Shortcuts through Colocation Facilities

Vasileios Kotronis
FORTH, Greece
vkotronis@ics.forth.gr

George Nomikos
FORTH, Greece
gnomikos@ics.forth.gr

Lefteris Manassakis
FORTH, Greece
leftman@ics.forth.gr

Dimitris Mavrommatis
FORTH, Greece
mavromat@ics.forth.gr

Xenofontas Dimitropoulos
FORTH, Greece
University of Crete, Greece
fontas@ics.forth.gr

ABSTRACT

Network overlays, running on top of the existing Internet substrate, are of perennial value to Internet end-users in the context of, e.g., real-time applications. Such overlays can employ traffic relays to yield path latencies lower than the direct paths, a phenomenon known as Triangle Inequality Violation (TIV). Past studies identify the opportunities of reducing latency using TIVs. However, they do not investigate the gains of strategically selecting relays in Colocation Facilities (Colos). In this work, we answer the following questions: (i) how Colo-hosted relays compare with other relays as well as with the direct Internet, in terms of latency (RTT) reductions; (ii) what are the best locations for placing the relays to yield these reductions. To this end, we conduct a large-scale one-month measurement of inter-domain paths between RIPE Atlas (RA) nodes as endpoints, located at eyeball networks. We employ as relays Planetlab nodes, other RA nodes, and machines in Colos. We examine the RTTs of the overlay paths obtained via the selected relays, as well as the direct paths. We find that Colo-based relays perform the best and can achieve latency reductions against direct paths, ranging from a few to 100s of milliseconds, in 76% of the total cases; ~75% (58% of total cases) of these reductions require only 10 relays in 6 large Colos.

CCS CONCEPTS

• **Networks → Network measurement**;

KEYWORDS

Overlay Network, Relay, Latency, Triangle Inequality Violation

1 INTRODUCTION

Every millisecond of Internet latency counts. A broker could lose $4 million with every passing millisecond (ms), if their electronic trading platform lags 5 ms behind the competition [37].

Overlay networks have historically been used to attach desirable properties to the classic best-effort Internet, including lower latency [29]; reliability [12], security [44], avoidance of certain areas via "detours" [36], and higher throughput [15] are only some of them. Operating over a stable IP-based underlay, overlays have revolutionized the way the Internet is used in the last decades. In particular, end-users, as well as their overlay application providers, have much to gain from low-latency overlay paths for real-time applications such as online gaming [41, 42], VoIP [11, 29], and financial transactions [33]. Such end-users typically reside in *eyeball* networks, namely access ISPs at the last mile [52].

Two important research questions that cut through most efforts studying overlay networks (see Section 4) are the following: "*What are the best locations to place overlay TIV relays, in order to improve performance or resiliency? What are the quantified benefits of choosing these relays instead of others?*". End-hosts in eyeball networks and dedicated servers in PlanetLab are common relay choices in real systems (such as p2p networks) and academic studies.

In this work, focusing on the latency-wise improvement of Internet paths, we examine the increasingly popular [17, 22, 24] colocation facilities (Colos) as relay sites. Colos provide space, power, cooling, and physical security for the server, storage, and networking equipment of colocated companies, connecting them to cloud/content providers, transit networks and eyeball ISPs, as-a-service, in multiple locations worldwide [4]. Thus, they host layer-2/3 interconnections (such as IXPs [24]), ranging from private to public multilateral peering setups among different ISPs. The pervasiveness of Colos on a global level has brought Internet organizations (and their users) closer to each other, driving Internet flattening [20]. In addition, the ecosystem of colocated companies has evolved in recent years to include many small and medium cloud providers that house their equipment (compute servers, storage, etc.) in the Colo. This change allowed for the first time in Internet's history third parties, such as end-users or application service providers, to easily rent Virtual Machines (VMs) in the largest Colos using the services of cloud providers [3].

Towards understanding the implications of this change for delay-sensitive overlay services, such as Skype and Hola, we investigate how the performance of Colo relays compares with other types of relays. Colos might be considered quite promising candidate

TIV relays due to their core networking location. However, it is not straightforward to expect that Colos always perform better; moreover, their exact benefit should be properly quantified. To investigate this further, we choose endpoints within eyeball networks, utilizing 3 different types of relays (see Section 2): Colo relays (interfaces located in facilities [24]), PlanetLab nodes (mainly at research institutes) and RIPE Atlas nodes (at eyeball and other networks). We simulate stitching of paths between endpoints through these relays, based on RTT measurements, to form single-relay TIV paths [25]. We compare the formed overlay paths with each other as well as with the direct BGP-derived paths, in terms of latency.

Based on a large-scale, 1-month measurement campaign, and after verifying eyeball networks (see Section 2.1) and Colo locations (see Section 2.2) for accuracy, we identify that the best locations for placing relays are actually the Colos (see Section 3). To the best of our knowledge, we are the first to study the impact of Colos w.r.t. latency at a large scale. We observe that Colo-relayed paths can yield median latency improvements of >10ms (with a 6% of the cases gaining >100ms) vs. the direct paths, and in contrast to the other relays (58% for RIPE Atlas, 43% for PlanetLab), improve 76% of the total studied cases. Interestingly, a relatively small number of Colos (~6) is required to achieve most of these gains (~75% of improved cases, 58% of total), while the rest of the studied overlays need to employ one order of magnitude more relays to reach their respective top performance. Moreover, relaying through different countries (compared to the endpoints) helps reduce latency, probably due to the forced discovery of alternate, non-inflated [51] BGP paths. We further show that our insights are consistent over time.

2 MEASUREMENT METHODOLOGY

The objective of our measurement methodology is twofold: (i) to study the path latency obtained by employing relays as intermediate nodes of overlay inter-domain TIV paths, and (ii) to assess the benefits of selecting vantage points at large Colos as Internet relays, vs. relays placed at the Internet's eyeball (and other) networks. To this end, we employ a real-world, Internet-wide testbed, comprising:

- **Endpoint nodes**: a set of globally distributed nodes, acting as source *src* and destination *dst* nodes of Internet inter-domain paths.
- **Overlay relays**: a set of relay nodes that are employed as intermediate hops within an inter-domain path between a *src* and a *dst*.
- **Inter-domain overlay links**: logical links that connect pairs of nodes (endpoints and/or relays) over the physical network of one or more intermediary ISPs. The underlying paths are typically derived by BGP.

We next describe how we select *src* and *dst* endpoints in eyeball networks (see Section 2.1), relays at colocation facilities (see Section 2.2) and relays at other locations (see Section 2.3). We further explain a strategy to limit the number of candidate relays based on their relative position to the endpoints in Section 2.4. Finally, we unfold our complete measurement framework, including the applied setup and workflow in Section 2.5.

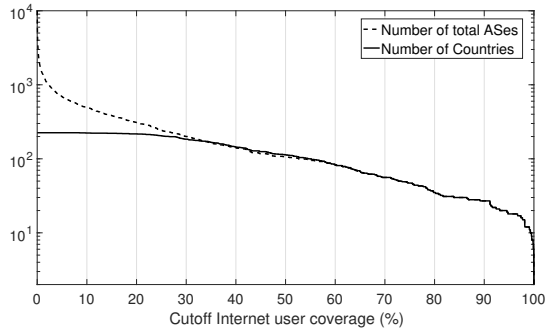

Figure 1: Number of covered ASes/countries (log-scale) worldwide *vs.* the cutoff Internet user coverage (coverage for each AS in its respective country of operation).

2.1 Selection of Endpoints at Eyeballs

To perform our measurement campaign, first we select pairs of endpoints (both in different countries), which communicate either directly or via relays. For this purpose, we use RIPE Atlas (RA) [9] nodes (i.e., probes and anchors), a globally distributed measurement infrastructure consisting of end-host devices, capable of conducting different types of data-plane measurements.

Since end-users primarily reside in eyeball networks, we want to select RA nodes located within eyeballs, hence close to the end-user. Our aim is not to exhaustively cover all eyeballs, but to find a set of ASes, with sufficient country/AS-level diversity, qualifying as such. To find these ASes, we utilize the results of the IPv6 measurement campaign by APNIC [13]. The dataset contains 19857 ASes from 225 countries. Besides IPv6 adoption and statistics related specifically to IPv6 users, APNIC estimates Internet user population percentages (i.e., user coverage) per AS per country for both IPv4 and IPv6 combined. These percentages drive the eyeball selection process as follows. The measured ASes face Internet users—browsing the web—but in order to characterize them as actual "eyeball" ISPs (and not e.g., enterprise networks), we also require a sufficient percentage of user population per country served by the ASes (i.e., a "cutoff" coverage). Note that a large cutoff coverage may support the eyeball characterization, but can exclude countries with fragmented eyeball ISP ecosystems (e.g., the US). In Fig. 1, we show the number of covered countries and ASes worldwide versus the cutoff threshold. If there is an AS present in a certain country, with a given coverage level, then the country is considered covered at this level. Almost all countries (223/225) as reported in APNIC's dataset, host at least one AS serving more than 10% of the respective country's user population; 494 ASes satisfy this threshold, offering relative diversity. Above ~30%, the 2 lines (see Fig. 1) converge, indicating that only 1 AS per country is present, yielding a low AS-level diversity. We validate if the 10% threshold is an appropriate lower bound for ASes to be considered as eyeballs (within their respective countries). We successfully verify all *494 ASes* by *manually* examining their official websites, and discovering Internet services provided to end-users (e.g., last-mile access).

We then select as endpoints RA nodes which belong to the verified (ASN, CC) tuples (CC = country code of AS; a single eyeball AS may be present in multiple countries). We consider only RA nodes that are: (i) running the latest RA firmware version to minimize

interference across measurements, affecting older versions [27], (ii) publicly available, (iii) connected and pingable, (iv) tagged with their geolocation coordinates, and (v) stable, connectivity-wise, during the last 30 days. This filtering yields ~1190 probes, associated with 141 ASes at 82 countries (where RA is present). For each measurement round (cf. Section 2.5) we perform sampling on this population by selecting randomly: (i) one eyeball AS per country, and (ii) one node from this AS as RA endpoint (*RAE*). This 2-step sampling limits the number of endpoints per round to a reasonable number of 82 *RAE*s on average, while preserving endpoint diversity and not incurring the bias of eyeballs with dense probe deployments. While this strategy may represent smaller vs. larger countries with the same number of *RAE*s, our goal is to achieve country-level instead of complete geographical/population diversity.

2.2 Selection of Relays at Colos

For this study, our objective is to use relays located at multiple Colos around the globe. We require pingable IPs, located at Colos, to use as ping targets. Root or user access e.g., to VMs hosted in colocated clouds is not necessary for the purposes of this latency-oriented study; sizable pools of colocated IPs, belonging either to routers or servers, remaining stable over time, suffice[1].

To generate such a pool of IPs, we use the publicly available dataset produced by Giotsas *et al.* [23, 24] in 2015. The authors identified facility crossings by applying their constrained facility search algorithm on extensive traceroute measurements, achieving an accuracy of over 90%, outperforming heuristics based on naming schemes and IP geolocation. The data provide IP addresses that belong to facility members and are present at a candidate set of facilities, together with their respective ASN and neighboring IXPs. However, due to the age of the dataset, we need to exclude stale information by applying, in-order, the following filters.

Single-facility & active PeeringDB presence. Preserve only the IP addresses for which the set of candidate facilities contains exactly one facility that is still present in PeeringDB [6] today[2]. **1008** out of the initial 2675 IP addresses pass this rule.

Pingability. Preserve only the IP addresses that are still pingable (after a period of almost two years). **764** out of the previous 1008 IP addresses pass this rule.

Same IP-ownership. Preserve only the IP addresses whose ASN is the same as given in the initial dataset [23], since the IP-to-ASN mapping needs to be consistent. We also check that this IP is not simultaneously advertised by multiple ASes (MOAS) to increase confidence in the dataset. To verify this, we use CAIDA's AS-to-Prefix dataset [16] to map IPv4 prefixes to ASNs. **725** out of the previous 764 IP addresses pass this rule.

Active Facility presence of ASN. We preserve the IP addresses whose verified ASN owner is still present at the candidate facility, according to PeeringDB data [6]. **725** out of the previous 725 IP addresses pass this rule.

RTT-based geolocation. First, we extract the facility's city from PeeringDB [6]. Our 725 candidate IP addresses are associated with 103 facilities present at 67 cities around the globe. We want to ensure that these IPs are located at the respective city of the candidate facility. Current IP-based geolocation services do not provide city-level accuracy [5, 45, 49], thus, to determine each IP location we use Periscope, a tool that utilizes available Looking Glass servers (LGs) [21] (1818 LGs at 526 cities at the time of measurement, i.e., between 1-6 April 2017). For each candidate IP, and for each set of LGs residing in the same city as this IP's facility, we measure the RTT from the LGs towards the IP address. Since Periscope currently supports only traceroute probes from LGs, we calculate the RTT as the one yielded on the last hop to the IP. We keep the minimum RTT for each IP as the primary indicator, to avoid RTT inflation effects affecting other LGs. We consider only IPs for which Periscope measurements are available and for which the minimum RTT does not exceed a threshold of 1ms [50].

The rule-checking process yields **356** IP addresses mapped to 58 facilities in 36 cities around the world (US, Europe, SE Asia and Australia). During each measurement round we select randomly 1 to 3 IPs per facility to both cover all available facilities and account for variance within facilities, thus working with a sampled population of 129 IP addresses on average, used as Colo relays (*COR*).

2.3 Selection of Relays at Other Locations

Except for overlay relays residing at Colos, we consider relay nodes hosted at other locations as alternative Internet vantage points. To this end, we use publicly available nodes from PlanetLab [18] (see Section 2.3.1) and RIPE Atlas [9] (see Section 2.3.2).

2.3.1 PlanetLab Relays. We extract the first set of relays from PlanetLab [18], a global research network that numbers ~1.4k nodes (at 717 sites), mostly located in research and academic institutions. Having allocated 500 nodes from 62 sites as candidate relays out of this set, we select randomly 1 to 2 nodes per site that are consistently accessible and pingable before each measurement round, thus working with an average sample of ~59 PlanetLab relays (*PLR*)[3].

2.3.2 RIPE Atlas Relays. We employ two independent sets of RA relays (*RAR*), the one from nodes at eyeball networks (*RAR_eye*) and the other from nodes at networks that have not been verified as such (*RAR_other*), potentially in core locations [10].

Eyeball Networks. To generate the eyeball relay set, we follow the methodology of Section 2.1. We then sample 82 relays (as many as the available countries) on average for each measurement round.

Other Networks. For the other relays, we use all the rest of the available (ASN, CC) tuples. Out of ~2500 remaining relays, we randomly select one relay per country, gathering 102 relays on average per measurement round.

2.4 Choosing Feasible Relays

Not all available relays are useful for a certain pair of endpoints. Some of them, even if used under ideal conditions within a "speed-of-light" Internet [14], still yield larger latency than the observed direct path. Thus, to exclude such relays, we follow a simple approach based on the geolocation information of the involved nodes. Given a certain pair of endpoints (n_1, n_2), we compute the geographical

[1] However, relay implementations could be hosted on colocated clouds (cf. Table 1).

[2] The facility mapping algorithm of [24] may yield more than one facility for a single IP, due to inability to converge; therefore we select only active single-facility mappings to eliminate the possibility of using the wrong facility.

[3] Despite the number of sampled *PLR* being smaller than the corresponding *COR* samples due to issues with the availability of functional PlanetLab nodes, both relay sets have geo-presence at a comparable number of sites (~60).

distance $d(n_1, n_2)$ between them and then the propagation delay $t(n_1, n_2) = d(n_1, n_2)/(c * \frac{2}{3}$, for the speed of light in an optical fiber [50]). If $RTT(n_1, n_2)$ is the measured RTT between the two endpoints, we keep only the *feasible* relays f that satisfy:

$$2 * [t(n_1, f) + t(f, n_2)] \leq RTT(n_1, n_2)$$

2.5 Measurement Framework

We measure RTT as the metric for inter-domain path latency, using pings between the following pairs of nodes.

Endpoint-to-endpoint. Pairs of *RAE* nodes to measure latency of direct, BGP-derived, Internet paths.

Endpoint-to-relay. Pairs of *RAE* and relay nodes to measure latency on overlay links, which may be stitched to form alternative paths from/to endpoints traversing a relay (i.e., relayed paths). Relays can be *COR*s, *PLR*s, and *RAR*s.

First, measuring the RTTs via pings between each pair of nodes for both directions, we observed that the direction of the ping does not affect the RTT. For example, for ~80% of the *RAE2RAE* cases, the difference between initiating the ping from one node instead of its counterpart does not exceed 5%, while it is averaged out to ~0% due to our randomized pair selection strategy.

We base our measurements on the following principles: (i) work under the RA measurement constraints [7], (ii) amortize timing differences between pseudo-parallel measurements, due to self or external interference [27] and lack of synchrony, via randomized setups, use of median[4] values, and long repeated experiments. We thus schedule measurements between endpoints as well as endpoints and feasible relays, via the RIPE Atlas API [8], repeating a 4-step workflow (*round*) every 12 hours (20 April - 17 May 2017) to account for diurnal patterns. The basic measurement pattern lasts 30 minutes; this window was chosen large enough to account for RTT variability, and small enough to encapsulate a sufficiently correlated batch of measurements. During this window, pings in 5-minute intervals are sent between each pair of nodes, generating an adequate number of measurements (6 per pair) to properly evaluate the associated median RTTs. The workflow steps are:

(1) Select the *RAE* set (see Section 2.1).
(2) For each possible *RAE* pair, measure the RTT on the direct path via single-packet pings. We repeat this process 6 times as mentioned above, and calculate the median RTT per *RAE* pair. In a time slot of 30 minutes, we send 6 consequent ping packets per pair with a time interval of 5 minutes.
(3) Select a set of feasible relays per type. We apply the selection methodology of Section 2.2 for *COR*, 2.3.1 for *PLR*, and Section 2.3.2 for *RAR*. To find only the feasible relays per (RAE_1, RAE_2) pair we use the methodology of Section 2.4, based on the median RTTs on the (RAE_1, RAE_2) paths calculated during Step (2).
(4) For each (RAE_1, RAE_2) pair extracted from the *RAE* set of Step (1), and using the relays from Step (3), we measure the RTT between (RAE_1, RAE_2)[5], $(RAE_1, Relay)$, and $(RAE_2, Relay)$ pairs, via single-packet pings. We repeat this process 6 times, with a time interval of 5 minutes, and calculate the median RTT per pair, based on at least 3 valid RTTs within a measurement window; thus allowing for meaningful median values. To infer the median RTT of a relayed path $(RAE_1, Relay, RAE_2)$ we stitch the associated median RTTs of $(RAE_1, Relay)$ and $(RAE_2, Relay)$.

[4]Since RA is a best-effort measurement infrastructure w.r.t. synchrony and concurrency [27], we use batches of measurements and search for representative pairwise RTTs. To avoid distorting the results with heavy outliers (which exist), we use the median, instead of the average, as a robust metric to represent each batch.

[5]The latency measurement for direct and relayed paths should be in sync. Thus, for each (RAE_1, RAE_2) pair we recalculate the RTT on the corresponding direct path.

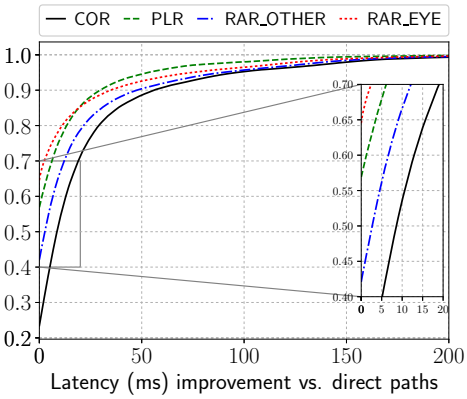

Figure 2: CDF of latency differences (RTT) *vs.* direct paths for the best relays (inducing minimal latency) per type per *RAE* pair. Improvements between 1 and 200ms are shown (83% of total cases). A few cases can reach up to 660ms.

3 MEASUREMENT RESULTS

We ran the measurement workflow of Section 2.5 for 45 rounds (20/4-17/5/2017), sending ~8.7M pings in total. We found ~84% of the destinations of the involved node pairs to be responsive with ≥3 ping replies per round. Next, we describe the most important insights related to the latency-wise performance of ~29 million studied relayed paths vs. ~90K direct paths.

Latency Improvements per Type. Fig. 2 displays the CDF of the latency differences (in ms) of the best-performing (i.e., inducing the least latency) relays per type, vs. the direct paths over the entire set of measurements. We show the improved cases (83% of total), where the relays yield lower-latency paths. We note that *COR* paths perform better than direct in 76% of the total cases, *RAR_other* in 58%, *PLR* in 43% and *RAR_eye* in 35%. The latency improvements range from 1 to 200ms. A few outliers, such as communications involving very distant countries can witness even larger improvements[6]. Median improvements range between 12 and 14ms for all types. *COR* and *RAR_other* yield improvements >100ms (which are critical for e.g., application service providers [37]) in 6% of the improved cases (5% of total). These gains stem solely from the discovery of fast TIV-enabled paths, and do not consider other sources of latency that cut through the network stack [14]. Note that *RAR_eye* and *PLR* have very similar (low) performance, while *RAR_eye* and *RAR_other* differ significantly; the latter supports our intuition of differentiating between the two *RAR* types. The difference between the best (*COR*) and the second-best overlay (*RAR_other*) does not surpass 5-10ms, for the cases where both perform better than the direct paths. Therefore, the most important

[6]E.g., a path involving Colombia and Slovakia, observed reductions of 660ms when relayed through large European Colos.

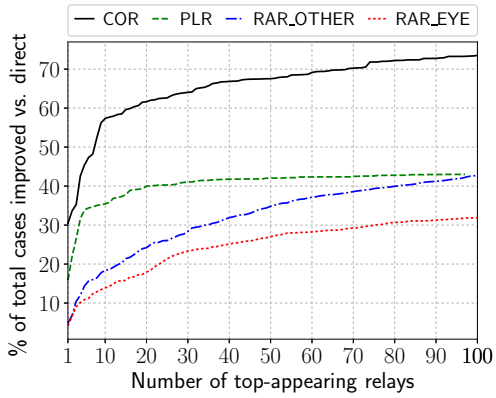

Figure 3: % of total cases (pairwise communications) where relayed paths improve latency against direct paths, *vs.* number of top relays (cut at top-100 relays for clarity).

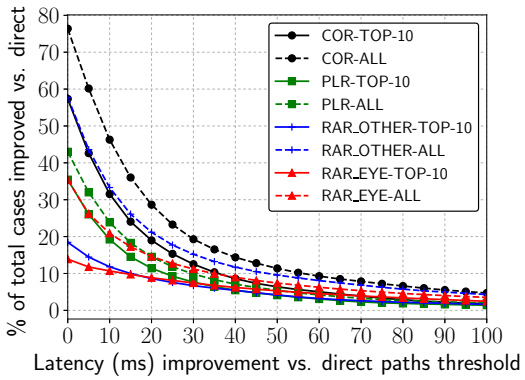

Figure 4: % of total cases (pairwise communications) where relayed paths improve latency against direct paths (top-10/all relays), *vs.* improvement threshold (cut at 100 ms). The best performance of each relay set is considered per case.

difference w.r.t. performance is the percentage of cases where relays actually yield better-than-direct paths, while the improvement itself does not vary significantly across types. We further calculate a median of 8 *COR*, 3 *PLR*, 2 *RAR_other* and 2 *RAR_eye* relays that yield improvements for each (RAE_1, RAE_2) pair, indicating a high redundancy of *COR* relays.

How Many Relays are Enough? We next show what is the maximum benefit we can achieve per relay type, for a given number of relays. Fig. 3 shows the percentage of improved (RAE_1, RAE_2) pairs (out of the total cases) vs. the number of top relays (ranked according to their frequency of improvement) employed to achieve those improvements. The number of improved pairs increases rapidly with the first few *COR* and *PLR* relays; in particular, the top *COR* relays are extremely beneficial (heavy hitters). In contrast, the number of improved pairs increases more smoothly for *RAR* relays, which require >>100 relays to yield their top improvements (58% of total cases, beyond x-axis bounds of Fig. 3). Overall, *COR* improve many more pairs with fewer relays. Specifically, 10 *COR* relays (in 6 Colos) with latency improvement in ~75% of the improved cases (58% of total) match the second-best performance (58% of total cases for *RAR_other*), which requires though >>100 *RAR_other*. After the first 10 relays, the incremental benefit of additional *COR* is decreasing fast. Fig. 4 shows the percentages of improved pairs (out of the total cases) vs. the threshold of latency reduction that they surpass, when employing the top-10 and all relays of each type, respectively. The best performance of each relay set is considered per case. We see that the top-10 *COR* perform better than the top-10 relays of all the other types, and follow closely the performance of all *RAR_other* relays (second-best performance after all *COR*). The gaps between top-10 and all relays differ per type; e.g., for *PLR*, this gap is minimal (~5%), indicating very few well-performing relays among the relay set. Interestingly, using only the top-10 *COR* relays, ~20% of all pairs witness latency reductions larger than 20ms; the cost of this limited selection is only 30% less –relative to all *COR*– pairs surpassing 20ms (10% being the exact percentage of "missing" total pairs). We next examine the features of the top COR relays.

Features of Top Facility Relays. Table 1 shows the facilities hosting the 20 top-appearing *COR* relays, i.e., the ones with the highest frequency of improvement vs. direct paths. We augmented the facility data using information from PeeringDB [6]. Interestingly, we note that only 10 facilities actually contain the top-20 relays, and 4 of them are in the top-10 of PeeringDB w.r.t. the number of colocated networks they host. All of them are colocated with at least 2 IXPs, with one facility (Equinix Frankfurt) connected with 11 IXPs. At least 22 networks (ISPs, content/cloud providers, etc.) are colocated at each facility, with one (Telehouse North London) hosting 361 networks. All top facilities are either offering cloud services themselves, or are colocated with cloud providers; these data centers could be used e.g., to host VM-based relays [15, 43]. In addition, the top facility relays reside in large metropolitan centers (and Internet hubs), mainly in Western Europe and North America.

Changing Countries and Paths. Path inflation [51], induced by BGP policies, can lead to increased inter-domain latency between remote endpoints (residing, e.g., in different countries). We expect that this effect may prevent relays that are located close to the endpoints (e.g., same country) from using alternate, non-inflated paths. To verify this assumption, we compare relayed paths when the relay changes (and does not change) country w.r.t. the endpoints; we consider the min-latency relays per case. We observed that when the relay is in a different country than both endpoints[7], latency is lower than the direct path in 75% of the cases for *COR*. In contrast, this number drops to 50% if the relay is in the same country as one of the endpoints. Similar remarks apply for the other types, albeit with lower percentages. Also, out of the totally studied (~90K) *RAE* pairs (and related direct paths), 74% are inter-continental, indicating a set conducive to path inflation. Indeed, a significant fraction (19%) of the total direct paths, turn up with RTTs that exceed 320ms (considered as threshold for poor VoIP performance [19, 28]). These values are in line with the work of Jiang *et al.* [29]. By employing only *COR* relays, the fraction of paths over 320ms falls to 11%.

Stability over Time. Regarding the evolution of the results *in time*, we observed a consistent pattern, with *COR* finding lower-latency paths in >75% of the cases, *RAR_other* in >50%, and *RAR_eye* and *PLR* having a positive impact in less than 50% of the cases in every measurement round. In general, in ~50% of the cases we found a 1-20ms improvement for *COR*. The cumulative insights

[7] Endpoints are located in different countries based on the selection of Section 2.1.

Table 1: Facilities of top-20 Colo relays (ranked according to their frequency of presence in improved paths), and their location and connectivity characteristics.

Facility Name (PDB ID)	% of Improved Cases	City (Country)	#Nets	#IXPs	Cloud Services	PDB top-10
1) Telehouse North (34)	47	London (GB)	361	6	✓	✓
2) Equinix-AM7 (62)	46	Amsterdam (NL)	184	4	✓	✓
3) Nikhef (18)	34	Amsterdam (NL)	151	6	✓	✗
4) Equinix-FR5 (60)	30	Frankfurt (DE)	235	11	✓	✓
5) Telehouse West (835)	29	London (GB)	89	5	✓	✗
6) Digital Realty Telx (125)	29	Atlanta (US)	125	2	✓	✗
7) Incolocate (105)	29	Hamburg (DE)	22	3	✓	✗
8) Interxion (68)	27	Brussels (BE)	58	3	✓	✗
9) Digital Realty Telx (10)	22	New York (US)	112	5	✓	✗
10) Equinix-LD8 (45)	21	London (GB)	208	4	✓	✓

from Fig. 2 seem to apply consistently in time. In fact, to further investigate the temporal stability of our observations, we calculated the Coefficient of Variation (CV) for all the direct and relayed pairs for all measurements, as the standard deviation of the median RTTs of each pair divided by the pair's average of medians over time. We observed that the CV ranges from 0% to 40%, and is less than 10% in 90% of the cases. This indicates stable, usable overlays.

4 RELATED WORK

Researchers have tinkered with the idea of exploiting TIVs to improve inter-domain routing for the last two decades. After the early pioneers with the Detour Framework [47, 48], Andersen *et al.* [12] introduce Resilient Overlay Network (RON) to form resilient and—potentially faster—paths compared to the default BGP paths. Similar insights remain timely [26], albeit by exploiting inter-continental cloud-terminated paths.

VIA [29] aims at improving Internet telephony by employing classic overlay techniques to relay calls. They show that an oracle-based overlay can potentially improve up to 53% of calls whose quality is impacted by poor network performance. Their relay selection strategy uses call history information, and is based on the empirical observation that even though a prediction-based approach may not identify the optimal relay, it is likely to exist in the top few predicted relays.

Regarding the number of relays per relayed path, useful insights are provided by Han *et al.* [25] and Le *et al.* [34]. Both works support our approach to consider only 1-relay paths as adequate to reduce latency, compared to N-relay paths ($N \geq 2$).

ARROW [44], an inter-domain routing approach based on waypoints (i.e., relay routers within ISPs), allows users to set up reliable and secure e2e tunnels. While for about 20% of the ARROW cases, the e2e latency increases (up to 20%), the performance of most paths may actually be improved when routed via ARROW waypoints. On the other hand, Lumezanu *et al.* [39, 40] analyze TIVs, concluding that even though faster inter-domain paths exist, their utilization can be prevented by business drivers of the ISPs themselves.

MeTRO [43] aims to offer QoS between endpoints, using virtual routers hosted in Amazon EC2 [1] and Bright Box [2] data centers as cloud relays. Latency improvements exist for 58% of the cases, while the best performing relays are close to large IXPs. In contrast to our work, no extensive comparison of overlay positioning is performed, to understand the location impact on the relay selection strategy. Similarly to MeTRO, Cai *et al.* [15] propose cloud-routed overlay networks (CRONets), to maximize throughput. Results show that CRONets consistently help for 78% of the cases. While MeTRO and CRONets relays are cloud-hosted, one of our goals is to suggest a large-scale methodology for measuring inter-domain paths passing through *diverse* relay types. To this end, we exploit pingable IP addresses of interfaces located in Colos, and we concatenate the latency of individual hops (endpoint to relay to endpoint). Since these interfaces do not have to be under our administrative control, they are not associated with any costs, therefore our methodology can scale seamlessly.

In summary, we identify a tendency towards inter-domain overlay networks, using relays in data centers [15, 29, 34, 43], ISPs [25, 44], or at the last mile [12, 26]. By exploiting TIVs [39, 40] to reduce inter-domain latency, results show an improvement of latency metrics when overlay paths are employed, as compared to direct BGP-based paths. It is worth mentioning that the use of overlays requires a delicate balance between overlay-based optimization and policy-driven TE (e.g., on the enterprise level [35]), to avoid potential policy conflicts [31, 38, 46] with monetary impact. However, our work focuses on strategically constructing and evaluating relayed paths for end-users and application providers; in particular, employing relays at Colos, not explored in previous works.

5 CONCLUSIONS & FUTURE WORK

The Internet is changing. Requirements for low-latency video distribution have driven the flattening of the Internet topology in recent years, resulting in short distances and dense fabrics at interconnection facilities [32]. This effect coupled with the emergence of numerous cloud providers, residing at Colos, is opening up the largest Colos of the Internet to end-users and application service providers, who can now easily host their services there. In this paper, we ask how this *democratization* of large Colos affects latency for services that use relays. We performed an Internet-wide measurement study, spanning 1 month, employing different types of relays which serve endpoints located at the last mile. We showed that Colos are useful locations to host relays, taking advantage of their high connectivity and core locations to discover low-latency TIV paths that are faster than the direct ones. A few Colo-based relays are found to improve many more (>20%) of the studied cases than one order of magnitude more PlanetLab or eyeball relays.

Future Work. We plan to investigate the following:

(i) The key factor(s) due to which Colos perform so well as relays. Even though a preliminary analysis has already been conducted in this work, the exact root-causes remain subject to further research.

(ii) The underlying reasons for the relatively good performance of *RAR_other* relays. RIPE Atlas is known to have a significant deployment even in commercial (core) networks. We plan to further examine the networks where these nodes are present.

(iii) Regional effects uncovered via traceroute measurements. For example, we intend to investigate potential correlations between the characteristics of the countries traversed by relayed paths and the achieved latency, as well as between the latency and the proximity of endpoints/relays to submarine cable landing points [53].

Software and Datasets. The software used to run, analyze and visualize the measurements presented in this paper is publicly available, together with the collected measurement data [30].

Acknowledgements. This work has been funded by the EU Research Council Grant Agreement no. 338402.

REFERENCES

[1] Amazon EC2. https://aws.amazon.com/ec2. Accessed: 15.03.2017.

[2] Brightbox. https://www.brightbox.com/. Accessed: 15.03.2017.

[3] DigitalOcean. https://www.digitalocean.com/. Accessed: 15.03.2017.

[4] Equinix Global Data Centers & Colocation Services. http://www.equinix.com/locations/. Accessed: 15.03.2017.

[5] Maxmind GeoIP2 City Accuracy. https://goo.gl/JFgv9r. Accessed: 15.03.2017.

[6] PeeringDB. https://www.peeringdb.com/. Dataset collected on: 28.03.2017.

[7] RIPE Atlas - User-Defined Measurements. https://atlas.ripe.net/docs/udm/. Accessed: 24.03.2017.

[8] RIPE Atlas API Reference. https://atlas.ripe.net/docs/api/v2/reference/. Accessed: 24.03.2017.

[9] RIPE ATLAS Home. https://atlas.ripe.net/. Accessed: 15.03.2017.

[10] RIPE Atlas: List of Anchors. https://atlas.ripe.net/anchors/list/. Accessed: 15.03.2017.

[11] Amir, Y., Danilov, C., Goose, S., Hedqvist, D., and Terzis, A. An overlay architecture for high-quality VoIP streams. *IEEE Transactions on Multimedia 8*, 6 (2006), 1250–1262.

[12] Andersen, D. G., Balakrishnan, H., Kaashoek, M. F., and Morris, R. The case for resilient overlay networks. In *Proceedings of the Eighth Workshop on Hot Topics in Operating Systems* (2001), IEEE, pp. 152–157.

[13] APNIC. IPv6 Measurement Campaign. https://stats.labs.apnic.net/v6pop. Measurement Methodology: https://labs.apnic.net/measureipv6, Dataset collected on: 31.03.2017.

[14] Bozkurt, I. N., Aguirre, A., Chandrasekaran, B., Godfrey, P. B., Laughlin, G., Maggs, B., and Singla, A. Why Is the Internet so Slow?! In *International Conference on Passive and Active Network Measurement* (2017), Springer, pp. 173–187.

[15] Cai, C. X., Le, F., Sun, X., Xie, G. G., Jamjoom, H., and Campbell, R. H. CRONets: Cloud-Routed Overlay Networks. In *36th International Conference on Distributed Computing Systems (ICDCS)* (2016), IEEE, pp. 67–77.

[16] CAIDA. IPv4 Prefix-to-AS Dataset. http://data.caida.org/datasets/routing/routeviews-prefix2as/. Dataset collected on: 12.03.2017.

[17] Chatzis, N., Smaragdakis, G., Feldmann, A., and Willinger, W. Quo vadis Open-IX? *ACM SIGCOMM Computer Communication Review 45*, 1 (2015), 12–18.

[18] Chun, B., Culler, D., Roscoe, T., Bavier, A., Peterson, L., Wawrzoniak, M., and Bowman, M. Planetlab: an overlay testbed for broad-coverage services. *ACM SIGCOMM Computer Communication Review 33*, 3 (2003), 3–12.

[19] Cisco. Quality of Service for Voice over IP. https://goo.gl/eC5Qnb. Accessed: 15.05.2017.

[20] Dhamdhere, A., and Dovrolis, C. The Internet is flat: modeling the transition from a transit hierarchy to a peering mesh. In *Proceedings of the 6th International COnference* (2010), ACM, p. 21.

[21] Giotsas, V., Dhamdhere, A., and Claffy, K. C. Periscope: Unifying looking glass querying. In *International Conference on Passive and Active Network Measurement* (2016), Springer, pp. 177–189.

[22] Giotsas, V., Dietzel, C., Smaragdakis, G., Feldmann, A., Berger, A., and Aben, E. Detecting Peering Infrastructure Outages in the Wild. In *Proceedings of ACM SIGCOMM* (2017), ACM, pp. 446–459.

[23] Giotsas, V., Smaragdakis, G., Huffaker, B., Luckie, M., et al. Mapping peering interconnections to a facility - Supplemental Material. https://goo.gl/4JnzV7. Accessed: 15.03.2017, Original dataset collected in 2015.

[24] Giotsas, V., Smaragdakis, G., Huffaker, B., Luckie, M., et al. Mapping peering interconnections to a facility. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies* (2015), ACM, p. 37.

[25] Han, J., Watson, D., and Jahanian, F. Topology aware overlay networks. In *Proceedings of IEEE INFOCOM* (2005), vol. 4, IEEE, pp. 2554–2565.

[26] Haq, O., Raja, M., and Dogar, F. R. Measuring and Improving the Reliability of Wide-Area Cloud Paths. In *Proceedings of the 26th International Conference on World Wide Web* (2017), International World Wide Web Conferences Steering Committee, pp. 253–262.

[27] Holterbach, T., Pelsser, C., Bush, R., and Vanbever, L. Quantifying interference between measurements on the RIPE Atlas platform. In *Proceedings of the Internet Measurement Conference* (2015), ACM, pp. 437–443.

[28] ITU. G.114: ITU Recommendation of One-way Transmission Time. https://www.itu.int/rec/T-REC-G.114/en. Accessed: 15.05.2017.

[29] Jiang, J., Das, R., Ananthanarayanan, G., Chou, P. A., Padmanabhan, V., Sekar, V., Dominique, E., Goliszewski, M., Kukoleca, D., Vafin, R., et al. Via: Improving internet telephony call quality using predictive relay selection. In *Proceedings of ACM SIGCOMM* (2016), ACM, pp. 286–299.

[30] Kotronis, V., Nomikos, G., Manassakis, L., Mavrommatis, D., and Dimitropoulos, X. Shortcuts through Colocation Facilities Project webpage, with links to software and datasets used in the paper. http://inspire.edu.gr/shortcuts_colocation_facilities/.

[31] Koutsoupias, E., and Papadimitriou, C. Worst-case equilibria. *Computer science review 3*, 2 (2009), 65–69.

[32] Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., and Jahanian, F. Internet inter-domain traffic. In *ACM SIGCOMM Computer Communication Review* (2010), vol. 40, ACM, pp. 75–86.

[33] Laughlin, G., Aguirre, A., and Grundfest, J. Information transmission between financial markets in Chicago and New York. *Financial Review 49*, 2 (2014), 283–312.

[34] Le, F., Nahum, E., and Kandlur, D. Understanding the Performance and Bottlenecks of Cloud-Routed Overlay Networks: A Case Study. In *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking* (2016), ACM, pp. 7–12.

[35] Lee, G. M., and Choi, T. Improving the interaction between overlay routing and traffic engineering. In *International Conference on Research in Networking* (2008), Springer, pp. 530–541.

[36] Levin, D., Lee, Y., Valenta, L., Li, Z., Lai, V., Lumezanu, C., Spring, N., and Bhattacharjee, B. Alibi routing. In *ACM SIGCOMM Computer Communication Review* (2015), vol. 45, ACM, pp. 611–624.

[37] Linden, G. Amazon Found Every 100ms of Latency Cost Them 1% in Sales. https://goo.gl/wO6d7X. Accessed: 15.03.2017.

[38] Liu, Y., Zhang, H., Gong, W., and Towsley, D. On the interaction between overlay routing and underlay routing. In *Proceedings of IEEE INFOCOM* (2005), vol. 4, IEEE, pp. 2543–2553.

[39] Lumezanu, C., Baden, R., Spring, N., and Bhattacharjee, B. Triangle inequality and routing policy violations in the Internet. In *International Conference on Passive and Active Network Measurement* (2009), Springer, pp. 45–54.

[40] Lumezanu, C., Baden, R., Spring, N., and Bhattacharjee, B. Triangle inequality variations in the Internet. In *Proceedings of ACM SIGCOMM* (2009), ACM, pp. 177–183.

[41] Ly, C., Hsu, C.-H., and Hefeeda, M. Improving online gaming quality using detour paths. In *Proceedings of the 18th ACM international conference on Multimedia* (2010), ACM, pp. 55–64.

[42] Ly, C., Hsu, C.-H., and Hefeeda, M. IRS: A detour routing system to improve quality of online games. *IEEE Transactions on Multimedia 13*, 4 (2011), 733–747.

[43] Makkes, M. X., Oprescu, A.-M., Strijkers, R., de Laat, C., and Meijer, R. MeTRO: low latency network paths with routers-on-demand. In *European Conference on Parallel Processing* (2013), Springer, pp. 333–342.

[44] Peter, S., Javed, U., Zhang, Q., Woos, D., Anderson, T., and Krishnamurthy, A. One tunnel is (often) enough. *ACM SIGCOMM Computer Communication Review 44*, 4 (2015), 99–110.

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

[46] Qiu, L., Yang, Y. R., Zhang, Y., and Shenker, S. On selfish routing in Internet-like environments. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications* (2003), ACM, pp. 151–162.

[47] Savage, S., Anderson, T., Aggarwal, A., Becker, D., Cardwell, N., Collins, A., Hoffman, E., Snell, J., Vahdat, A., Voelker, G., et al. Detour: Informed Internet routing and transport. *Ieee Micro 19*, 1 (1999), 50–59.

[48] Savage, S., Collins, A., Hoffman, E., Snell, J., and Anderson, T. The end-to-end effects of Internet path selection. 289–299.

[49] Shavitt, Y., and Zilberman, N. A geolocation databases study. *IEEE Journal on Selected Areas in Communications 29*, 10 (2011), 2044–2056.

[50] Singla, A., Chandrasekaran, B., Godfrey, P., and Maggs, B. The Internet at the speed of light. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks* (2014), ACM, p. 1.

[51] Spring, N., Mahajan, R., and Anderson, T. The causes of path inflation. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications* (2003), ACM, pp. 113–124.

[52] Sundaresan, S., Feamster, N., and Teixeira, R. Home network or access link? locating last-mile downstream throughput bottlenecks. In *International Conference on Passive and Active Network Measurement* (2016), Springer, pp. 111–123.

[53] TeleGeography. Submarine Cable Map. https://www.submarinecablemap.com/. Accessed: 11.09.2017.