A First Look at Public-cloud Inter-datacenter Network Performance

Valerio Persico*, Alessio Botta*,⋄, Antonio Montieri⋄, Antonio Pescapé*,⋄
*University of Napoli Federico II (Italy) and ⋄NM2 srl (Italy)
{valerio.persico, a.botta, pescape}@unina.it, montieri@nm-2.com

***Abstract*—Public-cloud providers do not disclose quantitative information about the performance of their inter-datacenter networks in spite of their importance and of the growing interest they are attracting. In this paper we propose an analysis of the inter-datacenter network of the two leading providers—*Amazon Web Services* and *Microsoft Azure*—only leveraging active monitoring approaches and thus not relying on information restricted to providers.**

Our results show that Azure inter-datacenter infrastructure performs better than Amazon's in terms of throughput (+52%, on average). On the other hand, the performance of the two providers is comparable in terms of latency, with the exception of isolated cases. Counterintuitively, lower performance may be even related to higher costs for the customer. Network management policies that may severely impact both the performance perceived by the customers and the results of the measurement activities have been observed and characterized. Finally, a comparison with previous works shows that TCP throughput has not improved recently.

I. Introduction

Companies more and more leverage cloud solutions to supply services across the Internet and a growing number of applications is now delivered through cloud-based infrastructures. Accordingly, top players have made huge investments in *networks of datacenters* that host on-line services and cloud platforms to cope with this increasing demand. Networking costs amount to around 15% of a datacenter's total worth and are roughly equal to power expenses [12], with wide-area transit bandwidth that costs more than building and maintaining datacenter internal networks. The expensive investments in this regard are justified by traffic trends recently estimated [7]: traffic between datacenters is expected to grow faster than either traffic to end-users or traffic within datacenters, with a 23-percent compound annual growing rate from 2013 to 2018. The rapid growth of this traffic is due to the proliferation of cloud services, the need to shuttle data between clouds, and the increasing volume of data that needs to be replicated across datacenters. The effects of this interesting trend can also be spotted in the scientific literature with novel solutions that leverage the high network performance offered by public-cloud inter-datacenter WANs [8], [9], [24]. This recent literature further extends the range of typical usages of public-cloud inter-datacenter networks, that include transfer of bulk-data or on-line contents (e.g., from and to storage buckets).

While the complexity of these wide-area inter-datacenter network infrastructures interconnecting geographically distributed datacenters is completely transparent to cloud customers, the performance achievable by final consumers is deeply affected by it and customers could significantly benefit from details about the Quality of the Service (QoS) offered or guaranteed [4]. Unfortunately, cloud providers often provide no more than qualitative information about the performance a customer should expect from the cloud network or its design, mainly due to security and commercial reasons [5], [17], [21]. In particular, very little information is available about the performance figures offered by public-cloud networks connecting datacenters placed in different geographic regions: the information provided by public-cloud monitoring services [1] currently does not include inter-datacenter performance, while, to the best of our knowledge, the scientific community did not focus on the problem yet, and the poor preliminary results cannot be considered exhaustive.

To fill this gap, we have experimentally evaluated the performance of the inter-datacenter network of the two leading cloud providers: *Amazon Web Services* and *Microsoft Azure* (hereafter simply referred to as Amazon and Azure) [15]. We did not rely on providers' support, adopting the point of view of a generic cloud customer, i.e. all our experimentations are based on *non-cooperative* methodologies [20]. In details, we have collected performance data about network paths interconnecting public-cloud datacenters, leveraging active monitoring approaches for more than 300 hours, taking into account a set of geographic regions hosting datacenters for both the providers. Our work depicts a clear picture of the inter-datacenter network performance in terms of throughput and latency for the two leading public-cloud providers mentioned above, also considering the impact of several configuration factors under customer control, such as the geographical region in which the datacenter is placed or the communication protocol adopted. In addition, we provide insights into the communication infrastructure leveraged by cloud providers, showing the existence of phenomena generated by their management strategies, which may impact both the performance experienced by customers and the results of research works investigating these networks. Compared with previous works, our results highlight the changes in terms of performance figures and analyze the trend that these infrastructures have been subjected to in the last years.

The paper is organized as follows: Sec. II positions the paper according to the related literature; Sec. III details the methodology adopted for the analysis and the collected dataset; Sec. IV shows the most interesting results of our work;

Sec. V ends the paper with the concluding remarks.

II. Related Work

Most of the works in the literature aimed at providing a broad characterization of the performance of public clouds, thus not directly focusing on network performance. Some of them evaluated intra-datacenter network performance—i.e. the performance of the network interconnecting cloud resources deployed within the same geographical site—with different purposes [21], [14], [13], [22], [23], [18], [19].

A very limited number of works took into account the performance of inter-datacenter networks. Chen *et. al* [6] performed a passive analysis on Yahoo! network flow dataset and focused on the interplay of multiple datacenters. Li *et al.* [16] also benchmarked cloud inter-datacenter networks, but only considering TCP throughput performance achievable between two US datacenters. They found that the throughput across datacenters is much smaller than the one within the datacenter (both Amazon and Azure showed median values larger than 200 Mbps and a higher variability with respect to the intra-datacenter performance). Feng *et al.* [9], [8] performed an experimental evaluation of Amazon network paths interconnecting seven different datacenters to support their study on a set of algorithms and protocols to minimize operational costs of inter-datacenter video traffic. They monitored network performance for only 3 minutes and revealed very different throughput values, ranging from 9.6 Mbps up to 545.1 Mbps. Values of end-to-end latency measured were lower than 587.3 ms for the 90% of the paths, while the average was about 349.1 ms [9]. Finally, Garcia-Dorado and Rao [11] conducted experimentations on both Amazon and Azure measuring TCP bandwidth and latency performing 2-minute-long measurements during one day in order to evaluate their overlay-network approach. They found low variation of throughput values, especially in paths exposing better performance.

Our study significantly differs from the others in recent literature dealing with the performance of the inter-datacenter networks. In spite of the analysis presented in [6], this work provides a more neutral point of view, not relying on provider-restricted information and being completely based on active measurements performed from the cloud customer angle. Differently from the analysis in [16], our work explicitly focuses on the performance of inter-datacenter networks, deepening the aspects strictly related to the measurement process. Thanks to this, we investigate also the interesting traffic engineering practices and their impact on both the measurement process and the QoS perceived by cloud customers. Finally, with respect to the measurement data presented in [8], [9], and [11], our work is more systematic, details a repeatable methodology, compares the performance of multiple providers, and takes into account the specific traffic management strategies they enforce.

III. Methodology and Dataset

In this section, we first detail the reference architecture,

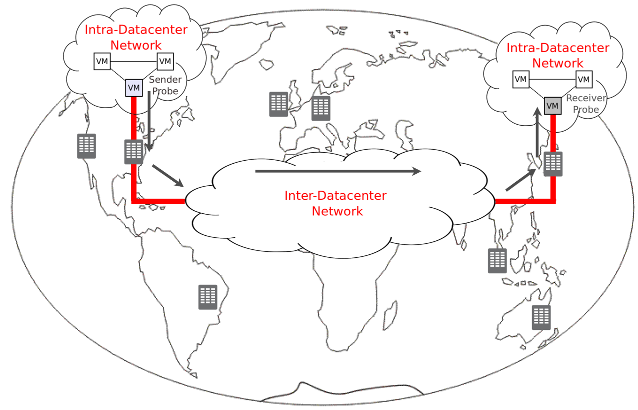


Fig. 1: Reference architecture. Before being delivered to the receiver VM, synthetic traffic traverses different layers, i.e. the intra-datacenter network at sender side, the inter-datacenter network, and the intra-datacenter network at receiver side.

the factors potentially impacting network performance, the settings, and the tools we have adopted, in order to unambiguously identify the scenarios considered. We then describe the dataset collected.

A. Methodology

In accordance with recent studies [15], a few global providers dominate the Infrastructure-as-a-Service (IaaS) market among a larger number of offers. In this work we take into account the IaaS offers of the two major providers: *EC2* for Amazon [2] and *virtual machines* for Azure [3]. The former is the clear market leader while the latter is the only clear challenger.

More specifically, our work aims at measuring the network performance of paths interconnecting virtual machines (or simply VMs) deployed in public-cloud datacenters geographically distributed. Looking at the reference architecture in Fig. 1, traffic directed from one side to the other of the communication traverses different, distinct layers. In particular, traffic generated by a VM normally traverses (i) the intra-datacenter, high-performance network at sender side. Then, it enters and traverses (ii) the inter-datacenter WAN, and it finally passes through (iii) the intra-datacenter network at receiver side, before being delivered to the receiver VM. Note that the internals of both intra- and inter-datacenter networks are out of our knowledge, as we adopt the point of view of the general customer. In fact, our approach is aimed at measuring the performance experimented by real customers' traffic.

In our work, the inter-datacenter WAN is assumed to be the bottleneck of the communication due to practical, technological, and physical limitations. Our results have therefore to be intended as related to these networks, if not stated otherwise. For both providers we took into consideration extra-large VMs. This is to avoid the limitations possibly imposed by the intra-datacenter network observed adopting smaller VM-sizes [18], [19]. This choice is also supported by preliminary experimentations conducted—not reported here—that have shown that

VM-size has a negligible impact in this context, differently from what found for intra-datacenter performance. Note that both providers provide details regarding RAM, CPU, and storage. However, much less or none information is reported about the network performance: Amazon only provides a qualitative description of the expected performance, whereas Azure provides no information.

We have picked a region per continent in order to ensure geographical diversity to our dataset. Specifically, we have identified the following regions hosting datacenters for both providers: Ireland (EU in the following), North Virginia (US), Sao Paulo (SA), and Singapore (AP). We have investigated the network performance of all the paths interconnecting these regions[1]. After selecting a region, Amazon customers can further choose an *availability zone* (i.e. which specific, independent, and isolated location inside the chosen region). In our study we have also taken the impact of the availability zones into account.

VMs have been instrumented with Ubuntu 14.04 operating system and the tools needed for measuring the network performance. We have used the network measurement tool named *nuttcp* [10] to inject synthetic traffic into the network as already done in previous works [18], [19]. Using nuttcp we have been able to measure the raw UDP and TCP throughput and latency.

B. Experimental Dataset

This analysis is based on data collected between March and November 2015. The collecting process required more than 300 hours of traffic generation. We considered the 12 combinations of the four regions selected for each provider and repeated 5-minute-long experiments in the same conditions, equally spaced in 24-hour intervals.

Our experimentations have been subjected to providers' fees, and according to their terms of service, inter-datacenter traffic is subjected to volume-based charging. Therefore the number of experimental runs was limited by budget constraints (especially for UDP because of the high volumes of traffic transferred). We publicly release the entire dataset, to foster further analyses and replication, and to support longitudinal studies[2].

IV. Inter-datacenter Network Performance

In this section we discuss the most interesting results of our analysis. Firstly, we provide an assessment of the performance of the network interconnecting geographically distributed cloud sites in terms of throughput and latency. Secondly, we deepen the analysis for interesting cases, providing insights into the communication infrastructures used by the providers for inter-datacenter communications.

[1]Hereafter, we will adopt the notation $A \rightarrow B$ to refer to the path from region A to region B. $A \leftrightarrow B$ will be used to refer to both directions.

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

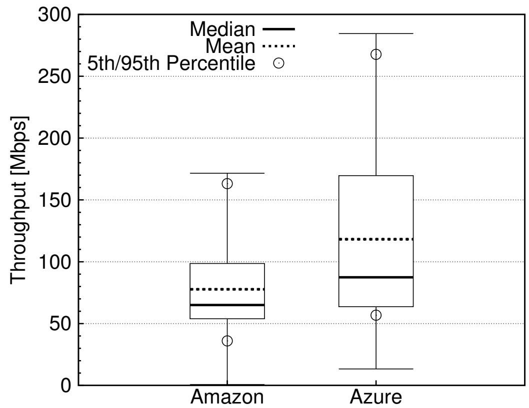


Fig. 2: TCP throughput distribution across different regions. Each sample represents the mean of a 5-minute-long experiment. Azure performs better on average (+52%).

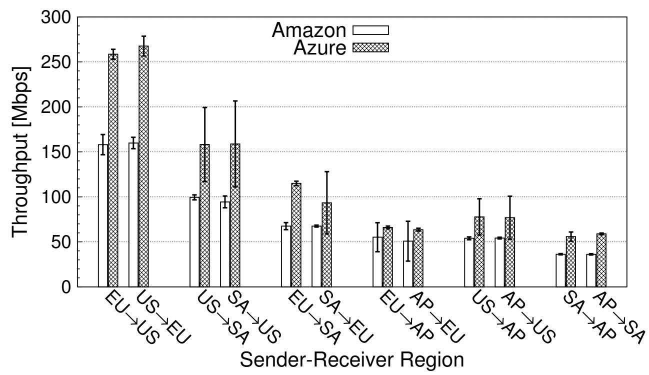


Fig. 3: TCP throughput breakdown on different region pairs (mean and standard deviation) for different providers.

A. Network throughput

Our analysis reveals that Azure inter-datacenter network performs better than Amazon's in terms of throughput. Fig. 2 reports an overall picture of the throughput for both providers in all the experiments. Each sample in the plot represents the mean of a 5-minute-long TCP experiment. It is worth noting that even values far from the global average of Fig. 2 well represent the samples collected during that particular 5-minute measurement. In fact, the *coefficient of variation*[3] (CoV) within each experiment is very low: the 95th percentile of its distribution along all the experiments is about 0.2. Azure achieves TCP throughput values 52% larger than Amazon (77.8 Mbps vs. 118.2 Mbps, on average). A slightly larger proportion (+65%) is obtained when considering the maximum value (284.5 Mbps vs. 171.6 Mbps). The interquartile range is smaller for Amazon, but throughput values as small as 1 Mbps have also been observed for this provider. Conversely, the throughput values for Azure are never smaller than 13 Mbps. Only 25% of samples collected for Amazon expose values larger than 99 Mbps, while 95% of samples

[3]$CoV(X) = \frac{\sigma}{\mu}$, where X is a set of experimental samples, σ is its standard deviation, and μ is its mean value.

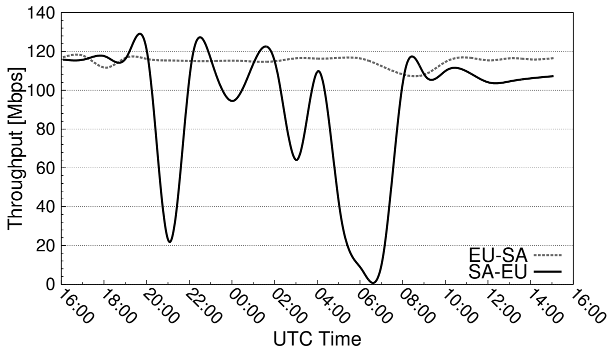


Fig. 4: An example of performance asymmetry for different directions (SA↔EU Azure).

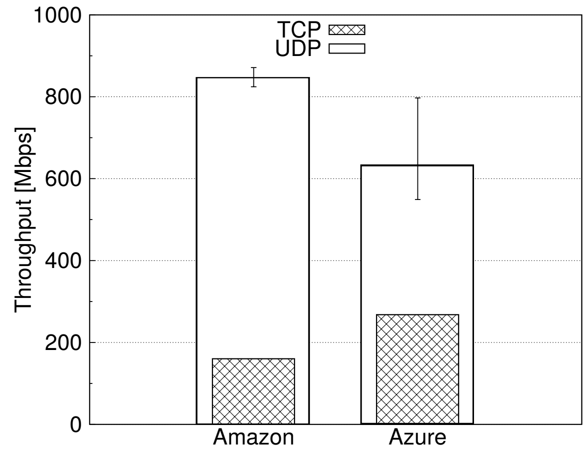


Fig. 5: TCP and UDP inter-datacenter average throughput for US→EU (whiskers report maximum and minimum). UDP is able to reach better end-to-end performance, giving evidence of path capacities as large as more than 800 Mbps.

collected on Azure's infrastructure have values larger than 57 Mbps. Fig. 3 provides a breakdown of the performance obtained with TCP. Mean and standard deviation across different regions are reported for the two providers. A significant difference of performance can be observed across different regions (up to about 80% in the worst case). Considering the ranking of the region pairs based on their related average throughput we obtain the same order for the two providers, with the only exception of US↔AP pair which performs better than EU↔AP pair for Azure. The variability within a region pair is normally very low, although some Azure region pairs show a larger standard deviation value (e.g., SA→US, US→SA, and SA→EU). Our results show that worst performance is typically related to two regions: AP and SA. On top of this, data transfer from AP and SA is subjected to higher costs with respect to EU and US regions. In fact, the increase of data-transfer expenses related to AP and SA amount to $8\times$ and $4.5\times$ for Amazon, and up to $3.2\times$ and $2.3\times$ for Azure, respectively. These two regions thus represent unfavorable choices for cloud customers. Performance figures appear to be roughly symmetric in the majority of the cases. However, we also encountered severe degradations involving only one direction of the communication. Fig. 4 reports an example for these interesting cases. As shown, intermittent but heavy performance degradations have been observed, with throughput values settling down to less than 10 Mbps.

The performance assessment presented above carries advantageous information to customers willing to draw upon public clouds to deploy their distributed architectures. In the following we provide a discussion of the obtained results comparing also with existing literature. Our analysis quantified the network performance discrepancy existing among regions and allows customers to wisely select among them. Also, comparing the two providers we found that Azure performs better on average, while asking higher costs for the VMs and for the data transfer. Due to this trade-off, the choice of the provider can be tailored according to the regions of interest and should also be driven by the specific characteristics of the application. Moreover, performance symmetry may also be taken into account, in order to properly place nodes in the different regions, also according to the specific application the inter-datacenter network is leveraged for and to the different roles of the counterparts involved in a communication.

In general, we observed TCP throughput values smaller than those reported in previous works. Authors of [16] reported TCP (median) throughput values larger than 200 Mbps for both Amazon and Azure. Unfortunately, their results are hard to interpret (and compare with) because no information is disclosed about experimental conditions. Moreover, the experimentation in [16] is restricted to pairs of datacenters placed in the same continent (US). Several interpretations are therefore possible for this discrepancy. For example, the different performance figures may be explained by the fact that experimentations in [16] involved datacenters separated by a smaller distance and hence backed by infrastructures implementing technologies having different performance. Another possible cause is related to the presence of less competing traffic across the inter-datacenter networks at the time when experimentations conducted in [16] were performed, i.e. around 5 years before ours [7]. Interestingly, the performance reduction observed after these 5 years is larger for Amazon than for Azure. It could be further justified by the impact of the larger number of customers Amazon has on TCP congestion-control dynamics. Finally, authors of [16] also found throughput variability markedly higher than the one we saw. This conclusion holds although the analysis in [16] refers to data collected over a more limited observation period (one single day) and is related to measurements between two datacenters both placed in the US. The reduced variability is in line with the increase of the competing traffic already hypothesized above.

UDP throughput values proved to be significantly larger than TCP ones, for all the source-destination pairs considered. We have also seen cases in which UDP inter-datacenter throughput durably reaches the intra-datacenter performance figures reported in [18] and [19]. In this case, therefore, the bottleneck is not the WAN but rather the limits imposed by

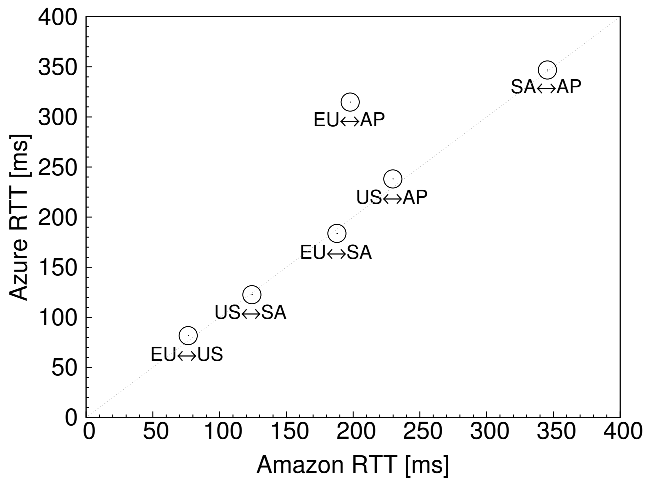


Fig. 6: Comparison of average inter-datacenter latencies experienced when relying on different providers. Latency across homologous pairs resulted to be comparable except than for EU↔AP.

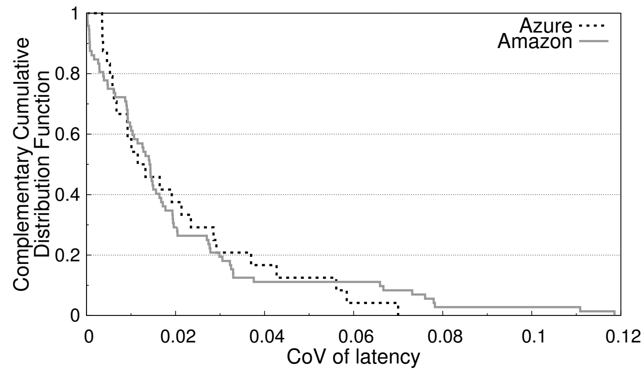


Fig. 7: CoV distribution of latency (RTT) across different 5-minute-long experiments. Both providers expose little variation for latency.

providers at source side [18], [19].

In detail, Fig. 5 compares UDP and TCP average throughput obtained between the pair of regions with the best performance for both providers: US and EU. UDP throughput reaches much larger values, with maximum values compatible with intra-datacenter limitations imposed by providers at source side [18], [19]. These results suggest that the worse TCP performance is determined by network congestion across datacenters. On the other hand, the better performance of UDP gives evidence of the network capacity of the inter-datacenter paths. We can find further justifications for this empirical result considering the impact of the higher number of customers Amazon has on TCP congestion control dynamics. Network congestion represents the main bottleneck when relying on TCP, although VMs are allowed to inject traffic into the inter-datacenter network at a very high rate and the inter-datacenter network is able to deliver traffic at such high speed. This result is generalizable across different regions, even if actual UDP throughput values change from case to case.

B. Network latency

Experimental data shows that latency and throughput are in general not highly correlated. While high throughput generally implies lower latency, low throughput does not necessarily imply high latency. Therefore observed throughput degradations (reported above) are not associated to latency increase.

As expected, latency values appear to be symmetric, in spite of non-negligible differences across different regions. The region with the smallest average latency towards the others is US, whereas AP exposes way larger values.

The mean latency values between homologous regions for Amazon and Azure are compared in Fig. 6. The Average latency is equal to 193.94 ms and 214.61 ms for Amazon and Azure, respectively. While experimented RTT is similar across providers for five out of six region pairs, EU↔AP shows a markedly higher latency for Azure (201 ms vs. 315 ms). The latency measured for EU↔AP for Azure is almost equal to the sum of the latencies measured for EU↔US and US↔AP. This result suggests that Azure traffic management policies may route traffic from EU to AP through US, thus inflating the length of the path and the perceived latency.

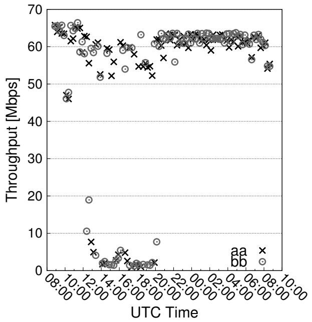
Limited latency variability has been observed for both providers. Fig. 7 shows the complementary cumulative distribution of the CoV of the RTT for different experiments. For both providers we observed CoV values higher than 0.05 for less than 15% of the 5-minute-long experiments. Interestingly, the worst-performing region pair in terms of latency (i.e. SA↔AP) has the stablest performance over time.

C. Impact of the availability zone

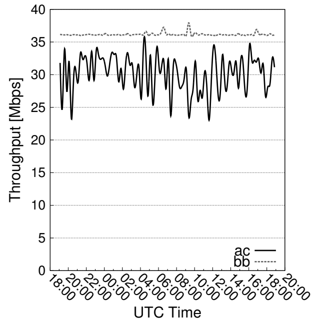
In our study we have also taken into account the impact of selecting different availability zones (AZs) inside a region, i.e. we investigated how performance varies when choosing one of the isolated locations made available inside a region by Amazon. The main outcome of this analysis is that the AZ does not clearly impact the achievable throughput. In general, the *coefficient of variation of the root mean square error*[4] (CV_{RMSE}), which gives indication about the difference of throughput performance perceived along paths between different AZs, has been lower than 0.2 for 90% of the samples.

In a limited number of cases however, severe performance degradations lasting for several hours have been identified, where throughput dropped down to values smaller than 5 Mbps. An example for AP→EU is reported in Fig. 8a. Pairs of homologous samples report different throughput values for different AZs. However, in the period between 14:00 and 20:00 all the tested logical paths connecting disjoint AZ sets (namely `aa` and `bb`) show a severe degradation of performance. This example shows how AZs although guaranteeing site isolation, revealed to be not completely independent from the network point of view. We have observed a few cases similar to the one described. On the other hand, we have also observed cases in which the degradation involves only one AZ pair and not

[4] $CV_{RMSE}(X,Y) = \frac{\sqrt{E[(X-Y)^2]}}{E[E[X],E[Y]]}$ where X and Y are the empirical distributions of the throughput values collected considering two distinct pairs of availability zones.

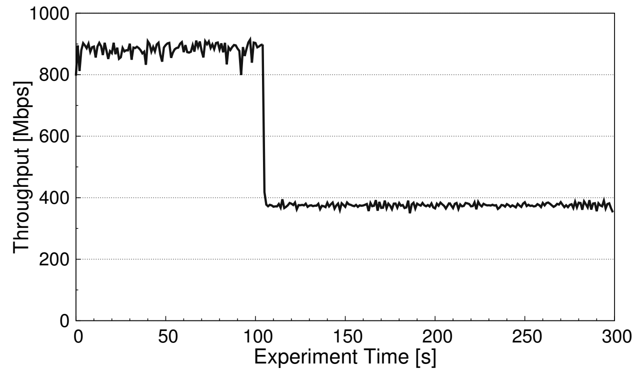


(a) Performance degradation involving all the AZs (AP→EU).

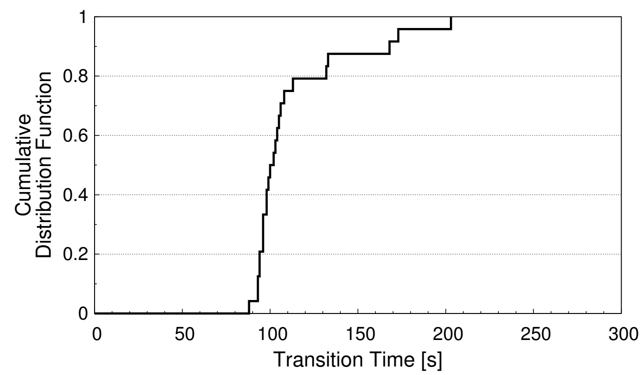


(b) AZs having consistently worse performance (AP→SA).

Fig. 8: Examples for the interesting cases.



(a) Typical instantaneous throughput evolution of a 5-minute-long experiment.



(b) Distribution of the high-to-low transition time.

Fig. 9: Azure, US→EU. UDP throughput within 5-minute-long experiments typically switches from a high to a low stable value of 400 Mbps (a). The transition typically happens around 100 s (b). The transferred traffic volume ranges from $8,000$ to $10,000$ MB for 75% of the cases.

the others. This happened for a single pair of regions in our dataset (AP→SA). Fig. 8b shows how the throughput values for the AZ pair identified by `ac` are consistently smaller than the ones of the homologous AZ identified by `bb` during the entire 24-hour-long observation period.

Extending latency considerations to Amazon AZs, we can point out some other interesting patterns: different AZ pairs present consistently but slightly distinct latency values. This latency information is useful to cloud customers to identify the actual AZ assigned inside a region and can be leveraged to set up resources into the most convenient AZ, according to potentially existing performance discrepancies found.

D. *Impact of traffic management policies enforced*

Providers may impose restrictions along the path, at one of the several layers traversed by the traffic, none of which is under the direct control of the cloud customer. Our experimentations revealed some interesting cases in which the measured throughput is not stable over time. These cases are particularly evident at the high rates attainable with UDP protocol. These phenomena inflate the variance of the throughput, as also reported by the larger error bars in Fig. 5.

An example of this phenomenon regards the paths interconnecting Azure VMs between US and EU regions, where the variability is due to performance variation *within* each of the 5-minute-long experiments. Experimental evidences are in shown Fig. 9. In detail, Fig. 9a shows how the throughput typically varies within each experiment: interestingly, two well-defined throughput values can be easily identified. All the experiments between these two regions show the same pattern: the throughput dramatically switches from a high value (around 850 Mbps) to a low one (around 400 Mbps). Also the stability of the throughput samples significantly changes, as the CoV is almost halved after the transition. It is worth noting how the dramatic throughput variation described above happens when a communication is active. Since the high-to-low transition may happen at differing points in time for different experiments, different average values have been observed, which generate the larger variability range in Fig. 5. The distribution of the transition time is in shown Fig. 9b. Interestingly, in more than 75% of the cases, the transition happens at around 100 s, thus exposing a certain deterministic behavior.

This empirical result gives evidence of mechanisms that restrict the maximum capacity available to a customer possibly based on the traffic volume previously generated. The volume of the transferred traffic before the transition ranges from $8,000$ to $10,000$ MB for about 75% of the cases. No variation of the actual traffic injected in the network by the sender VMs has been identified for the cases discussed above. Therefore, the differing performance levels identified are not caused by the traffic generation capabilities of the VMs. Accordingly, lower throughput values are reflected by a proportionally higher packet loss. This further suggests that the observed phenomenon is the consequence of traffic management policies enforced by providers along the paths that interconnect one region to the other. It is worth noting that this phenomenon may impact both the results of the measurement process and the user experience: (i) measurements shorter than 100 s can not spot the throughput transition; (ii) longer measurements could lead to misinterpret performance variability if not associated to a deeper analysis; (iii) the dramatic throughput drop observed (more than -50% in the example proposed) heavily impacts the perceived QoS, causing non-negligible troubles to customers.

V. Conclusion

Cloud inter-datacenter network performance is gaining more and more interest. In this paper we provided an assessment of these networks for the two leading providers (Amazon and Azure) in terms of throughput and latency, deepening the impact of both specific choices made by the customers and management policies implemented by providers.

We found that Azure performs better in terms of maximum achievable throughput ($+52\%$, on average), possibly because of the smaller number of users and consequently the more limited cross traffic. Counterintuitively, regions with worse performance are associated to higher costs for customers. Network latency is comparable for the two providers when considering homologous regions, with the remarkable exception of the path interconnecting AP to EU exposing higher latency for Azure. Finally, the traffic management policies are also enforced by providers. We have studied and characterized them and observed how they can heavily impact both the QoS experienced by cloud customers and the results of the measurement results.

Acknowledgments

This work is partially funded by art. 11 DM 593/2000 for NM2 srl (Italy). The experimental work in this paper was also supported by a grant provided by Amazon AWS in Education.

References

[1] CloudHarmony. https://cloudharmony.com/.

[2] Amazon Web Services website. http://aws.amazon.com/, Apr. 2016.

[3] Microsoft Azure website. http://azure.microsoft.com/, Apr. 2016.

[4] G. Aceto, A. Botta, W. de Donato, and A. Pescape. Cloud monitoring: Definitions, issues and future directions. In *Cloud Networking (CLOUDNET), 2012 IEEE 1st International Conference on*, pages 63–67, Nov 2012.

[5] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron. Towards predictable datacenter networks. In *Proceedings of the ACM SIGCOMM 2011 Conference*, SIGCOMM '11, pages 242–253, New York, NY, USA, 2011. ACM.

[6] Y. Chen, S. Jain, V. Adhikari, Z.-L. Zhang, and K. Xu. A first look at inter-data center traffic characteristics via yahoo! datasets. In *INFOCOM, 2011 Proceedings IEEE*, pages 1620–1628, April 2011.

[7] Cisco. Cisco global cloud index: Forecast and methodology, 2013-2018. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html, 2013.

[8] Y. Feng, B. Li, and B. Li. Airlift: Video conferencing as a cloud service using inter-datacenter networks. In *Network Protocols (ICNP), 2012 20th IEEE International Conference on*, pages 1–11. IEEE, 2012.

[9] Y. Feng, B. Li, and B. Li. Jetway: minimizing costs on inter-datacenter video traffic. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 259–268. ACM, 2012.

[10] B. Fink and R. Scott. nuttcp, v6.1.2.

[11] J. Garcia-Dorado and S. Rao. Cost-aware Multi Data-Center Bulk Transfers in the Cloud from a Customer-Side Perspective. *IEEE Transactions on Cloud Computing*, 7161(c):1–1, 2015.

[12] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel. The cost of a cloud: research problems in data center networks. *SIGCOMM Comput. Commun. Rev.*, 39(1):68–73, 2008.

[13] Z. Hill, J. Li, M. Mao, A. Ruiz-Alvarez, and M. Humphrey. Early observations on the performance of windows azure. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pages 367–376. ACM, 2010.

[14] K. LaCurts, S. Deng, A. Goyal, and H. Balakrishnan. Choreo: Network-aware task placement for cloud applications. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 191–204, New York, NY, USA, 2013. ACM.

[15] L. Leong, D. Toombs, and B. Gill. Magic quadrant for cloud infrastructure as a service, worldwide. http://www.gartner.com/technology/reprints.do?id=1-2G2O5FC&ct=150519&st=sb, 2015.

[16] 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. ACM.

[17] J. C. Mogul and L. Popa. What we talk about when we talk about cloud network performance. *SIGCOMM Comput. Commun. Rev.*, 42(5):44–48, Sept. 2012.

[18] V. Persico, P. Marchetta, A. Botta, and A. Pescapé. Measuring network throughput in the cloud: the case of amazon ec2. *Elsevier, Computer Networks, Special Issue on Cloud Networking and Communications II.*

[19] V. Persico, P. Marchetta, A. Botta, and A. Pescapé. On network throughput variability in microsoft azure cloud. In *GLOBECOM, 2015 Proceedings IEEE*, 2015.

[20] V. Persico, A. Montieri, and A. Pescapé. CloudSurf: a platform for monitoring public-cloud networks. In *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI) (IEEE RTSI 2016)*, Bologna, Italy, Sept. 2016.

[21] C. Raiciu, M. Ionescu, and D. Niculescu. Opening up black box networks with cloudtalk. In *Proc. 4th USENIX Conference on Hot Topics in Cloud Computing*, page 6, 2012.

[22] J. Schad, J. Dittrich, and J.-A. Quiané-Ruiz. Runtime measurements in the cloud: Observing, analyzing, and reducing variance. *Proc. VLDB Endow.*, 3(1-2):460–471, Sept. 2010.

[23] R. Tudoran, A. Costan, G. Antoniu, and L. Bougé. A performance evaluation of azure and nimbus clouds for scientific applications. In *Proceedings of the 2nd International Workshop on Cloud Computing Platforms*, page 4. ACM, 2012.

[24] F. Wang, J. Liu, and M. Chen. Calms: Cloud-assisted live media streaming for globalized demands with time/region diversities. In *INFOCOM, 2012 Proceedings IEEE*, pages 199–207. IEEE, 2012.