# Beating BGP is Harder than we Thought

Todd Arnold[†] Matt Calder[‡] Italo Cunha[♮†] Arpit Gupta[♯†]
Harsha V. Madhyastha[★] Michael Schapira[◇] Ethan Katz-Bassett[†]

[†] Columbia University [‡] Microsoft [♮] Universidade Federal de Minas Gerais [♯] UC Santa Barbara
[★] University of Michigan [◇] Hebrew University of Jerusalem

## ABSTRACT

Online services all seek to provide their customers with the best Quality of Experience (QoE) possible. Milliseconds of delay can cause users to abandon a cat video or move onto a different shopping site, which translates into lost revenue. Thus, minimizing latency between users and content is crucial. To reduce latency, content and cloud providers have built massive, global networks. However, their networks must interact with customer ISPs via BGP, which has no concept of performance.

The shortcomings of BGP are many and well documented, but in this paper we ask the community to take a step back and rethink what we know about BGP. We examine three separate studies of performance using large content and cloud provider networks and find that performance-aware routing schemes rarely achieve lower latency than BGP. We lay out a map for research to further study the idea that beating BGP may be more difficult than previously thought.

## CCS CONCEPTS

• **Networks** → **Routing protocols**; **Control path algorithms**; **Network performance analysis**;

## KEYWORDS

BGP, performance, traffic engineering, content delivery

**ACM Reference Format:**
Todd Arnold, Matt Calder, Italo Cunha, Arpit Gupta, Harsha V. Madhyastha, Michael Schapira, and Ethan Katz-Bassett. 2019. Beating BGP is Harder than we Thought. In *HotNets '19: ACM Workshop on Hot Topics in Networks, November 13–15, 2019, Princeton, NJ, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3365609.3365865



*HotNets '19, November 13–15, 2019, Princeton, NJ, USA*

ACM ISBN 978-1-4503-7020-2/19/11…$15.00
https://doi.org/10.1145/3365609.3365865

## 1 INTRODUCTION

Online services deeply care about offering low latency access to their clients, as studies have found that latency strongly correlates with clients' QoE [17]. Therefore, many large content and cloud providers have infrastructure spanning the globe with Points of Presence (PoPs) at many locations, interconnected by a private Wide Area Network (WAN). These providers also connect with many Autonomous Systems (ASes) at each PoP in order to ensure rich connectivity with the public Internet.

To make the most of such a global deployment, optimizing any client's perceived performance when communicating with a service depends on several decisions that the service provider makes regarding routing traffic from the client to the service and back. These decisions include:

- Given PoPs in many locations, the provider must determine which PoP it should direct any particular client's request to.
- At the PoP from which the service's response egresses the provider's network, given the connectivity to many ISPs at each PoP, the service provider must select which among multiple routes from that PoP to the client the response should be forwarded along.
- Lastly, in cases where the client is distant from the PoP at which her request is served, the provider must choose whether to route the response across its private WAN to egress at a PoP near the client, as opposed to relying on the public Internet to forward the response back from the serving PoP.

By default, these decisions are driven by the Border Gateway Protocol (BGP), the Internet's de-facto routing protocol. However, BGP's strategy for selecting from a set of routes is performance-oblivious. As prior studies have noted [18, 28], none of the criteria BGP uses for selecting among paths (*e.g.,* prefer peering over transit, prefer paths with fewer AS-level hops, do hot potato routing, *etc.*) directly correlate with performance. Many studies from over a decade ago explored the inefficiency of BGP routing. These revealed that an alternate path had lower latency than the one chosen by BGP 30–80% of the time [22], with a reduction of 20% on average, and substantially more in many cases [1, 2, 20–22].

Over the last several years, many large content and cloud providers (such as Facebook, Google, and Microsoft) have put in significant effort to route traffic from/to clients in

a manner that offers better performance than what BGP would by default. Examples include Facebook's Edge Fabric for selecting among egress routes at any of their PoPs [25]; Akamai and Google using DNS to direct clients to a nearby PoP [11, 32], rather than relying on BGP anycast [6]; and Google's use of Espresso [31] to route cloud traffic via its private WAN as opposed to the public Internet.

While descriptions and evaluations of some of these efforts to outdo BGP have been described in separate full-length papers [6, 11, 24, 25, 31], our goal in this paper is to point out a rather surprising takeaway that is easy to miss when looking at each of these papers in isolation: across settings, BGP does not fare that badly after all in terms of selecting low latency paths. We rely upon three large-scale studies that, in three different settings, isolate the benefit over BGP of performance-aware routing systems developed by some of the largest content providers on the Internet. We conducted two of the studies in collaboration with the providers and published them previously [6, 24], with the results in this paper representing a small fraction of the papers' analyses. We conduct a third supplemental study.

The three studies exercise different stages of content serving, and our overarching observation is that, in all cases, performance-aware routing provided *little benefit over BGP* in terms of latency. Indeed, put together, the results show that, although the aforementioned routing control systems have some benefit, latency with BGP-selected routes is good enough for the most part. *In hindsight*, and in combination with the hypotheses we posit and explore in this paper, these results may appear obvious. However, they are *not* what we would have *predicted* in advance; in all three cases, we expected to see more advantage to performance-aware decisions over BGP, based on earlier literature and our own understanding of the setting, honed both from our work and from collaborations with large content providers.

In light of the above, we call for the community to revisit the common assumption that one must necessarily work around BGP in order to optimize performance over the Internet. Instead, we ought to ask: despite BGP's path selection being performance-oblivious, why have large content and cloud providers found it hard to beat the performance that BGP offers?

Our goal here is not to present new results to answer this question; most of the results we present are ones already included in papers that describe the routing control systems we reference. Each of the earlier studies goes deeper into architectures and/or analysis of an individual setting, and the contribution of this paper is in considering them holistically. This paper is a first step in enumerating several hypotheses for explaining BGP's admirable ability to offer good performance. For some of these hypotheses, we can tell whether they are true or not based on the characteristics of the settings in which content providers have attempted to outdo BGP. For other potential explanations, we describe open questions, future studies that must be conducted to address them, and associated challenges.

## 2 SETTING THE STAGE

To improve performance for end-users of their services, large content and cloud providers have built out private infrastructure. To serve clients over short geographic distances and limit propagation delay, they host servers at locations worldwide and strive to serve clients from nearby [4]. They build out their own private WANs to interconnect their server locations and to extend their network to maintain control of traffic for much of its route towards end-users [31]. At each location, they interconnect with many networks to establish route diversity and short routes to many end-users [8, 25, 30].

However, building this infrastructure alone does not guarantee low latency performance. The infrastructure provides options (of routes, of servers, of ingress/egress locations into the WANs on paths between clients and servers), and good use of the options is required for low latency.

### 2.1 BGP's Route Selection is Performance-Oblivious

A challenge with making good use of the options is that Internet routing is not performance-aware. First, no one entity has complete control or visibility over the end-to-end route. Second, BGP, the Internet's inter-domain routing protocol, is oblivious to performance and performance changes [25]. BGP's obliviousness creates well-known challenges to each choice of how to use provider infrastructure. BGP anycast can be used to steer a client to one of the many server locations, but it is known to not always pick nearby servers [18]. BGP can choose circuitous routes between the edge of the provider's network and the client [28]. A router at the edge can select a path that is persistently bad or experiencing transient problems or congestion [25].

### 2.2 Providers Build Performance-Aware Control Systems

To sidestep the performance-obliviousness of BGP routing, content providers built sophisticated control systems that use performance measurements to serve clients via low-latency options. Akamai [11], Google [4], Facebook [25], and other providers use unicast to route a client's request to a particular serving site, rather than using anycast and relying on BGP to pick. For example, the provider can measure performance from clients to different servers by spraying background requests [7], then use their authoritative DNS servers to resolve a client lookup to the unicast address of a server found to perform well for that client [7]. Providers like Google built controllers that direct outgoing responses to particular egress routers and routes, using performance measurements

to inform the decisions [31]. Coupled with Google's worldwide WAN, such a system can supplant BGP's performance-oblivious routing with performance-aware routing for much of the path from server to client.

## 2.3 Evaluating Performance-Aware Routing

To isolate the benefit of performance-aware routing over BGP, we rely upon datasets from three settings where we evaluated performance-aware routing in different stages of content serving. We describe the settings here and defer descriptions of datasets and analyses to §3.

*2.3.1 Performance-aware route selection at each PoP.* We first consider Facebook, a content provider with dozens of PoPs around the world. Most clients are close to a PoP; based on geolocation of clients, half of all traffic is to clients within 500km of the serving PoP, which translates to as little as 5ms RTT, and 90% is to clients within 2500km and on the same continent [24]. For most clients, the PoP serving the client has at least three routes to the client's prefix: routes announced by one or more peers, and routes announced by two or more transit providers. Facebook employs a traffic monitoring and management system to enable performance-aware routing, which may override the performance-agnostic routing of BGP [25].

*2.3.2 DNS redirection to outperform anycast.* Second, we consider BGP's performance for an anycast Content Delivery Network (CDN), where it must select between multiple sites, some of which may be near and others far. We examine the dataset from an earlier paper that instrumented Microsoft's CDN in 2015, when it had a few dozen front-end server locations [6]. Microsoft uses anycast in production, announcing the same prefix from all locations, and BGP steers a client request to a particular front-end location. Most clients have several nearby front-ends: the median distance of the nearest front-end is 280 km, of the second nearest is 700 km, and of fourth nearest is 1300 km. The earlier study injected measurements into Bing results directing clients to fetch objects from multiple unicast server locations. It used these measurements to compare the performance of anycast to measurement-driven schemes that use DNS to direct clients to the best performing unicast server.

*2.3.3 Use of private WAN to beat the public Internet.* Lastly, we report on a recent measurement study we conducted to evaluate Google's tiered networking services, which gives us a unique opportunity to quantify how the performance of routes across the public Internet compares with ones that utilize the cloud provider's private WAN. Google offers two networking tiers for its cloud services: (1) Premium Tier, in which it uses its WAN to ingress/egress traffic near to the client, and (2) Standard Tier, in which it forces traffic to ingress/egress near the cloud data center and use the public Internet the rest of the way [15].

# 3 WHEN AND WHY IS BGP HARD TO BEAT?

In all three of the settings above, measurements summarized in this section reveal that the efforts to outperform BGP showed little latency benefit. Next, we examine the datasets and consider several potential explanations. Where available, we discuss results which either support or rule out certain explanations. In other cases, we describe open questions that should be addressed to test the hypotheses we present and the challenges involved in conducting these studies.

## 3.1 Performance-Aware Routing Provides Little Benefit over BGP (at a PoP)

To analyze the first setting, in which performance-aware routing is performed independently at every PoP (§2.3.1), we borrow a dataset and analysis from a recent measurement study of traffic at Facebook [24]. The dataset contains ten days of data from load balancers at all of Facebook's PoPs around the world. A sampled subset of client HTTP sessions are sprayed across different egress routes, including BGP's most preferred, second-most preferred, and third-most preferred path that a PoP has to each client prefix. These measurements let us compare the performance of BGP's preferred route versus an omniscient performance-aware route controller that always uses the path with the best instantaneous performance. The dataset contains TCP's MinRTT measurement for hundreds of trillions of HTTP sessions from billions of unique client IP addresses spread across more than 200 countries. Within each 15 minute window, we group the measurements by ⟨PoP, prefix, route⟩ to find the median MinRTT for each route and weigh the results by total traffic volume (bytes transferred).

The route selected by BGP performs the best for most traffic. Facebook's standard (performance-agnostic) BGP routing policy prefers private peers with dedicated capacity first, then public peers, and finally transit providers; and chooses shorter paths over longer ones [24]. The line in Figure 1 shows the difference in performance of the route preferred by this BGP policy to the performance of the best performing alternate route, and the shaded region shows the distribution of the lower and upper bounds of the confidence intervals around the performance difference. Values $< 0$ indicate a preferred BGP route that outperforms alternate routes. For the vast majority of traffic, BGP performs better than or roughly as well as the best alternative. Median MinRTT can be improved by 5ms or more for only 2–4% of traffic. We find qualitatively similar results for bandwidth (not shown).

*3.1.1 BGP is good enough when all route options degrade together.* From a specific PoP to a client prefix, a performance-aware routing control system can do better than BGP either temporarily (*e.g.,* when the path chosen by BGP is congested) or always (*e.g.,* BGP selects a path that traverses a longer distance). To tease apart these two cases, for each ⟨PoP, prefix⟩

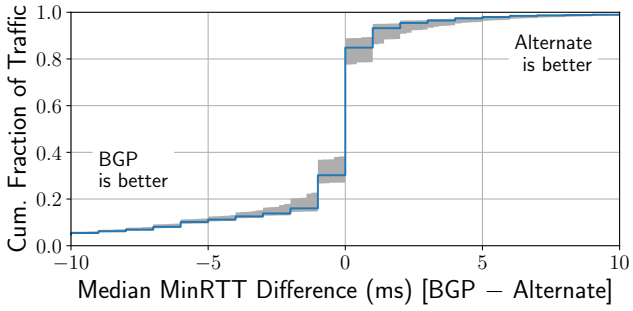


**Figure 1: Possible median latency improvement by routing traffic over alternate routes. Positive values mean the best-performing alternate path has lower latency than BGP's preferred path.**

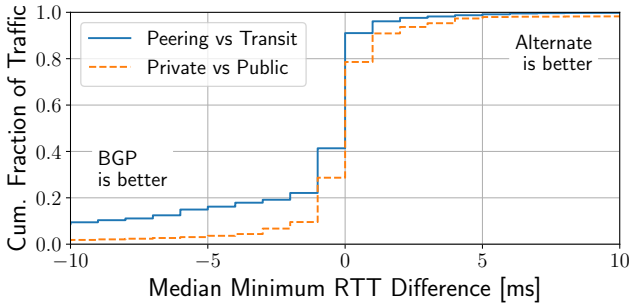


**Figure 2: Performance difference between peer and transit (solid blue line), and between private and public exchange peers (dashed orange line). Usually, transits have performance similar to that of peers, and routes via public exchange have performance similar to those via private interconnections.**

pair, we identify the duration for which performance-aware routing significantly improves performance over BGP. This viewpoint reveals the following:

- First, as seen in Figure 1, alternate paths usually offer the same latency as the path chosen by BGP.
- Second, periods of performance degradation on paths preferred by BGP (relative to a path's baseline performance) are more prevalent than opportunities to improve performance by routing over alternate paths.
- Third, most alternate paths which do beat BGP are consistently better all the time.

Although these observations were made in isolation in the prior work [24], we note that they collectively suggest the following takeaway: for many destinations, whenever the path chosen by BGP experiences congestion, so do other alternative routes. For example, when the destination network is congested, there are no alternate performant paths, and it is impossible for dynamic traffic engineering to improve performance relative to BGP for the (usually few) clients experiencing congestion in the affected networks.

*3.1.2 Direct peering does not fully explain why BGP performs well.* Modern content providers peer widely with ASes hosting many of their clients, allowing them to route much of their traffic over private network interconnects (PNIs) with dedicated capacity directly into these "eyeball' ' ASes [8, 25]. These direct peering paths are preferred over all others by standard BGP policy. One might hypothesize that the reason BGP performs well is because these direct paths make it easy to choose the best-performing path, as the content provider can avoid congesting the dedicated interconnection [25, 31], and the traffic quickly enters a network capable of routing it the whole way to the client. Such routes let a provider invest (in PoPs and in paid peering) to align policy, capacity, and performance. Figure 2 compares the difference in median MinRTT between peer vs transit (solid line) and between private vs public peers (dashed line). The results show that (less preferred) alternate routes usually have performance similar to that of BGP-preferred paths. Capacity limitations aside, this result suggests that BGP would often perform roughly as well even if the direct path was not an option.

*3.1.3 Open question: What is the impact of a reduced peering footprint?* If less preferred paths often perform as well as more preferred ones, a content provider may be able to drastically reduce its number of peers without impacting latency. Such a reduction in peerings could provide operational benefits, as peering with smaller networks can contribute an outsized amount of headaches due to generally lower operational sophistication (*e.g.,* more frequent route leaks and outages, less responsive NOCs). Even with access to measurements from a large content provider, it is not realistic to conduct such a study by temporarily shifting traffic away from a large number of peerings, as peers would complain if they suddenly received traffic via costly upstream providers rather than via the PNI they invested in. A study in emulation would need to properly account for the reduced peering capacity and accompanying increased likelihood of congestion as the number of route options is reduced. A challenge common to this and many of the open questions we see is that they are hard to address from outside of a large content provider. The PEERING [23] and CloudLab [9] testbeds could help some, but the community may need new testbeds, new methodologies, and increased data sharing from industry to fully understand this important setting.

## 3.2 BGP's effectiveness is not limited to settings where all paths are short

In the data for Figure 1, half the traffic travels an "as the crow flies' ' distance of at most 500km, which could be covered in as little as 5ms RTT. In such a setting, paths may traverse multiple inter-AS links without leaving the metropolitan area or sometimes even the colocation facility. A possible explanation for BGP's good performance is that it does well when most of its options cover very short geographic distances, but may not otherwise.

To evaluate this explanation, we now consider data from Microsoft's analysis of whether DNS-based redirection does

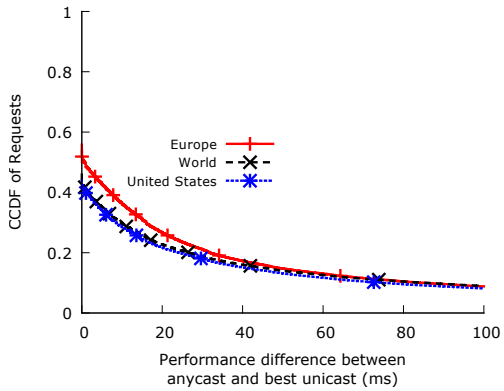


**Figure 3: The fraction of search results where a unicast front-end outperforms anycast.**

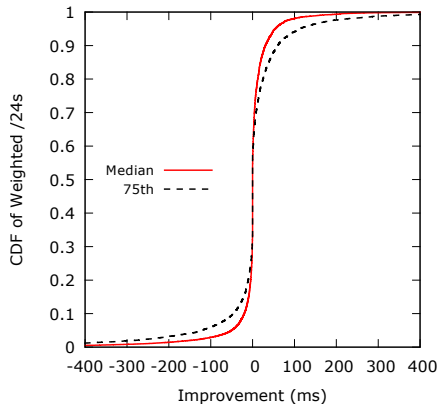


**Figure 4: Improvement over anycast from using DNS redirection to override poor performing anycast.**

better than BGP anycast (§2.3.2). This earlier work instrumented millions of Bing search results with JavaScript to measure from the client to both the anycast address and to a number of nearby unicast addresses [6]. The authors used the measurements to unicast addresses to compare anycast (BGP) performance versus the optimal front-end location and versus a DNS-redirection scheme.

*3.2.1 Performance-aware routing provides little benefit over BGP anycast.* The first takeaway is that, most of the time, anycast performs as well as the best possible unicast front-end, but performance is poor in the tail. Figure 3 shows that, globally, anycast is within 10ms of the best unicast for 70% of all requests. However, the best unicast is at least 100ms faster for nearly 10% of requests. This figure demonstrates that while anycast performs well most of the time, there is an opportunity for improvement using DNS-based redirection for some clients. However, to realize the full performance improvement seen in Figure 3, a DNS-based redirection system will need to be oracular and overcome well-known limitations of DNS redirection [5].

In practice, the performance gap is modest. In general, DNS redirection systems cannot see the IP address of the requesting client, only of client's local resolver (LDNS), limiting decisions to a per-LDNS granularity. EDNS Client Subnet was designed to overcome this limitation [11], but its adoption by ISPs is virtually non-existent (< 0.1% of ASes) outside of public resolvers [5]. To understand the achievable benefit of using unicast when anycast directs clients to suboptimal front-ends, the earlier study mapped each LDNS to either the best performing unicast front-end or anycast, whichever earlier measurements predict is better for clients of the LDNS. The LDNS-predicted optimal and anycast are then measured side-by-side from Bing clients. Figure 4 shows the difference in latency between the best predicted (anycast or unicast) versus anycast for median and 75th percentile. The median results show latency improvement for 27% of queries but the prediction did worse than anycast for 17% of queries.

This work shows that anycast directs most clients to the lowest latency front-end location. While anycast is at least 25ms slower for 20% of requests, DNS redirection schemes also struggle to direct clients to optimal server locations, performing worse than anycast (BGP) nearly as often as they beat it.

*3.2.2 Open question: Nature vs. nurture: does BGP perform well because of infrastructure investments that align policy with performance, or because of (manual) route optimization?* Why is it that performance-oblivious BGP anycast performs roughly as well as a measurement-driven, performance-aware redirection strategy? And how do we square these results with a recent study claiming that anycast poorly routes requests to nearby DNS root replicas [18]?

We hypothesize that a content provider is much more invested in achieving low latency than a DNS root operator (where availability is highest priority) and has more resources to put into optimizing latency, both in terms of capital investment and operational effort. CDN operators can manually "groom" their anycast routing by tweaking their BGP announcements (*e.g.*, prepending to a particular peer at a particular location, or adding a BGP community to control propagation) and by reaching out to other ASes making poor choices. Sections 3.1 and 3.1.1 suggest little opportunity for dynamic performance-aware routing, partly because most paths to a destination tend to degrade together. These results suggest that optimization and grooming of routes can provide benefit even when done at human timescales. The measurement study of Microsoft's CDN captures the performance at a single snapshot of the infrastructure, and after a certain amount of grooming efforts.

To what degree does each matter—does the performance stem from the "nature" of the infrastructure—the PoP locations and their interconnections—or from the "nurture" of operators identifying and fixing poor anycast routes over time? An interesting future direction is to disentangle these effects and to develop best practices for grooming anycast. When designing or expanding a CDN, how should a provider

decide where to locate PoPs and who to peer with? How well can the impact of adding a site be predicted? How quickly does benefit diminish when adding PoPs [10]? As PoPs are added, the chance of anycast picking a suboptimal one increases, but the number of reasonably performing ones increases. How do those factors relate, and how does grooming change with PoP density [26]? What is the performance of an ungroomed prefix versus a groomed one? What are the best ways to detect routes where opportunity for grooming exists? If an AS has groomed one prefix, does that carry over to newly announced prefixes and simplify the process of grooming them? What techniques exist for grooming, how effective are they, and how well can their impact be predicted before announcements are adjusted?

## 3.3 BGP often works as well as a private WAN, even over long distances

One might think that BGP works well for an anycast CDN because its job is easy: it only has to choose a short route and avoid longer ones. However, even in settings where BGP has to choose between multiple long paths (§2.3.3), it can fare admirably well.

We conduct measurements to compare the performance of Google's Premium Tier versus Standard Tier networking on routes between Google's US Central data center and vantage points around the world. The Premium Tier uses Google's private WAN to ingress/egress near the vantage points, whereas the Standard Tier ingress/egresses near the data center and traverses the public Internet the rest of the way. This setting would seem to create a challenge for BGP, which often selects more circuitous routes than those that traverse only a single WAN [28]. We further challenge BGP by restricting ourselves to vantage points whose route to the Standard Tier includes at least one intermediate AS between the vantage point's AS and Google, but whose route to the Premium Tier enters Google directly from the vantage point's AS, meaning Google can optimize the Premium Tier route up to the edge of the vantage point's AS.

To find such vantage points, we create VMs in Google's US Central data center, one each using the Standard and Premium Tier, and issue measurements to the VMs from Speedchecker[1] vantage points [27]. Speedchecker exposes an API to issue measurements (*e.g.,* ping, traceroute, HTTP GET, *etc.*) based on credits, similar to RIPE Atlas. Our credits allow us to issue one traceroute and five pings to each of the VMs 10 times a day from 800 vantage points, which we select daily to rotate across ⟨City, AS⟩ locations over time. We repeated the measurements over a period of 10 months. The traceroutes display a stark contrast in the use of Google's WAN depending on the networking tier: traceroutes from 80% of vantage points enter Google's network within 400km

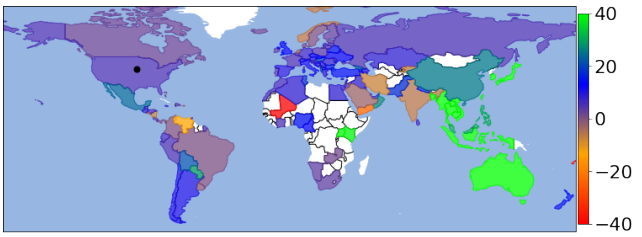


**Figure 5: Difference in median latency to Standard Tier minus Premium Tier to VMs hosted in the US-Central data center (the black dot), from vantage points in ASes that peer directly with Google on the Premium Tier but not on the Standard Tier. A positive value (green/blue color) indicates that the Premium Tier (private WAN) performed better, and a negative value (red/orange color) indicates that the Standard Tier (BGP on public Internet) performed better. Countries without measurements are colored white.**

of the vantage point when using the Premium Tier, whereas only 10% do when using the Standard Tier.[2] We selected the subset that met our criteria (direct Premium path, intermediate AS on Standard path), resulting in 7.5 million pings from vantage points spread across 17,000 ⟨City, AS⟩ locations, in ASes hosting 61% of Internet users according to APNIC user population estimates [3].

*3.3.1 BGP routes to Google cloud perform similarly to Google's private WAN.* This data reveals that the performance of routes for Standard Tier and Premium Tiers are usually comparable. Figure 5 shows the difference in median ping latency per country. Most countries in North America, South America, and Europe have little (+/- 10ms) difference in median latency when routed via Google's WAN versus when routed via BGP on the public Internet. Some countries in the Middle East and South America have better performance for Standard Tier (public Internet) routes. Routes that utilize Google's private WAN have better performance for most countries in Asia and Oceania.

*3.3.2 Open question: Do private WANs struggle to beat BGP routes primarily when BGP routes behave like a single WAN?* BGP routes on the public Internet consistently outperform Google's private WAN for measurements from vantage points in India. Traceroutes in our dataset, corroborated by another study [16, 29], reveal why. Google's WAN carries traffic from India east across the Pacific Ocean to reach North America. In contrast, BGP routes on the public Internet enter a Tier 1 network close to the vantage points, and the Tier 1 network carries the traffic the whole way west via Europe until it enters Google's network near the datacenter.

This case study suggests that BGP may perform best when it selects routes that spend much of their journey on a single large provider, which can optimize routing in a similar way to (or better than) a cloud provider on its WAN. The

[1]Speedchecker is a global measurement platform deployed in home routers, PCs, and wireless devices.

[2]We locate the ingress if we can find a RIPE Atlas probe with a ping RTT of at most 1ms to the border router, which we can for 72% of our traceroutes.

concentration of users in certain population centers and the benefits of colocation in facilities with many other ASes may lead to large WANs having similar footprints.

A number of questions seem worth investigating. How similar are the footprints of large providers, and how has this changed over time? Do Internet paths perform best when they spend a larger fraction of their journey on a single network, and can regional performance differences be predicted from topological differences in WANs? Does the public Internet performance observed to Google cloud data centers depend on Google paying Tier 1 providers for high-end service, or do we observe similar performance to other destinations? For example, do the Tier 1 networks use late-exit routing for Google but early-exit routing for others? While Google has resources and incentives to pay for the highest level of service, it is also possible that a route will often stay on a single large network for most of the way towards Google simply as an artifact of standard valley-free BGP policy: no Google peer will announce a Google prefix to a Tier 1, and so a Tier 1 will only receive the prefix directly from Google or from a provider of Google. If it receives the prefix from Google, the Tier 1 will have to carry the traffic to near the data center where Google announces the prefix. If it receives the prefix from a provider of Google, the provider likely connects to the Tier 1 in many locations around the world, and so early exit from the Tier 1 will result in the other provider carrying the traffic most of the way.

## 4 BEYOND MEDIAN PERFORMANCE

The three settings we discussed all found limited latency penalty to letting BGP policy select routes, and similar results hold for throughput [24].[3] Does this mean that large content and cloud providers can afford to simply rely upon BGP, and it is not worth the trouble for them to build sophisticated routing control systems? To the contrary, it is crucial to consider that many other factors are at play when evaluating the quality of routes used to serve clients.

First, although the measurements found that BGP performed well in the median, BGP underperformed in a small fraction of cases. The 2–4% of cases when performance-aware routing could improve performance by 5+ ms represent hundreds of billions of HTTP sessions (§3.1). Performance-aware routing or hybrid approaches [6] may be necessary to claim this "lost" performance. It is a business (and not technical) assessment of whether this benefit is worth the cost of building and maintaining a performance-aware system to replace battle-tested BGP routing, but the research community would benefit from a richer understanding of how latency impacts user experience and user actions [19].

[3]We used Speedchecker to measure goodput of 10MB downloads from Google's Premium and Standard Tiers and saw little difference. Results omitted for space.

Second, end-to-end latency and throughput are not the only (or even most important) metrics. Availability is the primary concern of content and cloud providers. Understanding the impact on availability of the factors discussed in this paper is a promising area for future research. A private WAN likely conveys availability advantages over the public Internet, and Google provides a more expansive availability service-level specification for its WAN-based Premium Tier networking than its Standard Tier. Anycast provides resilience against site outages [7, 13] and avoids availability problems that can be induced by DNS caching [14]; understanding how to trade this benefit off with its more limited control is an area of ongoing research, as is understanding how best to design hybrid approaches with the benefits of both anycast and DNS redirection.

The impact on availability of increased/reduced peering (§3.1.3) also deserves future study. Adding peering increases route diversity, which may add resilience. However, small peers may be less reliable and cause more issues (§3.1.3), plus a larger fraction of the capacity to a small peer may be concentrated on a single interconnection or router as compared to the redundant capacity to large providers [25], and so a failure can have an outsized impact. Further, if most problems are on the last mile or other shared portions of routes (§3.1.1), route diversity does not help to avoid them. However, adding peering can also increase capacity; paths that perform well when carrying a fraction of the traffic to an ISP may not be able to handle the full load served from a large content provider.

Finally, splitting TCP connections provides latency benefits over long distances [12]; an interesting area for study is how this benefit varies if the backend of the split connection is over a private WAN [12] versus the public Internet, as it traditionally was for Akamai before its recent WAN buildout.

## 5 CONCLUSION

This work re-examined BGP's reputation for poor performance on the Internet by highlighting results from three recent studies of large content provider networks that demonstrate only minor latency gains from using sophisticated policies and traffic engineering compared to standard Internet routing. Building on results contrasting BGP and performance-aware approaches such as PoP selection for content delivery, egress traffic engineering, and private WAN, we presented hypotheses for why it is difficult to outperform BGP on today's Internet. We outlined a research agenda to drive a deeper understanding of Internet routing.

**Acknowledgements.** We appreciate the valuable feedback from the HotNets reviewers. We acknowledge the contributions of our collaborators on studies that informed this paper. We thank Speedchecker, especially Janusz Jezowicz, for providing us access to their measurement platform. This work was partly funded by NSF awards CNS-1835253, CNS-1835252, CNS-1413978, and CNS-1563849.

## REFERENCES


[1] Aditya Akella, Jeffrey Pang, Bruce Maggs, Srinivasan Seshan, and Anees Shaikh. 2004. A Comparison of Overlay Routing and Multihoming Route Control. In *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '04)*.

[2] David Andersen, Hari Balakrishnan, Frans Kaashoek, and Robert Morris. 2001. Resilient Overlay Networks. In *Proceedings of the Eighteenth ACM Symposium on Operating Systems Principles (SOSP '01)*.

[3] APNIC. 2019. Visible ASNs: Customer Populations (Est.). (2019). https://stats.labs.apnic.net/aspop/.

[4] Matt Calder, Xun Fan, Zi Hu, Ethan Katz-Bassett, John Heidemann, and Ramesh Govindan. 2013. Mapping the Expansion of Google's Serving Infrastructure. In *Proceedings of the ACM Internet Measurement Conference (IMC '13)*.

[5] Matt Calder, Xun Fan, and Liang Zhu. 2019. A Cloud Provider's View of EDNS Client-Subnet Adoption. In *Proceedings of the Network Traffic Measurement and Analysis Conference (TMA '19)*.

[6] Matt Calder, Ashley Flavel, Ethan Katz-Bassett, Ratul Mahajan, and Jitendra Padhye. 2015. Analyzing the Performance of an Anycast CDN. In *Proceedings of the ACM Internet Measurement Conference (IMC '15)*.

[7] Matt Calder, Ryan Gao, Manuel Schröder, Ryan Stewart, Jitendra Padhye, Ratul Mahajan, Ganesh Ananthanarayanan, and Ethan Katz-Bassett. 2018. Odin: Microsoft's Scalable Fault-Tolerant CDN Measurement System. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI '18)*.

[8] Yi-Ching Chiu, Brandon Schlinker, Abhishek Balaji Radhakrishnan, Ethan Katz-Bassett, and Ramesh Govindan. 2015. Are We One Hop Away from a Better Internet?. In *Proceedings of the ACM Internet Measurement Conference (IMC '15)*.

[9] CloudLab. 2019. (2019). http://www.cloudlab.us/.

[10] Ricardo de Oliveira Schmidt, John Heidemann, and Jan Harm Kuipers. 2017. Anycast latency: How many sites are enough?. In *International Conference on Passive and Active Network Measurement (PAM '17)*.

[11] Marcelo Torres Fangfei Chen, Ramesh K. Sitaraman. 2015. End-User Mapping: Next Generation Request Routing for Content Delivery. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '15)*.

[12] Tobias Flach, Nandita Dukkipati, Andreas Terzis, Barath Raghavan, Neal Cardwell, Yuchung Cheng, Ankur Jain, Shuai Hao, Ethan Katz-Bassett, and Ramesh Govindan. 2013. Reducing Web Latency: The Virtue of Gentle Aggression. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '13)*.

[13] Ashley Flavel, Pradeepkumar Mani, David Maltz, Nick Holt, Jie Liu, Yingying Chen, and Oleg Surmachev. 2015. FastRoute: A Scalable Load-Aware Anycast Routing Architecture for Modern CDNs. In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI '15)*.

[14] Ashley Flavel, Pradeepkumar Mani, and David A Maltz. 2014. Re-evaluating the responsiveness of dns-based network control. In *IEEE 20th International Workshop on Local & Metropolitan Area Networks (LANMAN '14)*.

[15] Google. 2019. Google Network Service Tiers. https://cloud.google.com/network-tiers/. (2019).

[16] Archana Kesavan. 2019. Comparing the Network Performance of AWS, Azure and GCP. (2019). https://pc.nanog.org/static/published/meetings/NANOG75/1909/20190218_Kesavan_Comparing_The_Network_v1.pdf.

[17] Farhan Khan. 2015. The Cost of Latency. (2015). https://www.digitalrealty.com/blog/the-cost-of-latency

[18] Zhihao Li, Dave Levin, Neil Spring, and Bobby Bhattacharjee. 2018. Internet Anycast: Performance, Problems, & Potential. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '18)*.

[19] Greg Linden. 2006. Make Data Useful. http://sites.google.com/site/glinden/Home/StanfordDataMining.2006-11-28.ppt. (2006).

[20] Hariharan Rahul, Mangesh Kasbekar, Ramesh Sitaraman, and Arthur Berger. 2005. *Towards Realizing the Performance and Availability Benefits of a Global Overlay Network*. Technical Report. Massachusetts Institute of Technology.

[21] Hariharan Rahul, Mangesh Kasbekar, Ramesh Sitaraman, and Arthur Berger. 2006. Towards Realizing the Performance and Availability Benefits of a Global Overlay Network. In *Proceedings of International Conference on Passive and Active Network Measurement (PAM '06)*.

[22] Stefan Savage, Andy Collins, Eric Hoffman, John Snell, and Thomas Anderson. 1999. The End-to-end Effects of Internet Path Selection. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '99)*.

[23] Brandon Schlinker, Todd Arnold, Ítalo Cunha, and Ethan Katz-Bassett. 2019. PEERING: Virtualizing BGP at the Edge for Research. In *Proceedings of Conference on emerging Networking EXperiments and Technologies (CoNEXT '19)*.

[24] Brandon Schlinker, Ítalo Cunha, Yi-Ching Chiu, Srikanth Sundaresan, and Ethan Katz-Bassett. 2019. Internet Performance from Facebook's Edge. In *Proceedings of the ACM Internet Measurement Conference (IMC '19)*.

[25] Brandon Schlinker, Hyojeong Kim, Timothy Cui, Ethan Katz-Bassett, Harsha V Madhyastha, Italo Cunha, James Quinn, Saif Hasan, Petr Lapukhov, and Hongyi Zeng. 2017. Engineering Egress with Edge Fabric. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*.

[26] Pavlos Sermpezis and Vasileios Kotronis. 2019. Inferring Catchment in Internet Routing. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)* (2019).

[27] Speedchecker. 2019. http://probeapi.speedchecker.com/. (2019).

[28] Neil T. Spring, Ratul Mahajan, and Thomas E. Anderson. 2003. The Causes of Path Inflation. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '03)*.

[29] ThousandEyes. 2018. Public Cloud Performance Benchmark Report. (2018). https://marketo-web.thousandeyes.com/rs/thousandeyes/images/ThousandEyes-2018-Public-Cloud-Performance-Benchmark-Report.pdf.

[30] Florian Wohlfart, Nikolaos Chatzis, Caglar Dabanoglu, Georg Carle, and Walter Willinger. 2018. Leveraging Interconnections for Performance: The Serving Infrastructure of a Large CDN. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '18)*.

[31] Kok-Kiong Yap, Murtaza Motiwala, Jeremy Rahe, Steve Padgett, Matthew Holliman, Gary Baldus, Marcus Hines, Taeeun Kim, Ashok Narayanan, Ankur Jain, et al. 2017. Taking the Edge off with Espresso: Scale, Reliability and Programmability for Global Internet Peering. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*.

[32] Kyriakos Zarifis, Tobias Flach, Srikanth Nori, David R. Choffnes, Ramesh Govindan, Ethan Katz-Bassett, Zhuoqing Morley Mao, and Matt Welsh. 2014. Diagnosing Path Inflation of Mobile Client Traffic. In *International Conference on Passive and Active Network Measurement (PAM '14)*.