Characterizing Cloud-to-user Latency as perceived by AWS and Azure Users spread over the Globe

Fabio Palumbo*, Giuseppe Aceto*,◇, Alessio Botta*,◇,
Domenico Ciuonzo*, Valerio Persico◇, and Antonio Pescapé*,◇
*University of Napoli Federico II (Italy) and ◇NM2 srl (Italy)
{fabio.palumbo, giuseppe.aceto, alessio.botta, domenico.ciuonzo, valerio.persico, pescape}@unina.it

***Abstract*—With the growing adoption of cloud infrastructures to deliver a variety of IT services, monitoring cloud network performance has become crucial. However, cloud providers only disclose qualitative info about network performance, at most. This hinders efficient cloud adoption, resulting in no performance guarantees, uncertainties about the behavior of hosted services, and sub-optimal deployment choices. In this work, we focus on cloud-to-user latency, i.e. the latency of network paths interconnecting datacenters to worldwide-spread cloud users accessing their services. In detail, we performed a 14-day measurement campaign from 25 vantage points deployed via Planetlab infrastructure (emulating spatially-spread users) and considering services running in distinct locations on the infrastructures of the two most popular public-cloud providers, namely Amazon Web Services and Microsoft Azure. Our experimentation allows to provide an in-depth performance characterization (based on multiple probing methods and fine-grained sampling rate) of such networks as perceived by users spread worldwide. Results show the presence of both spatial and temporal latency trends. Finally, by evaluating the advantages of multi-cloud deployments, our results also provide useful guidelines to cloud customers.**

***Index Terms*—public-cloud networks; Amazon Web Services; Microsoft Azure; network measurements; network performance.**

I. Introduction

Thanks to the remarkable techno-economical benefits achievable, public clouds have seen increasing adoption during the last years.[1] In line with the wider spectrum of applications currently deployed onto the cloud, resulting in diverse service-level requirements, a fine-grained characterization of cloud-performance has become a key factor. Hence, measurement activities aiming at monitoring the performance of cloud networks have raised a growing interest in both providers and customers, and have thus become the workhorse to both operate and capitalize cloud services [1, 2]. Sadly, cloud providers rarely provide guarantees or disclose details on network performance [3]. Hence, *non-cooperative approaches* [4, 5] have emerged as a viable alternative to gain visibility about cloud-network performance "building blocks": (i) intra-datacenter, (ii) inter-datacenter, and (iii) cloud-to-user networks.

This work focuses on the performance of *cloud-to-user* network (i.e. the set of paths interconnecting users to the set of pooled resources composing the cloud), which is usually *harder to be monitored* and accurately predicted than that of the intra- and the inter-datacenter networks [4, 6]—also from the cloud provider viewpoint. In fact, the actual service performance experienced by the users is impacted by their location with respect to the cloud resources, and thus depends on network segments not under the direct provider control. This has led providers to distribute datacenters (and their offered services) geographically, so as to reduce propagation delays to users (by shorter distances) and improve the QoS.

Equally important, in order to get the "whole picture" of cloud performance, different network metrics are worth to be considered, with their relevance varying with the specific cloud application to deploy. Among these, the *latency* perceived by users is a critical parameter in several applications (e.g. real-time video processing, cloud gaming [7] or ultra-reliable and low-latency communications services in 5G [8]) requiring low latency, low latency variation, or both. These requirements have also led to the emergence of edge-cloud architectures, where computing resources are pushed towards the end-users to reduce the overall latency. However, while these novel paradigms represent the cutting edge of research and technology, only a limited set of customers can already leverage edge-computing services (often to *integrate* cloud-based services rather than to replace them) as they imply huge infrastructural investments. Thus, general users mostly rely on the cloud paradigm today whose performance evaluation is still expected to be of critical importance in next years.

This paper investigates the performance of the cloud services of the two main public-cloud providers, namely *Amazon Web Services (AWS)* and *Microsoft Azure*, currently retaining $\approx 50\%$ of the market share[2] and often used together in case of multi-cloud deployments.[3] In detail, we have conducted an extensive, 14-day long experimental campaign, monitoring the cloud-to-user latency for both providers at *high frequency* and with *multiple active methods* (i.e. relying on different functions of the TCP/IP stack and counterparts at cloud side), so as to

This work has been partially funded by GRISIS project (CUP: B63D180002800079), DD MIUR prot.368 of 24/10/2018, Programma Operativo FESR Campania 2014-2020.

[1]https://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html#~mobile-forecast.

[2]https://www.srgresearch.com/articles/leading-cloud-providers-increase-their-market-share-again-third-quarter.

[3]https://www.kentik.com/blog/report-multi-cloud-cost-containment-world/.

978-1-7281-0962-6/19/$31.00 ©2019 IEEE

enable a *fine-grained analysis*. Also, to investigate cloud-to-user network performance vs. the geographical position, we leveraged Planetlab [9] research infrastructure *deploying* 25 *Vantage Points (VPs)* to monitor the network latency perceived by cloud users towards cloud services deployed in 8 *distinct datacenters* (4 per provider) within different continents.

Our campaign, compared to related works (cf. Tab. I), *represents an unmatched investigation to date*, thanks to VPs spread with higher density and latency monitored with higher frequency and diversity. Thanks to the above campaign, we are able to investigate how latency varies according to the provider (for the two most popular ones), the impact of different probing methods on its estimates, the region in which the cloud services are hosted and the users' location, as well as over the time with a fine granularity. As a result, we provide interesting guidelines for both cloud customers (serving final users) and providers, with the goal of both supporting performance assessment and making deployment decisions, including the adoption of *multi-cloud*. Finally, to promote reproducibility and open research, our collected dataset is *publicly released*.

The rest of the paper is organized as follows. Sec. II reviews the literature on network performance of public-cloud providers; Sec. III describes the methodology underlying the experimental campaign provided to measure the considered public-cloud services, while Sec. IV discusses the results obtained; Sec. V ends with conclusions and future directions.

II. Related Work

Several recent works have investigated different aspects of the performance of the public-cloud networks, focusing on a number of evaluation metrics obtained via different measurement techniques [4, 10–13]. For example, the *goodput* home-users can achieve when accessing contents on AWS S3 storage-as-a-service is investigated in [10], while a characterization of the achievable intra-datacenter *throughput* of AWS EC2 Infrastructure-as-a-Service (IaaS) is given in [4]. Differently, the authors of [11] have estimated *available bandwidth* in public clouds, showing the impact of traffic shaping policies and virtualization. Focusing on *availability* and *reliability* instead, Hu et al. [12] have shown that their estimates may differ when considering probes at different levels of the communication stack, highlighting the need to make realistic requests at the application layer to obtain accurate results. Karacali et al. [13] have considered both *throughput* and *delay* between node pairs within cloud networks, evaluating different traffic patterns.

Referring to works specifically targeting *latency* in cloud networks, the authors of [6] provide a characterization for the latency in *inter-datacenter* networks. The study in [14] measures the *intra-datacenter* and *inter-datacenter latency*, and is aimed at network troubleshooting from the provider's viewpoint. Concerning *cloud-to-user latency*, the study in [15] leverages collected latency measurements to evaluate the deployment of *hypothetical* cloud services in different geographical locations. On the other hand, the work in [7] provides a study of latency in actual cloud infrastructures in the context cloud gaming applications, highlighting the need to expand the infrastructure at the edge to satisfy the stringent requirements of this scenario. Authors in [16] present *Claudit*, a platform for collecting latency measurements from distributed VPs, considering Azure as provider. Latency (more specifically, round trip time) is measured at different layers, including TCP `SYN`/`SYN ACK`, and `HTTP GET` requests (a) to a web server, or (b) implying additional queries to an auxiliary database. The same paper also provides a collected data overview, regarding the latency values experienced from the different clients. Also, in the same work and in [17], a methodology for the detection of suspicious events is presented, leveraging the multi-dimensional data collected. Claudit was then expanded to (additionally) collect measurements towards AWS; these data are then leveraged in [18] to evaluate a benchmarking methodology for cloud providers. Such methodology allows to compare cloud providers through user-defined metrics (e.g. mean latency, standard deviation, coefficient of variation). The work, however, does not provide an in-depth evaluation of the methodology, but simply applies it to a restricted scenario. Equally important, data from multiple source points are aggregated, not investigating per-VP (or per-region) results. The same authors in [19] leverage the collected measurements in order to detect anomalies via unsupervised learning.

TABLE I
Taxonomy of works on cloud-to-user latency monitoring.

Work	# of providers	# of probe types	# of layer-4 ports	Probing period (minutes)	# of VPs	# of CRs
[16]	1	4	1	3	5	2
[17]	1	4	1	3	6	4
[18]	2	3	1	4	4	4
[19]	2	4	1	4	6	4
This	**2**	**4**	**2**	**1**	**25**	**4**

Summarizing, all the above literature on cloud-to-user latency is mostly based on the data collected via Claudit platform, but each work focuses on a peculiar slice of the whole dataset, either considering different: providers, number of probe types, period between each measurement, number of VPs and Cloud Regions (CRs). To this end, in Tab. I we categorize the aforementioned works according to these features.

Compared to the above works, our analysis considers (other than the same number of CRs and both AWS/Azure providers) a *higher number of VPs* (i.e. 25 VPs as opposed to only 6 deployed by Claudit[4]), covering a *larger* geographical area. Secondly, we use the same probing methods included in Claudit with the *addition* of HTTP and TCP testing non-standard ports, thus allowing to investigate the presence of different enforced policies based on the transport-layer port used for communication. Thirdly, we measure latency with a *finer granularity* (1 min.) w.r.t. previous works (see Tab. I).

III. Methodology

Herein we describe the whole experimental procedure used to measure cloud-to-user latency, i.e.: (a) the public-cloud

[4]Note that this number would be higher even if we included the secondary and backup nodes deployed in the platform, reaching a total of 15 VPs.

providers and the CRs considered; (b) the geographically-spread VPs emulating cloud users; (c) the probing methods employed; (d) details of implementation and reproducibility.

Public Cloud Providers and Cloud Regions (CRs). Current cloud market is dominated by a few global providers, with *Amazon* being the clear leader (1M+ active customers in 190+ countries), and *Azure* representing the only clear challenger[5], and both are steadily expanding their global infrastructure. Hence, in this work we considered the IaaS of these two cloud providers, i.e.: *EC2* for Amazon and *Virtual Machines* for Azure. Also, to explore *spatial diversity*, we have identified *four* regions in distinct geographic continents (hereinafter CRs), where both providers have deployed their datacenters: `Ireland` (Europe), `Virginia` (North America), `Sao Paulo` (South America), and `Singapore` (Asia-Pacific).

Vantage Points (VPs). To deploy the source nodes for our campaigns, we leveraged the open platform *Planetlab* [9] for emulating cloud users spread worldwide. Specifically, we relied on 25 Planetlab VPs acting as probing sources and instructed to measure the latency perceived by users with different probing methods by means of probing bulks sequentially issued with 1-min. sampling rate. This rate is higher than that adopted in similar works [12, 16] and thus allows a finer-grained analysis. VPs have been placed in the *same four regions* as the CRs according to node availability, with the following distribution: Asia-Pacific (`AP`), 8 VPs; Europe (`EU`), 6; North America (`NA`), 10; South America (`SA`), 1.

Probing methods. In our experimental campaign, we adopted *active* probing methods, i.e. that inject probing traffic in the network to estimate the latency via Round Trip Time (RTT). We highlight that the measured RTT includes processing time at the end-host, as well as queueing, transmission, and propagation delays along the whole network path. The latter term, depending on the geographical distance between the VP and the CR, imposes a lower-bound on the latency due to physical constraints. On the other hand, DNS resolution has no impact on the estimated RTT, since cloud resources are addressed leveraging numeric IP addresses. Also, in line with recent works [16, 18], in our campaign we adopted *multiple* active methods. Precisely, the adopted probing methods (a) take advantage of communication mechanisms at different TCP/IP stack levels and (b) possibly rely on different counterparts at cloud side (i.e. servers). The probing methods used in our work are *(i)* ICMP, *(ii)* TCP, *(iii)* HTTP, and *(iv)* HTTP-DB.

ICMP probing relies on the `echo request`/`echo reply` messages. It operates at the network layer and does not require a specific setup at server side but the running host (i.e. the virtual machine running via the IaaS paradigm). Nonetheless, Hu et al. [12] suggested that ICMP probing should be used with caution as it may be unsuitable for measurements involving cloud environments. Differently, *TCP probing* takes advantage of the `SYN`/`SYN ACK` messages which provide RTT measurements as perceived by data-transfer protocols (instead of being related to ICMP control messages). It however requires a TCP server running on the cloud host. *HTTP probing* uses the `HTTP GET`/`200 OK` messages. It evaluates the time to download a few-byte resource from the cloud via HTTP. While also in this case the transmission delay is negligible (due to the size of the downloaded contents), HTTP probing requires a TCP connection to be established, thus resulting in at least 2×RTT. Moreover, a (negligible) processing on the cloud is implied to serve each request. Finally, *HTTP-DB probing* similarly uses `HTTP GET`/`200 OK` messages. Differently from HTTP probing, it relies on a web server that interacts with a database running onto another cloud VM (i.e. an auxiliary server), thus emulating a three-tier application, with latency impacted also by intra-datacenter contribution (between the web server and the database). To evaluate the potential impact of *preferential traffic policies* by both cloud and network providers, TCP and HTTP probing use both well-known (80) and non-standard (54321) destination ports. No method implements application-level retransmission.

Reproducibility and Open Research. To summarize, we measured the latency in cloud-to-user networks from 25 VPs at 1-min. rate for 14 days (since 1st Jun. '16). Measurements were run towards cloud datacenters in *four* distinct regions and operated by *two* different providers, for a total of 200 measured *paths*. As each path is monitored via multiple probing methods, our dataset results in 1100 distinct timeseries[6] with ≈ 14k samples each. Specifically, we employed *HPing3* for TCP and ICMP probing methods, and *HTTPing* for both *HTTP* and *HTTP-DB* probing methods. Also, we run *MySQL* database on the auxiliary server for HTTP-DB. Finally, to support open research via reproducibility of our characterization and fostering further advances on public cloud services assessment, the *dataset is publicly released* for research purposes.[7]

IV. Results

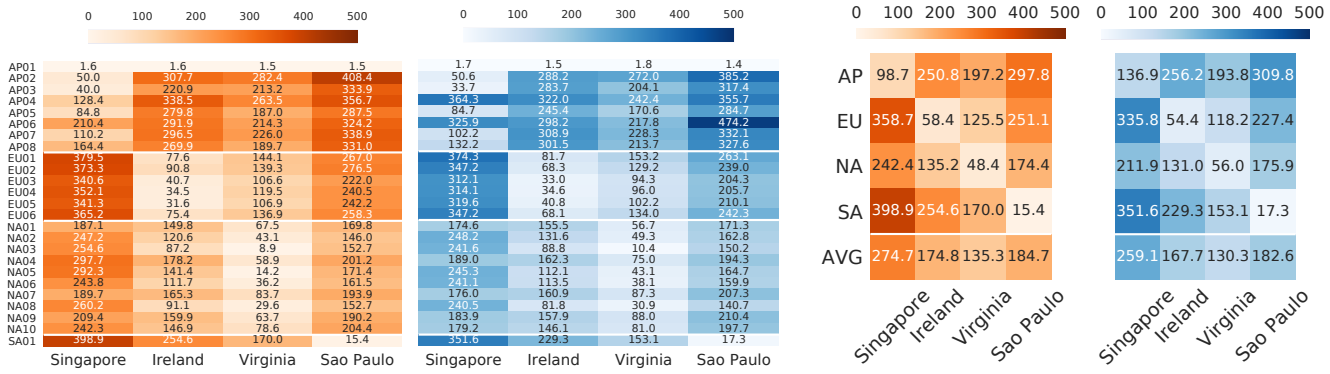
Henceforth we report and discuss the results of our experimental campaigns. Specifically, we first provide a high-level assessment of the overall campaign, and compare the performance observed for the two providers. Then, we deepen the performance variability over time, and detail how variability is perceived from different VPs. Also, we delve into latency dependence on different probing methods considered. Finally, we investigate the benefits of multi-cloud deployments.

Overall view and Comparison between Providers. We first provide a high-level view of the cloud-to-user latency considering each `(VP, CR)` pair *separately*. To this aim, Figs. 1a and 1b report the average latency (TCP probing, port 80) experienced from each VP when targeting the four CRs for AWS and Azure, respectively. First, results show that latency values (intuitively) grow with the distance between the VP and the CR (lower values are observed for paths connecting VPs and datacenters within the same geographic region). Interestingly, this finding does not hold when considering other metrics in analogous contexts, e.g. network throughput [10].

[5]https://www.bmc.com/blogs/gartner-magic-quadrant-cloud-iaas/.

[6]Note that ICMP probing was not suitable for Azure datacenter due to traffic-filtering policies implemented.

[7]http://traffic.comics.unina.it/cloud.



(a) AWS, detailed. (b) Azure, detailed. (c) AWS, aggregated. (d) Azure, aggregated.

Fig. 1. Average latency [ms] (14-day span, TCP probing method, port 80). (a) and (b) report detailed results at `(VP, CR)` pair granularity for AWS and Azure, respectively. (c) and (d) report results aggregated (average) by VP region for AWS and Azure, respectively. `AVG` reports the CR-average.

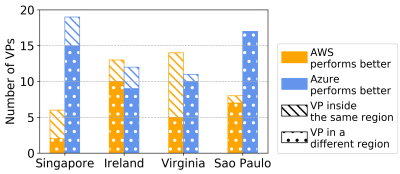


Fig. 2. Comparison of the two providers in terms of latency, according to the Wilcoxon signed-rank test (TCP probing method, port 80).

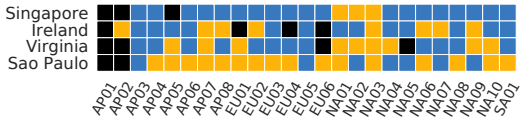


Fig. 3. Comparison in terms of standard deviation, according to Levene test (14-day span, TCP probing method, port 80). Orange and blue color report cases where AWS and Azure show lower variability, respectively. Black color highlights no significant difference between the providers.

Results highlight that neither provider always outperforms the other, as the outcome varies with the specific `(VP, CR)` pair. Secondly, Figs. 1c and 1d report the previous results after aggregating VPs by region. The figures (beyond expected lower values on the main diagonal, corresponding to latencies measured within the same region) show how the `Singapore` CR is the one with highest total-average latency, for both providers, with the VPs in `SA` representing the worst case. Differently, it can be derived that the deployments in `Virginia` offer, for both providers, the *lowest total-average latency*, namely considering all the VPs across the world (with VPs in `AP` being the more penalized). Hence, by supposing a cloud customer wants to deploy an application leveraging a single CR (e.g. for budget constraints), and considering potential users to be scattered around the globe, leveraging `Virginia` datacenters would be the most suitable choice.

In order to deepen the *comparison between providers*, we then compare the measured latency for AWS and Azure over time. In detail, we use the *Wilcoxon signed-rank test* to assess a statistically-significant[8] difference in latency timeseries for any `(VP, CR)` pair. We underline that such test is *non-parametric*: this confers robustness to deviations of the measured latency from Gaussianity and, also, to *outliers*. Fig. 2 reports the outcome, with a per-CR barchart highlighting for how many VPs each provider performed better on the 14-day span, based on Wilcoxon test.[9] Then, statistically-significant comparisons are broken down by (i) *intra-region* cases (VP and cloud datacenter in the same region) and (ii) *inter-region* cases (VP and cloud datacenter in different regions). Results show that the *best-performing provider changes with the CR considered*. For instance, for services delivered via `Ireland` and `Virginia` CRs, AWS reports better performance for more VPs (13 and 14 out of 25 VPs, respectively). Differently, Azure performs better for `SaoPaulo` and `Singapore` CRs (17 and 19 out of 25 VPs, resp.). Also, by limiting the analysis to *intra-region* cases, AWS *always outperforms* Azure, especially in `Virginia` CR (e.g. 9 out of 10 VPs deployed in `NA` experienced lower latency towards AWS). Lastly, an opposite trend is seen for *inter-region* cases with Azure (save `Ireland` CR).

Latency variability over Time. Beyond desirable low latency values, a wide range of applications also demand its small variability over time [7, 16]. Hence, the (non-parametric) *Levene test* is used herein to assess whether there is a statistically-significant difference in the latency variability of the two providers for the same `(VP, CR)`, i.e. to test the equal-variance hypothesis for the two timeseries. Fig. 3 reports the comparison (over the 14-day span) of latency variability expressed as the variance, with a row for each CR and a column for each VP.[10] First, a non-negligible amount of `(VP, CR)` pairs with no significant difference in latency variability between providers is observed (black color), with up to 4 VPs out of 25 toward `Ireland` and `Virginia` CRs, in contrast with Fig. 2. Interestingly, this implies that in a number of cases *lower latency does not imply also reduced variability*.

[8]In this work a conservative p-value of 0.01 was chosen for all the tests.

[9]We highlight that, to assess a statistically-significant lower latency of either Azure or AWS, the sign of the Wilcoxon statistic was taken into account.

[10]The transformed response variables of Levene statistic were used to assess a statistically-significant lower latency variability of either Azure or AWS.

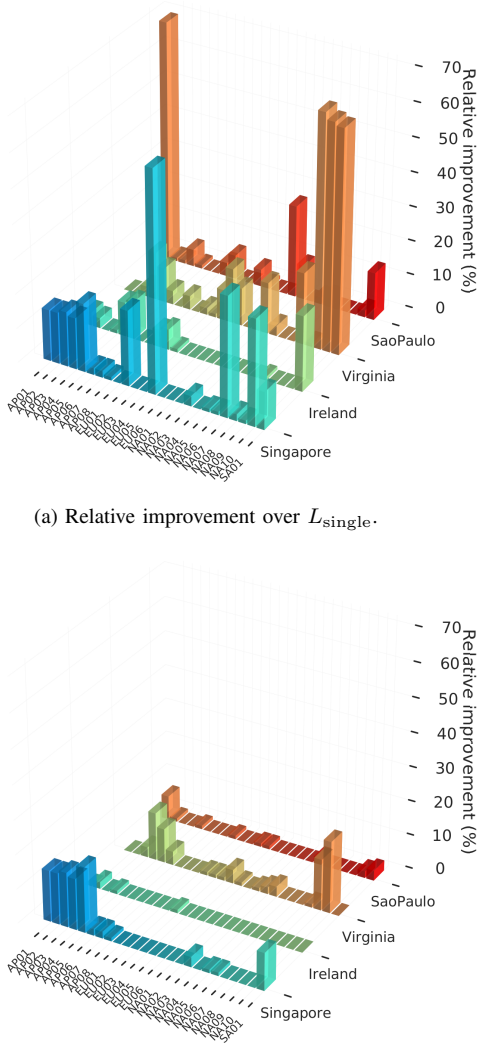
Differently, focusing on statistically-significant comparisons, the result is also in this case influenced by the specific CR, with `Singapore` (resp. `SaoPaulo`) leading to a lower variance for Azure (resp. AWS) in most cases. For other CRs, the comparison is more balanced and depends on VP.

Latency Variability over Space. To deepen how measured latency varies with VPs and CRs, we have analyzed the correlation between the latency timeseries measured between VP pairs to any CR and for both providers. Results (not shown for brevity) have highlighted that (i) there is low correlation in the majority of the cases, even considering paths from different VPs to the same CR (for a given provider) and (ii) only few negative correlations coefficients appear, in case of VP pairs located very far from each other. Nonetheless, in one specific case a *notable exception was observed*: timeseries related to latency measurements towards the `SaoPaulo` Azure datacenter from all but 3 VPs (i.e. `AP01`, `AP02`, and `EU05`) show significantly-positive correlation (> 0.3 for 25.6% of the cases, according to Pearson correlation coefficient). This leaves room to speculate that the correlated variations of latency may be possibly due to: network congestion (a) within the Azure intra-datacenter network in `SaoPaulo` or (b) at the network provider connecting this datacenter to the Internet.

Impact of Probing Methods. As discussed in Sec. III, cloud-to-user latency can be measured through different probing methods, possibly requiring different configuration at server side. Our results report that probing methods adopting mechanisms implemented at different levels of the TCP/IP stack may report different latency estimates, as discussed hereinafter.

Specifically, the Wilcoxon signed-rank test applied to compare 14-day timeseries obtained through the TCP probing with port 80 and port 54321, (for any `(VP, CR)` pair and both providers), reports that for 111 (resp. 89) pairs using TCP probing with standard port returns values statistically lower (resp. higher) than the same probing method with port 54321. In detail, for only 6 VPs (4 for Azure and 2 for AWS) the probing method reporting lower latency is the same for all the CRs. Nonetheless, in general, no clear pattern (i.e. dependence on either the VP or the CR) emerged. Results for HTTP probes using ports 80 and 54321 are also statistically different: for a total of 5 VPs (3 of which for AWS and 2 for Azure) HTTP port 80 experienced a lower latency towards every CR; while for 7 VPs (2 for AWS and 5 for Azure) experienced a lower latency with HTTP port 54321 regardless the cloud region.

As expected, for the HTTP-DB probing methods, a higher latency was found compared to HTTP. This result was statistically confirmed for every VP and cloud region: on average, HTTP-DB experiences around 9 ms higher latency than HTTP. This overhead results from the latency of the intra-datacenter network and the processing time at the auxiliary server. Finally, comparing TCP (port 80) and ICMP probing (only applicable for AWS provider, due to network-configuration constraints imposed by Azure), the former reports a (statistically significant) lower latency in the majority (almost 60%) of the cases. While investigating the root cause of these discrepancies is out of the scope of this work, results suggest that these aspects should be taken into consideration when designing non-cooperative methodologies for monitoring public-cloud networks, being provided that different probing methods may return results that differ up to 198 ms, on average. For instance, this is evident for the `AP01` VP, where the presence of a TCP proxy along the path towards the cloud causes the monitored latency to be heavily underestimated when using TCP probing with destination port 80. Concerning ICMP, our results are in line with what observed in [12] for service availability measurements: although ICMP is widely adopted—as it does not require particular instrumentation at the targeted cloud node—its results can differ from latency experienced by upper layer protocols, possibly leading to both an underestimation and (more often) an overestimation of the observed latency.



(a) Relative improvement over L_{single}.

(b) Relative improvement over L_{best}.

Fig. 4. Improvements achievable with multi-cloud deployments w.r.t. the *globally-better provider* (a) and the *locally-better provider* (b).

Evaluating the Benefits of Multi-cloud Deployments. Multi-cloud architectures (based on the concomitant use of services of two or more cloud providers) are increasingly adopted by enterprises, so as to exploit the flexibility deriving from

multiple cloud offerings, thus achieving cost reduction and increased reliability[3]. Hereinafter we focus on the potential gains customers could achieve when adopting multi-cloud architectures in terms of *improved network performance*. In detail, we evaluate the upper-bound of the cloud-to-user latency reduction w.r.t. *two baseline* cases: (i) the adoption of a sole cloud provider for all the users (in this case we consider the adoption of the provider with better performance, on average, on a global scale, i.e. Azure according to previously shown results), denoted with L_{single}; (ii) the adoption of the best-performing provider on a `(VP, CR)` basis (i.e. for each `(VP, CR)` pair we consider to *statically* adopt the provider with better performance on average, based on previously-discussed results, denoted with L_{best}). These two baselines are compared to the ideal performance obtained with a multi-cloud architecture, i.e. at each instant in time the user is served by the provider reporting the best performance (say L_{MC}). Notably, this ideal case is representative of an architecture either (a) leveraging a system predicting which provider offers better performance at each time, or (b) duplicating the resources and properly managing the redundancy.

The results when considering the two baselines, focusing on TCP probing (port 80), are reported in Figs. 4a and 4b. The former reports for each `(VP, CR)` the relative improvement with respect to L_{single}, (i.e. $\frac{L_{\text{single}}-L_{\text{MC}}}{L_{\text{single}}} \times 100$), while the latter the relative improvement with respect to L_{best}, (i.e. $\frac{L_{\text{best}}-L_{\text{MC}}}{L_{\text{best}}} \times 100$). Our results show how multi-cloud deployments achieve better performance when compared to a deployment relying on the provider performing better on a global scale (performance improves more than 5% in 29% of the cases and up to 70.8%, cf. Fig. 4a) but also when compared against the locally-better provider (performance improves $> 5\%$ in 7% of the cases and up to 21.3%, cf. Fig. 4b).

V. Conclusions and Future Directions

Since both customers and providers may suffer poor visibility in cloud-to-user network, this work assessed the latency performance of these networks for AWS and Azure. We measured the latency for a 14-day long timespan as perceived by globally-spread users, using different active probing methods and collected a dataset publicly-released to the community.

Aiming at an in-depth characterization, we first provided a latency overview, adopting statistical tests to compare the performance of the two providers for each `(VP, CR)` pair. Results have shown that the best-performing provider changes with the specific CR considered. Moreover, variability observed from different VPs is uncorrelated for most of the cases, except for a single datacenter (Azure's `Sao Paulo`) thus suggesting congestion in the cloud access network to be a potential root cause. Also, the analysis of latency measured by different probing methods shows that although some reasonable findings are observed, such as (i) higher latency for HTTP-DB w.r.t. HTTP (ii) lower latency for TCP w.r.t. ICMP, in general no clear pattern has emerged, thus highlighting their diversity and the need for all these probing methods in a complete characterization. Lastly, it has been shown that non-negligible latency gains are guaranteed in 7% of the cases with an ideal multi-cloud deployment, with a relative gain up to 21.3%, also w.r.t. the (*locally-better*) single-cloud deployment.

Such results provide interesting guidelines for both *cloud customers* and *providers*. The former can choose where to deploy their applications based on latency requirements and deployment cost. The latter can assess the performance as perceived by users and to identify bottlenecks impacting the performance of the offered services. Future works will leverage the collected data to develop prediction techniques, to allow proactive management of the cloud infrastructure and less-invasive probing by adaptive methods.

References

[1] M. Kwon, Z. Dou, W. Heinzelman, T. Soyata, H. Ba, and J. Shi, "Use of network latency profiling and redundancy for cloud server selection," in *IEEE CLOUD'14*.

[2] M. Menzel and R. Ranjan, "CloudGenius: decision support for web server cloud migration," in *ACM WWW'12*.

[3] J. C. Mogul and L. Popa, "What we talk about when we talk about cloud network performance," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 5, pp. 44–48, 2012.

[4] V. Persico, P. Marchetta, A. Botta, and A. Pescapé, "Measuring network throughput in the cloud: The case of Amazon EC2," *Elsevier Computer Networks*, vol. 93, pp. 408–422, 2015.

[5] A. Li, X. Yang, S. Kandula, and M. Zhang, "Cloudcmp: comparing public cloud providers," in *ACM IMC'10*.

[6] V. Persico, A. Botta, P. Marchetta, A. Montieri, and A. Pescape, "On the performance of the wide-area networks interconnecting public-cloud datacenters around the globe," *Computer Networks*, vol. 112, pp. 67–83, 2017.

[7] S. Choy, B. Wong, G. Simon, and C. Rosenberg, "The brewing storm in cloud gaming: A measurement study on cloud to end-user latency," in *IEEE/ACM NetGames'12*.

[8] A. Ksentini, P. A. Frangoudis, P. C. Amogh, and D. Nikaein, "Providing low latency guarantees for slicing-ready 5G systems via two-level MAC scheduling," *IEEE Network*, vol. 32, no. 6, pp. 116–123, 2018.

[9] A. C. Bavier et al., "Operating systems support for planetary-scale network services," in *USENIX NSDI'04*.

[10] V. Persico, A. Montieri, and A. Pescapé, "On the network performance of Amazon S3 cloud-storage service," in *IEEE Cloudnet'16*.

[11] P. Ha and L. Xu, "Available bandwidth estimation in public clouds," in *IEEE INFOCOM WKSHPS'18*.

[12] Z. Hu et al., "The need for end-to-end evaluation of cloud availability," in *PAM'14*.

[13] B. Karacali, J. M. Tracey, P. G. Crumley, and C. Basso, "Assessing cloud network performance," in *IEEE ICC'18*.

[14] C. Guo et al., "Pingmesh: A Large-Scale System for Data Center Network Latency Measurement and Analysis," *ACM SIGCOMM Computer Communication Review*, pp. 139–152, 2015.

[15] Y. A. Wang, C. Huang, J. Li, and K. W. Ross, "Estimating the performance of hypothetical cloud service deployments: A measurement-based approach," in *IEEE INFOCOM'11*.

[16] O. Tomanek, P. Mulinka, and L. Kencl, "Multidimensional cloud latency monitoring and evaluation," *Elsevier Computer Networks*, vol. 107, no. Part 1, pp. 104–120, 2016.

[17] P. Mulinka and L. Kencl, "Learning from Cloud latency measurements," in *IEEE ICCW'15*.

[18] V. Uhlir, O. Tomanek, and L. Kencl, "Latency-based benchmarking of cloud service providers," in *IEEE/ACM UCC'16*.

[19] P. Mulinka, P. Casas, and L. Kencl, "Hi-Clust: Unsupervised analysis of cloud latency measurements through hierarchical clustering," in *IEEE CloudNet'18*.