Inferring the Deployment of Top Domains over Public Clouds using DNS Data

Quentin Jacquemart
CNRS, I3S, Université Côte d'Azur
quentin.jacquemart@unice.fr

Clément Pigout
Université Côte d'Azur, I3S
clement.pigout@etu.unice.fr

Guillaume Urvoy-Keller
Université Côte d'Azur, I3S
urvoy@i3s.unice.fr

***Abstract*—Cloud technologies are becoming pervasive and available for private companies or public institutions in different flavors, mostly public cloud or private clouds. Our focus in this work is on the usage of public cloud technologies by the most popular sites in the Internet. While some studies have described the nascent landscape of public cloud computing 5 years ago, surprising little effort has been put to study the recent evolution of this domain.**

Using DNS data that enables us to map domains (e.g., netflix.com) and their subdomains (e.g., api.netflix.com) with the cloud providers actually used, we refresh our understanding of this ecosystem. We focus on the dominant four cloud providers, namely Amazon Web Services, Microsoft Azure, Google Cloud Computing and IBM Bluemix. We demonstrate that cloud penetration has clearly increased since 2013, reaching almost 50% of the top 1000 domains, from the Alexa list. Furthermore, a significant fraction of domains use multiple cloud providers simultaneously. Still, domain owenrs remain cautious when it comes to choose which subdomain is actually hosted in the cloud and only 17.8% on average of the subdomains are actually hosted in the cloud. In terms of performance, preliminary results indicate that hosting a subdomain in the cloud pays off as compared to private hosting with a decrease of application level latency of 28%.

I. Introduction

The public cloud market is dominated by a few players. Amazon Web Services is known to hold the leading position with over 34% of market share, followed by Microsoft Azure with 11%, and Google Cloud Computing with 5% [1], [2]. IBM Bluemix announces that it accounts for 8% of the market, with an activity more focused on private cloud services. The portfolio of services offered by public cloud providers is increasing at a steady pace. Amazon Web Services itself offers over 90 services including computing, storage, networking, database, analytics, mobile, developer tools, and tools for the Internet of Things to name a few.

In this work, we focus on the usage of those dominant public cloud providers by the most popular sites in the Internet. While it only reflects a specific type of activity ran over these public clouds, the actual usage of public cloud technologies can have a direct impact on the end user. This impact can be positive, e.g. with a better Quality of Experience (QoE), perceived thanks to the high availability and adaptivity-to-the-demand enabled by cloud technologies. It can also be negative in case of large scale outages or attacks, possibly affecting a large fraction of sites at once due to the higher centralization and higher inter-dependencies among cloud services [3].

A number of studies have tackled these issues in 2013 and 2014 through active and passive measurements [4], [5]. Surprisingly, little efforts have been put since then by the community to assess how the picture has evolved. In the present work, we mine DNS records to uncover the extent to which popular websites are deployed over the public cloud, and their modes of deployment.

Our findings are as follows.

- The usage of public cloud services has increased from below a percent in 2013 to close to 10% in 2018 when the considering the 1M top sites. (Section III-A).
- There does not appear to be an exclusive approach to cloud deployment. While a small set of subdomains is hosted in the cloud, the majority remains hosted outside the cloud. This is notably the case for the website front end (either `domain.com` or `www.domain.com`) (subsections III-A and III-B).
- Multi-cloud strategies are observed in about 20% of cases for the 1k most popular domains, where the company hosts some of its subdomains in different cloud providers (Section III-B).
- We adopt a European-centric approach to study service latency, and observe that it pays off to host services in the cloud. Results however depend on the exact cloud provider, due to the location of its data center. However, the effect of geographical distance can be dampened by other factors, such as the number of peering ASes with the data center.

II. Methodology and Datasets

We purposely follow a methodology close to the one used by He et al. [4] – the first paper seeking to understand the use of the public cloud infrastructure by the most popular web domains – to allow an easy comparison of our works' results, five years apart. Obviously, during those five years, the IaaS landscape evolved. For this reason, not only do we focus on the deployment of web domains on Amazon Web Services and Microsoft Azure like in [4], but also include Google Cloud Computing and IBM Bluemix. We further attempted to include VMWare vAir, but failed to secure any authoritative information to help us isolate their infrastructure. These represent the dominant public IaaS vendors, with Amazon Web Services known to be the largest platform [6], a trend we also observe.

978-3-903176-17-1 / © 2019 IFIP

A. Domain Popularity Lists

	Alexa vs. Majestic	Alexa vs. Umbrella	Majestic vs. Umbrella
Top 100	33	19	34
Top 500	147	60	73
Top 1000	302	147	162
Top 1M	235,502	138,498	121,512

TABLE I
NUMBER OF COMMON DOMAINS AMONG THE ALEXA 1M, MAJESTIC MILLION, AND UMBRELLA LISTS

There are multiple rankings to the "most popular" web domains: The Majestic Million list [7], the Cisco Umbrella list [8] and, of course, the Alexa list, the most used list in academic research [9], and used by [4]. It is difficult to assess which list reflects the true popularity, as the methodology to construct them is secret [9]. Moreover, the popularity of a site also depends on the considered part of the globe.

We compared the contents of these lists as of April 13th, 2018. The Umbrella list is the only one to include both domains (e.g. netflix.com) and subdomains (e.g. api.netflix.com), while the two others only list domains. We show the intersection of the lists at different levels of popularity in Table I. The Alexa 1M and the Majestic Million lists have the largest overlap. However, even when the two lists agree on the fact that a domain should be part of the n-th top, the actual rank might differ a lot, as illustrated in Figure 1.

We eventually ruled out the Umbrella list that significantly differs from the two other ones and opted for the Alexa top 1M [10], so as to keep close to [4]. Between the 2013 Alexa 1M list used in [4] and the 2018 version, there are 64 common domains in the top 100, 460 in the top 1k, and almost 230k in the top 1M.

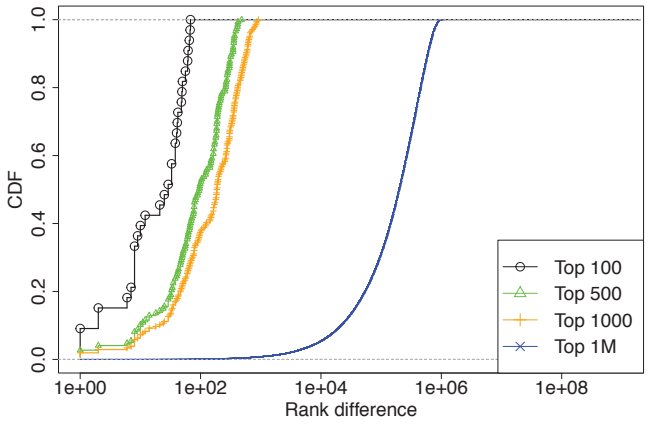

Fig. 1. Difference of rank between the Alexa 1M and Majestic Million lists

Beyond the domain front-end itself, it is also important to look at the deployment of subdomains, because, for strategic or technical reasons, a domain owner may decide to deploy a set of subdomains in the cloud, or on different clouds. For this reason, we create a list of most-likely subdomains relying on the *Knock* word-list [11], which contains around 2k prefixes. We prepend these prefixes to the Alexa list in order to create the top list of domains and subdomains. Please note that, unlike with the original (2013) Alexa list, this extended version does not contain the complete list of most popular subdomains.

B. IP Ranges of Public Clouds

We say that a (sub)domain d is *hosted* in the cloud if, during the DNS resolution for d, we end up with an IP address that belongs to the IP ranges of Amazon Web Services [12], Microsoft Azure [13], Google Cloud Computing [14], or IBM [15]. We once more emphasize that these represent over 90% of the public cloud market [6].

C. Methodology

Our study is carried out in two phases. In the first phase, we consider all the domains listed in Alexa top-1M, to assess the prevalence of cloud deployment on front-end domains. In the second phase, we consider the domains for the Alexa top 1k list, as well as their subdomains. As we will show in Section III-A, the cloud usage of the first top 1k, is representative, motivating our decision.

We rely on `dig` to carry out the DNS resolutions of the 1M domains and of the 2M potential subdomains. Based on these DNS queries results, we filter out the subdomains that cannot be resolved. We ended up with 473k valid subdomains for the Alexa top 1k domains. Out of these, 54k subdomains that are hosted in our list of public cloud providers. Furthermore, only 3 subdomains are hosted on IBM Bluemix. This lead us to discard them from any further analysis.

We also distinguish between the types of services offered by each cloud providers, and follow the taxonomy introduced in [4]. In particular, we distinguish between:

- a *IaaS back-end*, where the domain owner directly manages VMs;
- *load balancers*, where the (sub)domain relies on a load balancing services offered by the cloud provider;
- a *PaaS back-end*, where the (sub)domain owner pushes its content/code to a service offered by a cloud provider or a third party (e.g. Heroku in the case of AWS)
- *CDN*, where the (sub)-domain owner relies on content-delivery networks offered by cloud providers.

We note that these services are not offered by all cloud providers. Their identification can be done either at the IP level, if the cloud provider is known to separate specific services in distinct IP ranges (e.g. AWS), or based on string identification in the (intermediate) CNAMEs resulting from DNS queries. We report in Table II the method used to identify the services we consider. This methodology is derived from [4], and has been updated and extended to match today's behaviors.

The PaaS category only represents around 2k cloud-based (sub)domains, which corresponds to only 3.5%. Compared to the 28% isolated in [4], this is a steady decrease. This result

	IaaS	LB	PaaS	CDN
AWS	IP	If CNAME contains `elb.amazonaws.com`	If CNAME contains `elasticbeanstalk` or `heroku.com` or `herokuapp` or `herokudns` or `herokussl`	CloudFront IP range
Azure	If CNAME contains `cloudapp.net`			If CNAME contains `azureedge.net`
Google Cloud	IP		If CNAME contains `appspot.com`	If CNAME contains `googleapis.com`

TABLE II
IDENTIFICATION METHODS FOR CLOUD SERVICES

can be explained by the larger dataset used by [4], or by the fact that CNAMEs may not be as discriminative as in 2013. We decided to discard them from our further analyses.

III. RESULTS

A. Top 1M domains

Out of the one million domains included in the Alexa list, 967k correctly resolved through a DNS query. Table III displays a breakdown of these results in terms of cloud providers, and also considering different levels of popularity. As Amazon Web Services enables to detect the usage of its CDN through the analysis of the IP address, we also report in this table the number of CloudFront (CF) IPs.

We make two key observations. First, the overall adoption is close to 10%, which is way higher than what was observed in [4] (below 1%). Second, this value of 10% remains quite constant when considering different ranges of popularity (first 1k, 10k, etc), with a peak at 13% for the top 1k domains.

The 967k resolved domains map to 1.2M distinct IP addresses, because some domains are deployed over more than one server, letting the local DNS resolver perform load balancing over these addresses. However, there are also cases where a single IP address is shared by several domains. This is typical for CloudFront as can be seen in Figure 2 where

	Amazon		Azure	Google	Other
Rank	EC2	CloudFront			
100	3 (3%)	0 (0%)	4 (4%)	0 (0%)	93 (93%)
1,000	88 (8.8%)	22 (2.2%)	7 (0.7%)	11 (1.1%)	872 (87.2%)
10,000	916 (9.16%)	209 (2.09%)	95 (0.95%)	123 (1.23%)	8,661 (86.61%)
100,000	8,281 (8.28%)	1,364 (1.36%)	1,089 (1.08%)	1,319 (1.31%)	87,987 (87.98%)
415,275	63,487 (15.28%)	13,475 (3.24%)	2,107 (0.50%)	124 (0.03%)	336,082 (80.92%)
947,506	55,748 (5.75%)	5,070 (0.52%)	9,804 (1.01%)	18,371 (1.89%)	879,397 (90.85%)

TABLE III
BREAKDOWN OF DOMAINS PER CLOUD PROVIDERS AND POPULARITY

a single CloudFront CDN server can serve, on average, 9.5 (sub)domains.

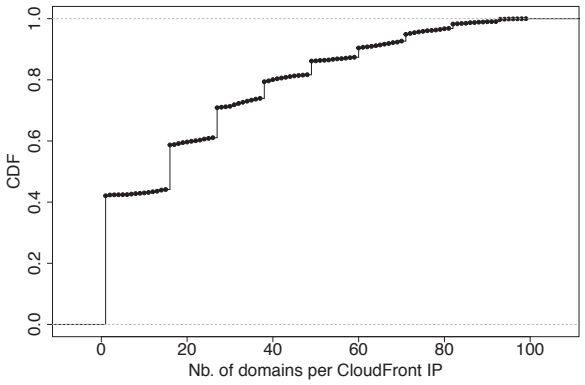

Fig. 2. Number of domains per CloudFront IP

We observed only a negligible number of cases where a domain is hosted on multiple clouds, i.e. resolved to IPs belonging to distinct cloud providers. When a domain has multiple IPs and is hosted in the cloud, it usually implies that some addresses correspond to a single cloud provider, while the others are to be part of the domain's own network. This is an apparent contradiction with [6], since, according to this study, multi-cloud deployments are stated as "popular". However, this can be explained by the fact that we only consider front-end domains here. The picture will become clearer when further considering subdomains (Section III-B).

We next turn our attention to the physical data center (DC) that hosts a specific service. In the case of Amazon Web Services and Microsoft Azure, the observed IP address can be mapped to an individual DC. For Amazon Web Services (Table IV), we observe that over 40% of the domains are served by a single data center located in the USA. This should be put in perspective with the 80% of traffic served by AWS data centers in the US, as observed by [5]. While the two figures are expressed in different units (domains vs. bytes), they indicate that the bias observed in 2014 is still present in 2018. Unlike [5], we observe that the Irish AWS data center accounts for about 20% of domains. The picture for Microsoft Azure (Table V) is different, with the majority of domains being hosted in a European data center. Please note that the difference in the total number of domains between Table IV, Table V and Table III are due do domains hosted on different data centers of the same cloud provider.

B. Top 1k Domains and Their Subdomains

We now focus on the Alexa 1K list, alongside with their generated subdomains (see Section II-A). The 473k resolved (sub)domains map to 947k IP addresses. For some domains, all queried subdomains do exist, which may seem odd given that some subdomains have no relation with the domain (e.g. `msn.com` and `counterstrike.msn.com`). In fact, we discovered that 208 domains rely on wildcard DNS records,

Region	Location	#Domains	#IPs
us-east-1	North Virginia (US)	23,520	39,597
us-east-2	Ohio (US)	766	970
us-west-1	California (US)	1,886	2,365
us-west-2	Oregon, (US)	6,629	9,151
us-gov-west-1	USA	21	32
ap-northeast-1	Tokyo (JP)	2,052	3,089
ap-northeast-2	Seoul South (KR)	357	542
ap-south-1	Mumbai (IN)	978	1,216
ap-southeast-1	Singapore (SG)	2,170	2,780
ap-southeast-2	Sydney (AU)	1,818	2,412
ca-central-1	Montreal (CA)	124	158
cn-north-1	Beijing (CN)	80	110
cn-northwest-1	Ningxia (CN)	1	1
eu-central-1	Frankfurt (DE)	2,837	3,795
eu-west-1	Ireland	10,906	15,938
eu-west-2	London (UK)	495	620
eu-west-3	Paris (FR)	574	597
sa-east-1	Sao Paulo (BR)	1,126	1,428
	Total	56,340	84,801

TABLE IV
NUMBER OF IPS AND DOMAINS IN EACH REGION FOR AMAZON WEB SERVICES

Region	Location	#Domains	#IPs
Central US	Iowa (US)	465	466
East US	Virginia (US)	1,564	1,567
East US 2	Virginia (US)	462	462
US Gov Iowa	Iowa (US)	0	0
US Gov Virginia	Virginia (US)	0	0
N. Central US	Illinois (US)	344	466
S. Central US	Texas (US)	695	695
W. Central US	Wyoming (US)	16	16
West US	California (US)	924	927
West US 2	Washington (US)	82	82
Canada East	Quebec (CA)	63	63
Canada Central	Toronto (CA)	91	91
Brazil South	Sao Paulo (BR)	228	230
North Europe	Ireland	1,331	1,379
West Europe	Netherlands	2,380	2,387
France Central	Paris (FR)	5	5
UK West	Cardiff (UK)	65	65
UK South	London (UK)	181	181
Southeast Asia	Singapore (SG)	326	328
East Asia	Hong Kong	223	224
Australia East	New S. Wales (AU)	184	184
Australia SE.	Victoria (AU)	110	111
Central India	Pune (IN)	109	110
West India	Mumbai (IN)	9	11
South India	Chennai (IN)	83	86
Japan West	Osaka (JP) Japan	33	33
Japan East	Tokyo (JP) Japan	49	49
Korea Central	Seoul (KR)	10	10
Korea South	Busan (KR)	8	8
	Total	10,040	10,236

TABLE V
NUMBER OF IPS AND DOMAINS IN EACH REGION FOR MICROSOFT AZURE

leading to a positive query response to any subdomain, and mapping to the same end IP address. As an example, for Reddit, the wildcard record points to the front-end webserver, which redirects a visiting browser to the subreddit (i.e. forum) of the same name as the prefix. Overall, these 208 domains account for 398k subdomains. Out of the 947k resolved IP addresses, only 21k are unique. This is a consequence of both the 208 domains mentioned before and also the use of CDNs that map several services by IP (see Figure 4).

To better quantify the relationship between subdomains and IP addresses (that should correspond to the number of VMs), we report in Figure 3 the number of subdomains per domain and the corresponding number of IPs. If one discounts the domains that use the wildcard technique, the coefficient of correlation between number of IPs and subdomains is 84%, meaning that the more subdomains a domain has, the more IP addresses it uses.

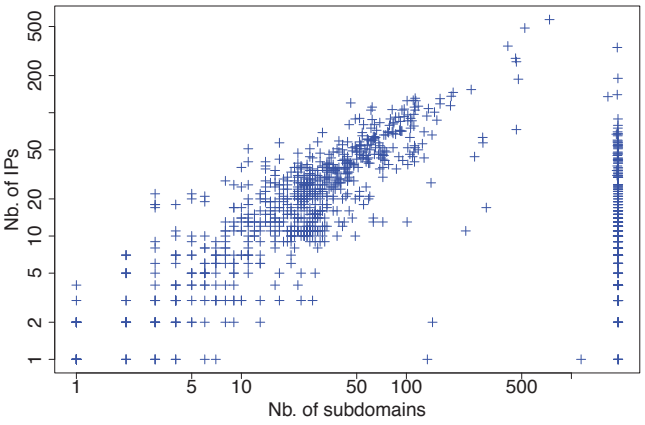

Fig. 3. Number of subdomains and IPs per domains

Out of the 473k (sub)domains, 54k are hosted on the cloud, mapping to 3k distinct IP addresses. Here again, wildcard DNS record are the culprit behind the order of magnitude difference. The 54k cloud-using subdomains correspond to 508 domains. This means that 50% of the top 1k domains rely on cloud deployment. In other words, 50% of the top 1k domains deploy *at least* one subdomain in the cloud.

The breakdown per provider is the following: 381 domains use EC2, 47 domains use Azure, and 80 use Google. Furthermore, almost 20% of them are multi-cloud: 28 use Azure and EC2; 56 Google and EC2; 7 use Google and Azure; and 6 use Azure, Google, and EC2. This is a significant increase when compared to the 0.7% of mixed EC2-Azure deployments observed in 2013 [4] for the 1 million top sites. This is also inline with the study in [6], which states that multi-cloud hosting is popular, now that we do not consider domains (as in Section III-A) but also subdomains.

The most popular prefixes of subdomains hosted in the cloud are: `api` (78); `www` (71); `support` and `blog` (67); `help` (65); `developer` and `b` (60); `c` (57); `a` (55); `m` (51). Consequently the website of the domain (`www`) is not the most likely to be hosted in the cloud. This is in line with our previous observations laid out in Table III, as, due to the popularity of the web, the `www` subdomain is often an alias of the domain name (without any subdomain).

Possible reasons behind this strategy can be that the domain owners (i) prefer to host their front-end website on their own

infrastructure (e.g. for better control), or (ii) delegate content to a CDN provider, e.g. Akamai. While discovering the reasons behind this strategy is out of the scope of this work, the key takeaway here is that domains owners follow a mixed strategy, with a part of their subdomains hosted in the cloud, and another part elsewhere.

We quantify this trend through Figure 4, which plots the number of subdomains in the cloud against the number of subdomains elsewhere for each domain. The domains relying on DNS wildcards are visible as points in the high values on either axis. The top dots on the y-axis imply that all subdomains are not hosted in the cloud, and the dots on the right side of the x-axis imply that these subdomains are indeed in the cloud. In both cases, the number of mapped IP addresses is very small, often a single one. Moreover, we observe that domains tend to have in general more subdomains out of the cloud than in the cloud. If we exclude the domains using DNS wildcards, we obtain, on average, 34.7 subdomains outside the cloud, and 8.13 subdomains in the cloud. This is also illustrated by Figure 5, where we report the ratio of subdomains in the cloud for the domains that have a presence in the cloud. We observe that the median stands below 20%, while only a handful of domains have all their subdomains in the cloud.

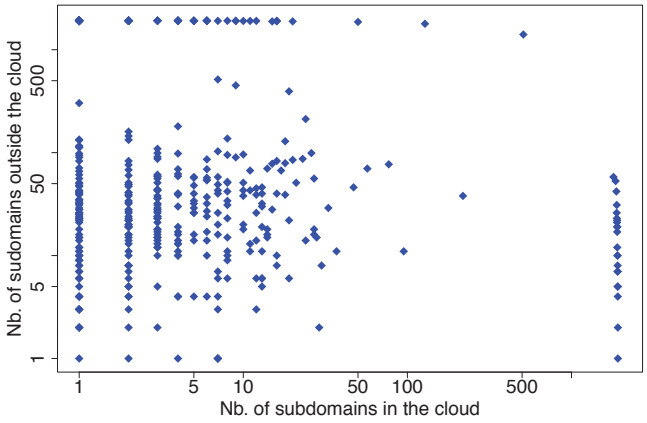

Fig. 4. Number of subdomains in/outside the cloud for each domain

C. Analysis of Top 10 Sites

We present in Tables VI, VII, VIII the top 10 customers of, respectively, Amazon Web Services, Microsoft Azure and Google Cloud Computing ranked according to their Alexa popularity. Regarding Amazon Web Services, each site relies heavily on the IaaS service (EC2). Only GitHub uses the PaaS service, which in line with its business. When the Elastic Load Balancing (ELB) service is used, e.g. by Netflix, the ELB CNAMEs map to way more ELB IP addresses as the client may choose how many front end load balancers it wants for a given service (subdomain).

Similarly, on Microsoft Azure, we observe a significant use of the IaaS service per client. MSN offers a clear example

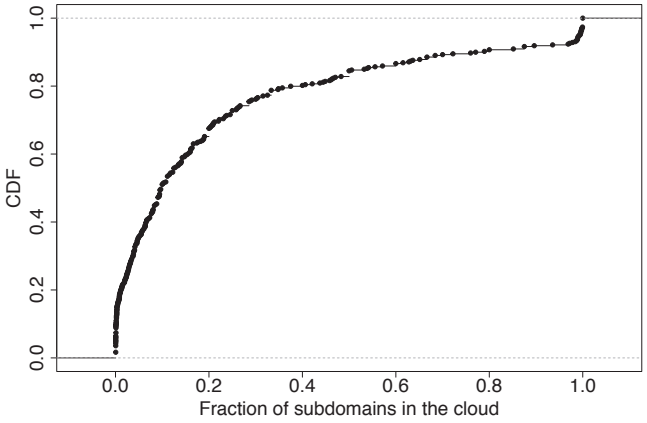

Fig. 5. Fraction of subdomains in the cloud for each domain with a presence in the cloud

of a domain configured with a wildcard DNS record, mapped behind the scene to a single IP address. Finally, on Google Cloud Computing, the deployment appears to entail far less observable front-end VMs as compared to Amazon Web Services and Microsoft Azure. Moreover, less subdomains are hosted on Google Cloud Computing per domain, as compared to Amazon Web Services and Microsoft Azure. A clear exception here is Spotify.

D. Application-level Performance

We now focus on the response time of the servers for all (sub)domains that we discovered to be in the cloud. To this end, we use `httping` to connect to each of the (sub)domains on port 80, and send an HTTP request. If a web server is on the other end, it will reply with a set of HTTP headers, but no page content. HTTP timeout is set to 3 seconds, meaning that we will wait for a successful response from the remote server for 3 seconds after the TCP handshake has been acknowledged. We hardcoded the IP addresses of all the (sub)domains in the host file of the machine so that DNS resolution times do not affect httping results. We queried all the discovered (sub)domains during several consecutive days from our lab and aggregated the results. The good connectivity of our lab enables us to assess the performance a typical French user with good network connectivity could experience.

We opted to send a `HEAD` request instead of a `GET` HTTP request, as it would only have additionally returned the code for the front web page associated to the (sub)domain, thereby increasing the observed time. Similarly, we avoided TLS connections due to the overhead needed to establish the TLS session itself. Comparing the response times we obtain to the perceived quality of experience (QoE) through a web browser is complex. Indeed, the full page response time is a function of the complexity of the page itself (i.e. number of objects, their size, the number of other web servers to contact, dynamic scripts to evaluate, time to render the page, etc). In addition, it is a complex task to select an appropriate QoS metric to

Alexa Rank	Domain	#cloud subdom	Front-End			ELB IPs	Use CDN
			VM	PaaS	ELB		
7	reddit.com	2	5	0	0	0	0
11	amazon.com	9	6	0	0	0	11
27	linkedin.com	1	1	0	0	0	0
32	netflix.com	27	47	0	15	42	8
34	pornhub.com	1	3	0	0	0	0
35	twitch.com	12	23	0	5	20	15
38	microsoft.com	1	0	0	1	2	0
56	imgur.com	2	5	0	0	0	0
57	github.com	6	0	23	1	2	0
63	imdb.com	1	1	0	0	0	4

TABLE VI
CLOUD DEPLOYMENT OF THE TOP SITES ON AMAZON WEB SERVICES.

Alexa Rank	Domain	#cloud subdom	Front-End	
			CS	TM
14	live.com	20	22	3
36	office.com	12	20	2
38	microsoft.com	77	38	14
44	microsoftonline.com	6	8	0
47	bing.com	3	3	0
49	msn.com	1,803	17	1
72	diply.com	6	3	1
124	dailymotion.com	2	3	0
138	chase.com	2	2	0
163	flipkart.com	1	1	0

TABLE VII
CLOUD DEPLOYMENT OF THE TOP SITES ON MICROSOFT AZURE.

Alexa Rank	Domain	#cloud subdom	Front-End VM
1	google.com	1	1
35	twitch.com	1	1
45	ebay.com	1	1
73	googleusercontent.com	1	1
107	nytimes.com	1	1
115	soundcloud.com	1	1
124	dailymotion.com	1	1
129	utorrent.com	1	1
132	spotify.com	14	14
149	mozilla.com	1	1

TABLE VIII
CLOUD DEPLOYMENT OF THE TOP SITES ON GOOGLE CLOUD COMPUTING.

assess the QoE of a site. For instance, the total download time is not considered a representative enough metric [16] of the user perceived QoE. Consequently the time we measure is to be seen as the minimal application latency, a lower bound to the time-to-first-byte (TTFB) metric [16].

Figure 6 reports the cumulative distribution functions (CDFs) for EC2, CloudFront, Azure, and Google, of 5 complete measurement campaigns. We further added Akamai to the picture for the (non cloud-hosted) domains that were using it. We first observe that it pays off to use a CDN service, with Akamai offering better performance than CloudFront, quite likely due to its larger and more distributed infrastructure. As for the IaaS/PaaS services, Azure and Google appear slightly better than EC2, which we attribute to the ubiquitousness of the Virginia data center in EC2 deployments.

Figure 6 also includes the response times of approximately 800k domains not hosted in the cloud, and not using Akamai. The comparison of the "cloud" and "non-cloud" domains highlights that being hosted in the cloud improves latency by 25% on average, and 14% for the median.

Apart from geographical distance, another important factor that will affect throughput is the number of upstream (peering) links each data center has. We investigate this aspect relying on `traceroute`, and map each hop to its autonomous system (AS) by enabling the `-A` option. Unfortunately, Microsoft Azure blocks all passing ICMP packets as a policy, so we do not receive any response to our measurements. For Amazon Web Services, we present in Table IX the number of AS peers to the Amazon AS. We also report the results obtained by [4], which are, as depicted, identical or larger.

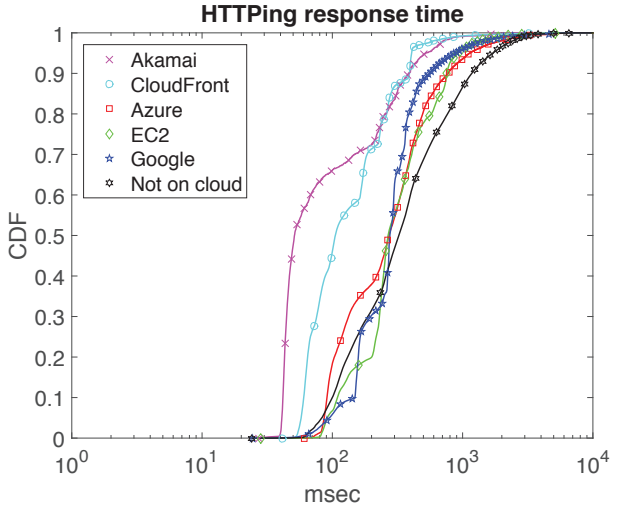

Fig. 6. Minimal application latency response time obtained with httping and HEAD method

IV. DISCUSSION

In this Section, we take the opportunity to tackle possible bias and limitations resulting from the methodology of our

Region	#Peer AS	#Peer AS (in 2013, [4])
us-east-1	36	36
us-east-2	34	19
us-west-1	18	19
us-west-2	23	
ap-northeast-1	16	9
ap-northeast-2	16	
ap-south-1	8	
ap-southeast-1	9	12
ap-southeast-2	12	4
ca-central-1	37	
eu-central-1	46	
eu-west-1	36	
eu-west-2	41	
eu-west-3	45	

TABLE IX
NUMBER OF PEERING LINKS OF AMAZON WEB SERVICES DATA CENTERS

measurements.

First, using a single vantage point to collect the DNS records used in our analysis might prevent us from witnessing dynamic configurations for subdomains, such as: (i) situations where domain owners redirect some of their subdomains to different data centers of the *same* cloud provider, depending on the localization of the client; and (ii) situations where domains owners rely on *distinct* cloud providers for different regions of the world. Indeed, for instance, Netflix hosts catalog.netflix.com in different data centers across the globe, and redirects its customers to the geographically closest instance. Clearly, the geographical location should be the major factor explaining these cases.

Looking at Table IV and Table V, we observe that the fraction of domains relying on this strategy shall not be large. Indeed, in our measurements, 15k out of the 56k domains hosted on Amazon Web Services are located in Europe, while our DNS resolver is located (in the South-East of France). The fraction is higher for Microsoft Azure, however the total number of domains is smaller for Microsoft Azure than for Amazon Web Services.

To further investigate on these deployments, we used 15 virtual machines, one each in each of the Amazon Web Services locations, to perform DNS resolutions for the top 100 domains (and their subdomains) of the Alexa list. Out of these 100 domains, 35 use a cloud provider to host at least one of their subdomains on a cloud provider, amounting to a total of approximately 900 subdomains. We excluded the subdomains of msn.com from the analysis due to their use of DNS wildcard configuration, which adds noise to our data.

Within those 900 subdomains, the DNS resolution of approximately 200 subdomains appears to change depending on the location of the DNS resolver. These subdomains only correspond to 13 domains, 6 of them being owned by Microsoft, and being geographically distributed on their platform, Microsoft Azure. Out of these 13 domains, 12 exhibit a dynamic configuration of type (i). In other words, the cloud provider (either Amazon Web Services or Microsoft Azure) remains the same for the 12 domains, but the location of the service instance changes. One single subdomain (and, hence, domain), appeared to be deployed on Amazon Web Services, and on Google Cloud depending on the resolving location.

In summary, we observe a number of cases of type (i) and a single case of type (ii). As we can reasonably expect that the complexity of the DNS deployment lessens with the Alexa ranking, we believe that the picture drawn in our work indeed reflects the current deployment situation for public cloud-hosting usage.

Second, out httping measurements were also collected from a single vantage point. Performing these measurements from multiple locations could indeed be beneficial. However, accurately measuring metrics relative to the real end-user experience is a complex task as it relies on the ability to capture the diversity of situations in which end users will be using the service: connected at home, at work, on a wired or wireless medium, on a mobile device, etc. In contrast, our objective is to provide qualitative results that underline the relative latencies typically experienced by end users. In Figure 6, we show that domains serviced by CDN servers will served fastest, and that domains hosted on cloud providers offer better latency than the majority of privately-hosted domains. This results from the fact that distance is the key factor in latency. CDN providers (e.g. Akamai) have numerous points of presence at the edge of the network, even within ISP premises. On the other hand, public-cloud providers currently exhibit tens of points of presence around the globe, and hence, located further away from almost all users on stub autonomous systems.

V. RELATED WORK

The first work seeking to understand the use of the public cloud infrastructure by the most popular web (sub)domains was carried out in 2013 by He et al. [4]. Back then, 4% of the Alexa 1M list relied on either Amazon EC2 or Microsoft Azure to deploy at least one sub-domain. Within those, only a limited number of cases were relying on advanced services such as load balancers. As stated in Section II, we purposely followed a methodology similar to the one used by He et al. [4] to allow for an easy comparison of our results: relying on DNS queries for the most popular web domains in order to detect the use (or lack thereof) of a public IaaS cloud. We put our results in perspective, so as to highlight the main differences between 2013 and 2018, a time during which reliance on public IaaS has grown. We also included "new players" in the IaaS space, such as Google Cloud. Section III consistently underlined the similarities, differences, and evolutions based on our experiments.

Contemporary and parallel work to [4] are few. First, Wang et al. [17] perform longitudinal web-server probing in order to understand the long-term infrastructural churn related to web-service deployment. Their platform, called WhoWas, relies on timely probing the full IP space allocated to EC2 and Azure for web services. They identify clusters of websites based on content clustering, and show that deployments are quite static in terms of IP addresses, i.e. websites – which mostly rely on a single IP address – do not appear to move.

Second, Bermudez et al. [5] rely on a combination of passive and active measurement resulting from 2-year traces of 50K Italian ISP users to understand the prevalence of Amazon AWS traffic, from a user perspective. They highlight the skewness of traces towards a single AWS data center located in Virginia, USA, and the progress made over one year to improve the connectivity of AWS data centers by increasing the number of peering agreements, thereby reducing the AS-level path length, and consequently increasing bandwidth. Moreover, they show that the usage of CloudFront, the CDN service of AWS, significantly improves performance. We made similar observations for the today's cloud landscape.

More recently, a number of studies have focused on other types of applications relying or hosted in the cloud. For example [18] seeks to measure the extent of email deployment on the cloud based on end-user email headers and the complete MX records of the three largest generic TLD zones. Other works focus on the interplay between mobile applications and cloud infrastructure, in terms of privacy, e.g. [19], or performance, e.g. [20].

Finally, Dell'Amico et al. [21] focus on existing dependencies among cloud services in order to uncover chains possibly leading to widespread outages, such as the results of the 2016 DDoS attack on DynDNS, and the 2017 Amazon S3 outage. The authors rely on a large set of passive DNS data in order to build a directed graph in which edges represent dependencies between domains. They then analyze the resulting topology for the Alexa list, and uncover a surprisingly large number of (chained) dependencies. However the results in terms of *dependencies* are not straightforward to map to the results in terms of *deployment* that we put forward in this paper. Indeed, a IaaS-based website may be affected negatively by the failure of other services hosted on the same infrastructure, but there are also techniques to make resilient deployment in cloud environments that are not taken into account in [21], e.g. availability zones.

VI. Conclusion

Cloud computing, be it public or private, has significantly reshaped the way companies manage their IT infrastructure. In this paper, we have revisited a topic that has been left aside for over 5 years: how the most popular domains rely on public cloud providers for hosting their main domain and their domains.

We demonstrated through active DNS measurements that the market share of public cloud providers has significantly increased since 2013, reaching about 10% of the top 1M domains and almost 50% of the top 1k domains including their subdomains. This is a significant increase when compared to the 4% of 2013. We further observed that the domain owners use public cloud with some kind of caution, migrating only a modest fraction (with a median at 20%) of their subdomains to the cloud.

In terms of performance, our results indicate that the performance achievable with public cloud hosting lies in between the one offered by typical CDNs and private hosting.

As future work, we intend to automate our measurement process so as to perform a longitudinal study over a longer period of time that should encompass all the top 1M domains and their subdomains. We also want to better assess the performance achieved by client by going beyond latency, and with a large set of vantage points.

Acknowledgment

This research has received funding from the European Commission's H2020 Framework for Research and Innovation, grant agreement #732339: PrEstoCloud.

References

[1] G. Rama, "Report: AWS market share is triple azure's," https://awsinsider.net/articles/2017/08/01/aws-market-share-3x-azure.aspx?m=1, Jan 2017.

[2] Synergy Research Group, "The leading cloud providers continue to run away with the market," https://www.srgresearch.com/articles/leading-cloud-providers-continue-run-away-market, Jul 2017.

[3] J. Brodkin, "Amazon cloud sputters for hours, and a boatload of websites go offline," https://arstechnica.com/information-technology/2017/02/amazon-cloud-sputters-for-hours-and-a-boatload-of-websites-go-offline/, Feb 2017.

[4] K. He, A. Fisher, L. Wang, A. Gember, A. Akella, and T. Ristenpart, "Next stop, the cloud: Understanding modern web service deployment in ec2 and azure," in *ACM IMC*, 2013.

[5] I. Bermudez, S. Traverso, M. M. Munafò, and M. Mellia, "A distributed architecture for the monitoring of clouds and cdns: Applications to amazon AWS," *IEEE Transactions on Network and Service Management*, vol. 11, no. 4, 2014.

[6] S. of the Cloud Report (2018 Data to Navigate Your Multi-Cloud Strategy), "Top cloud providers 2018: How aws, microsoft, google cloud platform, ibm cloud, oracle, alibaba stack up," https://www.rightscale.com/lp/state-of-the-cloud?campaign=7010g0000016JiA, 2018.

[7] M. Million, "Top 1 million list," https://majestic.com/reports/majestic-million, 2018.

[8] Cisco, "Umbrella top 1 million list," http://s3-us-west-1.amazonaws.com/umbrella-static/index.html, 2018.

[9] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A long way to the top: Significance, structure, and stability of internet top lists," in *ACM IMC*, 2018.

[10] Alexa, "Top 1 million list," https://www.alexa.com/topsites, 2018.

[11] Knock, "Knock word-list," https://github.com/guelfoweb/knock/tree/4.1/knockpy/wordlist, 2018.

[12] Amazon, "Ec2 ip range," https://docs.aws.amazon.com/general/latest/gr/aws-ip-ranges.html, 2018.

[13] Microsoft, "Azure ip range," https://www.microsoft.com/en-us/download/details.aspx?id=41653, 2018.

[14] Google, "Google-cloud ip range," https://cloud.google.com/, 2018.

[15] IBM, "IBM cloud IP ranges," https://console.bluemix.net/docs/infrastructure/hardware-firewall-dedicated/ips.html, 2018.

[16] E. Bocchi, L. De Cicco, and D. Rossi, "Measuring the quality of experience of web users," in *Internet-QoE*, 2016.

[17] L. Wang, A. Nappa, J. Caballero, T. Ristenpart, and A. Akella, "Whowas: A platform for measuring web deployments on iaas clouds," in *ACM IMC*, 2014.

[18] M. Henze, M. P. Sanford, and O. Hohlfeld, "Veiled in clouds? assessing the prevalence of cloud computing in the email landscape," in *TMA*, 2017.

[19] M. Henze, J. Pennekamp, D. Hellmanns, E. Mühmer, J. H. Ziegeldorf, A. Drichel, and K. Wehrle, "Cloudanalyzer: Uncovering the cloud usage of mobile apps," in *MobiQuitous*, 2017.

[20] F. Michelinakis, H. Doroud, A. Razaghpanah, A. Lutu, N. Vallina-Rodriguez, P. Gill, and J. Widmer, "The cloud that runs the mobile internet: A measurement study of mobile cloud services," in *IEEE Infocom*, 2018.

[21] M. Dell'Amico, L. Bilge, A. Kayyoor, P. Efstathopoulos, and P. Vervier, "Lean on me: Mining internet service dependencies from large-scale DNS data," in *33rd Annual Computer Security Applications Conference*, 2017.