, especially in densely populated areas. Figure 5 shows the distribution of edge servers to end-users when each end-user is assigned to the closest edge server. In total, out of 6,341 edge servers, only 2,855 are closest for at least one end-user while the rest have longer latencies to end-users. We observe a power-law distribution where 1,624

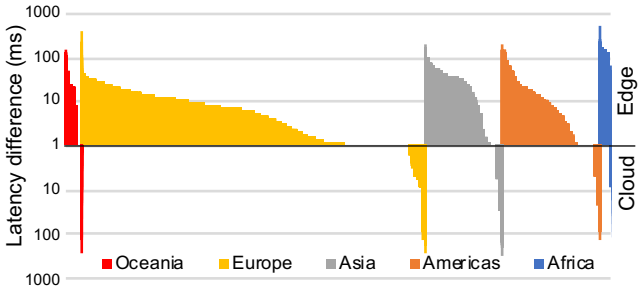

Fig. 4: Latency comparison of best performing cloud provider C1 and edge servers. The x-axis shows the end-users partitioned by continent and ranked by the latency difference.

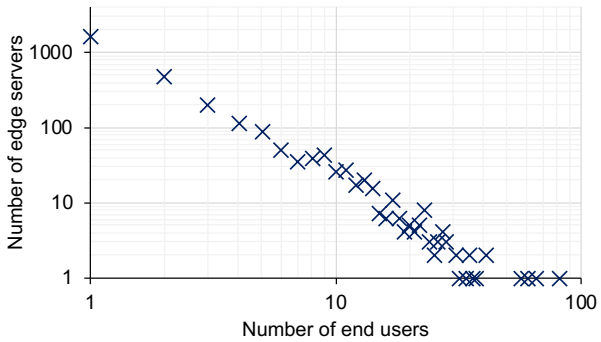

Fig. 5: Distribution of end-users to edge servers

edge servers are the closest to only one end-user (i.e., x=1) while an edge server is the closest to 81 end-users. Power law exponent of α = 2.78 falls in the typical ranges for scale-free networks. As a result, some end-users might be directed to a more distant edge server (or a cloud location) when the closest edge server does not have sufficient resources available for the user workload.

To analyze the impact of limited resources of edge servers under different user workloads, we used a public Google cluster workload trace dataset [4]. The dataset previously used in [20] to represent edge node capacities, [19] used the dataset to evaluate COSTA a task offloading model in mobile-edge computing, and [8] used it to evaluate Dedas online deadline-aware dispatching and scheduling algorithm for edge computing. It is important to note that while server specifications and workload characteristics of Google cloud traces could potentially be different from genuine edge server properties and edge user workload. Yet, this emulation is valuable to assess the impact of capacity limitations under high traffic scenarios as, to best of our knowledge, it is the most relevant public dataset that provides both user workloads and server capacities measured from a real system.

Anonymized and normalized traces of the dataset contain server specifications (i. e., CPU, RAM, and disk capacity) for 12,583 servers and 437,377 user requests (i.e., CPU and RAM requirements, arrival time, and job duration) which we will also refer as user workload, for over a month. Servers in the Google cluster traces are located in one particular datacenter. To resemble the edge scenario and distribute the centralized servers, we assigned CPU and memory capacity from the traces to each edge server (i. e., Akamai server). Then we assigned each user request from traces to one end-user (i.e., Ripe Atlas node).

The workload can be increased by dividing the start time of the user requests. Since each user request is mapped to one end-user, the number of active end-users at a given time will also increase. Note that, the start time and the duration of a user request are provided by the traces. We considered low and high workload cases in which low workload has less than 10,000 active user requests at a given time and high workload has more than 100,000 user requests. We believe that a high workload can be expected in metro cities with a large number of end-users. Figure 6 presents user latency coverage when an edge is complemented with cloud (i. e., Edge with C1) and when it is not supported with cloud (i. e., Limited Edge) thus, have limited capacity. While 59% of user requests can be served with a latency of less than 10 ms with the support of C1 under low workload, the ratio falls below 40% as the workload intensifies to higher levels. When there is no cloud support, the proportion of the covered users does not reach to 100% as some users cannot be scheduled to any edge server in their vicinity. This implies that supporting edge servers with resource-rich cloud locations can be a viable option to address the resource limitations of edge servers.

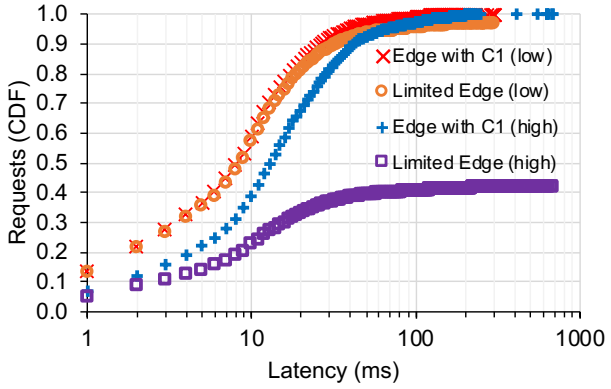

Fig. 6: User latency coverage under low and high workloads

Combined with the result from Figure 5, where we observed a power-law distribution of edge server reachability from end-users, we can conclude that it is necessary to consider edge server capacity and intensity of user demands to mitigate over-subscription of edge servers. Another approach to address congested edge servers is to complement them with cloud locations that have a much higher CPU and memory resources.

V. Related Work

There have been measurement studies focusing on the latency of communication of remote computing services. [1] comparatively analyzed latency toward four cloud providers from 200 PlanetLab nodes and observed that the number of data centers and their locations play a crucial role in latency. [2] showed that Akamai redirections overwhelmingly correlate with network latency on the paths between clients and the Akamai servers. [24] analyzed the benefits of switching from a single-cloud to a multi-cloud deployment and showed that 20-50% of IP prefixes would reduce their latency to the closest data center by more than 20%. Choy et al. [15] investigated the impact of augmenting the cloud infrastructure with servers located near the end-users in the on-demand gaming industry. They showed that adding a small number of servers in new locations increases user coverage significantly. Authors found that Amazon AWS EC-2 is capable of providing a median latency of 80 ms or less to fewer than 70% of end-users. By adding 300 servers at different networks, they showed that the ratio of covered users increases by 28%. Similar study was conducted by Zhang et al. [21], in which they propose EdgeGame that offloads the video rendering of the mobile games to the edge servers instead of cloud. In evaluation, authors use servers deployed in the central part of China as cloud data center and servers deployed in the same city with end users as edge nodes. The experiment results show that EdgeGame can reduce the average network delay by 50 percent and improve the user's QoE by 20 percent. Finally, Wang et al. [5] compares the Cloud Content Delivery Networks in terms of the performance and cost for video streaming. They use PlanetLab to conduct measurements and evaluate three cloud CDNs including Amazon Web Service (AWS) CloudFront, Microsoft Azure Verizon CDN, and Google Cloud CDN. The results show that cloud vendors vary in providing

QoE across regions, the QoE provided by one CDN can change over time, cloud CDNs vary in scalability, and finally some cloud CDN is more economical than others given certain cache hit rate.

Chang et al. [6] propose an edge cloud to augment cloud data centers with service nodes placed at the network edge to improve the performance of latency-sensitive applications. The edge cloud helped 3D indoor localization and video surveillance applications by yielding better latency and bandwidth utilization. Yi et al. proposed LAVEA, an edge computing platform for low-latency video analytics [17]. The results show that client-edge configuration has led up-to 4x speedup compared to client-cloud configuration. Furthermore, they investigate three task placement schemes for inter-edge collaboration and observed that *the shortest scheduling latency first* has the best overall task placement performance compared to *the shortest transmission time first* and *the shortest queue length first* schemes. Our latency analysis sheds further light on how much latency improvement can be enhanced by selecting the closest edge device in a global deployment. Zhuo et al. [23] analyze the performance of seven applications in terms of latency when they are offloaded to edge computing platforms with different configurations. They showed that offloading to cloud incurs an additional 100 to 200 ms latency compared to offloading to a nearby edge device.

While previous works focused on latency analysis mostly on specific domains, technology, or use cases, this study presents a comprehensive measurement from around 8,500 end-users toward 6,300 destinations; provides detailed analysis of the latency observed by end-users on different regions and with genuine user resource request data; and explores the impact of limited capabilities of edge platforms compared to the resource-rich cloud under different user workloads.

VI. Conclusion and Future Work

In this paper, we performed a large-scale measurement to compare the latency from end-users to edge and cloud providers. Our results confirm that edge servers can provide considerably lower latency than cloud locations for the significant majority of the end-users. The results, however, also indicate that solely latency-oriented assignment of user requests to edge servers leads to power-law load distribution. Few of the edge servers are closest to many of the end-users, while a large number of edge servers are located nearby to only a couple of end-users. We further extended our study by analyzing the impact of the edge server capacity limitations on observed latency using a sample from Google cloud workload traces. Results indicate that edge servers either need to be provisioned to handle increasing user workloads in certain hot spots or backed by the cloud to ease user over-subscription.

As future work, we plan to develop efficient resource allocation algorithms for edge platforms, which may be supported by cloud locations. We also plan to develop a brokerage service which can direct user request to available edge server or cloud location based on the QoS metrics such as latency, bandwidth, power, and cost. Such a service will enable edge providers to market their excess capacity and allow users to access resources within expected QoS conditions.

References

[1] Ang Li, Xiaowei Yang, Srikanth Kandula, and Ming Zhang, "Cloudcmp: comparing public cloud providers." SIGCOMM IMC, p: 1-14, 2010.

[2] Ao-Jan Su and Aleksandar Kuzmanovic, "Thinning Akamai", SIGCOMM IMC, pP 29–42. ACM, 2008.

[3] B G Chun, S Ihm, P Maniatis, M Naik, and A Patti, "Clonecloud: elastic execution between mobile device and cloud." EuroSys'11 Proceedings of the sixth conference on Computer systems April 2011 P 301–314

[4] C Reiss, J Wilkes, and J L. Hellerstein. "Google cluster-usage traces: format schema." Technical report, Google Inc., Nov 2011. https://github.com/google/cluster-data.

[5] C. Wang, A. Jayaseelan and H. Kim, "Comparing Cloud Content Delivery Networks for Adaptive Video Streaming," 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), San Francisco, CA, 2018, pp. 686-693.

[6] H. Chang, A. Hari, S. Mukherjee and T. V. Lakshman, "Bringing the cloud to the edge," INFOCOM WKSHPS, Toronto, 2014, pp. 346-351

[7] I. Grigorik "High Performance Browser Networking: Understanding the latency benefits of multi-cloud webservice deployments." O'Reilly September, 2013 Chapter 1

[8] J. Meng, H. Tan, C. Xu, W/ Cao, L. Liu, B. Li, "Dedas: Online Task Dispatching and Scheduling with Bandwidth Constraint in Edge Computing", 2019 IEEE INFOCOM

[9] K. Bakhshaliyev, M. A. Canbaz, and M. H. Gunes, "Investigating Characteristics of Internet Paths," ACM Transactions on Modeling and Performance Evaluation of Computing Systems, 4(3):1-24, 2019.

[10] M. H. U. Rehman, C. Sun, T. Y. Wah, A. Iqbal and P. P. Jayaraman, "Opportunistic Computation Offloading in Mobile Edge Cloud Computing Environments," 2016 17th MDM, pp. 208-213.

[11] M. Pathan and R. Buyya, "A Taxonomy and Survey of Content Delivery Networks.", The Univ. of Melbourne, GRIDS-TR-2007-4, 2007

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

[13] Pierr Johnson "With The Public Clouds Of Amazon, Microsoft And Google, Big Data Is The Proverbial Big Deal" June 15 2017 https://www.forbes.com/sites/johnsonpierr/#1e4e4b3a33b0

[14] "The Akamai Intelligent Edge Platform", Accessed: Feb 22, 2020. https://www.akamai.com/us/en/what-we-do/intelligent-platform/.

[15] S. Choy, B. Wong, G. Simon and C. Rosenberg, "The brewing storm in cloud gaming: A measurement study on cloud to end-user latency," 2012 11th NetGames, Venice, 2012, pp. 1-6.

[16] S. Ramesh K, K. Mangesh, L. Woody, J. Manish, "Overlay networks: An akamai perspective", 2014 Advanced Content Delivery, Streaming, and Cloud Services, volume 5, pp: 305-328

[17] S. Yi, Z. Hao, Q. Zhang, Q. Zhang, W. Shi, and Q. Li. "LAVEA: latency-aware video analytics on edge computing platform.", SEC '17: October 2017 Article 15 Pages 1–13

[18] S. Yi, Z. Hao, Z. Qin and Q. Li, "Fog Computing: Platform and Applications," HotWeb Workshop, Washington DC, 2015, pp. 73-78.

[19] T. X. Tran, K. Chan, D. Pompili, "COSTA: Cost-aware Service Caching and Task Offloading Assignment in Mobile-Edge Computing" 2019 IEEE SECON

[20] X. Deng, J. Li, E. Liu, H. Zhang, "Task allocation algorithm and optimization model on edge collaboration" Journal of Systems Architecture Vol 110, Nov 2020, 101778

[21] X. Zhang et al., "Improving Cloud Gaming Experience through Mobile Edge Computing," in IEEE Wireless Communications, vol. 26, no. 4, pp. 178-183, August 2019.

[22] Yuhao Zhu, Gu-Yeon Wei, and David Brooks. "Cloud no longer a silver bullet, edge to the rescue." CoRR abs/1802.05943 2016.

[23] Z. Chen, W. Hu, J. Wang, S. Zhao, B. Amos, G. Wu, K. Ha, K. Elgazzar, P. Pillai, R. Klatzky, D. Siewiorek, M. Satyanarayanan, "An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance." SEC '17: Article No 14 pp 1–14

[24] Z. Wu and H.V. Madhyastha, "Understanding the latency benefits of multi-cloud webservice deployments", ACM SIGCOMM CCR, 43(2):13–20, 2013.