arTs Università degli Studi di Trieste
Archivio della ricerca – postprint

Five Years at the Edge: Watching Internet from the ISP Network

Martino Trevisan
Politecnico di Torino
martino.trevisan@polito.it

Danilo Giordano
danilo.giordano@polito.it
Politecnico di Torino

Idilio Drago
idilio.drago@polito.it
Politecnico di Torino

Marco Mellia
Politecnico di Torino
marco.mellia@polito.it

Maurizio Munafo
Politecnico di Torino
maurizio.munafo@polito.it

ABSTRACT

The Internet and the way people use it are constantly changing. Knowing traffic is crucial for operating the network, understanding users' need, and ultimately improving applications. Here, we provide an in-depth longitudinal view of Internet traffic in the last 5 years (from 2013 to 2017). We take the point of the view of a national-wide ISP and analyze flow-level rich measurements to pinpoint and quantify trends. We evaluate the providers' costs in terms of traffic consumption by users and services. We show that an ordinary broadband subscriber nowadays downloads more than twice as much as they used to do 5 years ago. Bandwidth hungry video services drive this change, while social messaging applications boom (and vanish) at incredible pace. We study how protocols and service infrastructures evolve over time, highlighting unpredictable events that may hamper traffic management policies. In the rush to bring servers closer and closer to users, we witness the birth of the sub-millisecond Internet, with caches located directly at ISP edges. The picture we take shows a lively Internet that always evolves and suddenly changes.

CCS CONCEPTS

• **Networks** → **Network performance analysis**; **Network measurement**;

KEYWORDS

Passive Measurements; Broadband Traffic Characterization.

ACM Reference Format:
Martino Trevisan, Danilo Giordano, Idilio Drago, Marco Mellia, and Maurizio Munafo. 2018. Five Years at the Edge: Watching Internet from the ISP Network. In *The 14th International Conference on emerging Networking EXperiments and Technologies (CoNEXT '18), December 4–7, 2018, Heraklion, Greece.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3281411.3281433

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CoNEXT '18, December 4–7, 2018, Heraklion, Greece
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6080-7/18/12…$15.00
https://doi.org/10.1145/3281411.3281433

1 INTRODUCTION

Measurements have always been among the best ways to understand complex systems. Not surprisingly, measurements are the key means to gather information about the overall status of the Internet, identify eventual issues, and ultimately improve its design [30, 36, 40]. Being the Internet an evolving system, novel measurement mechanisms are continuously devised to extract information about applications, protocols, deployments, etc. However, having a long-term picture on how the Internet is evolving is a rather challenging task. Researchers often design new tools and approaches that focus on specific phenomena, which are observed and described in details for limited time. It is rare to find works that offer a longitudinal view on systems over time.

In this paper, we offer such longitudinal view of the Internet in the past 5 years. We rely on a humongous amount of data collected from a nation-wide Internet Service Provider (ISP) infrastructure. We focus on broadband Internet access via ADSL and FTTH technologies. We instrument some of the ISP aggregation links with passive monitoring probes. By observing packets flowing on links, our probes extract detailed per flow information, that we collect and store on a centralized data lake. Keeping the pace with Internet evolution during 5 years is per se a challenging task. We rely on custom designed software probes that have been constantly updated during the monitoring period to account for and report information about new protocols and services.

Technically, we follow a well-established approach. Passive measurements are popular among researchers since early 2000 [2, 9], with current tools able to process several tens of Gb/s on commodity hardware [31]. Extracting information from packets is possible thanks to Deep Packet Inspection (DPI) techniques [1], while the availability of big data solutions [11, 41] makes it possible to store and process large volumes of traffic with unprecedented parallelism.

Here, we dive into this data, depicting trends, highlighting sudden changes and observing sudden infrastructure upgrades. Instead of focusing on a specific angle, we aim at offering examples of trends on the Internet evolution. The Internet indeed rapidly evolves: Services get popular and other get abandoned; Users change habits; New protocols change the way information is carried. Observing such trends is vital to understand the Internet, the users, and the systems.

First we give an overview of users' habits over 5 years, assessing the costs of broadband customers to the ISP in terms of traffic consumption. We observe for example that the traffic per broadband customer has increased at a constant rate over the years, with a growth of heavy users, i.e., those who exchange tens of GB per

day. When comparing service usage between ADSL and FTTH customers, we see that the larger capacity offered to FTTH customers has a moderate impact on per customer data consumption.

Next, we turn our attention to the traffic loads imposed by web services to the ISP. We quantify the rise (and death) of services in terms of traffic volumes as well as popularity among customers. Here we confirm and precisely quantify some well-known trends: video content – no longer accessed via peer-to-peer systems – drives the bandwidth demand. Yet, users of modern social messaging systems such as Instagram (accessed from mobile phones) consume more and more traffic. Indeed, the traffic of each Instagram user is already comparable to the traffic of video-on-demand users, such as Netflix or YouTube.

Finally we study how changes in the infrastructure and protocols impacted the ISP network. For example, we detail the (slow) migration of services to HTTPS and several (sudden) deployments of custom protocols by large companies that may hamper traffic engineering and troubleshooting of ISPs. We testify the growth in the infrastructure of popular services, and show how services are more and more deployed close to users, with caches deployed at the first aggregation point at the ISP, in an effort to cut off the latency to reach the Internet contents.

Despite our dataset being limited to one country and focused on broadband Internet (thus missing mobile networks), we believe the information we offer is key to understand trends and inform researchers and practitioners about recent changes on Internet infrastructure and users' behavior.

The paper is organized as follows: Section 2 presents our monitoring infrastructure and the analyzed dataset. Section 3 investigates traffic demand of ISP customers, while Section 4 illustrates trends of services in terms of traffic volume and popularity. Section 5 analyses protocol usage and episodes of unpredictable traffic variations, whereas Section 6 shows notable trends in Big Players' infrastructure. Section 7 summarizes the related work. Finally, Section 8 concludes the paper.

2 MEASUREMENT METHODOLOGY

We now describe the measurement methodology and tools used to collect the data.

2.1 Measurement architecture

We build on data collected by the passive monitoring infrastructure of a nation-wide ISP in Italy that captures and analyses in real-time traffic from vantage points located at the edge of the ISP network. A schematic view of the infrastructure is depicted in Figure 1. We process traffic directly in the ISP Points-of-Presence (PoPs). Exploiting router span ports or optical splitters (depending on the link rates), we mirror the traffic to the monitoring probes. Both uplink and downlink streams are exposed to the probes. Since probes are deployed in the first level of aggregation of the ISP, no traffic sampling is performed. Customers are assigned fixed IP addresses, that the probes immediately anonymize in a consistent way.

Each probe is equipped with multiple high-end network interfaces. Packets are captured using the Intel Data Plane Development Kit (DPDK) [23] that allows line-rate capture even for multiple

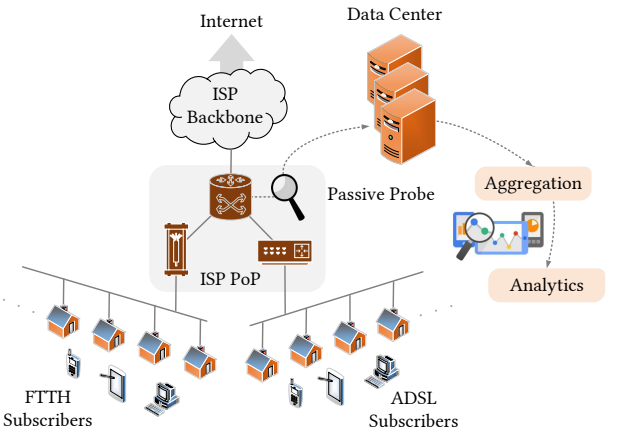


Figure 1: Measurement infrastructure and processing steps.

10 Gbit/s links. Traffic is then processed by our custom-made passive traffic analyzer, called Tstat [39].

Each probe exports only flow records, i.e., a single entry for each TCP/UDP stream with per-flow statistics.[1] Each record contains classical fields on flow monitoring [22], such as IP addresses, port numbers, packet-wise and byte-wise counters. Advanced analyzers extract some few fields from packet payloads, such as information seen in the Application-Layer Protocol Negotiation (ALPN) fields of TLS handshakes, which allows us to identify HTTP/2 and SPDY flows, and fields from QUIC public headers. Tstat also exports the *domain name* of the contacted servers, exchanged in clear in HTTP `Host:` headers, or requested in the TLS Server Name Indication (SNI) within TLS Client Hello messages. For flows missing such information, Tstat exports the host name the client resolved via DNS queries prior to open the flow.[2] This mechanism, called DN-Hunter, is explained in details in [4]. Such hostnames extracted from the different sources are vital to associate traffic flows to web services [38].

For the analysis of server infrastructure (Section 6), we rely on the estimation of RTT provided by Tstat for TCP flows [29]: It searches for acknowledged TCP segments, registering the time from the observation of the TCP segment and its acknowledgment. For each flow, Tstat exports the minimum, average and maximum RTT estimation, as well as the number of RTT samples. Notice that this metric represents the RTT from the probe to servers, missing the delay from clients to the probes. Thus, in our deployment we ignore the access delay, since probes are deployed at the ISP's PoPs.

Among the vantage points, here we consider the traffic of two PoPs, covering more than 10 000 ADSL and 5 000 Fiber-To-The-Home (FTTH) subscribers, all located in the same city in Italy, and active since 2013. ADSL downlink capacity varies from 4 Mbit/s up to 20 Mbit/s, with uplink limited to 1 Mb/s. FTTH users enjoy 100 Mb/s downlink, and 10 Mbit/s uplink. Each subscription refers to an ADSL or FTTH installation, where users' devices (PCs,

[1] Streams are expired either by the observation of particular packets (e.g., TCP packets with RST flag set) or by timeouts. See http://tstat.polito.it/measure.shtml.
[2] Our vantage points observe all DNS traffic directed to any resolver.

Table 1: Examples of domain-to-service associations.

Domain	Service
`facebook.com`	Facebook
`fbcdn.com`	Facebook
`^fbstatic-[a-z].akamaihd.net$` (RegExp)	Facebook
`netflix.com`	Netflix
`nflxvideo.net`	Netflix

smartphones, tablets, smart TVs etc) connect via WiFi and Ethernet through a home gateway. ADSL customers are mainly residential customers (i.e., households), whereas a small but significant number of business customers exist among the FTTH customers.

During the 5 years of measurements we observed a steady reduction on the number of active ADSL users and an increase in FTTH installations. The ISP has confirmed these trends are due to churning and technology upgrades. To compensate for such changes, we will report statistics aggregating measurements and normalizing numbers according to the number of active users per day.

2.2 Data storage and processing

Flow records are created, anonymized and stored on the local probe disks. Daily, logs are copied into a long-term storage in a centralized data center and discarded from the probes.

By the time of writing, the considered dataset covers 5 years of measurements, totaling 31.9 TB of compressed and anonymized flow logs (around 247 billion flow records). To process this deluge of data, we use a Hadoop-based cluster running Apache Spark. This structure allows us both to update predefined analytics continuously, as well as to run specific queries on historical collections.

Our analytics methodology follows a two-stage approach: firstly data is aggregated on a per day basis, secondly, advanced analytics and visualizations are computed. In the aggregation stage, queries compute per-day and per-subscription aggregates about traffic consumption, protocol usage, and contacted services. This round requires processing of millions of raw flow records.

Special attention is needed for identifying the services used by subscribers. Content providers are known to rely on large infrastructure and/or Content Delivery Networks (CDNs), which make the association between flow records and services tricky. For this step, we rely mostly on the server domain names. Examples of the association domain-service are provided in Table 1. Flexible matching based on regular expressions is allowed.[3] Along the years, our team has *continuously* monitored the most common server domain names seen in the network, maintaining the list of domains associated with the services of interest. For ambiguous cases [38], e.g., domains used by multiple services, we rely on heuristics, mostly based on traffic volumes, to decide whether a subscriber actually contacted a particular service (see Section 4.1). This methodology thus allows on-the-fly and historical classification of services. Once such aggregated dataset is available, flexible analytics perform the analysis and visualization of the data.

[3]The full list of rules to classify services is found at https://smartdata.polito.it/five-years-at-the-edge-watching-internet-from-the-isp-network/.

2.3 Challenges in long-term measurements

Several challenges arise when handling a large-scale measurement infrastructure. Network probes are the most likely point of failure, as they are subject to a continuous and high workload. During the period considered in the paper, probes suffered few outages, lasting from few hours up to some months (when severe hardware issues arose). As such, the results we present have missing data for those periods.

A second issue arises from the evolution of network protocols and service infrastructure. Large content providers have the power of suddenly deploying new protocols leaving passive monitors and ISPs with few or no documentation to handle them. We incurred several cases, and report our experience in addressing them.

Third, the domain-to-service associations need to be continuously updated. Also in this case, there is no public information to support this operation, so that our team has to manually define and update rules, often by running active experiments to observe new patterns.

At last, users' privacy must be preserved. For this, we carefully limit the collected information and always consider only aggregated statistics. Customers' IP addresses and server names are the most privacy-sensitive information being collected. The former gets immediately anonymized by probes, while the latter is used to derive aggregate statistics on per-service basis. Importantly, all data collection is approved and supervised by the responsible teams in the ISP.

3 THE COST OF A USER

We first characterize the amount of traffic consumed by subscribers in the last 5 years. This analysis is instrumental to understand costs of ISPs in terms of capacity and forecasting trends.

For the results that follow, we consider only *active subscribers*. Subscribers are considered *active* if they have generated at least 10 flows, downloaded more than 15 kB and uploaded more than 5 kB.[4] This simple criterion lets us filter those cases where only background traffic is present, e.g., generated by the access gateway, or by incoming traffic (due to, e.g., port scans). On average we observe about 80% subscribers active each day, with respect to the total number of subscribers observed in the whole trace.

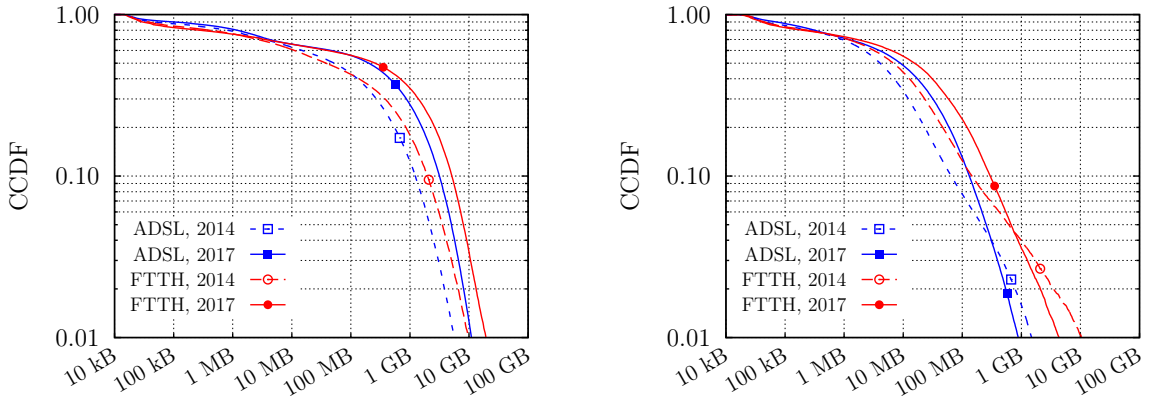
Notice that these percentages are actually a lower-bound given churning (see Section 2.1). Notice also that smartphones contribute to make subscribers active in more days.

3.1 How much you eat: Consumption per day

Figure 2 depicts the empirical Complementary Cumulative Distribution Function (CCDF) of daily traffic consumption of active subscribers in the ISP. In other words, for each day, we compute the overall traffic each active subscriber exchanges. We report the CCDF of all measurements as seen in April 2014 and 2017. Figure 2 depicts CCDFs separately per access-link technology and down/up links. Log scales are used.

Observe the bimodal shape of the distribution. In about 50% of days, subscribers download (upload) less than 100 MB (10 MB) – i.e., days of light usage. However, a heavy tail is present. For more than 10% of the days, subscribers download (upload) more than

[4]These thresholds have been determined by visually inspecting knee points in the distributions of daily traffic per user.



(a) Download volumes **(b) Upload volumes**

Figure 2: CCDF of per active subscriber daily traffic for April 2014 and 2017.

1 GB (100 MB) – i.e., days of heavy usage. Manual inspection shows that many different subscribers present days of heavy usage, often alternating between days of light and heavy usage.

Comparing 2014 (dashed lines) with 2017 (solid lines), we notice an increase in daily traffic consumption. The median values have increased by a factor 2 for both ADSL and FTTH installations, and for both upload and download. This behavior highlights an increasing trend in average per-subscriber traffic volume, that we examine more in depth later in this section.

We observe no differences for the days of light usage when contrasting ADSL (blue curves) and FTTH installations (red curves). Instead, during heavy usage days, FTTH users download about 25% more data than ADSL users – a moderate increase given they enjoy 5-20 times higher capacity. The differences are higher considering upload traffic: ADSL users are indeed bottlenecked by the 1 Mb/s uplink, thus FTTH subscribers upload twice as much per day.

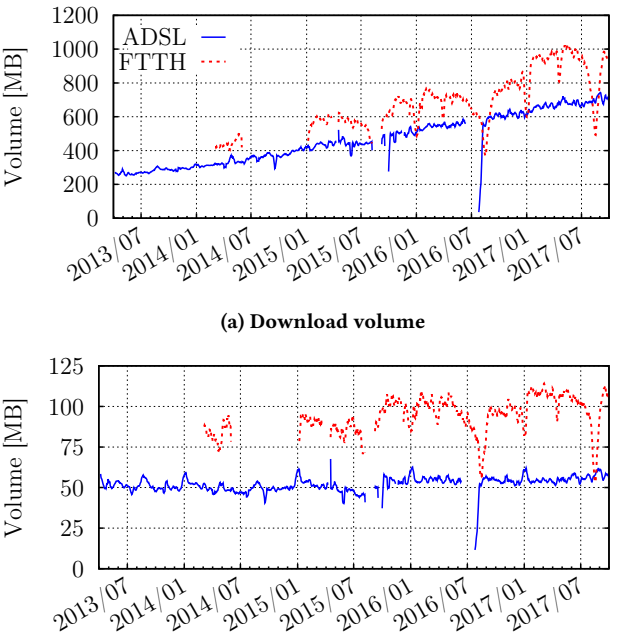
At last, we witness an interesting effect in uploaded traffic: Even if traffic volume increased in median between 2014 and 2017, the tail of the distributions in Figure 2b decreased. Notice the clearly visible bump in the tails present in 2014, which disappeared in 2017. This trend is rooted in the decline of Peer-To-Peer (P2P) traffic, both in volume and popularity, as we will show in Section 4.

3.2 Eager and Eager: Trends on traffic consumption

Figure 3 illustrates per subscriber average traffic consumption over time. The x-axis spans over the 54 months of the dataset, y-axis shows the average byte consumption over monitored subscriptions, separately per access technology and down/up link. Curves in the figure contain interruptions caused by outages in monitoring probes, without affecting trends.[5]

Considering the average amount of data downloaded daily, illustrated in Figure 3a, a clear increasing trend emerges. For ADSL subscribers, average daily traffic increased at a constant rate – from

[5]FTTH figures are noisier than the ADSL ones due to the smaller numbers of FTTH customers. Some drops in FTTH curves are visible during summer and holiday breaks, thanks to the low number of customers and their profiles (e.g., business customers).



(a) Download volume

(b) Upload volume

Figure 3: Average per-subscription daily traffic.

300 MB in 2013 up to 700 MB in late 2017. FTTH subscribers consume on average 25% more traffic, topping to 1 GB per day on average in 2017. Interesting, very similar slow increasing trends have been reported 10 years ago [7].

When considering uploads (Figure 3b), we confirm that the higher uplink capacity lets FTTH users to upload more with respect to ADSL. The latter are bottlenecked and thus the average amount of data remains constant. FTTH subscribers show a modest increase in average uploaded traffic over time. This modest increase

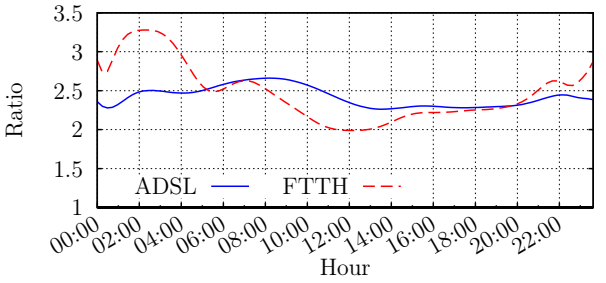


Figure 4: Ratio of traffic consumption between April 2017 and April 2014 for download.

is due to two factors. At the one hand, P2P uploads have decreased significantly in recent years. On the other hand, this decrease has been compensated by a significant increase in the upload of user-generated content to the cloud, including to cloud storage services (e.g., iCloud or Dropbox) as well as to social networks and video providers (e.g., YouTube and Instagram).

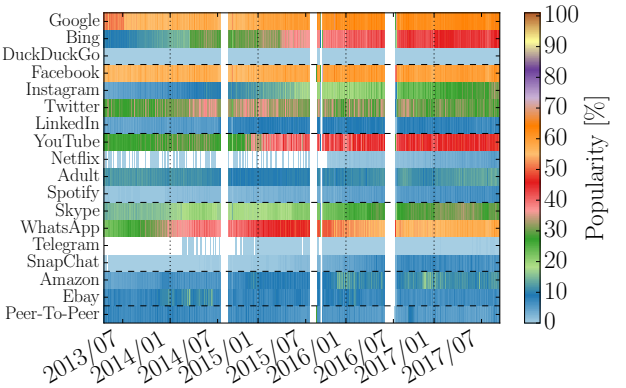
To check whether the increase observed in Figure 3 is homogeneous during the hours of the day, we consider the downloaded volume in each 10 minute-long time interval. We then average all values seen for the same time bin in all days of a month. At last we compute the ratio between April 2017 and April 2014. Figure 4 shows results (curves are smoothed using a Bezier interpolation). It confirms that the average amount of traffic consumed in 2017 is more than 2 times larger than 2014. Interestingly, the increase is higher during late night hours, possibly due to the automatic download of updates of apps and other machine-generated traffic, such as from home IoT devices. FTTH users exhibit also a higher increase during prime time, which we confirm to be associated to the consumption of video streaming content.

4 THE COST OF SERVICES

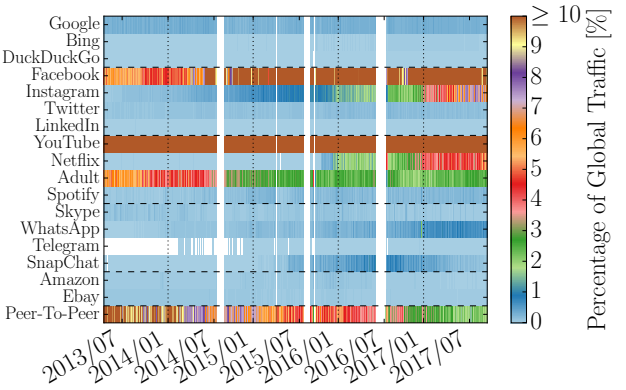
4.1 Give me that: Service popularity

The changes in the per-subscriber traffic volume can be due to changes in the users' habits (e.g., people using different services), or changes in the services (e.g., high definition videos being automatically served). In this section, we analyze in details how popular and bandwidth demanding services evolved throughout years. We again focus on *active subscribers*, observing the fraction of them that accessed a given *service* on a daily basis.

Notice that selecting subscribers that contacted a service is not trivial. Indeed popular services may be *unintentionally* contacted by users. Consider for example Facebook. Its social buttons are embedded in websites and generate traffic to the same Facebook domains as an access to `facebook.com` services. To coarsely distinguish these cases, we have inspected the distribution of daily traffic per subscriber for each considered service. Not reported here for brevity, we manually set per-service thresholds to separate (i) subscribers with at least one visit to the target service (moderate to large traffic volumes), and (ii) subscribers which unintentional contacted domains due third party objects (negligible volumes).



(a) Popularity



(b) Downloaded bytes

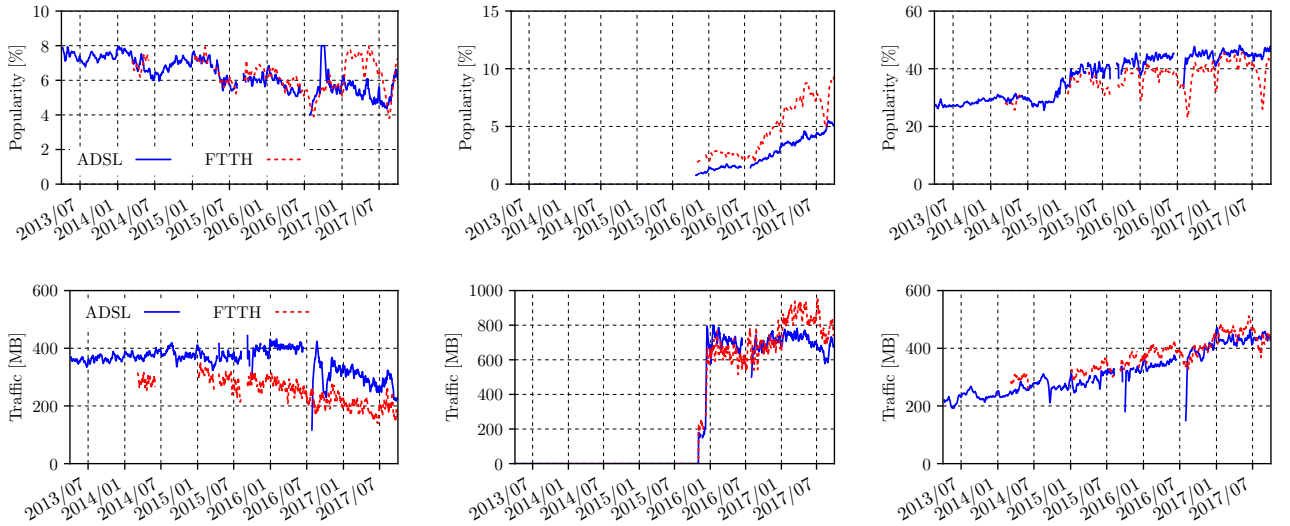
Figure 5: Popularity and percentage of downloaded bytes for selected services over time.

We start by providing a coarse picture about service popularity over time.[6] Figure 5a shows per-day percentage of active users that access popular services. We depict the ADSL data only, since FTTH results in similar figures. The multi-color palette highlights changes in the popularity of services, which are coarsely sorted by type. For instance, Google search engine is accessed regularly by about 60% of active users on a daily basis, and this pattern is rather constant over time.[7] On the contrary, Bing shows a constant growth, moving from less than 15% to about 45% of active users that contacted it at least one time per day in 2017. This pattern is likely a consequence of Windows telemetry which uses `bing.com` domains. Sadly, DuckDuckGo, a privacy respecting search engine, is used only by few tens of users (less than 0.3% of population).

Figure 5b depicts a similar picture for the percentage of downloaded bytes for each service in the ISP traffic mix. The multi-color

[6]Data tables used to generate these figures, including popularity of services and bytes per user per day, can be downloaded from https://smartdata.polito.it/five-years-at-the-edge-watching-internet-from-the-isp-network/.

[7]Some fluctuations are due to changes in Google domains that have taken time to be identified and updated in probes.



(a) Peer-To-Peer **(b) Netflix** **(c) YouTube**

Figure 6: Popularity (top) and volumes (bottom) for P2P and 2 popular video streaming services.

palette is set to 10% to improve the visualization. We can observe how services have changed their contributions to the traffic mix during the monitored period. Notice, for instance, how services such as Facebook, Instagram, WhatsApp and Netflix have increased traffic share throughout the years. Others, such as SnapChat have gained momentum only during a limited period.

Overall, we observe a continuously changing picture, with services showing an increase in popularity and traffic share, some of which with remarkable growth, while others that struggle to gain grounds. Next, we dive into some interesting use cases.

4.2 The downfall of Peer-To-Peer - finally

It is no news that P2P is no longer among the preferred means to download content. Here we quantify this phenomenon showing the popularity of P2P applications over the years. Figure 6a details the percentage of active users using a P2P service (Bittorrent, eMule and variants) (top plot) and the average P2P traffic volume per user (bottom plot). We still observe a hardcore group of users that exchange about 400 MB of P2P data daily. At end of 2016 the traffic volume they generate starts to decrease. Interestingly, FTTH subscribers start abandoning P2P applications earlier in terms of volume. Based on findings of previous studies [18, 28], a conjecture to explain this decline is that the availability of cheap, easy and legal platforms to access content is finally contributing to the downfall of P2P. In the following we explore this conjecture.

4.3 The usual suspects: YouTube and Netflix

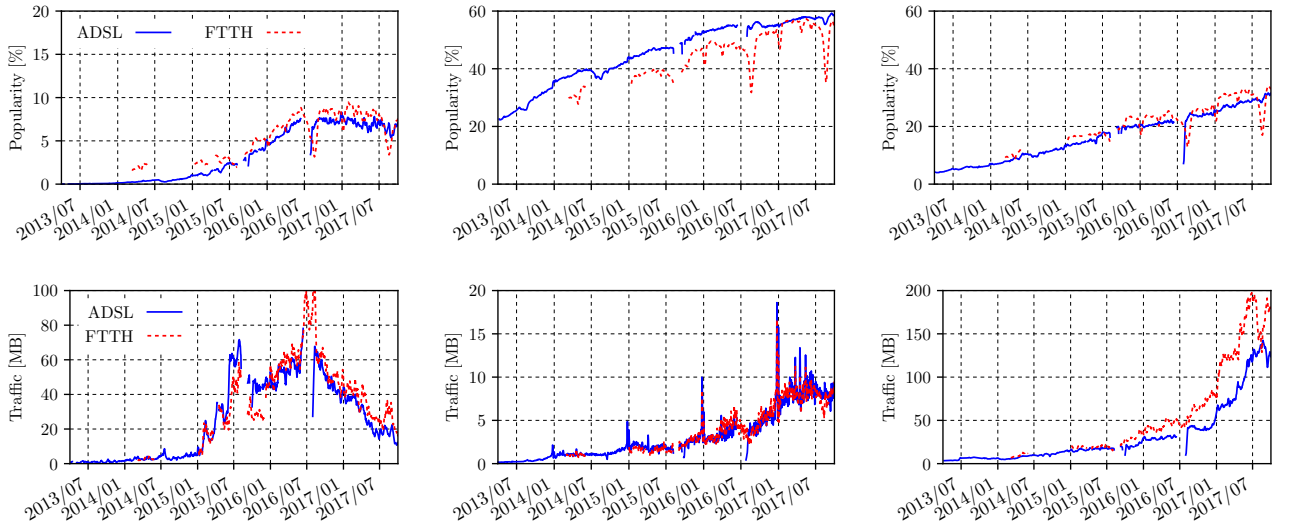
We now consider popular video streaming services. Figure 6b shows the percentage of active users accessing Netflix (top) and the average per-user daily traffic (bottom). Netflix has gained momentum since the day it started operating in Italy. FTTH subscribers have been eager to adopt it, with about 10% of the monitored ISP customers using it on a daily basis at the end of 2017. Considering weekly statistics, we see that more than 18% (12%) of FTTH (ADSL) subscribers access Netflix at least once in 2017. Considering the amount of traffic they consume (bottom plot), we see no major differences between ADSL and FTTH subscribers up to end of 2016. Since October 2016, Netflix started offering Ultra HD content. This is reflected into each active FTTH subscriber downloading close to 1 GB of content on average per day. ADSL subscribers instead cannot enjoy it, or are not willing to pay the extra fee.

Next, we evaluate YouTube (Figure 6c). The figure shows a consolidated service, that is accessed regularly by users, who are consuming more and more content: more than 40% of active subscribers access it daily, and download more than 400 MB (about half of Netflix volume per subscriber). Interestingly, no differences are observed between ADSL and FTTH subscribers – hinting that YouTube video works similarly on FTTH and ADSL.

4.4 The new elephants in the room: Social messaging applications

We now study usage patterns for social messaging applications, namely SnapChat, WhatsApp and Instagram. All are popular applications accessed mostly on smartphones, whose traffic we observe once connected via WiFi from home. As before, we consider popularity and daily traffic consumption per active subscriber (recall Section 4.1), depicted in top and bottom plots in Figure 7.

Interesting trends emerge in the rise and fall of social networking apps. Observe first Snapchat (Figure 7a). It enjoyed a period of notoriety starting from 2015, topping in 2016 when it was adopted by around 10% of subscribers. Each active subscriber used to exchange up to 100 MB of data daily! Starting from 2017, the volume of data starts to decrease, with active subscribers that nowadays exchange



(a) SnapChat **(b) WhatsApp** **(c) Instagram**

Figure 7: Popularity (top) and volumes (bottom) for 3 popular social messaging services.

less than 20 MB per day. Popularity is mostly unaffected, suggesting that people keep having the Snapchat app, but hardly use it.

The decline of SnapChat coincides with the growth of other social apps. See WhatsApp in Figure 7b: Its popularity is indisputable, with a steady growth in adopters that has almost reached saturation. Observe instead the growth in daily volume per active subscriber. Each subscriber exchanges around 10 MB daily, pointing to the intensive use of the app for sharing multimedia content. Note also the large peaks in the figure, corresponding to Christmas and New Year's Eve, when people exchange wishes using WhatsApp.

Finally, considering Figure 7c (Instagram), we see a constant growth in popularity and, more impressive, a massive growth in traffic volumes. Each active subscriber exchanges on average 200 MB and 120 MB per day, for FTTH and ADSL respectively. This is almost a quarter of the traffic of the active customers contacting Netflix! Recalling that Instagram, Snapchat and WhatsApp are predominantly used from mobile terminals, these figures point to a shift on traffic of broadband users, with mobile terminals taking a predominant role even when people are at home.

5 WEB TRENDS, AND SURPRISES

In this section, we study how web protocols usage varied across the last 5 years. We show in particular events associated with the slow migration of services towards newer standard web protocols, and sudden relevant changes on the traffic matrix caused by experiments of big players with custom protocols.

In its early life, the Web was predominantly plain Hyper Text Transfer Protocol (HTTP) traffic. It is by now known that most of the web traffic is running encrypted [32], first with the deployment of HTTPS, followed by the push towards HTTP/2 [3] (which relies on TLS) and more recently QUIC [26]. We here want to document to what extent these protocols have been adopted in the Internet.

Figure 8 answers this question. It shows the traffic share of the several Web protocols observed in the network over time. Five years ago, in 2013, only the two "classic" web protocols were observed, with the majority of traffic served by clear-text HTTP, and only around 13% of the web traffic due to TLS/HTTPS. Then, several notable changes happened, which are marked with letters in the figure:[8]

A) January 2014: YouTube starts serving video streams over HTTPS. The migration has taken Google several months in 2014, in which we can see a steady change in the mix of HTTP and HTTPS traffic. HTTPS share tops to 40% at the end of 2014 already, and it is mainly driven by YouTube traffic.
B) October 2014: After announcing it in 2013, Google starts testing QUIC in the wild deploying its Chrome Web browser. Web traffic carried by QUIC (over UDP) starts growing steadily.
C) June 2015: We update our probes to explicitly report SPDY protocol (previously generically labeled as HTTPS). We discover 10% of traffic carried by an experimental protocol.
D) December 2015: Google disables QUIC for security issues [26]. Suddenly 8% of the traffic falls back to TCP and HTTPS/SPDY. Around a month after, the bug is fixed and QUIC is suddenly back.
E) February 2016: Google migrates traffic from SPDY to HTTP/2, slowly followed by other players.
F) November 2016: Facebook suddenly deploys "FB-Zero", a protocol with a custom 0-RTT modification of TLS used from the Facebook mobile app only.[9] Suddenly, 8% of web traffic moves to this new protocol. More than a half of Facebook

[8]These events have been confirmed manually throughout the years while upgrading the software of our probes to keep-up with protocols evolution.
[9]Zero protocol would be announced only in January 2017 – https://goo.gl/vuQ1Jy

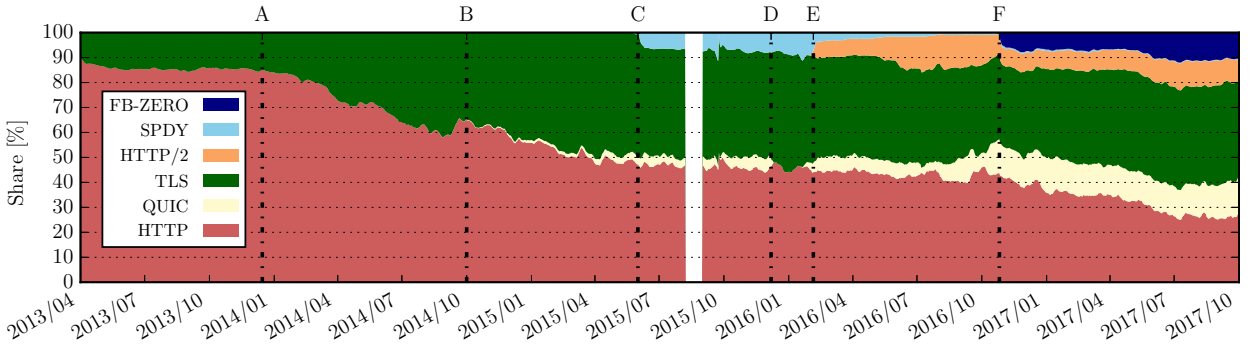


Figure 8: Web protocol breakdown over 5 years. Sudden changes and custom protocols deployment in the wild are highlighted.

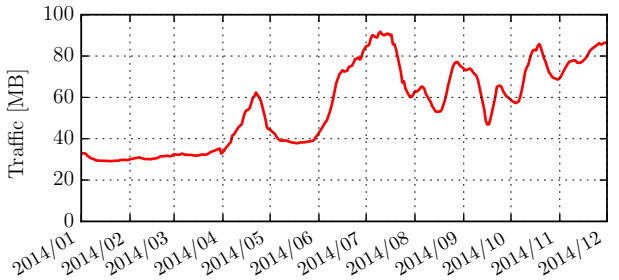


Figure 9: Facebook average daily per-user traffic before and after automatic video play.

traffic is now carried by Zero, showing that mobile app traffic surpassed website, even for fixed ADSL installations.

At the end of 2017, HTTP is down to 25%, with HTTP/2 that is slowly gaining momentum. QUIC and Zero together carry 20–25% of web traffic. Both are yet to be standardized protocols, showing how giants like Google and Facebook are free to deploy experiments on the web, since they own both server and client applications. Such experiments may create issues for ISP network administrators, e.g., making network proxies and firewalls suddenly inefficient, or creating issues with home gateway.

Finally, we illustrate in Figure 9 another interesting episode of sudden traffic changes. Around March/April 2014, Facebook started enabling video auto-play for its applications. The immediate effect on ISP traffic is striking. Figure 9 illustrates the daily average traffic per subscriber towards Facebook. Starting in March 2014 traffic has grown from around 35 MB to around 70 MB in a month. After an apparent pause in the deployment of the feature during May, the service enabled video auto-play again. In July, the daily traffic per subscriber was around 90 MB on average, 2.5 times higher than the rate observed in March 2014!

This figure illustrate once more how the big players controlling key client software and servers can deploy impactful changes in the Internet, complicating the planning and management of ISP networks.

6 WHERE ARE MY SERVERS?

In the previous section we have witnessed both slow and sudden changes due to overall trends, and big players migration policies. Here we go deeper into showing the impact of big players infrastructure changes over the years.

6.1 The birth of the sub-millisecond Internet

CDNs were born in the '90s to reduce both the load on centralized server and the delay to access the content. Nowadays shared and private CDNs are making it possible to scale Internet content distribution, allowing users to fetch content from nearby surrogate servers. Being delay one of the main parameters affecting users' Quality of Experience, we focus our attention on how it changed over years.

We consider the Round Trip Time (RTT) as performance index. Remind that probes measure RTT by matching TCP segments sent by clients with corresponding TCP ACKs sent by servers. We focus on the RTT from the probe to the server – excluding the access network delay. For all TCP connections to a given service, we extract the minimum per-flow RTT, and plot the corresponding CDF. By doing so for a long time interval and large sample of users, we can spot how the RTT distribution is composed. Thus, we focus on the body of the distribution of minimum per-flow RTT, ignoring samples in the tails of the distribution, which may be caused by queuing and processing delays.

Figure 10 shows the results contrasting measurements seen in April 2014 versus April 2017. We focus on Facebook and Google services as notable examples of big players that pay particular attention to speed up content delivery. Consider Instagram traffic (red curves) on Figure 10a. Dashed line refers to 2014 figures. At those time, there were already CDN surrogate nodes at just 3 ms RTT from the ISP PoP. However it served only 10% of flows. Other traffic was served by far away CDN nodes, with RTT of 10, 20 and 30 ms.[10] About 7% of flows was served by servers with RTT higher than 100 ms – a clear sign of intercontinental path. Facebook caches (blue curves) follows a very similar placement – with different share of traffic being served by different caches.

Consider now the 2017 CDF (solid lines). Results clearly show that many more requests are now served by close servers, with

[10]Fraction changes by hour. Figures refer to statistics collected on the whole month.

80 % of both Instagram and Facebook traffic that is served by the 3 ms far CDN nodes. As we will see later, this change is due to two factors: i) Facebook that deployed its own CDN; And ii) Instagram infrastructure being integrated into Facebook one.

Look now at Figure 10b which depicts the RTT CDF for Google web search servers and YouTube streaming servers. In 2014, 80 % of YouTube traffic (blue curves) was already being served by nodes that were just 3 ms far away from the ISP PoP. This is to guarantee the high volume due to video traffic. In 2017, this already marginal figure decreased even more – with the YouTube video cache now breaking the sub millisecond RTT. That is, YouTube now directly places video servers inside the PoP, at the first level of aggregation, going further towards a very distributed and pervasive infrastructure. Interestingly, Google search engine web servers (red curves) have not yet reached such a fine grained penetration. This is because they have to handle less traffic, and perform more complicated processing than YouTube video caches.

We have confirmed these findings by directly contacting the ISP staff, who reported the deployment of third-party CDN and cache nodes at the ISP first aggregation point.

We repeated the analysis for other services – not reported for the sake of brevity. With the only notable exception of WhatsApp, whose servers are still following a centralized approach with RTT in the 100 ms range, all services are exhibiting the same trend, with more and more CDN surrogate servers being placed closer and closer to the edge of the network.

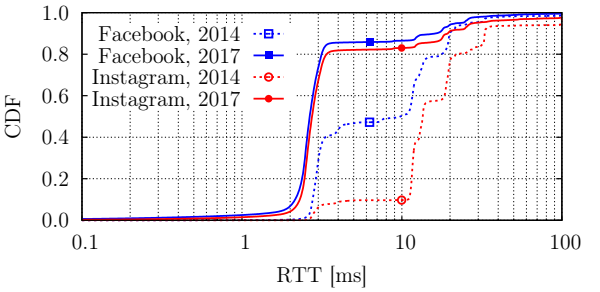
On the one hand, this proliferation of edge caches, and the delay of modern FTTH access network is leading us to the sub-millisecond Internet [37]. On the other hand, this poses new burdens on the ISPs, which have to host (and in some cases manage) infrastructure of different content and CDN providers inside their network. Network Function Virtualization (NFV) would possibly help in reducing this burden [21], allowing ISPs to host virtual CDN surrogates into their infrastructure.

6.2 The Internet of few giants

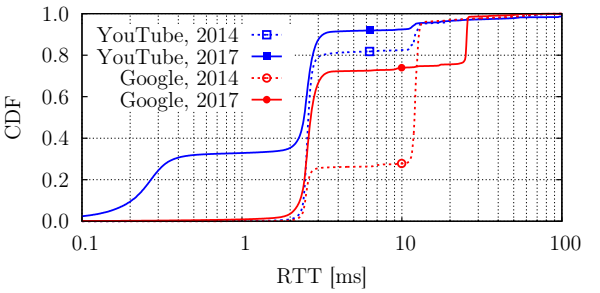
We now analyse the infrastructure of large content providers. Indeed, during the last 5 years most web services incurred restructuring, replacing servers, deploying their own CDN, etc.

Figure 11 depicts the evolution over time of the infrastructure of Facebook (left plots), Instagram (center plots), and YouTube (right plots). Top plots show the server IP addresses being active in each day, for the considered service. The y-axis represents a single server IP address, sorted in order of appearance. A red dot is present if for that day, the IP address was being used only for traffic of the considered service. A blue dot is present if that IP addresses served also content for other services. Finally, no dot is present if the IP address was not contacted in that day.

In all cases, we see that new IP addresses keep appearing over time, counting several tens of thousands unique IP addresses. Compare Facebook and Instagram in Figure 11a and Figure 11b, respectively. During 2013 and 2014, a good fraction of addresses were shared with other services. During the second half of 2015, we notice that both started having major changes, with i) a decrease in the number of servers being contacted, and ii) a specialization of servers that are not shared with any other services. In details, the



(a) CDF of RTT in 2014 and 2017 for Facebook and Instagram.



(b) CDF of RTT in 2014 and 2017 for YouTube and Google.

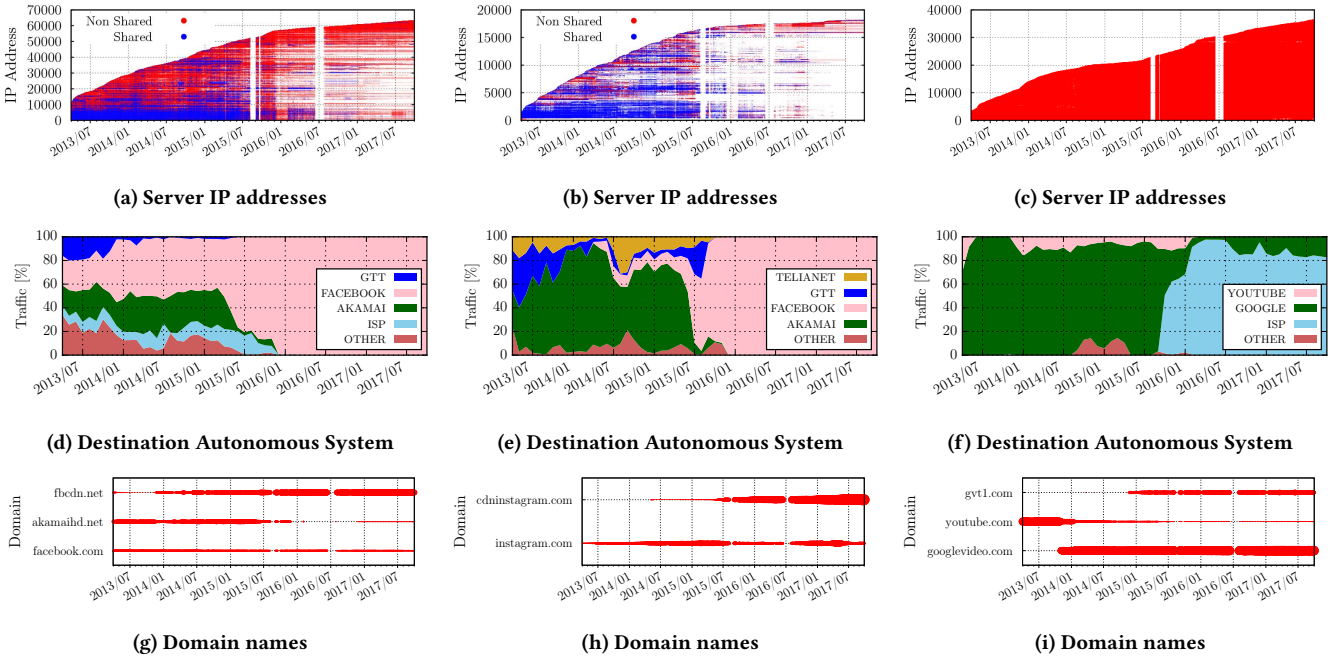
Figure 10: CDF of Round Trip time.

total number of IP addresses used daily by Facebook dropped from 3 800 to less than 1 000, out of which 700 are still shared. Since July 2016, shared IP addresses drop to very few.

To better understand the reason behind this major change, we analyze to which Autonomous System Number (ASN) each IP addresses belonged.[11] Middle plots in Figure 11 show the breakdown of the per-day contacted IP addresses over major ASNs. Figure 11d and Figure 11e show a migration from generic CDNs to the Facebook private CDN. In 2013, both services used third party CDNs, whose IP addresses where thus shared with other services. For Facebook, the migration started before 2013, and was completed by the end of 2015. For Instagram, the integration with Facebook infrastructure started in 2014 (Facebook acquired Instagram in April 2012), and was completed by end of 2015. This migration has two major effects: i) IP addresses are now dedicated to either Facebook, or Instagram; ii) the number of IP addresses contacted per day reduces. Indeed since 2016 only 1 000 IP addresses are used to serve Facebook traffic, and only 300 for Instagram. Contrasting these figures with Figure 10a, we notice that this change also benefited the RTT, which reduced significantly.

To better describe these changes, bottom plots in Figure 11 detail the traffic share served by most important second level domain names. The thicker is the line, the higher is the fraction of traffic served. For instance, Figure 11g confirms the migration from generic Akamai CDN to Facebook proprietary infrastructure. Even more evident is the migration for Instagram in Figure 11h.

[11] We use the Routing Information Base(RIB) for each month from a major vantage point in the Route Views project to map IP addresses to ASNs



(a) Server IP addresses (b) Server IP addresses (c) Server IP addresses

(d) Destination Autonomous System (e) Destination Autonomous System (f) Destination Autonomous System

(g) Domain names (h) Domain names (i) Domain names

Figure 11: Facebook (left), Instagram (center) and YouTube (right) infrastructure evolution over time.

Finally, we study the YouTube infrastructure evolution as a case of study of a very popular service with a massive infrastructure. From Figure 11c, it is already possible to see how different YouTube is with respect to the previous two cases. Indeed, YouTube always used a totally dedicated infrastructure to serve videos. Its infrastructure keeps growing until now, where 40 000 IP addresses are used daily. By looking at Figure 11f, we observe that starting from the end of 2015, the caches deployed in the ISP start serving most of YouTube traffic. This benefited RTT as previously shown. Regarding to the Domain names used by YouTube, Figure 11i shows three main changes: until January 2014, all the traffic was served by the *youtube.com* domain; In 2014 the *googlevideo.com* domain suddenly appeared, and immediately handled the majority of traffic; Finally, in 2015 YouTube introduced *gvt1.com*.

These results confirms the trend toward a consolidation of large services, which deploy their own infrastructure, in a more and more capillary way, reaching several tens of thousands of IP addresses. Furthermore, these infrastructure undergo sudden and undocumented changes that have impact on the traffic monitoring and management of IPSs and corporate networks.

7 RELATED WORK

Several works measured Internet traffic from different points of views. Gebert *et al.* [19] characterized the observed traffic mixtures in an ISP network during 14 days. Liu *et al.* [27] designed a large scale measurement infrastructure and deployed it in the core network of a cellular operator. Their focus is on the architecture, not on measurements. Authors of [17] reported their experience on operating a monitoring infrastructure in ISP networks during 20 months in 2013, describing how protocols and services are typically consumed from such networks. Muhammad *et al.* [35] analyzed a week-long traffic trace collected from a tier-1 cellular network, showing how machine-to-machine traffic is different from human-generated traffic. All these works cover a relatively short period, which prevent them to evaluate how the identified phenomena have evolved over time.

Some works provided longitudinal views on Internet evolution. The authors of [12] analyzed a dataset of BGP measurements that covers 12 years, showing how the BGP ecosystem has evolved. Authors of [15] presented one of the first longitudinal studies of Internet traffic, covering the period of 1998–2003. Authors of [6] evaluated 7 years of MAWI traces, summarizing the evolution of Internet traffic in Japan. In [5] authors evaluated 23 months of data collected from 53 k broadband installations, highlighting for instance the relation between capacity and demand.

Our work is similar to those efforts in terms of the employed methodology and general goals. Similar to [5, 6, 12, 15] we focus on long-term trends instead of exploring details of a measurement snapshot. We report statistics and trends about users' habits, usage of services and protocols, while also focusing on the infrastructure changes. More important, we show figures from a recent period, thus updating the knowledge about Internet usage.

Also in terms of methodology, we monitor close to end-users (e.g., similar to [6, 16, 28]) and not in the core (e.g., as in [25, 33]). This allows us to provide a comprehensive picture of users' data consumption, which is particularly relevant for ISPs.

Regarding our conclusions, we highlight many interesting facts about the Internet traffic mix. A number of recent studies also reported on Internet traffic mix using different vantage points. Authors of [33] reported the traffic observed in an IXP in 2013, comparing their findings to other vantage points [19, 28]. Labovitz *et al.* [25] analyzed two years of network measurements collected from several Internet backbones, illustrating how core Internet traffic is converging around few big players.

Our work updates these studies showing trends from 2013 onward. Similar to [20] and others, we present traffic mix focusing on services and the most popular application-layer protocols. Whereas our data would allow us to drill down on per-protocol breakdowns (e.g., as in [10]), these details are left out for the sake of brevity.

As said above, many of our conclusions validate results already identified in previous works. Examples of known results that are confirmed or extended by our measurements include: (i) the slow increasing trend on traffic per user [7]; (ii) the predominance of video traffic [1, 13]; (iii) the fast increase in HTTPS deployment [14]; (iv) the decline of P2P [18, 28]; (v) the concentration of Internet traffic around few big players [25]; (vi) the deployment of experimental protocols resulting in sudden changes in the traffic mix due to bugs and private tests by large companies [24, 26, 34].

In some other cases, our results add more data points to complement previous findings. For example, we could not find a clear general relation between the capacity customers and their demands as in [5]. However, for customers relying on particular services (like Netflix) these conclusions seem to hold true. Besides that, we also shed light on new aspects of the Internet evolution, such as the costs of services to providers, the usage dynamics of new social network services such as Instagram and Snapchat, among others.

Finally, some companies such Cisco periodically report traffic trends and forecasts [8], including predictions on connected devices, Internet usage and traffic nature. By reporting detailed statistics from measurements collected in operational networks, our work complements such studies and can contribute in gaining a better understanding of Internet traffic.

8 CONCLUSION

In this paper we evaluated the evolution of Internet traffic during 5 years (2013–2017). By processing large scale and longitudinal measurements from a national ISP in Italy, we characterized the traffic consumption of broadband subscribers, and the infrastructure web services deploy to reach customers. We observed subscribers' daily traffic that more than doubled in the analyzed period. We studied the typical loads imposed by popular and bandwidth hungry services. We testified the death of P2P in exchange for legal, cheap and easy to use video content, and the quick rise and sudden death of social messaging applications typically accessed via mobile phones, able to generate massive amount of data.

We observed the concentration of services within few big Internet providers, each deploying its own infrastructure, unrolling custom protocols, and penetrating more and more network boundaries. In the rush to bring servers closer and closer to users, we witnessed the birth of the sub-millisecond CDNs, where Internet giants like Google or Facebook are placing caches directly in the ISP PoPs. All such changes and their unpredictability complicate the planning and management of the networks, possibly calling for closer integration between content providers and operators.

We believe the figures we presented in this paper are vital to researchers, ISPs and even web service provider to better understand the liveliness of the Internet, which continuously changes, mixing slow and unpredictable changes.

ACKNOWLEDGMENTS

The research leading to these results has been funded by both the Vienna Science and Technology Fund (WWTF) through project ICT15-129 (BigDAMA) and the SmartData@PoliTO center for Big Data technologies. We would like to thank also the technical teams from the ISP providing data for this research for their continuous collaboration and support.

REFERENCES

[1] Vijay Kumar Adhikari, Sourabh Jain, and Zhi-Li Zhang. 2010. YouTube Traffic Dynamics and Its Interplay with a Tier-1 ISP: An ISP Perspective. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC'10)*. 431–443. http://doi.acm.org/10.1145/1879141.1879197

[2] Paul Barford and Mark Crovella. 1999. Measuring Web Performance in the Wide Area. *SIGMETRICS Perform. Eval. Rev.* 27, 2 (1999), 37–48.

[3] Mike Belshe, Roberto Peon, and Martin Thomson. 2015. *Hypertext Transfer Protocol Version 2 (HTTP/2)*. Technical Report 7540. RFC Editor.

[4] Ignacio Bermudez, Marco Mellia, Maurizio M. Munafò, Ram Keralapura, and Antonio Nucci. 2012. DNS to the Rescue: Discerning Content and Services in a Tangled Web. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC'12)*. 413–426. http://doi.acm.org/10.1145/2398776.2398819

[5] Zachary S. Bischof, Fabian E. Bustamante, and Rade Stanojevic. 2014. Need, Want, Can Afford: Broadband Markets and the Behavior of Users. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC'14)*. 73–86. http://doi.acm.org/10.1145/2663716.2663753

[6] Pierre Borgnat, Guillaume Dewaele, Kensuke Fukuda, Patrice Abry, and Kenjiro Cho. 2009. Seven Years and One Day: Sketching the Evolution of Internet Traffic. In *IEEE INFOCOM 2009 (INFOCOM'09)*. 711–719. https://ieeexplore.ieee.org/document/5061979

[7] Kenjiro Cho, Kensuke Fukuda, Hiroshi Esaki, and Akira Kato. 2008. Observing Slow Crustal Movement in Residential User Traffic. In *Proceedings of the 2008 ACM CoNEXT Conference (CoNEXT'08)*. 1–12. http://doi.acm.org/10.1145/1544012.1544024

[8] Cisco. 2017. Visual Networking Index. https://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html

[9] Kimberly C. Claffy. 2000. Measuring the Internet. *IEEE Internet Computing* 4, 1 (2000), 73–75.

[10] Jakub Czyz, Mark Allman, Jing Zhang, Scott Iekel-Johnson, Eric Osterweil, and Michael Bailey. 2014. Measuring IPv6 Adoption. *SIGCOMM Comput. Commun. Rev.* 44, 4 (2014), 87–98.

[11] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* 51, 1 (2008), 107–113.

[12] Amogh Dhamdhere and Constantine Dovrolis. 2011. Twelve Years in the Evolution of the Internet Ecosystem. *IEEE/ACM Trans. Netw.* 19, 5 (2011), 1420–1433.

[13] Jeffrey Erman, Alexandre Gerber, K.K. Ramadrishnan, Subhabrata Sen, and Oliver Spatscheck. 2011. Over the Top Video: The Gorilla in Cellular Networks. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC'11)*. 127–136. http://doi.acm.org/10.1145/2068816.2068829

[14] Adrienne Porter Felt, Richard Barnes, April King, Chris Palmer, Chris Bentzel, and Parisa Tabriz. 2017. Measuring HTTPS Adoption on the Web. In *Proceedings of the 26th USENIX Security Symposium (USENIX Security'17)*. 1323–1338. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/felt

[15] Marina Fomenkov, Ken Keys, David Moore, and Kimberly C. Claffy. 2004. Longitudinal Study of Internet Traffic in 1998-2003. In *Proceedings of the Winter International Synposium on Information and Communication Technologies (WISICT'04)*. 1–6. http://dl.acm.org/citation.cfm?id=984720.984747

[16] Chuck Fraleigh, Sue Moon, Bryan Lyles, Chase Cotton, Mujahid Khan, Deb Moll, Rob Rockell, Ted Seely, and Christophe Diot. 2003. Packet-level Traffic Measurements from the Sprint IP Backbone. *Netwrk. Mag. of Global Internetwkg.* 17, 6 (2003), 6–16.

[17] José Luis García-Dorado, Alessandro Finamore, Marco Mellia, Michela Meo, and Maurizio M. Munafò. 2012. Characterization of ISP Traffic: Trends, User Habits, and Access Technology Impact. *IEEE Trans. Netw. Service Manag.* 9, 2 (2012), 142–155.

[18] Arnau Gavalda-Miralles, David R. Choffnes, John S. Otto, Mario A. Sanchez, Fabián E. Bustamante, Luís Amaral, Jordi Duch, and Roger Guimera. 2014. Impact of Heterogeneity and Socioeconomic Factors on Individual Behavior in Decentralized Sharing Ecosystems. *Proceedings of the National Academy of Sciences* 111, 43 (2014), 15322–15327.

[19] Steffen Gebert, Rastin Pries, Daniel Schlosser, and Klaus Heck. 2012. Internet Access Traffic Measurement and Analysis. In *Proceedings of the 4th International Conference on Traffic Monitoring and Analysis (TMA'12)*. 29–42. http://dx.doi.org/10.1007/978-3-642-28534-9_3

[20] Alexandre Gerber and Robert Doverspike. 2011. Traffic Types and Growth in Backbone Networks. In *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference*. 1–3. http://www.osapublishing.org/abstract.cfm?URI=OFC-2011-OTuR1

[21] Bo Han, Vijay Gopalakrishnan, Lusheng Ji, and Seungjoon Lee. 2015. Network Function Virtualization: Challenges and Opportunities for Innovations. *IEEE Commun. Mag.* 53, 2 (2015), 90–97.

[22] Rick Hofstede, Pavel Čeleda, Brian Trammell, Idilio Drago, Ramin Sadre, Anna Sperotto, and Aiko Pras. 2014. Flow Monitoring Explained: From Packet Capture to Data Analysis with NetFlow and IPFIX. *Commun. Surveys Tuts.* 16, 4 (2014), 2037–2064.

[23] Intel. 2011. DPDK - Data Plane Development Kit. http://dpdk.org

[24] Arash Molavi Kakhki, Samuel Jero, David Choffnes, Cristina Nita-Rotaru, and Alan Mislove. 2017. Taking a Long Look at QUIC: An Approach for Rigorous Evaluation of Rapidly Evolving Transport Protocols. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC'17)*. 290–303. http://doi.acm.org/10.1145/3131365.3131368

[25] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. 2010. Internet Inter-Domain Traffic. In *Proceedings of the ACM Conference on Data Communication (SIGCOMM'10)*. 75–86. http://doi.acm.org/10.1145/1851182.1851194

[26] Adam Langley, Janardhan Iyengar, Jeff Bailey, Jeremy Dorfman, Jim Roskind, Joanna Kulik, Patrik Westin, Raman Tenneti, Robbie Shade, Ryan Hamilton, Victor Vasiliev, Alistair Riddoch, Wan-Teh Chang, Zhongyi Shi, Alyssa Wilk, Antonio Vicente, Charles Krasic, Dan Zhang, Fan Yang, Fedor Kouranov, and Ian Swett. 2017. The QUIC Transport Protocol: Design and Internet-Scale Deployment. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'17)*. 183–196. http://dl.acm.org/citation.cfm?doid=3098822.3098842

[27] Jun Liu, Feng Liu, and Nirwan Ansari. 2014. Monitoring and Analyzing Big Traffic Data of a Large-Scale Cellular Network with Hadoop. *Netwrk. Mag. of Global Internetwkg.* 28, 4 (2014), 32–39.

[28] Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman. 2009. On Dominant Characteristics of Residential Broadband Internet Traffic. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC'09)*. 90–102. http://doi.acm.org/10.1145/1644893.1644904

[29] Marco Mellia, Michela Meo, Luca Muscariello, and Dario Rossi. 2006. Passive Identification and Analysis of TCP Anomalies. In *Proceedings of the IEEE International Conference on Communications (ICC'06)*. 723–728. http://ieeexplore.ieee.org/document/4024214/

[30] David Moore, Colleen Shannon, Douglas J. Brown, Geoffrey M. Voelker, and Stefan Savage. 2006. Inferring Internet Denial-of-service Activity. *ACM Trans. Comput. Syst.* 24, 2 (2006), 115–139.

[31] Victor Moreno, Javier Ramos, Pedro M. Santiago del Río, José Luis García-Dorado, Francisco J. Gomez-Arribas, and Javier Aracil. 2015. Commodity Packet Capture Engines: Tutorial, Cookbook and Applicability. *Commun. Surveys Tuts.* 17, 3 (2015), 1364–1390.

[32] David Naylor, Alessandro Finamore, Ilias Leontiadis, Yan Grunenberger, Marco Mellia, Maurizio Munafò, Konstantina Papagiannaki, and Peter Steenkiste. 2014. The Cost of the "S" in HTTPS. In *Proceedings of the 10th ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT'14)*. 133–140. http://doi.acm.org/10.1145/2674005.2674991

[33] Philipp Richter, Nikolaos Chatzis, Georgios Smaragdakis, Anja Feldmann, and Walter Willinger. 2015. Distilling the Internet's Application Mix from Packet-Sampled Traffic. In *Proceedings of the 16th International Conference on Passive and Active Measurement (PAM'15)*. 179–192. https://doi.org/10.1007/978-3-319-15509-8_14

[34] Jan Rüth, Ingmar Poese, Christoph Dietzel, and Oliver Hohlfeld. 2018. A First Look at QUIC in the Wild. In *Proceedings of the 19th International Conference on Passive and Active Measurement (PAM'18)*. 255–268. https://doi.org/10.1007/978-3-319-76481-8_19

[35] Muhammad Zubair Shafiq, Lusheng Ji, Alex X. Liu, Jeffrey Pang, and Jia Wang. 2012. A First Look at Cellular Machine-to-machine Traffic: Large Scale Measurement and Characterization. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'12)*. 65–76. http://doi.acm.org/10.1145/2254756.2254767

[36] Yuval Shavitt and Eran Shir. 2005. DIMES: Let the Internet Measure Itself. *SIGCOMM Comput. Commun. Rev.* 35, 5 (2005), 71–74.

[37] Ankit Singla, Balakrishnan Chandrasekaran, P. Brighten Godfrey, and Bruce Maggs. 2014. The Internet at the Speed of Light. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks (HotNets'14)*. 1–7. http://doi.acm.org/10.1145/2670518.2673876

[38] Martino Trevisan, Idilio Drago, Marco Mellia, and Maurizio M. Munafò. 2016. Towards Web Service Classification using Addresses and DNS. In *Proceedings of the 7th International Workshop on Traffic Analysis and Characterization (TRAC'16)*. 38–43. https://ieeexplore.ieee.org/document/7577030/

[39] Martino Trevisan, Alessandro Finamore, Marco Mellia, Maurizio Munafo, and Dario Rossi. 2017. Traffic Analysis with Off-the-Shelf Hardware: Challenges and Lessons Learned. *IEEE Commun. Mag.* 55, 3 (2017), 163–169.

[40] Carey Williamson. 2001. Internet Traffic Measurement. *IEEE Internet Computing* 5, 6 (2001), 70–74.

[41] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster Computing with Working Sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud'10)*. 10–10. http://dl.acm.org/citation.cfm?id=1863103.1863113