Internet Society Pulse

The 50/50 Vision for Internet Traffic

Version 2.0, January 2024

Introduction

The 50/50 Vision defines the Internet Society's strategy to rally multistakeholder efforts and international and national resources to ensure that at least 50% of all locally generated traffic in selected economies remains local by 2025. Reaching this ambitious target will strengthen Internet quality and reduce access costs for individuals.

A 2012 study established a baseline at Internet Exchange Point (IXP) hubs in Kenya and Nigeria, and a follow-up study in 2020 showed that levels of local traffic jumped from 30% to 70%. The effort helped increase understanding of the impact of peering on the local infrastructure. It was found that the increase in local traffic led to significant cost savings for participating networks and put these two countries in a stronger position to participate in the digital economy. Meanwhile, South Africa localized over 80% of its local traffic and now enjoys stable, resilient, high-quality, and affordable Internet.

The aim of the 50-50 vision study is to measure the percentage of Internet traffic that is being served from an in-country server or cache. This document describes the Internet Society's methodology to achieve the above goal.

Definitions

- Local traffic refers to traffic that stays local and does not leave the country.
- Local content: Internet content that is hosted within the country (e.g., news website, e-government services, etc.)
- RIPE Atlas is an Internet measurement framework operated by the RIPE NCC. It consists of probes (vantage points) and anchors (targets). The probes can be used to run several measurements (latency, traceroute, DNS, SSL, HTTP) from the edge.
- CDN: Content Delivery Networks are responsible for delivering content to the edge.

Internet Society Pulse

- Content cache is a content hosting equipment placed by a content provider close to the end-users.
- Vantage point: where the measurement is being conducted from.
- Edge network refers to the access network where eyeballs (consumers) are located.

Lack of Data on Traffic Volume

Measuring local vs non-local traffic levels for a country is not straightforward. Access providers are usually aware of the most used services and which ones are the biggest consumers of bandwidth within their network. Data about the most used services are rarely shared publicly. In some rare cases and based on prior agreement, ISPs provide s-flow traffic data (sample flows) from which the destination IP can be extracted. In the absence of any ground truth, we will use the percentage of websites served locally (natively or through a CDN cache) by looking at the Top 1000 websites in a country.

Google CrUX

We used the Top 1000 websites from the Google Chrome UX Report (CrUX). It is a public dataset and initiative by Google that aims to provide real-world performance metrics for how real users experience the web. It aggregates anonymized performance data from users who have opted in to share their browsing data with Google, primarily through the Chrome browser. CrUX produces a monthly report of the most popular websites by country.

How We Determine Content Locality

1. For each website in the Top 1000 list, we need to determine which CDN service it is using. For this, we scrape the website to extract the following:
    a. HTTPS Server Headers: headers such as "server" or "via" usually contain geohints
    b. CNAME records: A Canonical Name Record (CNAME RECORD) is a type of record in Domain Name Servers (DNS servers) that creates an alias from one domain to another. By observing a CNAME record, we can see the domain name linked to a CDN, before directing to the desired resource.
    c. WHOIS data: Using the IP address of the domain, we will do a WHOIS lookup and infer usage of CDNs.
    d. Server header based: Some HTTP Response headers would provide geohints.
2. Using local vantage points from within the country, we run active measurements towards a selected set of targets (CDNs, applications, websites) to extract geohints and other geolocation information. Using the above information, we can confidently mark a website as local.

Internet Society
Pulse

3. For websites that are natively hosted, we use geolocation to determine the location of the IP address. The IP address is obtained by running the query from a local vantage point through a local DNS resolver.
4. We then calculate the ratio of locally-hosted websites vs. remotely-hosted websites.

Example

Below are the steps to calculate the percentage of websites hosted locally:

1. We take the Top 1000 website from Google CrUX for country X.
2. We categorized the websites by <CDN name> if hosted by a CDN or "Native" if the website is natively hosted.
3. We use IP geolocation to determine the location of the natively hosted domains.

CDN	Count	Location
Cloudflare	454	Local
Akamai	92	External
Cloudfront	55	Local
Google	51	Local
Fastly	25	External
Facebook	8	Local
Yahoo	6	External
Amazon AWS	5	External
EdgeCast	4	External
Level3	3	External
Natively hosted	250	Local: 100 External: 150

Table 1: Distribution of CDN usage for Top 1000 websites in country X

Internet Society
Pulse

4. The output is the ratio of local vs external websites.

Output

We plan to produce a detailed report on how much traffic is staying local based on the above methodology by the end of February 2024. We will also develop a dashboard on our Pulse platform to visualize the data in Q2 2024. We will split the Top 1000 websites into different datasets to present different perspectives on traffic locality: (1) top 1000 websites (2) top 10 applications (Netflix, Disney+, etc.) (3) top 100 websites using the country ccTLD (4) government websites (5) local news websites (6) major CDN players.

Internet Society
Pulse