Exploring and Analysing the African Web Ecosystem

RODÉRICK FANOU †★, CAIDA/University of California, San Diego (UCSD), USA
GARETH TYSON, Queen Mary University of London, United Kingdom
EDER LEAO FERNANDES, Queen Mary University of London, United Kingdom
PIERRE FRANCOIS, Independent Contributor, France
FRANCISCO VALERA, Universidad Carlos III de Madrid (UC3M), Spain
ARJUNA SATHIASEELAN, University of Cambridge, United Kingdom

It is well known that Internet infrastructure deployment is progressing at a rapid pace in the African continent. A flurry of recent research has quantified this, highlighting the expansion of its underlying connectivity network. However, improving the infrastructure is not useful without appropriately provisioned services to exploit it. This paper measures the availability and utilisation of *web infrastructure* in Africa. Whereas others have explored web infrastructure in developed regions, we shed light on practices in developing regions. To achieve this, we apply a comprehensive measurement methodology to collect data from a variety of sources. We first focus on Google to reveal that its content infrastructure in Africa is, indeed, expanding. That said, we find that much of its web content is still served from the US and Europe, despite being the most popular website in many African countries. We repeat the same analysis across a number of other regionally popular websites to find that even top African websites prefer to host their content abroad. To explore the reasons for this, we evaluate some of the major bottlenecks facing content delivery networks (CDNs) in Africa. Amongst other factors, we find a lack of peering between the networks hosting our probes, preventing the sharing of CDN servers, as well as poorly configured DNS resolvers. Finally, our mapping of middleboxes in the region reveals that there is a greater presence of transparent proxies in Africa than in Europe or the US. We conclude the work with a number of suggestions for alleviating the issues observed.

CCS Concepts: • **Information systems → World Wide Web**; • **Networks → Network measurement**; *Middle boxes / network appliances*; *Network structure*; • **Computer systems organization → Client-server architectures**;

Additional Key Words and Phrases: Content infrastructure; Measurements; DNS; Web.

ACM Reference Format:
Rodérick Fanou †★, Gareth Tyson, Eder Leao Fernandes, Pierre Francois, Francisco Valera, and Arjuna Sathiaseelan. 2018. Exploring and Analysing the African Web Ecosystem. *ACM Trans. Web* 1, 1, Article 1 (January 2018), 28 pages. https://doi.org/10.1145/3213897

†This work was mostly done while Rodérick Fanou was a PhD Student at IMDEA Networks Institute and Universidad Carlos III de Madrid (UC3M), Spain.
Manuscript received July 13, 2017; accepted for minor revisions December 6, 2017; revised February 28, 2018; accepted on April 16, 2018. ★ Corresponding author: Rodérick Fanou.
Authors' addresses: Rodérick Fanou †★, CAIDA/University of California, San Diego (UCSD), San Diego, USA, roderick@caida.org; Gareth Tyson, Queen Mary University of London, London, United Kingdom, gareth.tyson@qmul.ac.uk; Eder Leao Fernandes, Queen Mary University of London, London, United Kingdom, e.leao@qmul.ac.uk; Pierre Francois, Independent Contributor, Lyon, France, pfrpfr@gmail.com; Francisco Valera, Universidad Carlos III de Madrid (UC3M), Madrid, Spain, fvalera@it.uc3m.es; Arjuna Sathiaseelan, University of Cambridge, London, United Kingdom, arjuna.sathiaseelan@cl.cam.ac.uk.

1 INTRODUCTION

The Internet infrastructure in Africa is developing rapidly. It has been deploying fibre [59], Internet eXchange Points [3], and edge connectivity at a significant rate [1; 18; 40]. Despite this, Africa is far from achieving the online capacities enjoyed in the West. A prominent reason for this is the poor provisioning of content infrastructure in the region, which forces African clients to often fetch website content from the other side of the world [37]; however, there is little existing evidence to quantify this. Hence, we believe that researchers and engineers should begin to place more focus on *both* underlying connectivity and content infrastructure (*e.g.*, web servers, caches) in the region.

There have been a number of recent works measuring global web infrastructures [20; 31; 33; 37; 41; 47; 62; 69]. However, they have not (*i*) focussed on developing regions like Africa; or (*ii*) explored if worldwide results apply to these regions. This leaves critical questions unanswered, primarily driven by the unusual make-up of African Internet infrastructures. First, the Internet in Africa is at a very different stage of its evolution; sub-optimal topology and peering configurations can make communications (*e.g.*, protocol behaviour) very different [36; 69]. Second, common practices used for content delivery (*e.g.*, placement of caches at IXPs) are difficult due to the lack of IXPs that fulfill the requirements of content delivery networks (CDNs) [23; 24; 31]. Third, hosting services are not as ubiquitous in Africa, potentially making the management of web content much more complex [37]. Fourth, due to the lower level of Internet penetration and disposable incomes [35], there are fewer (medium term) business incentives for optimising web delivery. Again, the depth, veracity, and severity of this reasoning remain unproven. It is therefore essential to explore some of these factors, in an attempt to improve deployments.

This work aims to offer a thorough understanding of the web infrastructure serving Africa. We employ several measurement methodologies for exploring content provider and network operator configurations (§3). We start by analysing traffic from a large European IXP to quantify the amount of traffic failing to be localised in the continent (§5). We find that Africa performs poorly with this measure. Despite the geographical distance, significant amounts of African traffic are transited through Europe (even when the destination is another network in Africa). To help explain this, we focus on one of the largest web providers in the world: Google. After substantially improving our earlier geolocation methodology presented in [26], we show that Google has made notable deployments in the region (§6). However, unlike their operations in Europe and the United States (US) where 90% of caches have been mapped to Google's own AS [41], in Africa, they have primarily partnered with local network operators to deploy their caches. We find 1,067 functional caches in Africa hosted in 59 Autonomous Systems (ASes) and geolocated in 27 countries. Despite this achievement, roughly 48.3% of AFRINIC IPv4 prefixes still rely (exclusively or not) on North America for access to Google content. By measuring redirections, we discover that local network operators tend not to serve each other. Significant inter-AS delays (caused by poor peering) mean that it is often actually more efficient to contact North America or Europe. This is particularly the case for Central African countries, which contain no Google Caches (GGCs). We further investigate other reasons for sub-optimal performance to find that various ASes have inefficient DNS configurations, using distant public resolvers that introduce significant delays to web fetches because of sub-optimal redirects and high resolution delays (§7).

We then broaden our analysis to cover other popular global and regional web providers. Most are far behind Google in their support for African users (§8). Those popular providers, which include regional ones, have a very limited presence in Africa. Even the top local websites host their front-end services outside of the continent. Using network path measurements, we show that these decisions have severe performance implications for all web providers under-study. This leads us to explore the use of transparent web caches in the region. After applying standard detection techniques, we find that there is a higher propensity to deploy these proxies in Africa. We conjecture that caching is more valuable to these regions, where often origin web servers are distant. Finally, we conclude by highlighting key lessons learnt, as well as suggesting recommendations for improving future deployments (§9).

Before getting into the heart of the matter, it is essential to underline that this paper results from substantial extensions and improvements made to our prior work [26]. They consist of:

- A significantly improved geolocation technique (including multilateration geolocation and speed-of-light sanity checks) that addresses a number of limitations in our previous work. A large body of new measurements (§3) underpins this.
- The update of all previously published sections, figures, and tables to reflect the new geolocation technique used.
- The additional exploration of a large IXP dataset, which is used to quantify and understand a lower-bound of the amount of traffic that fails to be localised in Africa.
- The introduction of a new measurement methodology to collect data using the Hola peer-to-peer proxy network. It allows us to identify the use of web proxies/caches in the region, which obviously augment the infrastructure used by web providers.

2 RELATED WORK

Expanding Internet deployment in Africa has received a lot of attention recently [3; 36; 37; 66], mainly from organisations such as the African Union and the Internet Society. There has also been an expanding push from companies like Google and Facebook who see the economic potential of Africa. Of particular interest has been the use of Internet Exchange Points [23; 24; 31], which are seeing an expanding uptake. Further, Fanou *et al.* [24] have underlined, in their four-year study of the interdomain routing in Africa, the remaining reliance on ISPs based outside the region for serving intra-continental traffic in Africa. Moreover, they have revealed the increase over the last years in the number of local IXPs, as well as the positive impact of new IXPs on AS path lengths and delays. More recently, Formoso *et al.* [28] have revisited the African interdomain topology, using a commercial measurement network (Speedchecker) spanning the continent. Note, the spread of Speedchecker in the region is larger than that of the open measurements platform RIPE Atlas network adopted in [23; 24]. Their analysis of inter-country delays highlights a number of clusters, where countries have built up low delay interconnectivity. These confirm the positive results of local initiatives for increasing interconnection and IXPs setups in the region, highlighted above. Also, they noticed, similarly to [23; 24], that the main shortcoming of the infrastructure is an excessive reliance on intercontinental transit providers.

A range of performance studies has accompanied these works. For example, Chetty *et al.* investigated mobile performance, finding that it can often be superior to wireline [12]. Zaki *et al.* [69] focussed on web performance, highlighting that key bottlenecks include slow DNS resolution and a lack of local caching. They found that DNS caching, redirection caching, and the use of SPDY [19] can all yield substantial improvements to user-perceived latency. We take this as clear evidence of the limitations of solely provisioning better connectivity and not considering the upper layers. Next, Fanou *et al.* [21] investigated the prevalence, the causes, and the impacts of congestion on the African IXP substrate, using time-sequence latency probes (TSLP) measurements run for a whole year at selected local IXPs. They found no evidence of widespread congestion during their measurement period. Fanou *et al.* [22] then explored whether IXP interconnection would be possible in the said region to alleviate both the issues related to intra-African traffic and access to content, and estimated the best-case benefits that could be realised regarding traffic localisation and performance. They demonstrated that their distributed IXP layout, which notably parameterises external socioeconomic factors, doubles the percentage of continental intra-African paths, reduces their lengths, and drastically decreases the median of their RTTs, as well as RTTs to ASes hosting top global and regional Alexa websites.

A major theme of our work is understanding the use of web infrastructure in Africa. There have been a number of more general studies into content delivery infrastructures. Farahbakhsh *et al.* [27] depicted and analysed the global picture of the current Facebook network infrastructure, including native Facebook servers and Akamai nodes. Calder *et al.* [41] studied the infrastructure of Google. They enumerated the IP addresses of Google's infrastructure, finding their geographic locations, inspecting its growth, and matching users to clusters. Otto *et al.* [47] examined the role of DNS in the redirection process, exploring the potential of the EDNS client-subnet extension (ECS). Interestingly, by combining distributed DNS queries with ECS queries, we observe potential limitations of this past work. We note similar studies have been expanded to other CDNs such as EdgeCast and CacheFly [61].

Bischof *et al.* [10] explored the performance of end-users by analysing data collected from end-hosts and residential gateways in 160 countries. They provided insight into the impact of broadband service market characteristics, *e.g.*, connection capacity, pricing, cost of increasing capacity, and connection capacity on network usage. Our work is orthogonal to this, focussing on web infrastructure, rather than end-user choices. Prominent works have further analysed redirection strategies to understand how CDNs map users to edge caches. For example, Su *et al.* found that Akamai primarily redirects clients based on active network conditions [62]. More recently, Fan *et al.* [20] evaluated the dynamics of the mapping of network prefixes to front-ends from Google. They found high variance across the servers mapped to each location, with nearby clients often being redirected to clusters that are far apart. Further, Cicalese *et al.* [13] proposed a method for exhaustive and accurate enumeration and city-level geolocation of anycast instances, which requires only a handful of latency measurements from a set of known vantage points.

Our focus differs from these works in that we target web deployments in Africa. We also shed further light on the more general topic by improving existing methodologies through the combination of several measurement approaches. We take a broad perspective, looking at several different websites, content providers, and network operators. The rest of this paper explores this topic to understand the current state of content infrastructure in the African region.

3 MEASUREMENTS METHODOLOGY

We begin by presenting our methodology used to analyse the nature and availability of content infrastructure. It involves three essential steps: (*i*) collecting all IP prefixes for networks in Africa; (*ii*) discovering all the content servers/caches that serve these African networks; (*iii*) mapping the underlying path characteristics between users and the content infrastructure. All our measurement data is public and available at [25] with the corresponding dates of their collection from 2015 to 2016. We further augment this data collection with traces taken from a large European Internet eXchange Point (IXP).

3.1 AFRINIC Prefixes

To map content delivery infrastructure in Africa, it is clearly necessary to compile a comprehensive list of the IP addresses and networks within the continent. To achieve this, we parse the AFRINIC IPv4 assignment and allocation files from 2005 to 2015 [4]. These files gather the IP prefixes allocated by this Regional Internet Registry (RIR), as well as which countries they have been allocated to. By extracting these, we can discover the IP ranges of all networks in Africa. Among 3,488 available IPv4 prefixes, 3,082 of diverse lengths are assigned or allocated as of April 30, 2015. These are the prefixes we consider in this study; we term them *AFRINIC prefixes*. Since the adoption of IPv6 in the region has been recently found to be of only 20% of African ASes by [36], we believe that although our analysis does not involve IPv6 prefixes, this has little effect on the results presented in this work.

3.2 EDNS Client-Subnet (ECS) Probes

Next, we collect a list of content caches that serve these *AFRINIC prefixes*. Since it would clearly be impossible to discover *every* cache, we focus on Google Caches (GGCs). Note that www.google.com is the top-ranked website across the world and most African countries [6]. GGCs operate in a traditional CDN fashion: Whenever a client fetches a Google webpage, it is simply redirected, via DNS, to a nearby GGC.

To measure this, we use the EDNS0 [1] client-subnet extension [14; 47]. It has been developed to improve the accuracy of DNS-based redirections when a client is using a remote public resolver (*e.g.*, Open DNS). The extension allows clients to include their network prefixes in DNS queries (the recursive resolver determines the prefix length). By doing so, CDNs can redirect users to the correct server (rather than a location nearby to the public resolver).

We exploit this feature to launch EDNS Client-subnet (ECS) queries [14; 47] with the client-subnet set to each of the *AFRINIC prefixes* (following a similar methodology to [41]). Through this, we can collect information on the GGCs to which users from across Africa are redirected. We performed three ECS crawls for www.google.com, using a variety of resolvers. First, we sent, every hour on March 06, 2015, ECS queries through Google public DNS (8.8.8.8). Second, we directed our queries through their name servers ns1.google.com, ns2.google.com, and ns3.google.com (all support ECS) every hour on April 12, 2015. Third, we sent again ECS queries through ns1.google.com from April 23 to May 09, 2015 every hour. This revealed 3,011 unique GGC IP addresses, which we term the *ECS Probes* dataset.

3.3 RIPE Atlas DNS Probes

A limitation of the above methodology is that we cannot be sure that the results returned via EDNS Client-subnet (ECS) queries are equivalent to those that would have been returned to an actual client. To verify this, we augment our dataset with a second set of DNS measurements. We use the RIPE Atlas infrastructure, as it is currently the largest open measurement infrastructure in the region. As of June 5, 2017, it has 527 vantage points deployed in 231 ASes across 45 African countries (out of 58 African countries and neighbouring islands) [55; 57].

Next, we infer the network category of the ASes hosting the RIPE Atlas probes. To achieve this, we check within the description of each of those ASes fetched from the corresponding RIRs data, whether they contain any word in the lexicons academia or government (which we previously built). The remainder are considered as commercial networks. For example, the lexicon academia contains the words: laboratory, school, university, college, campus, institute, education, *etc.* As of February 14, 2018, Africa contains 189 online IPv4 probes in 174 ASes, and 52 online IPv6 probes in 26 ASes. We split the set of IPv4 ASes hosting a probe in Africa, finding 15.7% academic networks, 0.9% government networks, and 83.4% commercial networks. Meanwhile, the global RIPE Atlas network contains 7,566 online IPv4 probes in 2,881 ASes and 3,656 online IPv6 probes in 1,224 ASes. The set of ASes hosting the IPv4 probes is composed of 11.25% academic networks, 0.48% government networks, and 88.27% commercial networks. We conclude from these statistics that our results mainly depict the behaviour of commercial networks, as we adopted a measurements platform, which covers primarily such networks both in Africa and worldwide.

Using the RIPE Atlas measurements network, we then repeatedly launched, in parallel, six DNS requests of type A from all the available IPv4 probes in Africa to www.google.com. This was kept running for seven days (from March 24

[1] EDNS0 refers to the Extension mechanism for DNS (RFC6891) [17].

to March 30, 2015). The active probes performed the query three times each, roughly every 60s. We obtained 28,387,226 DNS queries.

Since not all the probes were online during the whole measurement campaign, our DNS lookups involve a total of 225 probes hosted in 38 African countries. AFRINIC has allocated 988 ASes as of May 07, 2015. After removing all the requests that have been performed by probes in Africa hosted in non-*AFRINIC prefixes*, our DNS probes cover 111 AFRINIC ASes (11.23%), and 146 *AFRINIC prefixes* (4.73%). This constitutes the widest vantage on the infrastructure of Google in Africa available yet. From this campaign, we obtained 1,917 GGCs IPs, which we term the *RIPE Atlas DNS* dataset.

3.4 Filtering Inactive Caches and private DNS resolvers

We discovered 3,428 GGC IPs via our RIPE Atlas DNS and ECS campaigns (some IPs were in the outputs of both methods). Following the above, we performed 10 ICMP pings to each discovered cache to verify that it was active. We also issued HTTP requests towards all GGCs to check which ones were alive. These tests have been performed from both Spain (ES) and the United Kingdom (UK) over multiple runs to ensure correctness (on March 09, April 09 and 13, as well as on May 18, 2015). We discard IPs, which did not respond to either pings or HTTP requests. 3,120 IPs remained. We call this set of IPs the *functional GGCs*. The RIPE probes also allows us to discover which DNS resolvers are used by African ISPs. We collect the IP addresses of all (239) default resolvers used by the probes. 70 are RFC1918 private addresses (*e.g.*, 10.0.0.1) [68]; we discard these for the rest of this paper.

3.5 Measuring path characteristics

The above provides a comprehensive set of GGCs and DNS resolvers in Africa. Alone, this does not give insight into the path cost for users though. We, therefore, launched from February 18 to May 22, 2015, a paris-traceroute campaign from all the RIPE Atlas probes in Africa to each of the GGCs IPs. A traceroute between each probe and each GGC IP is issued at five randomly defined timestamps during the said period. We use the UDP protocol [15]. The measurement campaign resulted in a total of 1,309,151 paris-traceroutes. Note that contrary to Gupta *et al.* [31] who performed traces towards GGCs in Kenya (KE), Tunisia (TN), and South Africa (ZA), our traces target all the GGCs worldwide, previously found to serve AFRINIC IP ranges. This provides a topology showing the routes and delays taken from African ISPs to the caches that serve them.

3.6 IXP Packet Traces

The previous measurements are all active and give little insight into the traffic generated by African users. To address this, we augment our data with packet traces collected from a large European IXP. We do this to explore (and exploit) the observation that large amounts of African traffic traverse European IXPs [23; 24; 31]. We wish to verify this claim and quantify the potential benefits from localising traffic within Africa. The collected traffic consists of almost 2 Terabytes of pcap captures from IPFIX records, covering five days worth of traffic (August 23 to 28, 2015). The IXP data is sampled 1 per 10,000 and an approximation of the total traffic observed is given by multiplying the number of bytes in a flow by the inverse of the sampling interval [32]. In total, over 15 billion flows are seen. We tag each flow with the specific Regional Internet Registries (RIRs) that assigned its source and destination IP address [4; 7; 8; 39; 58]. Before doing so, we remove duplicates and overlaps (which are due to prefixes transfer among RIRs or prefixes resales among operators [54]) by considering that a given prefix is only operated by the last RIR to assign it. Clearly, this vantage point only provides us with a subset of African and regional traffic and, therefore, offers a biased sample point. Most notably this is because of the geographical location of the IXP (Europe), as well as the existence of several other large-scale IXPs in the same region. Nevertheless, it still provides a lower-bound vantage into the need for African traffic localisation.

4 GEOLOCATION OF IP ADDRESSES

The previous section has detailed our methodology for data collection. Before carrying out any analysis, it is necessary to geolocate the positions of all IPs we witness (*e.g.*, GGCs and DNS resolvers). This is not trivial and is particularly difficult in Africa, which has seen less attention from mainstream geolocation research. Hence, we take a four-step approach to gain accurate location insight on all GGCs and DNS resolvers. The first step of our approach relies on the geolocation methods used in [24].

4.1 Geolocation Databases

We begin using the traditional approach of geolocation databases (DBs). To avoid problems found with individual geolocation databases [29; 34; 51], we use 10 different geolocation DBs from various sources to find the location associated with each IP. These are: OpenIPMap (*OIM*) [56] whose entries are based on crowdsourcing geolocation data from up to 25 operators; MaxMind GeoIP2City (*MM*) [42]; Team Cymru (*TC*) [64]; the RIRs' assignment and allocation files for AFRINIC DB (*AF*) [4], APNIC DB (*AP*) [7], ARIN DB (*AR*) [8], LACNIC DB (*LAC*) [39], RIPE NCC DB (*RP*) [58]; Whois (*WHOIS*); and Reverse DNS lookups (*RDNS*) from which we infer the geolocation of an IP based on country codes (CCs), cities/airports names, or airport codes embedded in the reverse names. 1,357 GGCs return a domain via a Reverse DNS lookup, whereas 103 DNS resolvers return a domain. Only 11.5% of the 3,120 GGC IPs had an airport or city code in their name. The rest (88.5%) contained no RDNS geolocation info and is composed of (*i*) 14.6% IPs with their names under the format of either cache.google.com or google.cache.com; (*ii*) 21.5% IPs that do not have any airport or city code in their name; (*iii*) and 63.8% IPs that have not been resolved.

When all the DBs with an available entry for an IP give the same result, we use that country code (CC). But when this is not the case, we choose five random RIPE Atlas probes in each of the possible countries and perform three user-defined ping measurements towards the considered IP. We assume that the IP is located in the country with the lowest round trip time (RTT). For 42% of GGC IPs, all the available DBs return the same country code. Amongst the remaining (1,812) IPs, only 1.1% show an inconsistency of three countries, while the rest have an inconsistency of two. The delay tie-breaking approach allows us to geolocate a further 57.6% of the GGCs. At the end of both steps, 99.5% of functional discovered GGCs are geolocated. As far as the DNS resolvers are concerned, all the available DBs return the same country code for only 15 IPs (9.5%). We applied the tie-breaking process for the rest, thereby geolocating 91.7% IPs. It is worth noting that to evaluate the accuracy of commercial and public geolocation databases, Huffaker *et al.* [34] and recently Gharaibeh *et al.* [29] have also adopted, among other techniques, the checks of the consistency of country-level resolution by a given database against the majority answers and the calibration of the IP geolocation against measured RTTs. However, our set of retained databases differs from theirs in that we have only used publicly available DBs.

We summarise the results of this first step in Table 1. The coverage column shows the percentage of IPs for which a DB has answered (*i.e.*, the DB has returned a valid CC). The Trust column shows the percentage of IPs for which the DB entry is equal to the country that we finally selected for that IP. Overall, the DBs are surprisingly accurate with many attaining a Trust above 0.9. That said, there are some significant outliers. *LAC* has no coverage, whilst some DBs such as *OIM*, *AP*, *RDNS*, *RP*, and *AR* have a very low coverage (*e.g.*, 10% and below). *RP* and *WHOIS* are particularly poor. We notice, for instance, that 16.8% of the answers from *RP* are "EU," while the final location is either in Ghana (GH), Tunisia (TN), or the Netherlands (NL). Similarly, although it has a high coverage (97.93%), over half of the geolocations provided by *WHOIS* are inaccurate. These results highlight a key point: using these DBs in isolation would be very unwise in Africa.

Table 1. *Comparison of Geolocation DBs for both GGCs' and DNS resolvers' IPs as of October 2015. N/A stands for Not Applicable.*

	3,105 GGCs IPs		**144 DNS resolvers**	
DB	**Coverage**	**Trust**	**Coverage**	**Trust**
OIM	0.45%	100%	0%	N/A
RDNS	8.27%	93.77%	0%	N/A
MM	98.29%	89.54%	100%	98.61%
RP	10.04%	75.32%	12.5%	88.89%
AF	35.81%	93.07%	81.25%	94.02%
AP	2.58%	100%	0.69%	100%
AR	10.66%	98.49%	22.91%	87.88%
LAC	0%	N/A	0%	N/A
TC	98.97%	90.34%	100%	95.13%
WHOIS	97.93%	47.41%	94.44%	8.82%

Combining the cross-checking of several DBs with latency-based measurements may not be sufficient to achieve accurate geolocation in this study that deals with the web infrastructure for which the addressing is different. Three more steps are, therefore, added to verify the accuracy of our results, thus leading to a four-step geolocation approach. These are (*i*) speed-of-light sanity checks, (*ii*) multilateration geolocation, and (*iii*) final speed-of-light filtering.

4.2 Speed-of-Light Sanity Checks

As a next step, we seek to filter any geolocations that show signs of discrepancies. We follow a similar strategy to [41] for filtering incorrect geolocations based of speed-of-light violations. Towards this end, we repeatedly launched from August 28 to October 18, 2016, (instantaneous) ping measurements from 100 RIPE Atlas probes randomly selected worldwide towards the geolocated GGC and DNS resolver IPs. Since the random sampling was repeated several times, the maximum number of unique probes involved in the measurements run towards a given IP is 230. In total, 2,217 IPs replied, resulting in 480,849 latency measurements. From these, we extract the lowest RTT for each probe-IP pair, termed $Measured_{RTT}$.

Knowing that the signal is transmitted at the speed of $\frac{2}{3}c$ through optical fibre [53], we compute the minimum possible delay Min_{RTT} from each probe to the IP location as $3D/2c$. Note, D is the great circle distance between the coordinates of the probe (in km) and the geolocated IP; and c is the speed-of-light in the vacuum (in km/ms). In cases where $Min_{RTT} > (Measured_{RTT}/2)$, we consider the IP wrongly geolocated. Otherwise, the geolocation is (potentially) correct. 454 GGC IPs and 8 DNS Resolvers IPs violated one or more of these speed-of-light checks, *i.e.*, about 20.8% of the probed IPs.

In 87% of the cases, the IPs whose geolocations were found to break/fail the speed of light test were geolocated during the first phase (*i.e.*, when all DBs agree on the same CC for a given IP). The most common error is incorrect geolocation in the US: 385 GGC IPs out of 454 are wrongly geolocated in the US, while the rest had been incorrectly geolocated in Mauritius (MU), the Netherlands (NL), or Great Britain (GB). Further, six DNS resolvers out of 8 are incorrectly geolocated in the US, and the rest in MU. These findings illustrate how selecting the only available country code for a given IP can also introduce discrepancies in the geolocation results.

4.3 Multilateration Geolocation

The previous section highlighted a number of IPs that could not be correctly geolocated using geolocation databases (as shown via the speed-of-light checks). We next apply multilateration with geographic distance constraints to address this [16; 30]. Multilateration is the technique adopted in the Global Positioning System (GPS), where satellites are used as landmarks. In our case, we consider all the RIPE Atlas probes (selected worldwide) involved in the previous latency measurements as landmarks, since we know their ground-truth locations: Although RIPE Atlas obfuscates these locations, we note that the amount of obfuscation (within 1 km of actual location) does not affect our results, which are based on a country-level geolocation accuracy (as detailed below).

For each IP, our dataset contains a total of landmarks (RIPE Atlas probes) for all of which the GPS coordinates are known. Since the maximum number of unique probes from which we run our latency measurements towards a given IP is 230 (as explained in §4.2), we varied the number of sampled landmarks from a low value (that we set to 15) to 230. For each number of considered landmarks, we obtained a possible geolocation of the targeted IP after applying the multilateration geolocation technique: we could, therefore, compare all these geolocations to make sure they are identical, regardless of the number of landmarks, before deducing that the corresponding IP is thus not an anycast IP. In other words, we could later check if the geolocation for each IP is the same by using $M = 15, 16, 17, \ldots, 230$ landmarks (randomly selected) to identify and remove cases of anycast IPs. We note that this anycast detection methodology was first proposed by [13]. We further report on the obtained results in the subsequent paragraphs. For all IP addresses, we compute the estimated physical distance D from each probe based on its measured RTT $Min_{RTTmeas}$. To this end, we use $(c \times Min_{RTTmeas})/3$. This produces an estimated radius, indicating the potential locations of the IP address (one radius per landmark). By then computing the centroid of the intersection of all radiuses from all landmarks, we can map the IP address to the corresponding country code [16; 30].

To perform this intersection and determine the geolocation of each IP, we first convert all landmarks' GPS coordinates into Earth-Centered, Earth-Fixed (ECEF) coordinates. This information is stored into an $M \times 3$ matrice, P. We then compute the estimated physical distance (D) from each landmark to the IP with which we populated the $M \times 1$ matrice $Dists$. Next, we compute the least Squares solution of this $M \times N$ system to obtain the centroid's ECEF coordinates [30]. After reconverting the ECEF coordinates into GPS ones, we can infer the country code of the IP.

To identify anycast IPs, we vary the number of landmarks M of each IP while running the computation mentioned above. Except for cases in which the IP is an anycast, or cases in which the intersection polygon is too big and covers many countries or islands, the CC obtained should be the same regardless of the number of landmarks. In cases where there is ambiguity, the IPs are removed from our data. 346 out of 2,217 IPs successfully pinged from our landmarks have been geolocated using this methodology: The non-geolocated IPs correspond to cases in which the positions of the landmarks are not suitable for the circles to intersect. Amongst those 346 IPs, 171 are geolocated in only one country, regardless of the number of landmarks. We also noticed that, for example, Google DNS IPs "8.8.8.8" and "8.8.4.4" (both located by all geolocation DBs as being in the US) have different geolocations given the number of landmarks used, highlighting the fact that they correspond to anycast IPs.

Through this methodology, we have found 175 cases of wrong geolocations; we, therefore, removed these since they correspond to anycast IPs. Also, we corrected 69 previous wrongly geolocated IPs. At the end of this step, we could geolocate 2,732 GGCs and 151 DNS resolvers IPs, corresponding to a total of 89.33% of the discovered IPs.

4.4 Final Speed-of-Light Filtering

As a final step, we repeated the speed-of-light checks using a separate testbed to identify any potentially erroneous geolocations from the previous section. We utilise three servers, known to be located in the US (California, San Diego), in Africa (South Africa, Johannesburg), and in Europe (Spain, Madrid). From these three machines, we ping thrice all discovered geolocated IPs. We registered a total of 15,626 measurements outputs (2,219 IPs replied to our pings). As a last cross-check, we then applied the same speed-of-light test like the one in §4.1. Next, we remove any GGCs and DNS resolvers that violate the new speed-of-light checks. 81 IPs are removed, leaving 2,654 GGCs and 148 DNS resolvers IPs. In total, we geolocated 86.8% IPs of the discovered online GGCs and public DNS resolvers IPs. In the rest of this paper, for any statistics related to only IPs and their ASes, we work with all (3,120 GGCs and 169 resolvers IPs) functional GGCs and DNS resolvers, while any statistics including geolocation results are computed for the portion of GGCs and DNS resolvers IPs that we could geolocate (2,654 GGCs and 148 resolvers IPs).

5 MEASURING TRAFFIC LOCALISATION IN AFRICA

Although there has been a wealth of studies looking at traffic from the vantage of European and US networks, we still know very little about the generation and treatment of African traffic. Thus, before diving into the nature of web infrastructure, we first inspect the *need* for improved Internet and web infrastructure in Africa by analysing the amount of traffic that leaves the continent as seen from the vantage of our European IXP data.

5.1 Does Africa have a traffic localisation problem?

Recent work has argued that a major problem in Africa is a lack of peering, and the subsequent need for (Africa-to-Africa) traffic to be routed via remote transit networks — usually through European IXPs [23; 24; 31]. This forms a large part of the motivation for our work. These studies, however, were performed using active traceroute measurements. We, therefore, begin by utilising our European IXP dataset to confirm the veracity of these assertions.

We compute, for comparison purposes, the total volumes of traffic exchanged between IPs allocated by each RIR as seen from this vantage point. Figure 1 shows the quantities of total traffic originated and destined to the same region traversing the IXP. This provides a crude measure of how efficient each region is at localising such traffic and avoiding intercontinental trombonning or remote peering [11]. Again, we emphasise that this is just a single vantage point and therefore our data can only be used as a lower bound.

Unsurprisingly, it can be seen that the greatest traffic volume is exchanged between RIPE NCC (European) prefixes. This is natural considering the physical location of the exchange. More unusually, we also observe a significant volume of ARIN-to-ARIN traffic (North America). Of more interest are the developing regions, AFRINIC, APNIC, and LACNIC — all of which can be seen to route non-negligible amounts of traffic through Europe in a circuitous manner (*cf.* Figure 1). These observations confirm that there is a significant need for greater traffic localisation in inter-African networking. We find that African networks (1,273 ASes as of February 2017 [4]) could offload from intercontinental links at least 0.66 Gigabits of traffic (on average) per second from this single IXP alone. This would lead to improved performance for end-users as well as significant transit costs savings, considering the expensive pricing of a 10 Gbps wavelength on major international routes linking Africa to Europe (US$112,500) compared to the pricing of those linking other continents [46; 52].

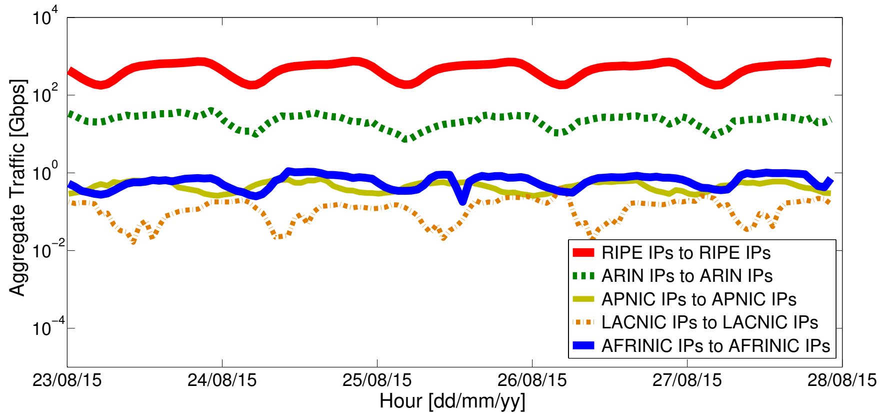


Fig. 1. *Volumes in Gigabits per second of total traffic originated and destined to (v4 and v6) IPs allocated by each RIR passing via the studied IXP.*

Our discussion aims at underlining that, given the pricing of the submarine links interconnecting Africa to Europe, such an amount of traffic going through an EU IXP seems disproportionate. Moreover, apart from underlining the need for greater traffic localisation in inter-African networking, Figure 1 also highlights the same need for other regions as well (*e.g.*, in absolute terms, more ARIN-ARIN traffic is routed through the IXP).

5.2 Where is inter-continental African traffic destined to?

The above shows that the amount of Africa to Africa traffic traversing the studied IXP is not negligible. Before continuing, it is essential to take a closer look at the destinations of traffic generated by AFRINIC prefixes. We next focus on the destinations of traffic originated and destined to IPs allocated by AFRINIC passing through the European IXP. We note that much of the physical cabling connecting Africa to the world runs up through Europe [44], so it is safe to assume that our dataset contains a reasonable amount of traffic leaving Africa.

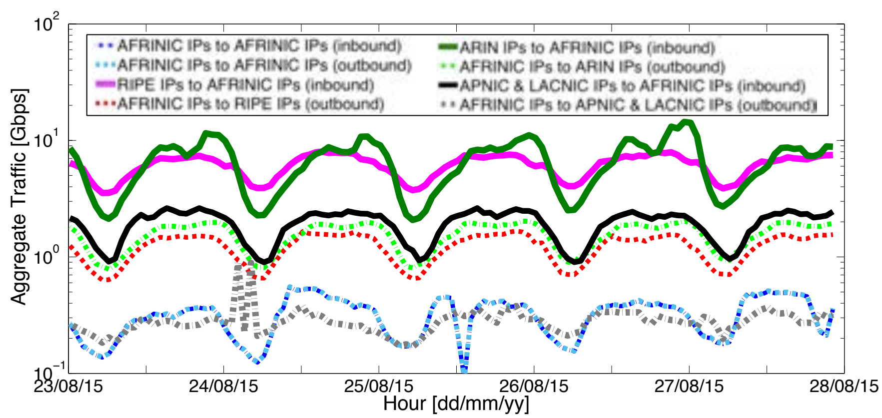


Fig. 2. *Volumes in Gigabits per second of total traffic originated by AFRINIC (v4 and v6) IPs and destined to (v4 and v6) IPs allocated by each RIR (and vice-versa) passing via the studied large European IXP. The inbound traffic is the total traffic traversing the reverse path, while the outbound traffic corresponds to that of the forward path.*

Figure 2 summarises the results across the duration of the IXP dataset. As shown in the figure, the total volumes of traffic originated from and destined to AFRINIC IPs and exchanged via the IXP can be classified from the highest to the lowest in the order of the following RIRs: ARIN, RIPE, APNIC or LACNIC, and AFRINIC IPs. The above shows that most traffic passing through the IXP (originated and destined to AFRINIC IPs) is exchanged with ARIN and RIPE IPs. Interestingly, despite the European location of the IXP, ARIN is actually the most popular destination. This is likely because of the bulk of web and service infrastructure hosted in the US [5]. Regardless, the analysis suggests that significant amounts of traffic and content consumed in Africa are sourced from outside of the continent. This observation, therefore, indicates that Africa could benefit significantly from more local hosting of content and services. The rest of this paper explores the current provisioning from an infrastructural perspective to understand the critical deficiencies.

6 EXPLORING GOOGLE IN AFRICA

Due to its scale and popularity, we start by mapping out the Google infrastructure used by African networks. The statistics presented in this section are computed based on the redirection of *AFRINIC prefixes* to any functional GGC from both the ECS and DNS campaigns.

6.1 Mapping Google Cache Locations

Overall we discover 3,120 functional GGCs serving Africa. However, when discussing country codes, we only use the 2,654 GGCs that we could correctly geolocate (contrary to the results we presented in our previous work [26]). We first investigate which countries these GGCs are located in, shown in Figure 3. We colour code the locations: yellow markers represent GGCs hosted in RIPE NCC ASes, red ones are in ARIN, blue markers are in APNIC, and green ones are in AFRINIC ASes. The size of the marker is proportional to the number of IPs geolocated at that position. Table 2 also lists the top 10 ASes and countries in terms of cache numbers. The percentage between parentheses indicates the fraction of GGCs located in either the corresponding AS or country.

A range of ASes can be seen hosting GGCs. We discover 80 ASes in total, most of which are not owned by Google. 70.2% of the ASes are allocated by AFRINIC, 22.6% by RIPE NCC, 5.9% by ARIN, and 1.1% are APNIC ASes. However, most GGC IPs are in ARIN and AFRINIC IP ranges: Indeed, 43.8% of the 3,120 functional GGCs belong to prefixes allocated by ARIN while 36.2% belong to AFRINIC. The rest (14.9% and 5.1% respectively) belong to prefixes allocated by RIPE NCC and APNIC. African deployments have, therefore, deviated heavily from Google's prior setup in developed regions, which has seen Google hosting most (90%) servers within its own networks [41]. From our traces, only 41.9% of GGCs are hosted in Google ASes: 37.2% in AS15169 (Google) and 4.7% in AS36040 (YouTube Google). All other caches are spread across third-party networks; prominently, AS4788 (TMNET-AS-AP) has 5.1%, and AS3356 (Level3) has 2.56%. All other ASes contain under 2.5% of the caches. We also find that many of the above ASes are based outside of Africa (≈30%).

Compared to our results presented in [26], our new geolocation technique reveals there is a higher proportion of GGCs in Africa than in North America, while the percentages of GGCs in Europe and Asia have slightly increased. Despite efforts, a large number of foreign caches are still relied upon though. 32% of the 2,654 geolocated functional caches are in the United States (US). As shown in Table 2, other prominent countries include the Netherlands (NL), Malaysia (MY), and Germany (DE). Overall, 47 countries host a GGC: 27 in Africa, 12 in Europe, 3 in Oceania (Australia, AU, New Polynesia, and New Caledonia), 2 in North America (US and Canada, CA), 2 in Asia (MY and Bahrain, BH), and 1 in South America (Peru, PE). Africa contains only 40.2% of all caches accessed by its users. Most are located

Fig. 3. *Geolocation of GGCs serving AFRINIC prefixes according to our refined geolocation methodology. The marker size is proportional to the number of IPs geolocated at that longitude and latitude.*

in South Africa (ZA), Egypt (EG), Mauritius (MU), Kenya (KE), and Nigeria (NG). An obvious reason for this setup is that Google's ASes seem to have only a marginal presence in Africa. We also highlight that, surprisingly, Africa is not particularly reliant on Europe for Google content. Only 21% of caches are based in Europe, despite the closer geographic proximity than the US.

We also note that there are *no* caches in most central African countries, *e.g.*, Democratic Republic of Congo (CD), Congo (CG), Gabon (GA), Central African Republic (CF). Instead, caches are mostly based nearer the edges of the continent (as shown in Figure 3). This is likely driven by the expanding number of coastal submarine cables (inland cabling is much more expensive) [45; 63]. That said, we find that even well-meshed countries such as Angola (AO) and Namibia (NA) [60] have no GGCs. It is worth noting that not only our ECS queries include all prefixes allocated by AFRINIC to the above listed countries, but also some of the RIPE Atlas probes from which we launched our DNS queries are hosted in networks operating in those countries.

Table 2. *Top 10 ASes and countries hosting GGC IPs serving AFRINIC prefixes extracted from both DNS and ECS methods. Parentheses contain the percentage of GGCs hosted.*

Rank	**AS (3,120 GGCs considered)**	**CC – Country (2,654 GGCs)**
1	GOOGLE, US (37.21%)	US – United States (31.84%)
2	TMNET-AS-AP, MY (5.13%)	MY – Malaysia (6.07%)
3	YOUTUBE GOOGLE, US (4.74%)	DE – Germany (5.54%)
4	LEVEL3, US (2.56%)	ZA – South Africa (5.24%)
5	MEO-INTERNACIONAL, PT (2.05%)	NL – Netherlands (4.89%)
6	RETN-AS, UA (1.98%)	EG – Egypt (4.48%)
7	ROSTELECOM-AS, RU (1.53%)	MU – Mauritius (2.82%)
8	ETISALAT-MISR, EG (1.51%)	IT – Italia (2.59%)
9	TELECOM ITALIA, IT (1.5%)	KE – Kenya (2.33%)
10	MTNNS-AS, ZA (1.47%)	NG – Nigeria (2.29%)

6.2 Mapping Redirections

We next explore which caches African users are redirected to. This is because the presence of caches in the US is not important if they are only used occasionally. Table 3 presents (*i*) the proportion of caches found in each continent; (*ii*) the percentage of countries that are served by various combinations of continents; and (*iii*) the percentage of *AFRINIC prefixes* served by various combinations of continents.

Table 3. Statistics on Google redirections from AFRINIC prefixes extracted from ECS and DNS query data

Combinations of continents found to serve our queries	% of functional GGCs of the dataset that are geolocated in §4.1	% of African countries served by each set of continents	% of AFRINIC IPs jointly served by each set of continents
AFRICA (AF)	40.2%	1.7%	32.4%
NORTH AMERICA (NAm)	32%	1.7%	31.2%
EUROPE (EU)	21%	3.4%	13.2%
ASIA (AS)	6.5%	—	≃ 0%
OCEANIA (OC)	0.2%	—	0.6%
SOUTH AMERICA (SAm)	0.2%	—	—
AF_NAm	—	6.9%	8.4%
AF_EU	—	—	5.2%
EU_NAm	—	5.2%	3.7%
OC_NAm	—	—	≃ 0%
AF_EU_NAm	—	65.5%	5%
AF_AS_EU	—	—	≃ 0%
EU_OC_NAm	—	1.7%	—
AF_EU_OC_NAm	—	12.1%	0.1%
AF_AS_EU_OC_NAm	—	1.7%	—

The first column of Table 3 shows, as stated previously, that a significant number of GGCs are deployed in Africa (40.2%). Nevertheless, 94.8% of African countries are served by the US at least once in our dataset. In fact, the second column of Table 3 highlights that 65.5% of countries spread their requests amongst Africa, Europe, and North America. This could be for many reasons, *e.g.*, using external caches to support "overflow," where demand exceeds local capacity. It also shows that 12.1% of countries are served by Africa, Europe, North America, and Oceania together. That said, we observe that 5.2% of countries are exclusively served by North America and Europe. In fact, Mayotte (YT), though being an island nearby Comoros and Madagascar, is solely served by North America, indicating that this is not caused by the need for an "overflow." In that case, YT does not host its own GGC, forcing it into using external caches. Ideally, end-users in that country would be redirected to other nearby African countries but, clearly, certain reasons (later explored) prevent this.

Comparing the second to the last columns of Table 3 also highlights some interesting properties. Whereas the bulk of requests on a per country basis are redirected to North America, Europe, and Africa, this is not the case on a per network basis. Only 1.7% of *countries* solely use North American caches. In contrast, 31.2% of *networks* solely rely on North America. Further, whilst *only* 1.7% countries are exclusively served by African caches, we find that 32.4% of networks are. In other words, redirection is primarily based on specific networks rather than countries. This means that many networks fail to gain access to African caches, even though others in their country can do so. Choosing the "right" ISP, therefore, seems particularly important in this region.

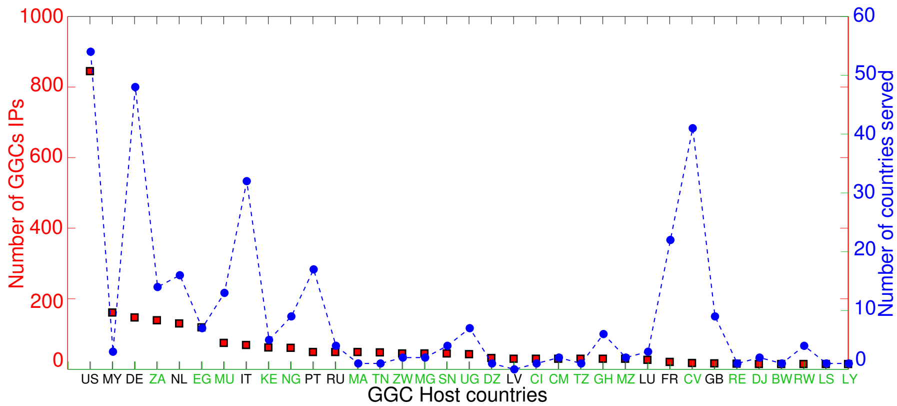


Fig. 4. *Distribution of Google caches (GGCs) serving AFRINIC prefixes across countries. It includes the percentage of other countries that the GGCs are shared with. We consider only the top 35 countries hosting a GGC. African GGCs host countries are in green, whilst GGCs host countries on other continents are in black.*

6.3 Cache Sharing

We next inspect in what circumstances countries and networks share their caches with others. It is particularly pertinent in Africa, as recent work has highlighted that network operators are often resistant to cooperate [31]. Note that sharing is a product of both individual network policy and redirection strategies employed by Google. Figure 4 compares the number of caches within each country against the number of African countries that use those caches. Theoretically, if cache deployment were ubiquitous, each country should only need to service requests from its residents. In such a case, the number of countries mapped to a GGC should always be 1 (*i.e.*, the blue line). Figure 4 shows, however, that this is not the case. In total, 60.6% of countries found to host GGCs share their caches with at least one other country. Indeed, 57.9% of African countries (hosting GGCs) share their caches with other countries, whilst this percentage is 81.8% for those outside Africa.

Unsurprisingly, the most extreme is the US (845 caches), which serves almost all African countries (54). US-based ASes of Google dominate this group of caches. Similarly, in Europe, 48 African countries are served by DE (147 caches). As shown by red squares in Figure 4, Italia (IT) serves 32 African countries with its 69 caches, while the Netherlands (NL) serves 16 countries with its 130 caches. Countries outside Africa share their caches, on average, with 15 other countries, compared to just the half by African countries. In Africa, sharing is largely performed by more developed states, *e.g.*, ZA (serves 14 countries with 139 caches), MU (serves 13 countries with 75 caches), and KE (serves five countries with 62 caches). In contrast, many less developed countries have very different trends. There are countries that host a large number of caches, yet only serve one other country, *e.g.*, Zimbabwe (ZW), which contains 45 caches, Mozambique (MZ) 30, and Cameroon (CM) 30. Meanwhile, in our collected dataset, countries such as MA, TN, Algeria (DZ), Tanzania (TZ), and the Ivory Coast (CI) never serve a user in another country.

Table 4 compares the percentage of GGCs in a country against the percentage of requests redirected to that country (last column). Given the fact that we suppressed any IP suffering from problematic geolocation in §4, 77.4% of the ECS probes and 49.3% of the DNS queries outputs are covered. A proportional and cooperative redirection strategy would result in these two percentages being identical. This, however, is not the case. Clear trends can be seen, with 31.8% of caches in the US receiving 33.6% of our requests from Africa when considering ECS probes. We notice that caches in DE (5.5%) receive 12.7% and 25.4% of requests for ECS probes and DNS queries, respectively. Caches in these countries,

Table 4. *Percentage of total redirections towards GGCs in top 10 countries hosting caches. It is computed based on outputs from ECS probes from all AFRINIC prefixes and DNS queries from RIPE Atlas probes (due to the suppression of any IP suffering from problematic geolocation in our geolocation methodology, 77.4% of the ECS probes and 49.3% of the DNS queries are covered).*

Rank	**CC**	**Country**	**% caches hosted**	**ECS queries**	**DNS queries**
1	US	United States	31.8%	33.6%	14.8%
2	MY	Malaysia	6.1%	0.07%	0.04%
4	DE	Germany	5.5%	3.6%	25.4%
5	ZA	South Africa	5.2%	12.1%	11.3%
3	NL	Netherlands	4.9%	1.9%	0.8%
6	EG	Egypt	4.5%	3.7%	0%
7	MU	Mauritius	2.8%	5.3%	2.1%
8	IT	Italia	2.6%	1.7%	4.8%
9	KE	Kenya	2.3%	3.5%	0.3%
10	NG	Nigeria	2.3%	8%	0.008%

therefore, serve a disproportionately large number of requests. In contrast, Seychelles (SC) and South Africa (ZA) are the only African countries that service about 10% of the requests. The rest service really low proportions (5.5% and below). Hence, despite wide deployment, African caches do not receive a fair proportion of requests.

Of course, the lack of sharing among caches in Africa while servicing requests from the continent that we highlighted above is actually driven by individual networks, rather than entire countries. 15.1% of the networks containing our RIPE Atlas probes host a cache. Only 63.1% ever share their caches with others. For instance, in the collected dataset, AS37183 Utande Internet Services (ZW), AS36914 Ubuntunet (TZ), AS21042 Gulfsat AS (MG), and AS24835 RAYA Telecom (EG) never serve other networks. It is impossible to concretely state the reason; however, we conjecture that it is a combination of both well reported inter-AS performance issues [23; 24; 41] and network operator policy. We explore the former in §6.5, but the latter highlights a critical problem faced in Africa, where it is often challenging to initiate cooperation across organisations and countries [9; 66].

6.4 Understanding Disincentives for Sharing

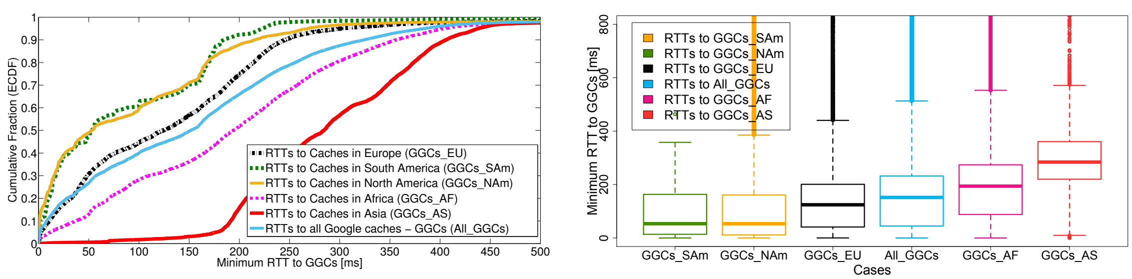
The above raises questions about *why* caches in Africa are not typically shared across networks. Our analysis suggests that a key reason is that many African networks still remain disconnected from regional Internet eXchange Points (IXPs) [23; 24]. Sharing cache capacity would, therefore, generate transit costs, suffer from high inter-AS delay and, consequently, reduce the probability of a CDN redirection algorithm selecting a non-peered neighbour. To explore this, we collect information on IXP peering from IXP websites, PeeringDB, and Packet Clearing House (PCH) [49; 50].

This reveals that most networks sharing caches are peered at IXPs. For example, 99.9% of the requests served by DE caches are redirected to networks peering at DE-CIX in Hamburg; all redirects to the UK go to Google's own AS peered at the LONAP IXP, and 99.7% of redirects to NL go to third-party networks peering at AMS-IX. Similarly, 99.9% of redirects to the US go to peers of one of 33 US IXPs. In these cases, sharing cache capacity is straightforward, as IXP membership allows low-delay, low-cost interactions between networks. To explore this in Africa, we use our paris-traceroute dataset to check if the African networks sharing their caches are peered at IXPs. We find that all African ASes connected to an IXP share their caches. The top two networks for sharing are in ZA (MWEB and Internet Solutions). Unfortunately, however, only 18.6% of African ASes that host a GGC are also peered at an IXP (within our dataset). This means that for the remainder, sharing their caches would generate transit costs. Further, the higher inter-AS delays would drive Google's redirection algorithms away from selecting non-peered networks. Nearly all redirects that stay within Africa are between networks peered together at an IXP. This strong correlation suggests that the main barrier to unlocking significant web performance improvements in Africa is actually to enable cache sharing via peering.

6.5 GGC Performance

Finally, we wish to quantify the performance of Google in Africa by measuring the delay between the RIPE probes and the GGCs (§3.5). In the UDP traceroutes dataset obtained by running a paris-traceroutes campaign from all the RIPE Atlas probes in Africa towards each GGC IP, we consider the last RTT values corresponding to the IPs of the GGCs. As three RTT values are recorded per measurement, we extract the minimum RTT for each probe for measuring the best case scenario. Figure 5(a) shows a CDF of the minimum RTTs to the GGCs measured over each probe in our dataset. Remarkably, the web requests to North American caches actually attain an average delay, with a mean RTT of 99.64 ms (median of 53.28 ms) compared to 223.7 ms for African caches (median of 193.9 ms) (see Figure 5(b)). RTTs to caches in South America have the lowest mean RTT of 89.9 ms (median of 53.38 ms). This is perhaps driven by the direct submarine cable between Africa and Brazil. These confirm that CDN redirection algorithms are right to avoid sending users to other African networks, regardless of their geographical closeness.

Delays to Europe are also high (with an average of 124.2 ms and a median of 137.2 ms), but lower than those to African caches. Only caches in Asia (284.1 ms average and 297.4 ms median) perform worse than those in Africa while serving African end-users. The key exceptions to these observations are African networks that host their own cache. These are reachable by their users with an average minimum RTT of 179.06 ms (median of 75.49 ms) compared to 251.45 ms for those without (median of 201.56 ms). This confirms that the sub-optimality found in African topologies [41] impacts the ability for caches to be locally used/shared within a reasonable delay bound.



(a) *CDF of the minimum RTT distribution*

(b) *Boxplot of the minimum RTT distribution*

Fig. 5. *Delay distribution from different sets of RIPE Atlas probes in African networks to serving GGCs. Minimum RTTs per probe are considered. The cases listed in figure (b) correspond to those in the legend of figure (a) and the colors are the same.*

7 DNS RESOLVER LOCATIONS AND DNS RESOLVER PERFORMANCE

A critical part of web behaviour is DNS (which is typically used by CDNs for redirection). Hence, we explore the DNS configurations used by African networks.

7.1 Mapping DNS Resolver Locations

The RIPE probes allow us to discover which DNS resolvers are used by African ISPs. We collect the IP addresses of all (239) default resolvers used by the probes. 70 are private IP addresses (*e.g.*, 10.0.0.1, 192.168.2.1, etc.); we discard these for the rest of this section. We then geolocate 87.6% of the remaining resolvers using our methodology presented in §4. This set of resolvers contains no anycasted IPs, as these have been removed above (in §4.3) after their detection by our speed-of-light checks. Our results show that the majority are based in Africa (as expected); however, 2.1% are located outside of the continent.

It has previously been found that non-local resolvers can adversely impact CDN performance [47]. In total, 64.04% of resolvers are hosted within the same network as the probe. This case is ideal for CDN redirection, as the CDN would be able to effectively locate the client (using the DNS resolver's IP address). However, 35.95% of unique resolvers are hosted within different networks. Moreover, 46.6% of all the probes share these resolvers located in different networks, showing that many ISPs utilise third-party resolvers by default. Furthermore, we observe that these ISPs use DHCP to automatically configure clients to use third-party resolvers (*i.e.*, this is a network rather than end user choice). As an example, 28.74% of the DNS queries are redirected to IPs identified in §4.3 as being anycasted. The reason for ISPs adopting this behaviour is generally easier management — undoubtedly attractive in Africa.

To explore DNS performance, we now consider all DNS resolvers (identified as anycasted or not during our application of the geolocation methodology). By using distant DNS resolvers, it is possible that significant start-up delays may be introduced for web fetches. Third-party resolvers hosted in other countries have an average delay of 129 ms compared to just 25 ms for resolvers hosted by the ISP. We split the DNS queries into two categories: those sent to anycast-based resolvers (28.74%) and those sent to non-anycast-based resolvers (71.25%).

The first category is composed of anycast-based DNS queries sent to (*i*) Open DNS resolvers (0.89%); (*ii*) open resolvers (6%); (*iii*) ISP resolvers (6%); and (*iv*) Google DNS (86%). We registered as average (median) response time per sub-category 179.3 ms (176.35 ms), 263.03 ms (270 ms), 11.41 ms (6.36 ms), and 114.36 ms (61.7 ms) respectively for ≈29% of RIPE Atlas probes. The second category is composed of non-anycast-based DNS queries sent to (*i*) Google DNS (3%); (*ii*) open resolvers (16%); and (*iii*) ISP resolvers (79%) for the remaining 79.8% of RIPE Atlas probes. The average (median) response time per sub-category is 117.01 ms (59.45 ms), 44.53 ms (6.64 ms), and 28.55 ms (4.67 ms) respectively. One can notice that the first category is dominated by queries redirected to Google DNS resolvers, while in the second one those sent to ISP resolvers are predominant.

Figure 6 presents the corresponding resolution delay distributions. The best performance is obviously attained by resolvers in the local ISP, whatever the category. In fact, ISPs using local resolvers have distances ranging from just 0.07 km to a maximum of 3,554 km (average 325 km). Marginally worse performance is provided by third-party resolvers, which we could geolocate in the same country. However, the most significant drop in performance is introduced by anycast-based public resolvers such as Google DNS. Although they are presented as methods to improve performance, this does not work in Africa due to the lack of public resolver infrastructure on the continent. We find, for instance, that in all the cases for which we could geolocate a Google DNS IP, our African queries are routed to US resolvers. Using these distant resolvers adds over 100 ms delay. In other words, some African operators are outsourcing not only the hosting of web content but also the operation of critical infrastructure such as DNS.

8 EXPANDING TO OTHER PROVIDERS

So far, Google has been focussed on. Next, we expand our analysis to a variety of other popular websites.

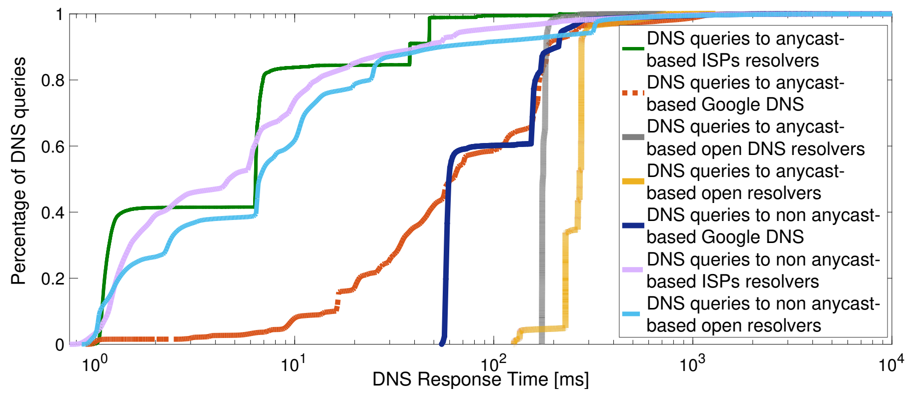


Fig. 6. *Cumulative distribution of DNS resolution delays.*

8.1 Measuring Top Websites

Table 5. *The sizes and locations of the infrastructures of the top 18 websites in Africa (Alexa & Afrodigit), and top 10 global sites (Alexa). We classify websites by their content type.*

Top 15 sites in Africa (by Alexa & Afrodigit)	**Type**	**#IPs caches**	**CCs host caches**	**ASes**	**Top 10 global web-sites** (by Alexa)	**Type**	**#IPs caches**	**CCs host caches**	**#ASes**
jumia.com.ng	E-commerce	1	DE	20546	**amazon.com**	E-commerce	4	US	2
konga.com	E-commerce	1	US	15169	**taobao.com**	E-commerce	15	ZA, UK, CN	4
bidorbuy.co.za	E-commerce	1	ZA	3741					
fnb.co.za	Financial services	1	ZA	17148	**qq.com**	Internet services	2	CN	2
gtbank.com	Financial services	1	US	26496					
absa.co.za	Financial services	1	ZA	3741					
standardbank.co.za	Financial services	1	ZA	10798					
almasryalyoum.com	News/media	5	NL, CR	13335	**google.com**	Search engine	924	18 (§ 6.1)	26
elkhabar.com	News/media	2	US	13335	**yahoo.com**	Search engine	4	US, UK	2
vanguardngr.com	News/media	1	US	14618	**baidu.com**	Search engine	1	HK	1
news24.com	News/media	1	ZA	10474					
punchng.com	News/media	1	IE	16509	**wikipedia.com**	encyclopedia	2	NL, US	2
iol.co.za	News/media	2	IE	16509					
ghanaweb.com	News/media	1	US	7859					
nairaland.com	Online community	5	US	13335	**facebook.com**	Social network	5	US, DE, NL	1
supersport.com	Sports	1	ZA	10474	**twitter.com**	Social network	7	US	2
alwafd.org	Politics	2	NL	13335					
iroking.com	Videos	2	IE	16509	**youtube.com**	Videos	41	SN, MU, US	3

To compile a list of popular websites, we take: (*i*) the global top 10 Alexa websites; (*ii*) the top 15 Alexa websites in Africa; (*iii*) the top 15 most popular websites in Africa listed by Afrodigit.com; and (*iv*) iroking.com, a well-known video content provider on the African continent. We include websites from Afrodigit, because we noted that the top Alexa websites were biased towards websites in certain countries (*e.g.*, South Africa, Nigeria, Egypt). We also added iroking.com to gain an understanding of video websites in Africa (because there are no local video websites in either the top Alexa or Afrodigit sites). Again, we utilise DNS to discover their front-end infrastructures. We concurrently issued DNS queries from RIPE Atlas probes to each of the domains over a four-day period on a per hour frequency (May 23–26, 2015). It allows us to observe the location of front-end servers hosting the websites using our method from §4. In total, 566,994 DNS queries were launched. Note that we only request the home domain of each website and, therefore, these fetches do not include other third-party domains.

Table 5 compares the sizes, the server geolocation, and the networks hosting the websites. Surprisingly, only five websites from the 18 regional ones actually operate their front-end servers in Africa. This is probably attributable to the more reliable and cheaper foreign hosting available [37]. It can also be explained by the significant inter-AS delays, due to which it is sometimes more efficient (in terms of delay/QoS but not in terms of cost) to contact North America or Europe. The five sites hosted in Africa are in ZA, within four ASes. The remainder are in the US or Europe, with common platforms like Amazon and CloudFlare dominating. As for hosting practices, all of the African websites we measured used a single AS to host their content (from the vantage of the 146 *AFRINIC Prefixes* hosting our probes).

In contrast, the top global Alexa websites seen from our probes have a more distributed infrastructure. The global Alexa websites are generally hosted across multiple countries and ASes. That said, we do not see any others achieving the distribution of caches that Google has in Africa. For instance, facebook.com only reveals five front-end IP addresses for our probes (all hosted in Facebook's AS). Unlike Google, Facebook does not host within African networks, instead placing their infrastructure at their own points of presence [33]. Similar results are found across all global Alexa websites. For instance, yahoo.com serves Africa from the GB and US (both hosted in Yahoo's AS); and amazon.com serves Africa from the US (via AS16509 Amazon and AS46475 LimestoneNetworks). That is, the deployment of Google in Africa is *not* the norm. A compelling case was taobao.com, which we found to serve our probes from 15 caches hosted in 3 countries, namely ZA, CN, and the UK.

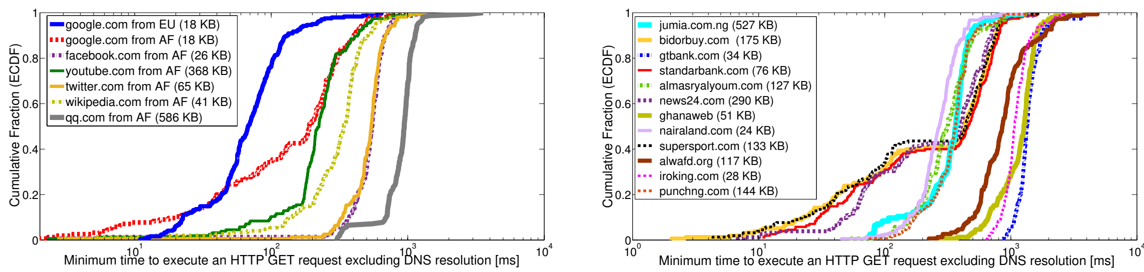
8.2 Website Performance

We next expand upon the previous delay measurements (§6.5), to explore the HTTP performance characteristics of all websites studied. To gain a comparative benchmark, we augment our African RIPE probes with 242 extra probes randomly chosen from Europe. We launched HTTP requests every 12 hours during the period June 2–5, 2015 from every probe to the homepage of every website. To reduce the impact of differences in page size and third-party objects, we only fetch the homepage HTML; we do *not* request images, adverts, javascript, etc. This results in a mean page size of 169 KB, with a standard deviation of just 166 KB (we include website size in the figures). Figure 7(a) shows the minimum time to fetch the global Alexa websites from each probe (measured by the length of the TCP connection). Again, we take the minimum to observe the best case scenario for each probe.

We first inspect Google, which obtains very different page loads across the probes. We see load times varying from 2 ms to 1,250 ms. The mean is 200.9 ms, whilst the interquartile range is 224.4 ms. This is partly caused by the existence of GGCs in a subset of the probes networks. The median load time in networks hosting a cache is just 148 ms compared to an overall median of 190.2 ms. Moreover, 60.7% of probes in ASes hosting GGCs have a delay that is below the average for the continent. However, overall, only 26.2% have a delay that is below that of the median seen in Europe (67.6 ms) and only 32% have an HTTP performance below its mean (84.6 ms). This is not simply caused by the high DNS resolution times previously reported. Even when ignoring the DNS resolution times, we see that only 35% of probes in Africa fetch google.com in under 100 ms; this value is 78% in Europe. Furthermore, the average of the HTTP performance from Europe to Google is more than twice the one experienced from Africa. For medians, it is thrice.

In comparison, the other websites seen from Africa on Figure 7(a) have a greater density around the mean (indicated by a sharp upturn in their CDF). This is because their infrastructures are not as well distributed in Africa as that of Google. Consequently, most African users have similar performance to each other. The median of the HTTP requests performed by the RIPE Atlas probes hosted in African networks is 223.8 ms towards youtube.com, 339.8 ms towards wikipedia.com, 540 ms towards twitter.com, 549.1 ms towards facebook.com, and 943.41 ms to qq.com.

Figure 7(a) can also be compared to Figure 7(b), which presents the same data for the top African websites (from Alexa and Afrodigit). We find that the top African websites get approximately equivalent performance to the global top websites, suggesting that these regional services have made little effort to optimise their local distribution on the continent. The regional websites on Figure 7(b) can also be separated into roughly three groups of varying load times. We note that the ones gaining highest performance are predominantly hosted in Africa, *e.g.*, supersport.com and standardbank.co.za, confirming the benefits that could be gained by services located themselves in Africa. In all cases, these websites are based in ZA, where infrastructure is well developed and affordable. Unfortunately, the worst performing local websites get even lower performance than the globally popular equivalents, indicating that they are not well provisioned. Unsurprisingly, they correspond to those that are based in either the US or Europe. An obvious take-home message is that these websites should aim to host their content locally. In the future, as inter-AS connectivity improves, the increase of sharing caches across networks (via IXPs) could hopefully incentivise this.



(a) *Distribution of minimum time to execute an HTTP GET request per probe (ms) from Europe (EU) and Africa (AF) to top global Alexa websites.*

(b) *Distribution of minimum time to execute an HTTP GET request per probe (ms) from Africa to selected top local Alexa and Afrodigit websites.*

Fig. 7. *HTTP fetch time for websites from RIPE Atlas probes. Website sizes are in parentheses.*

8.3 Transparent Caching

So far, we have exclusively explored servers operated by content providers. However, another major part of the Internet's web infrastructure is that of local network web caches/proxies. Considering the limited number of web servers in the region, we posit that these ISP caches may play a key role in the African web ecosystem. Unfortunately, RIPE Atlas does not allow us to run cache detection algorithms, and therefore we expand our methodology. We use a peer-to-peer proxy network, Hola [2], which allows us to proxy web requests through other users' machines. Using the Hola API, we selected peers in all African countries, alongside the US, the UK, DE, Denmark (DK), and Finland (FI). We use the latter five as benchmarks representing developed countries. Through these open Hola peers, we then proxied 142k web requests to an HTTP server that we control (hosting a “Hello World” webpage). Thus, each request received at our server was first forwarded via a Hola peer in one of the above countries. We subsequently checked for any changes to the HTTP headers in all successful requests at both the client and the server (we obviously controlled which headers were sent in both the request and response). Full details of the methodology can be found in [67].

Figure 8 presents the number of ASes we observe modifying request headers, whereas Figure 9 presents those modifying response headers. In both cases, we only show results for countries where we have a sample of at least 5 ASes. The X-axis is ordered by the countries with the largest fraction of ASes using proxies (shown by the green line). Unsurprisingly, the US provides the largest number of ASes in our dataset. Two African countries stand out in this regard too: ZA and NG. In terms of absolute numbers, they have by far the most African ASes present (62 and 42 respectively). That said, they are not the most prevalent in terms of HTTP proxies. Instead, less developed countries such as Mauritania (MR) and Ethiopia (ET) far outstrip them. This is quite surprising considering the number

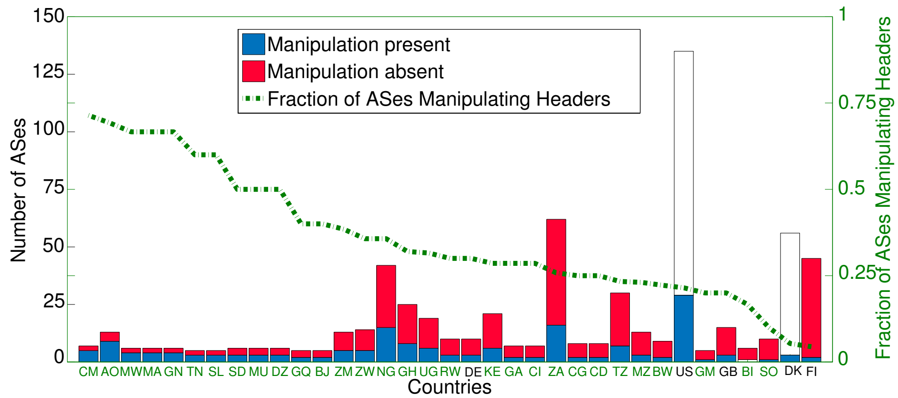


Fig. 8. *Number of ASes modifying HTTP request headers per country. Fraction of modifying ASes per country also shown. On the x-axis, country codes of African countries are in green, whilst those of countries on other continents are in black.*

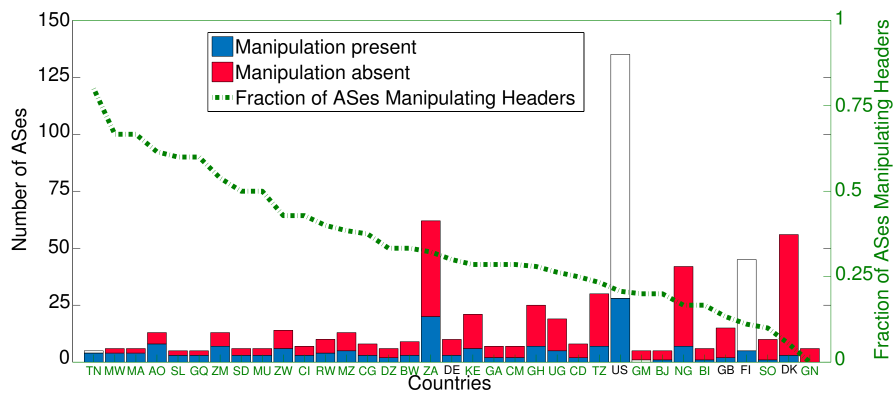


Fig. 9. *Number of ASes modifying HTTP response headers per country. Fraction of modifying ASes per country also shown. On the x-axis, country codes of African countries are in green, whilst those of countries on other continents are in black.*

of networks that outsource functions such as DNS to third-parties. Overall, 36% of the African ASes exhibit request manipulation, with 36% also manipulating responses (this can be compared to 8% in Europe).

The aforementioned results indicate that there is a greater presence of transparent proxies in Africa than in Europe or the US. We contacted several African and European network operators to better understand the causality. They confirmed the finding and offered insight into the reasons. The main reason listed by European operators was the progressive reductions in network transit prices, alongside greater peering via IXPs. In conjunction with higher line rates, this meant that such operators may actually have to pay more for running multi-Gbps web caches than simply contacting the origin via peering or transit. This, however, was not the case for African operators, who still complained of high transit costs and a distinct lack of peering. Another frequently cited reason by European operators was the deployment of dedicated provider-specific caches in their networks (*e.g.*, Google Caches, Netflix Appliances). As our findings confirm, the presence of these appliances in Africa is still limited. Of course, this means the performance benefits of local caching are increased too, although it is worth noting that multiple providers mentioned the challenges faced by the increasing proportion of encrypted web traffic [43]. This is perhaps evident in the earlier measurements

(Figure 7), where we still see poor performance when accessing popular pages (suggesting that transparent caching is not working well). Collectively, these reasons have meant that the business case for transparent caching in developed regions has reduced, whereas it is still strong in developing countries. This situation highlights how African operators have adapted to the surrounding web ecosystem.

9 CONCLUSION AND DISCUSSION

This paper has explored the deployment of web infrastructure in Africa. Whilst recent studies have measured the topology of the African Internet, we argue that this only addresses a subset of the challenges.

We have shown that Africa is far from being self-sufficient in terms of its hosting infrastructure. We began by inspecting packet traces from a large European IXP to witness notable amounts of traffic failing to be localised in Africa. This inspired us to study the deployment of Google, which we found routed significant amounts of Africa-destined traffic through Europe. Although we discovered caches across half of the African countries, we found that US infrastructure is regularly used. Unlike Google's global footprint, these African caches were largely based in third-party networks, which nearly always exclusively service their own subscribers. Only those connected via local IXPs (*e.g.*, JINX, CINX, TIX, or NAPAfrica) broke this trend. Due to poor peering, we find that, in many cases, reaching a geographically nearby African cache actually has a higher delay than contacting the US. As such, sharing cache capacity across networks can only work with improved operator cooperation [9; 66].

That said, we find that Google is considerably more developed in Africa than other providers. We analysed both global and regional websites to find that even local websites are hosted outside of the continent. In fact, only 5 out of the 18 regional website front-ends surveyed were hosted locally (all in ZA). The cheaper cost of hosting abroad and the significant inter-AS delays amongst African ASes are two possible reasons for this. In all cases, we find clear trends showing that these hosting decisions have negative implications for performance. We consistently observed higher HTTP load times for non-Google websites hosted outside of the continent. For those hosted within the continent, we see roughly consistent performance, although it is not yet equivalent to the performance seen in Europe. To complement these server-side studies, we also inspected the presence of transparent web proxies in the region. We found a greater propensity for web proxies in the region compared to more developed areas. Upon discussion with local network operators, this was driven by high transit costs and the aforementioned remote hosting of most content.

There are a number of key implications of our work. We have clearly shown that improving connectivity in Africa is only one part of the equation — it is also necessary to ensure that services are appropriately provisioned. Thus, content providers should begin to improve their presence there. Intuitively, popular regional providers should be the front-runners in this effort. Although perhaps not immediately financially beneficial, this could act as a powerful catalyst for Internet uptake, which will result in revenues in the future.

Combining the above, we can, therefore, propose some steps that should be taken by both network operators and web providers: (*i*) operators must improve peering between networks to enable cache capacity to be shared cheaply and with low delay; (*ii*) content providers must concurrently be encouraged to host caches at existing IXPs; (*iii*) network operators must correct their DNS configuration settings to rely on local DNS for resolution; and (*iv*) public DNS resolvers should be placed in Africa (*e.g.*, at some of the 42 active African IXPs as of April 2018 [38; 48; 65]) to reduce the overheads for clients that continue to use them. These steps are complementary, with the ability of all stakeholders to encourage each other. As an example, if Google were to redirect more clients to GGCs hosted in Africa, network operators would be encouraged to increase peering to reduce the cost of these redirections. Finally, future work may involve exploring these steps further and monitoring the evolution of web infrastructure in the region to constantly access the quality of service experienced by end-users while accessing the web.

ACKNOWLEDGEMENTS

We would like to thank Victor Sanchez-Agüero, Jose Felix Kukielka, Steve Uhlig, and the RIPE Atlas team. Special thanks are expressed to the large European IXP who agreed to share its data for research purposes. We are also grateful to the reviewers and editors for their insightful comments, which contributed to improve this manuscript. While conducting this study, Rodérick Fanou was supported by IMDEA Networks Institute, the National Science Foundation (NSF) CNS-1414177, and NSF OAC-1724853. Arjuna Sathiaseelan is funded by the EU H2020 (Grant agreement No. 644663). Gareth Tyson and Arjuna Sathiaseelan are supported through the EPRSC African Internet Measurement Observatory (AIMO) project, funded under the GCRF. Eder Leao Fernandes is supported by the EU H2020 ENDEAVOUR project (Grant agreement No. 644960). Francisco Valera is partially funded by the European Commission under FP7 project LEONE (Grant agreement No. FP7-317647).

REFERENCES

[1] Google Africa Blog. http://google-africa.blogspot.co.uk/, February 2018.
[2] Hola vpn. https://hola.org, April 2018.
[3] African Union. African Internet eXchange System (AXIS). https://au.int/en/axis, April 2018.
[4] AFRINIC. AFRINIC Database. ftp://ftp.afrinic.net/pub/stats/afrinic/, April 2018.
[5] B. Ager, W. Mühlbauer, G. Smaragdakis, and S. Uhlig. Web Content Cartography. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, pages 585–600. ACM, 2011.
[6] Alexa. Alexa Websites. http://www.alexa.com/topsites/, April 2018.
[7] APNIC. APNIC Database. ftp://ftp.apnic.net/pub/stats/apnic/, April 2018.
[8] ARIN. ARIN Database. ftp://ftp.arin.net/pub/stats/arin/, April 2018.
[9] J. Bezzina. Interconnection Challenges in a Converging Environment, the World Bank. Global Information and Communication Technologies Department, 2005.
[10] Z. Bischof, F. Bustamante, and R. Stanojevic. Need, Want, Can Afford - Broadband Markets and the Behavior of Users. In Proceedings of the 2014 Conference on Internet Measurement Conference, pages 73–86. ACM, 2014.
[11] I. Castro, J. C. Cardona, S. Gorinsky, and P. Francois. Remote Peering: More Peering Without Internet Flattening. In Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies, pages 185–198. ACM, 2014.
[12] M. Chetty, S. Sundaresan, S. Muckaden, N. Feamster, and E. Calandro. Measuring Broadband Performance in South Africa. In Proceedings of the 4th Annual Symposium on Computing for Development, page 1. ACM, 2013.
[13] D. Cicalese, D. Joumblatt, D. Rossi, M.-O. Buob, J. Augé, and T. Friedman. A fistful of pings: Accurate and lightweight anycast enumeration and geolocation. In Computer Communications (INFOCOM), 2015 IEEE Conference on, pages 2776–2784. IEEE.
[14] C. Contavalli, W. Van Der Gaast, D. Lawrence, and W. Kumari. Client Subnet in DNS Queries (No. RFC 7871). 2016.
[15] P. Cristel, C. Luca, V. Stefano, and R. Bush. From Paris to Tokyo: On the suitability of ping to Measure Latency. In Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC), pages 427–432, 2013.
[16] M. Crovella and B. Krishnamurthy. Internet Measurement: Infrastructure, traffic and applications. John Wiley Sons, Inc., 2006.
[17] J. Damas, M. Graff, and P. Vixie. Extension mechanisms for dns (edns (0)) (no. rfc 6891). 2013.
[18] Ekinops. Liquid Telecom Deploys New Optical Network in Africa Using Ekinops Long-Haul DWDM Technology. http://www.ekinops.net/en/press-releases/liquid-telecom-deploys-new-optical-network-in-africa-using-ekinops-long-haul-dwdm-technology, July 2017.
[19] Y. Elkhatib, G. Tyson, and M. Welzl. Can SPDY Really Make the Web Faster? In Networking Conference, 2014 IFIP, pages 1–9. IEEE, 2014.
[20] X. Fan, E. Katz-Bassett, and J. Heidemann. Assessing Affinity Between Users and CDNs sites. In International Workshop on Traffic Monitoring and Analysis, pages 95–110. Springer, 2015.
[21] R. Fanou, A. Dhamdhere, and F. Valera. Investigating the Causes of Congestion on the African IXP substrate. In Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC), pages 57–63. ACM, November 2017.
[22] R. Fanou, V. Francisco, P. Francois, and A. Dhamdhere. Reshaping the african internet: From scattered islands to a connected continent. Computer Communications, 113:25–42, September 2017.
[23] R. Fanou, P. Francois, and E. Aben. On the Diversity of Interdomain Routing in Africa. In International Conference on Passive and Active Network Measurement (PAM), pages 41–54. Springer, 2015.

[24] R. Fanou, P. Francois, E. Aben, M. Mwangi, N. Goburdhan, and F. Valera. Four Years Tracking Unrevealed Topological Changes in the African Interdomain. Computer Communications, 106:117–135, 2017.

[25] R. Fanou, G. Tyson, E. L. Fernandes, P. Francois, F. Valera, and A. Sathiaseelan. Technical Report: African Content Measurement Campaign. https://techrepwebinf:bRCA9hFZ@fourier.networks.imdea.org/external/techrep_web_infrastructure/index/, April 2018.

[26] R. Fanou, G. Tyson, P. Francois, and A. Sathiaseelan. Pushing the Frontier: Exploring the African Web Ecosystem. In Proceedings of the 25th International Conference on World Wide Web (WWW), 2016.

[27] R. Farahbakhsh, A. Cuevas, A. M. Ortiz, X. Han, and N. Crespi. How far is Facebook from me? Facebook network infrastructure analysis. IEEE Communications Magazine, 53:134–142, 2015.

[28] A. Formoso, J. Chavula, A. Phokeer, A. Sathiaseelan, and G. Tyson. Deep Diving into Africa's Inter-Country Latencies. In Proceedings IEEE International Conference on Computer Communications (INFOCOM), April 2018.

[29] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos. A Look at Router Geolocation in Public and Commercial Databases. In Proceedings of the ACM Internet Measurement Conference (IMC), 2017.

[30] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-based geolocation of internet hosts. IEEE/ACM Transactions On Networking, 2006.

[31] A. Gupta, M. Calder, N. Feamster, M. Chetty, E. Calandro, and E. Katz-Bassett. Peering at the Internet's Frontier: A First Look at ISP interconnectivity in Africa. In Passive and Active Measurement (PAM) Conference, pages 204–213. Springer, 2014.

[32] R. Hofstede, P. Čeleda, B. Trammell, I. Drago, R. Sadre, A. Sperotto, and A. Pras. Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX. volume 16, pages 2037–2064, 2014.

[33] Q. Huang, K. Birman, R. van Renesse, W. Lloyd, S. Kumar, and H. C. Li. An Analysis of Facebook Photo Caching. In Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, pages 167–181. ACM, 2013.

[34] B. Huffaker, M. Fomenkov, and K. Claffy. Claffy kc (2011) Geocompare: a comparison of public and commercial geolocation databases. In Network Mapping and Measurement Conference (NMMC). http://www. caida. org/publications/papers/2011/geocompare-tr, pages 1–12, 2011.

[35] Internet World Stats. Internet World Stats: Usage and Population Statistics. http://www.internetworldstats.com/stats.htm, April 2018.

[36] L. Ioana, E. Ahmed, and D. Amogh. Measuring ipv6 adoption in africa. In International Workshop on Internet Measurements Research in Africa, December 2017.

[37] M. Kende and K. Rose. Promoting Local Content Hosting to Develop the Internet Ecosystem. ISOC Report, 2015.

[38] Kyle Spencer. The African IXP Association. https://wp.internetsociety.org/afpif/wp-content/uploads/sites/26/2017/10/Africa-IXP-Survey-Report.pdf, April 2018.

[39] LACNIC. LACNIC Database. ftp://ftp.lacnic.net/pub/stats/lacnic/, April 2018.

[40] Liquid Telecom. Liquid Telecom network map. http://liquidtelecom.com/about-us/network-map, August 2017.

[41] C. Matt, F. Xun, H. Zi, E. Katz-Basset, H. John, and G. Ramesh. Mapping the Expansion of Google's Serving Infrastructure. In Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC), 2013.

[42] MaxMind. GeoIP. http://www.maxmind.com/en/geolocation_landing, April 2018.

[43] D. Naylor, A. Finamore, I. Leontiadis, Y. Grunenberger, M. Mellia, M. Munafò, K. Papagiannaki, and P. Steenkiste. The Cost of the s in https. In Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies, pages 133–140. ACM, 2014.

[44] Network Startup Resource Center (NSRC). Map of African Terrestrial and Undersea Fibre Networks. https://afterfibre.nsrc.org/, April 2018.

[45] Network Startup Resource Center (NSRC). Mapping Undersea and Terrestrial Fibre Optic Cables. https://afterfibre.nsrc.org/, April 2018.

[46] P. Okui. International Internet Bandwidth and Pricing trends in Africa (Telegeography). https://www.slideshare.net/InternetSociety/international-bandwidth-and-pricing-trends-in-subsahara-africa, August 2016.

[47] J. S. Otto, M. A. Sánchez, J. P. Rula, and F. E. Bustamante. Content Delivery and The Natural Evolution of DNS: Remote DNS Trends, Performance Issues and Alternative Solutions. In Proceedings of the 2012 ACM conference on Internet measurement conference, pages 523–536. ACM, 2012.

[48] Packet Clearing House (PCH). Internet Exchange Point Growth. https://prefix.pch.net/applications/ixpdir/summary/growth/, April 2018.

[49] Packet Clearing House (PCH). PCH IXP directory. http://prefix.pch.net/images/applications/ixpdir/ip_asn_mapping.txt, April 2018.

[50] PeeringDB. http://www.peeringdb.com/private/exchange_list.php, July 2017.

[51] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. IP Geolocation Databases: Unreliable? ACM SIGCOMM Computer Communication Review, 41(2):53–56, 2011.

[52] PriMetrica. TeleGeography Internet Exchange Map. http://www.internetexchangemap.com/, July 2017.

[53] R. Ramaswami, K. Sivarajan, and G. Sasaki. Optical networks: a practical perspective. Morgan Kaufmann, 2009.

[54] P. Richter, M. Allman, R. Bush, and V. Paxson. A Primer on IPv4 Scarcity. ACM SIGCOMM Computer Communication Review, pages 21–31, 2015.

[55] RIPE NCC. Global RIPE Atlas Network Coverage. https://atlas.ripe.net/results/maps/network-coverage/, April 2018.

[56] RIPE NCC. Infrastructure Geolocation: Plan of Action database. https://labs.ripe.net/Members/emileaben/infrastructure-geolocation-plan-of-action, June 2018.

[57] RIPE NCC. RIPE Atlas - Raw Data Structure Documentation. https://atlas.ripe.net/docs/data_struct/, April 2018.

[58] RIPE NCC. RIPE NCC Database. ftp://ftp.ripe.net/ripe/stats/, 2018.

[59] S. Song. African Undersea Cables. https://manypossibilities.net/african-undersea-cables/, 2018.

[60] S. Song. Mapping Terrestrial Fibre Optic Cable Projects in Africa. https://afterfibre.net/, April 2018.

[61] F. Streibelt, J. Böttger, N. Chatzis, G. Smaragdakis, and A. Feldmann. Exploring EDNS-client-subnet Adopters in Your Free Time. In Proceedings of the 2013 conference on Internet measurement conference, pages 305–312. ACM, 2013.

[62] A.-J. Su, D. R. Choffnes, A. Kuzmanovic, and F. E. Bustamante. Drafting Behind Akamai (Travelocity-Based Detouring). In ACM SIGCOMM '06 Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications, volume 36, pages 435–446. ACM, 2006.

[63] Submarine Telecoms Forum, Inc. Submarine Telecoms Industry Report 2017. Technical report, Terabit Consulting, 2017.

[64] Team Cymru. Team Cymru Services. https://www.team-cymru.com/, April 2018.

[65] The African IXP Association (Af-IX). List of Active Internet eXchange Points in Africa. http://www.af-ix.net/ixps-list, 2018.

[66] N.-J. Towela and B. Tesfaye. Internet Development and Internet Governance in Africa. Technical report, Internet Society (ISOC), 2015.

[67] G. Tyson, S. Huang, F. Cuadrado, I. Castro, V. Perta, A. Sathiaseelan, and S. Uhlig. Exploring http header manipulation in the wild. In WWW Conference, 2017.

[68] J. Weil, V. Kuarsingh, C. Donley, C. Liljenstolpe, and M. Azinger. IANA-reserved IPv4 prefix for shared address space. 2012.

[69] Y. Zaki, J. Chen, T. Pötsch, and T. A. Lakshminarayanan Subramanian. Dissecting Web Latency in Ghana. In Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC), 2014.