A Decade Long View of Internet Traffic Composition in Japan

Irina Tsareva*, Trinh Viet Doan*, Vaibhav Bajpai†
*Technical University of Munich, Germany
[irina.tsareva | trinhviet.doan]@tum.de
†CISPA Helmholtz Center for Information Security, Germany
vaibhav.bajpai@cispa.de

***Abstract*—The Internet traffic composition has changed considerably over the last decade and continues to evolve as a consequence of several factors. In this paper, we showcase this change taking a holistic view derived from the deployment of new and emerging protocols (QUIC, encrypted DNS), shifting user behavior towards a (more) secure Web, natural incidents (Covid-19 pandemic), and effects of peering with large content delivery hyper-giants (Google) that tend to have both short– and long–lasting impact on Internet traffic. Knowledge of such changing trends is essential for better management of networks and services. To this end, we analyze >6.6 TiB of data collected at a large backbone link in Japan (MAWI dataset) and investigate traffic composition across several months and years. We observe that IPv6 traffic in 2019 represents a volume comparable to that of the total traffic observed in 2007 and is (now) increasingly used to carry Web traffic. Over IPv4, we observe significant growth in encrypted Web traffic with HTTPS-to-HTTP ratio evolving to 2-to-1 in 2019 compared to less than 2% HTTPS share in 2007. Meanwhile, we witness (for the first time) an alteration of the traffic composition on the educational network as a result of the Covid-19 pandemic. We observe a vanishing weekday-weekend pattern and a shift towards increased usage of OpenVPN and `rsync` traffic due to increased remote work. Finally, we also study the impact of Google as a peering entity in the routing ecosystem and observe a significantly increased traffic share of QUIC over both address families, and overall a larger HTTP-to-HTTPS ratio in terms of traffic volume.**

I. Introduction

The Internet traffic composition constantly changes due to various factors. One factor is the increasing awareness of privacy [1] and security on the Internet. Much effort has been done to push towards a more secure Web, e.g., by Let's Encrypt or by Google. Consequently, HTTPS traffic has been steadily increasing, although, the adoption of encrypted Web traffic varies by country, mobile/desktop usage, and also the content of the website [2]. Another essential part of the Internet is the Domain Name System (DNS), yet, it does not guarantee confidentiality and integrity because of missing encryption. The IETF lately standardized DNS over TLS (DoT) [3], [4] and DNS over HTTPS (DoH) [5] as two secure alternatives for DNS. Traffic over DoH and DoT is expected to increase after Apple added support for DoT and DoH to MacOS/iOS devices [6] and public DNS services [7] added support for these new protocols. Another factor that contributes to changing traffic composition is emerging protocols such as QUIC [8]. Since Google first introduced QUIC in 2013, the protocol has been further developed and improved by Google and standardized [9] by the IETF. In Oct/20, >33% of Google's traffic was carried over QUIC [10]. Additionally, QUIC is the basis of HTTP/3 [11], the newest version of HTTP(s) that underpins the Web in its narrow waist; the recent standardization of DNS over QUIC [12]–[14] adds another secure DNS alternative. Further factors that can impact the Internet traffic are pandemics, which can change the Internet traffic significantly within a short period of time. For instance, with the COVID-19 pandemic, people are advised to work from home; thus, shifting face-to-face communication to online communication, e.g., via instant messaging, emails, or more bandwidth-intense video conferencing. As such, people find new ways to spend their time, such as with online gaming [15]. Consequently, the traffic workload has increased drastically: For example, the DE-CIX in Frankfurt reported a world record of >9 Terabits per second (Tbps) traffic exchanged on March 19, 2020 [16].

Previous studies [17], [18] have investigated some of these factors. Understanding the reaction to incidents as well as predicting Internet trends is crucial to allow ISP networks and content providers to adapt to such changes by reconfiguring networks or by increasing network capacities. Such insights are also relevant for the scientific community to understand the impact of such changes on the regular usage of the Internet. In this paper, we therefore ask the following questions: *How did the Internet traffic composition and application mix evolve over the last (2007→2019) decade? How and to what extent does peering with large content delivery hyper-giants (Google for example) change the traffic composition? How did the Covid-19 pandemic (2020−) impact the Internet?*

To answer these questions, we leverage the MAWI dataset collected by the WIDE project at a large backbone located in Japan. The MAWI dataset includes traces captured at two samplepoints: F (link to an upstream provider) and G (link to an Internet exchange point (IXP) in Tokyo). In this work, we study >6.6TiB dataset collected at Samplepoint−F and −G (§II). Our main findings are:

1. **Traffic volumes** (§III) — We find that the amount (by volume) of IPv6 traffic in 2019 is comparable to the total amount of traffic a decade ago (in 2007). There is a larger share of small packets over IPv4 in 2019, as well as an increase

ISBN 978-3-903176-57-7 © 2023 IFIP

in UDP traffic (in 2019) due to QUIC, although TCP flows are larger and TCP still remains the dominant transport layer protocol in terms of traffic volume.

2. **Application mixes** (§IV) — In 2007, >70% of traffic was HTTP, while the HTTPS share was less than 2%. We notice substantial growth in encrypted Web traffic in the last decade, with the HTTPS-to-HTTP ratio over IPv4 changing to 2-to-1 in 2019. At the same time, we observe a decline in P2P traffic. While, UDP traffic (in 2007) was mostly DNS, QUIC (<10%) starts to contribute to the mix. In 2019, we also observe DNS traffic over TCP, some of which is also DNS over (D)TLS. Over IPv6, traffic composition is increasingly over the Web, which was previously (in 2007) dominated by DNS and `rsync` in the last decade.

3. **COVID-19 pandemic** (§V) — The Covid-19 pandemic results in a significant alteration in the traffic composition. We observe decreased traffic volume (after schools and universities closed) as an emergency response. During the Japanese academic year, the daily traffic volume is now as low as the traffic volume seen over weekends in 2019. Consequently, a characteristic weekend-weekday traffic pattern has disappeared. This trend coincides with decreased HTTPS traffic (>3 times) with a simultaneous increase in usage of remote access protocols (`rsync`, OpenVPN, IPSec and NAT traversal methods) and file sharing applications (Dropbox) with less traffic exchanges with US telecommunication services.

4. **Peering with Google** (§VI) — We find that peering with Google largely increases traffic volume. At Samplepoint−G (peered with Google), the monthly traffic volume (in bytes) is significantly larger by a factor of 4.3 (2019) to 10 (2020) compared to that of Samplepoint−F. Further, QUIC traffic share is also higher, with an average share of >7% over IPv6, while at Samplepoint−F, IPv6 carries relatively less QUIC. The HTTP-to-HTTPS ratio is 9-to-5 in favor of HTTP. Lastly, we observe a large traffic volume originating from educational autonomous systems (ASes), while at Samplepoint−F, one third of the traffic originates from content-type ASes.

II. Datasets

MAWI Dataset — The WIDE backbone is "a mixture of commodity traffic and research experiments". It offers two active samplepoints at which traces are captured: One samplepoint monitors a Trans-Pacific link from the WIDE backbone to the upstream provider NTTC GIN (AS2914) (Samplepoint-F), and the other samplepoint (Samplepoint-G) collects traces at a link that connects the backbone with its main IXP (DIX-IE). This Internet exchange point includes 22 peers, one of which is Google LLC [19]. The traces on Samplepoint-F are collected daily from 14:00 to 14:15. Samplepoint-G records weekly traces on Wednesdays from 14:00 to 14:15. The traces do not contain any payload of the transport layer and IP addresses are anonymized. We requested non-anonymized data from MAWI for years 2019 (Jan-Dec) and 2020 (Jan-May) to specifically dissect traffic by TLS versions. Due to the size of the data, we only request traces of Wednesdays and Sundays for each of the months. This way, the non-anonymized data contains a representative weekday and day of the weekend. To pre-process the pcap-files, we use *YAF* in combination with *SiLK* [20] to aggregate the `pcap` files into bidirectional flows and IPFIX-like format.

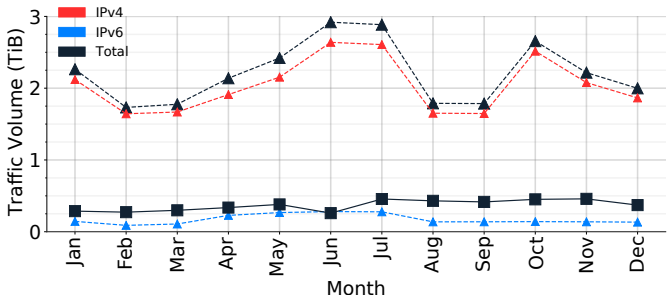


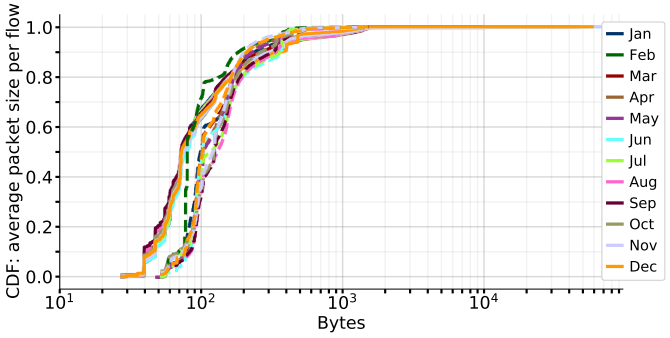
Fig. 1. *Monthly aggregated volume of byte traffic in 2007 (solid line) and 2019 (dashed line). For 2007, IPv4 and IPv6 traffic shares are not explicitly shown since the IPv6 share is negligibly small.*

Routeviews Archive — We also use BGP information for mapping IP prefixes to ASNs. We chose *routeviews.wide.routeviews.org*, as it is collected from the WIDE network in Tokyo, Japan. We use a service from *Team Cymru* [21] to get information about the ASNs. Further information about the ASes, such as type or peering partners, are retrieved from *PeeringDB*. For some ASes without entries in *PeeringDB*, we use *ipinfo.io* [22] to fill the gaps.

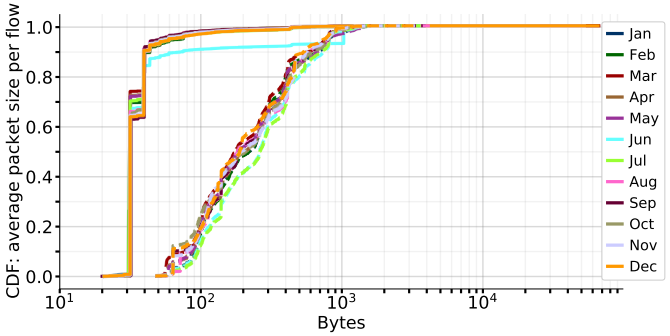
III. Traffic Shares

We compare traffic traces that are collected at Samplepoint-F between 2007–2019. The traffic volume increases greatly by all metrics (bytes, packets, flows). Fig. 1 visualizes the total, IPv4, and IPv6 monthly traffic volume in TiB. We see an expected decrease in traffic during the summer (August, September) and winter (February, March) academic breaks in Japan. During the semester courses, there is a visible weekday-weekend pattern, but during the other months, there is no significant volume increase on weekdays. Comparing 2007 and 2019, the average monthly byte traffic increases by 480%, from 387.1 GiB to 2.2 TiB. Similarly, the average monthly packet count increases by a factor of 5.5 (from 620M to 3.3B). Yet, as byte and packet traffic increase by roughly a similar percentage, the average packet sizes do not change significantly. The average monthly flow count is 23 times as high as in 2007. Therefore, in 2007, one flow contains on average 14 packets, while a flow in 2019 carries on average <4 packets. The smaller number in 2019 is due to the increased ICMP traffic in 2019. Between 2007–2019, the annual IPv6 [23] byte traffic increases by a factor of roughly 181; its volume is comparable to the total traffic volume in 2007.

We further show the average packet size per flow aggregated for each month in Figs. 2a and 2b. In 2007, the monthly largest average packet size per flow for IPv4 varies between 45,057 (most common) and 65,028 bytes; in 2019, it is between 14,648 and 65,028 bytes instead. It is noticeable that both years have a maximum average packet size per flow of 65,028 bytes. Further, the largest average packet size per flow in IPv6



(a) *2007*



(b) *2019*

Fig. 2. *CDF of the monthly aggregated average packet sizes per flow. The solid lines show IPv4 and the dashed lines IPv6.*

traffic for 2007 is roughly around 1,500 bytes for each month, while in 2019, it varies between 2,521 and 4,181 bytes, i.e., an increase of more than 68%. Compared to the average IPv6 packet sizes per flow in 2019, in 2007 IPv6 packets were smaller; all months have >80% of IPv6 packets having a size of less than 200 bytes. Contrary, in 2019, only 50% of the IPv6 traffic have an average packet size per flow of 200 bytes. This is likely due to a larger amount of Web traffic over IPv6 (see Fig. 3b). IPv4 traffic, on the other hand, has shifted from larger average packet sizes per flow in 2007 to smaller average packet sizes per flow in 2019. Even excluding ICMP traffic, the distributions show a sharp increase at around 40 bytes, indicating a large share of small packets over IPv4.

We further investigate the shift in the packet sizes per flow. In 2007, TCP is the dominant protocol in terms of bytes and packets; >95% of all bytes and 85% of all packet traffic is over TCP. In contrast, UDP has a byte share of 4% and an average packet share of 12%. The remaining percentage of non-TCP/UDP packets consist mostly of ICMP and GRE. However, in terms of flows, UDP has roughly the same percentage as TCP; both on average of 46%. As such, TCP flows are mostly larger than UDP flows, while UDP flows tend to be smaller. In 2019, TCP is still the dominant protocol, but has only 75% - 90% bytes and packet traffic share. Compared to 2007, the relative UDP byte traffic share grows for every month by up to 4 times. QUIC is majorly causing the increased UDP traffic. However, UDP share is less stable over the year. The average UDP-to-TCP byte ratio changes from 24-to-1 in 2007 to 12-to-1 in 2019 over IPv4 and 3-to-1 in 2007 to 20-to-1 in 2019 over IPv6. As such, UDP share reduces compared to that of TCP over IPv6 in 2019. One reason is that although the monthly IPv6 traffic volume increases up to more than 0.26 TiB in 2019, the absolute UDP share stays on average 7.1 GiB in 2019. The observations for the IPv6 protocol breakdown in terms of packets are similar: The protocol mix for 2007 is more diverse, and the relative UDP share is higher compared to 2019. This is not true for flows: The relative UDP IPv6 traffic share is more stable in 2019 than in 2007; in 2007, it varies between 46% and 92%, while in 2019, it stays at 85%. This is a contrast to the observation of IPv6 byte traffic: The IPv6 UDP share does not increase, although the overall byte traffic increases. Therefore, UDP flows over IPv6 are smaller in 2019 than in 2007. This observation is similar to Zhang *et al.* [24] who observe a trend towards smaller UDP flows in 2009, but do not differentiate between IPv4 and IPv6.

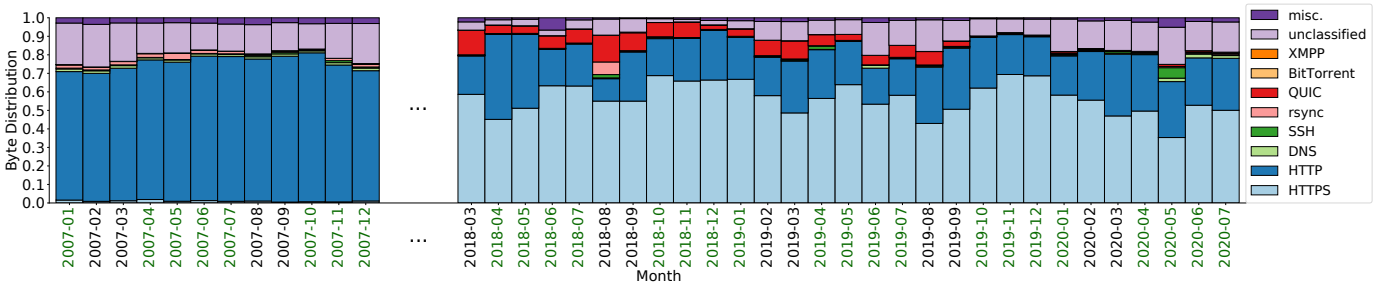
Takeaway: The annual traffic has increased significantly; IPv6 traffic in 2019 is now similar to the total traffic in 2007. IPv4 packets tend to be smaller, while the average packet size per IPv6 flow increased, due to increased Web traffic. QUIC is the main reason why UDP traffic increased in 2019, which could also be a reason for increased UDP flow size over IPv4. However, over IPv6, the flow size has decreased.

IV. Application Mix

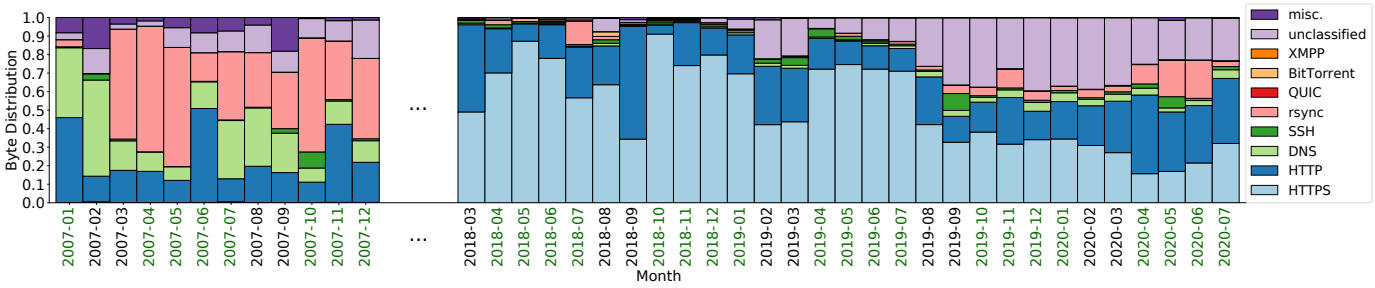
Figs. 3a and 3b show the application distribution for 2007 and 2018–2020. One decade later, a shift towards more encrypted Web is visible: The most used TCP port (destination and source) is assigned to HTTPS (443). In 2007, HTTP contributes mostly to the Web traffic over IPv4 with a share of >70%. Moreover, IPv4 byte traffic in 2007 consists mainly of HTTP, whereas HTTPS, DNS, SSH, `rsync`, and SMTP together have a share of <5% of the monthly byte traffic. In 2019, their shares decreased further, while the HTTP-to-HTTPS ratio increased to roughly 1-to-2. Simultaneously, in terms of flows, the shares are more equally distributed among the most observed port numbers (123, 53, 3283) of about 10%. In 2007, DNS contributes >20% to the overall flow count.

The IPv6 traffic shows a completely different application composition in 2007: HTTP (22%), DNS (20%), and `rsync` (36%) together have a share of >70%. One decade later, IPv6 carries mostly HTTPS and HTTP (together on >60%). There are also shares of DNS, SSH, and `rsync`, but they contribute to an average of <8% over the year. In terms of bytes, the top UDP source port is QUIC (443) and the top TCP ports are HTTPS (443) and HTTP (80). We observe that under the top 8 UDP destination and source ports, only one port had no official assignment. In contrast, in 2007, it is almost the half of the top eight ports. The unassigned ports in 2007 may indicate P2P traffic with randomly assigned high ports or DDoS attacks.

Fig. 3 also includes years 2018 and 2020 to provide a spatial view on the development of the application mix. Few



(a) *Timeline for IPv4 traffic.*



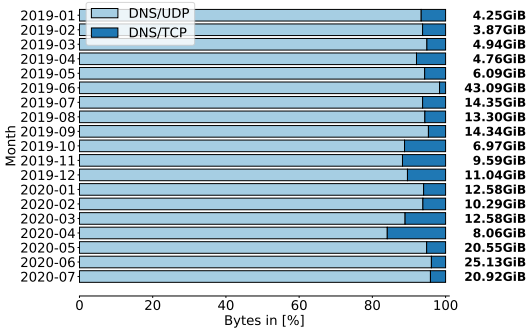
(b) *Timeline for IPv6 traffic.*

Fig. 3. *Monthly application distribution in 2007 and between 03/2018 and 07/2020 in terms of bytes for IPv4 traffic at the top and for IPv6 traffic at the bottom. Months with green labels represent the months during the semester (i.e., with classes) in Japan.*

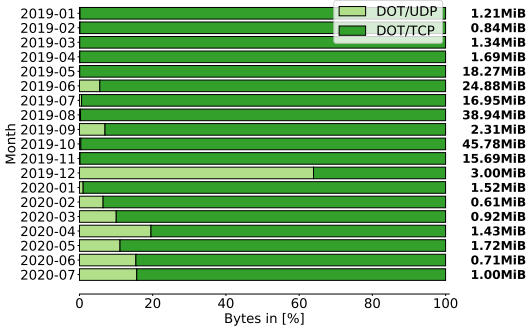
applications such as HTTPS, HTTP, DNS, `rsync`, QUIC, and SSH dominate the traffic over both IP protocols. *unclassified* traffic consists mainly of high port numbers (>30000) with no official assignment. Traffic of type *misc* includes the remaining known applications, such as FTP and DoT, that have smaller traffic shares. Note that the academic year is visible in the change of the application distribution between Mar/18 and Sep/19; the percentage of HTTPS increases significantly in academic cycles. Over IPv6, the pattern is even more visible: Up to Sep/19, during months of vacation, the percentage is on average 44%, while during almost all months of classes, it is on average 75%. However, the HTTPS traffic over IPv6 decreases after Sep/19 and stays put during the fall term, which differs from the pattern in the previous year. In addition, the *unclassified* traffic over IPv6 increases; in Oct/19, it increases by >138% and by >160% in Feb/20 compared to Aug/19. As a result, the relative share of HTTPS over IPv6 decreases starting from Oct/19. On the other hand, the HTTPS share in IPv4 traffic does not decline and the absolute *unclassified* traffic share remains constant during fall term 2019. From Apr/20 onwards, there is a decline in HTTPS traffic due to COVID-19 restrictions. Over IPv6, the decline is more significant than for IPv4: Starting from Mar/20, the HTTP share is larger than the HTTPS share, partially by >50%. In 2020, 15% more Web traffic is sent over IPv4 than over IPv6. Further, there is a trend towards more DNS traffic over IPv4: While in 2018 the IPv4 traffic share of DNS is on average <0.2%, it increases to an average of 0.4% (8.0 GiB) in 2019. In 2020, the relative DNS share stays high at on average 1.0% (10.6 GiB). Therefore, there is an increase in the relative as well as in the absolute DNS traffic share. A similar trend can be seen in IPv6: The share quintuples from an average of 0.67% in 2018 to an average of 3.6% (5.1 GiB) in 2020. Thus, the relative DNS share is higher over IPv6 than over IPv4. Another application having a larger relative share over IPv6 is `rsync`. While the relative share over IPv4 stays small 0.1%–0.7% (< 8.8GiB) over the years, over IPv6, there is a trend towards more `rsync` traffic. The `rsync` share increases from an average of 2% in 2018, over 2.3 % in 2019 (4.2 GiB), to 9% in 2020 (13.7 GiB).

In general, we do not observe a trend towards more IPv6 traffic. In 2018, the IPv6 share stays on average 7.7% and increases to 10.1%. After Jan/20, the IPv6 share increases to >12%. In May/20, it even reaches a percentage of 15.1%, followed by a steadily decrease in Jun/20 (10.7%) and Jul/20 (6.8%). Meanwhile, for QUIC as well, the data shows a trend towards less traffic over IPv4. While the QUIC share is constantly high of up to 14.5% in 2018, in 2019, it only reaches up to 9.8%. The QUIC percentage decreases to less than 0.6% (11.94 GiB) after Sep/19. In 2020, QUIC traffic stays at a similar share of 0.3% (2.54 GiB) to 1.06% (9.52 GiB). We suspect that the IPv4 QUIC traffic shifted from Samplepoint-F to Samplepoint-G (see §VI). Finally, the IPv6 QUIC share is negligibly small over the entire timeline 2018-2020 (on average 0.08%), and there is no trend towards increasing or decreasing IPv6 QUIC traffic volume.

We also do not observe any DoT share because it is negligible, and thus, included in *misc*. Figs. 4a and 4b show the total traffic volume relative to bytes of DNS (DoT excluded) and DoT only in more detail. DNS traffic (DoT excluded) is dominated by DNS/UDP. On average, >93% (10.7 GiB) of all DNS traffic (DoT excluded) in 2019 and on average



(a) *DNS (w/o DoT) traffic composition.*



(b) *DoT traffic composition.*

Fig. 4. *DNS (without DoT) and DoT traffic composition in terms of bytes between 2019–2020. The monthly sum of all traffic is 100%*

>92% (14.7 GiB) in 2020 is DNS/UDP, while the DNS/TCP [25] percentage is 6.9% (0.63 GiB) in 2019 and 7.5% (1 GiB) in 2020. The high average DNS/TCP share in 2020 is due to an exceptionally large increase of 16% in Apr/20. Similarly, DoT traffic is also mainly composed of DoT/TCP. While there is some DoT/UDP traffic (i.e., DNS over DTLS), its shares are hardly visible in 2019. In 2020, the relative share of DoT/UDP increases to an average of 11.3% (0.13 MiB). Simultaneously, the absolute DoT/TCP traffic decreases noticeably in 2020 and is less than the share of DoT/UDP of the top three months in 2019. We also study the TLS versions; however, out of all DoT/TCP flows, only 1.82 % (2020: 47.5%) have a TLS payload. Further, none of the dataframes contain a `SERVER HELLO`, which does not allow distinguishing TLS 1.2 and TLS 1.3. We observe that TLS 1.2 (+1.3) are the dominant protocol for both years. Most TLS traffic originates from or is sent to Quad9 and Cloudflare servers.

Takeaway: Compared to 2007, the application mix over IPv6 resemble now more the one over IPv4; For both the HTTPS traffic share increased to a significant percentage in terms of bytes. The Japanese academic cycles are clearly visible in the application mix mainly due to varying HTTPS traffic volumes. Over IPv6 DNS and `rsync` have higher relative shares compared to IPv4. Besides the typically used DNS/UDP and DoT/TCP, there is a slightly increasing percentage of DNS/TCP and DoT/UDP.

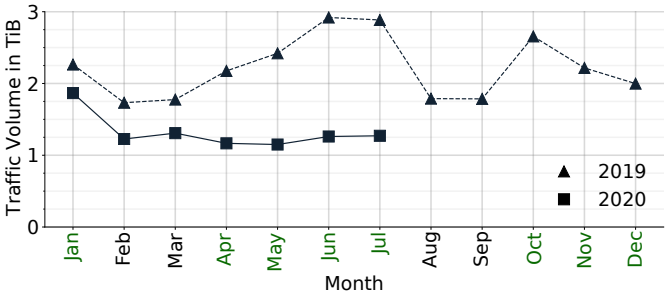


Fig. 5. *Aggregated traffic volume per month in 2019 and 2020. Months with green labels represent the months during the semester.*

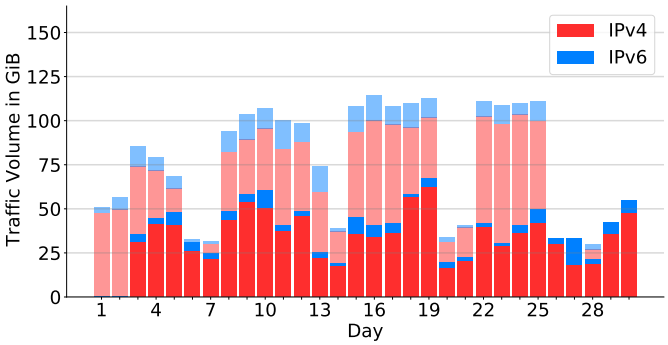
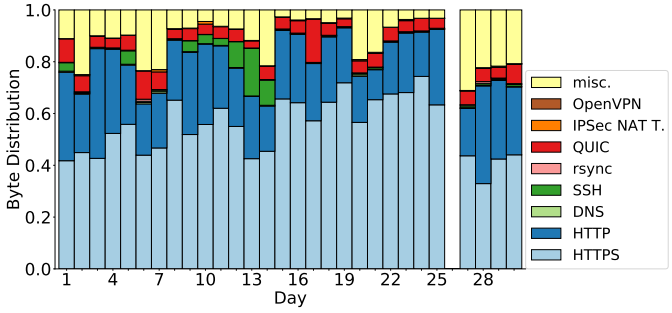


Fig. 6. *The transparent bars depict the monthly traffic volume in Apr/19 and the solid bars the one for Apr/20. The days for 2020 are shifted by 2 to the right to match the weekdays of Apr/19.*

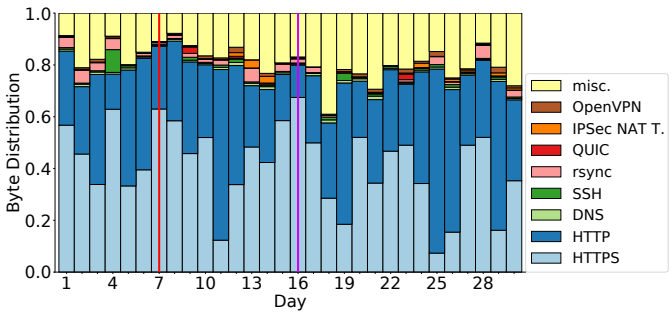
V. COVID-19 Pandemic

We investigate the first seven months of 2020 to understand a possible change since the beginning of the COVID-19 pandemic. The traces of Jan/19 until Jul/19 are used as reference. Fig. 5 shows the monthly byte traffic volume. It highlights the pandemic related byte volume decrease starting from Feb/20. In 2019, there is a continuous increase in monthly byte traffic starting from Apr 01. During the summer break from Aug 04–Sep 30, 2019, the monthly byte volume decreases. At the start of the fall term in Oct/19, the byte traffic increases rapidly to 2.7 TiB. In 2020, Jan/20 and Mar/20 have the highest traffic volume around 1.9TiB and 1.3TiB. Yet, in Apr/20 and May/20, the traffic volume declines to 1.16 TiB. During the spring term 2020, the byte volume is even lower than during the summer break 2019.

Apr/20 is the first month to show significant changes in the daily traffic volume compared to the previous year. The academic year in Japan starts on Apr 01; with the declaration of a nationwide state of emergency on Apr 16, universities closed and moved all their courses for the spring term online. The closure of universities in Apr/20 is visible in the dataset: Contrary to Apr/19, the traffic in Apr/20 neither increases on weekdays nor show a clear weekday-weekend pattern (see Fig. 6). The daily byte volume even decreases slightly compared to Mar/20. Also, in May/20, the daily traffic volume is as low as the traffic on weekends in May/19. Despite restrictions being lifted in Jun/20, the traffic volume does not increase in Jun/20



(a) *April 2019*



(b) *April 2020*

Fig. 7. *Daily application mix in Apr/19 (top) and Apr/20 (bottom). The red (first) and violet (nationwide) lines depict emergency declarations.*

and Jul/20, and stays as low as on weekends of the respective month in 2019, thus, indicating off-campus classes.

The missing weekday-weekend pattern starting from Apr/20 is well visible in the application distribution: While Fig. 7a shows a clear increase of the relative share of *misc* traffic on Saturdays and Sundays, in 2020, this relative share increases and decreases at no specific days (see Fig. 7b). The reason is not the increase of *unclassified* traffic, but the decline of the HTTP and HTTPS traffic on weekends. Since Web traffic carries high traffic volume, the decreased HTTP and HTTPS traffic is the main reason why significantly less traffic volume is captured on the link over weekends in 2019. In 2020, we observe that more remote access protocols such as Remote Shell and OpenVPN are used. In Apr/20, the daily share of OpenVPN increases on average by a factor of 67.2 to an average of 244.2 MiB (0.71%) compared to weekdays/weekends in Apr/19. Beyond that, the share of `rsync` increases for every month; at specific days, the share is even up to 28.4% high of the daily traffic. We also observe increased traffic of IPSec NAT Traversal: For instance, in Apr/20, the daily share increases to an average of 0.53% (189.0 MiB), while in Apr/19, it is on average 0.21% (143.2 MiB) of the daily byte traffic. In general, the relative daily SSH share is mostly around 0.2%–2.7% for both years. Additionally, the daily traffic volume of BitTorrent increases, e.g., in Apr/20, on average by more than 5.6 times to an average of 20.1 MiB (0.06%) compared to weekdays/weekends in Apr/19.

Due to the closure of universities, we also expect to see a shift in the source and destination ASes. Table I shows

TABLE I

Top 10 Destination ASes in Apr/19 and Apr/2020 relative to bytes. The ASes mentioned in the text are marked in green.

April 2019			April 2020		
Destination	ASes	Bytes[GiB]	Destination	ASes	Bytes[GiB]
AS4538	ERX-CERNET-BKB	119.91	AS17676	GIGAINFRA	121.40
AS4134	CHINANET-BACKBONE	88.86	AS4766	KIXS-AS-KR	78.48
AS2907	SINET-AS	88.21	AS1659	ERX-TANET-ASN1	55.50
AS2500	WIDE-BB	56.25	AS17816	CHINA169-GZ	37.34
AS5609	ASN-CSELT	52.98	AS17512	JAL	35.14
AS9462	BOLEH-NET-AP	52.42	AS8803	MIGROS	34.78
AS4637	ASN-TELSTRA-GLOBAL	50.64	AS4782	GSNET	32.34
AS17676	GIGAINFRA	49.20	AS2830	MCI-DUAL-HOMED-CUSTOMERS	30.38
AS9667	HOSTWORKS-AS-AP	44.35	AS4837	CHINA169-BACKBONE	28.91
AS55552	NETWORK-BOX-HK	44.27	AS2500	WIDE-BB	27.35

TABLE II

Top 10 Source ASes in Apr/19 and Apr/20 relative to bytes.

April 2019			April 2020		
Source	ASes	Bytes[GiB]	Source	ASes	Bytes[GiB]
AS714	APPLE-ENGINEERING	210.54	AS714	APPLE-ENGINEERING	75.94
AS701	UUNET	75.46	AS6319	MARRIOT-ASN	39.32
AS7018	ATT-INTERNET4	52.51	AS3462	HINET	37.29
AS16625	AKAMAI-AS	40.80	AS16625	AKAMAI-AS	30.01
AS10796	TWC-10796-MIDWEST	33.82	AS4662	QTCN-ASN1 GCNet	28.75
AS3320	DTAG	33.77	AS21928	T-MOBILE-AS21928	23.97
AS4134	CHINANET-BACKBONE	31.05	AS38676	FLEXNET-AS-KR	23.51
AS16509	AMAZON-02	29.82	AS4782	GSNET	21.84
AS20940	AKAMAI-ASN1	27.11	AS19679	DROPBOX	20.69
AS2500	WIDE-BB	27.02	AS16509	AMAZON-02	19.87

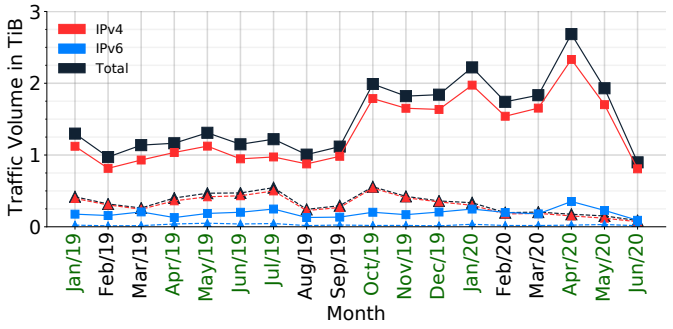
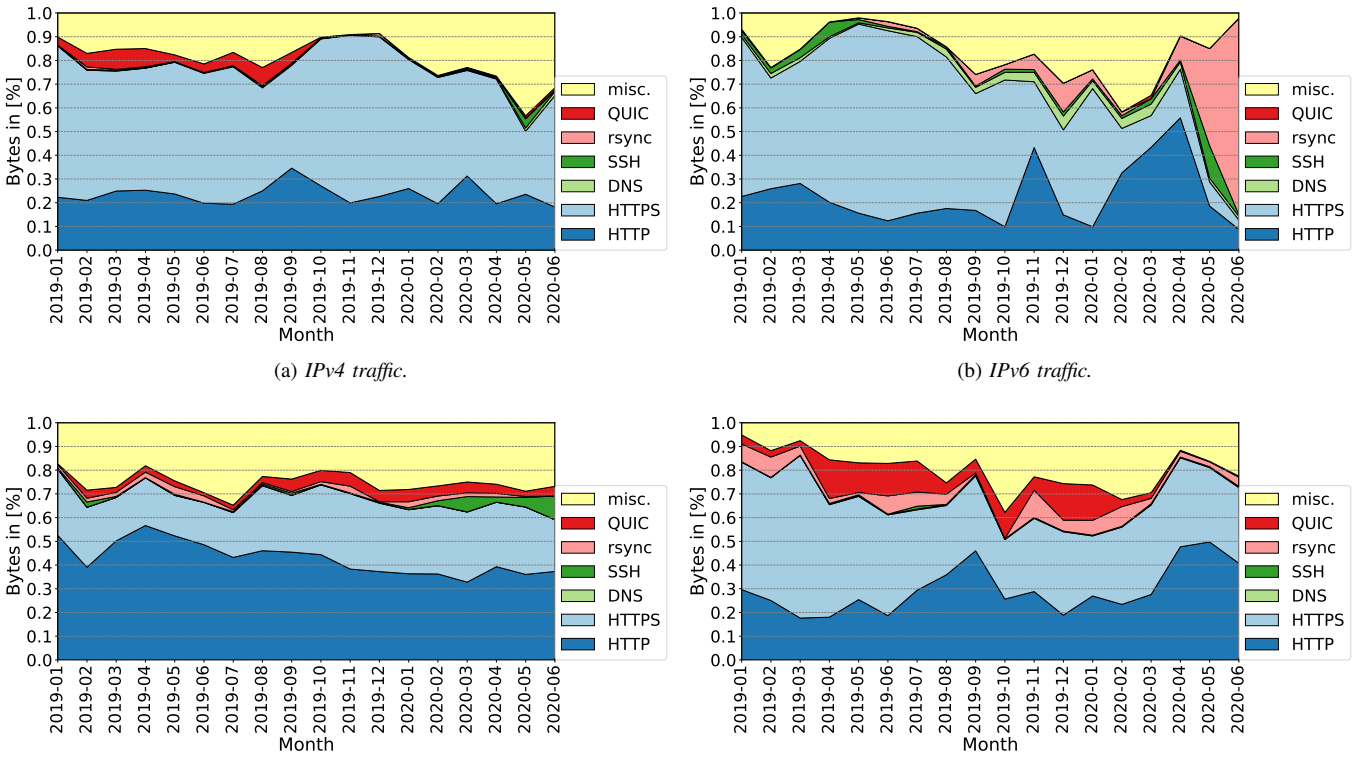


Fig. 8. *Monthly aggregated volume of the IPv4, IPv6, and total traffic at Samplepoint-G (solid line) and -F (dashed line). Samplepoint-G and -F both include only Wednesday traces for fair comparison. Since June 10, 2020, no further traces from Samplepoint-G are publicly available, therefore, we consider also only the first two Wednesdays of Samplepoint-F for Jun/20. Months denoted in green represent the months during the semester.*

exemplary top 10 destination ASes and Table II top 10 source ASes for Apr/19 and Apr/20. Between Apr/19 and Jul/19, the top 10 destination ASes receive in total 2.63 TiB of traffic volume. In 2020, the volume decreases by 30.4% to 1.83 TiB. In 2019, AS2500 (WIDE-BB), AS2907 (SINET-AS), and AS4538 (ERX-CERNET-BKB) have the largest shares belonging to educational institutions. In contrast, in 2020, AS2500 (WIDE-BB) and AS17676 (GIGAINFRA) dominate the traffic shares, while AS4538 is not under the top 10 anymore. Softbank C&S (GIGAINFRA) distributes Dropbox for Business in Japan, indicating an increased use of file hosting services. We also observe a higher decrease (>52%) in outgoing traffic volume compared to the incoming traffic volume. We observe AS714 (APPLE-ENGINEERING) as the top AS for each month in 2019 with 3 times higher traffic share than the second highest AS. AS714 is still under the top 10, but the traffic volume decreases significantly, e.g., in May/20 by a factor of nearly 9 compared to May/19. Thus,



(a) *IPv4 traffic.*

(b) *IPv6 traffic.*

(c) *IPv4 traffic.*

(d) *IPv6 traffic.*

Fig. 9. *Monthly aggregated application mix at Samplepoint-F (above) and Samplepoint-G (below).*

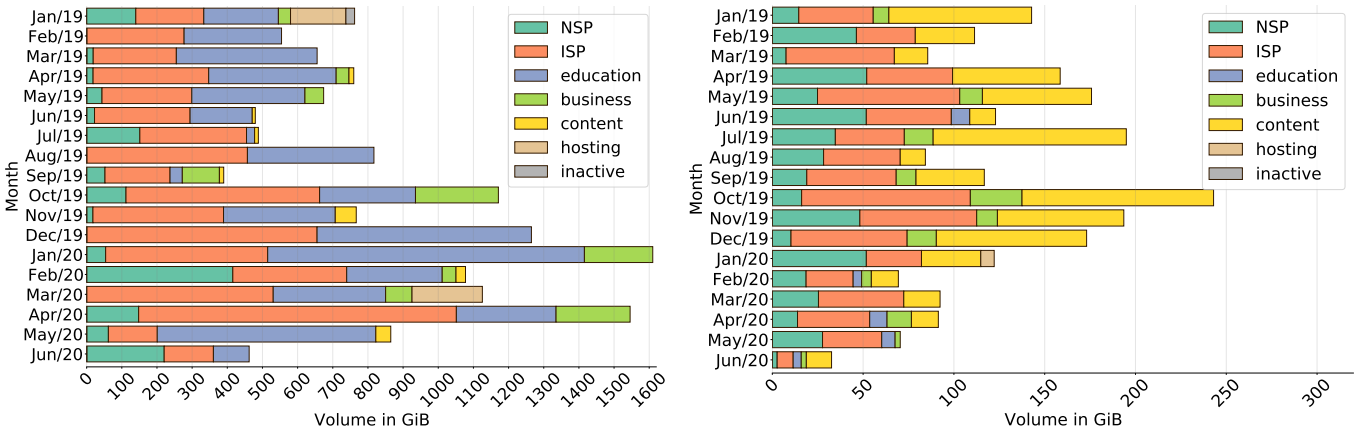
Apple services like iCloud or iTunes are less often used during the pandemic from the educational backbone. AS1221 (Telstra Corporation Ltd) and AS3462 (HINET) gain more traffic share, while AS7922 (COMCAST) and AS7018 (AT&T) lose traffic volume and are not under the top 10 anymore. Traffic from AS19679 (DROPBOX) also increases.

Takeaway: Due to universities closure, the byte volume during spring term 2020 is even lower than during the summer break 2019. In addition to that, the weekday-weekend pattern disappears and the daily traffic volume stays as low as on weekends, even after the restriction were lifted. OpenVPN and Remote Shell protocols are more frequently used. There is also a shift towards file hosting services like Dropbox, and overall less traffic from US telecommunication companies.

VI. Impact of Peering with Google

The MAWI backbone peers with Google; the link's traffic is collected at Samplepoint-G. Thus, we study the impact of peering with Google on the application mix. Fig. 8 shows the captured monthly traffic volumes on G and F. The monthly amount of traffic captured at Samplepoint-G in 2019 contains on average roughly 2.5 times (almost 5 times in 2020) more packets and 3.4 times (>10 times in 2020) more bytes, but roughly 5 times (6 times in 2020) less flows than the traffic collected on Samplepoint-F. Compared to the amount of traffic of all traces (not only traces collected on Wednesdays) at Samplepoint-F between 2019 and 2020, the monthly volume is at most 2.5 times lower. Although the traffic at Samplepoint-G is also dominated by IPv4, the absolute share of IPv6 is 7 times larger in 2019 (>8 times in 2020) than on Samplepoint-F. Thus, on a single Wednesday, more IPv6 traffic is captured than during a whole year on Samplepoint-F (only Wednesdays). Thus, peering with Google increases IPv6 traffic significantly. Both Samplepoint-F and -G show characteristic decreases around the summer break (Aug/19–Sep/19) and winter break (Feb/19–Mar/19). At Samplepoint-G, the traffic volume decreases in Nov/19–Dec/19 and increases to 2.68 TiB in Apr/20. This is contrary to the traffic captured on Samplepoint-F; Here, the traffic significantly decreases after Mar/20 due to Covid-19 regulations.

The average packet sizes per flow also differ. On Samplepoint-G, 80% of IPv4 flows have an average packet size of <100 bytes. Over IPv6, the flows are slightly larger; only 50% of flows have an average packet size of <150. In contrast, at Samplepoint-F the gap betweeen IPv4 and IPv6 is significantly larger: 90% of the IPv4 flows carry <100 bytes for both years. 23%–40% of IPv6 flows carry average packet sizes of <150 bytes and more than 40% of the IPv6 flows have sizes between 200–600 bytes. At Samplepoint-G, <28% are between 200–600 bytes, thus, having around 52%–80% of flows with packet sizes smaller than 200 bytes. As such, IPv6 carries larger packets per flow on Samplepoint-F. At Samplepoint-G,



(a) *Monthly top 10 source ASes at G.*

(b) *Monthly top 10 source ASes at F.*

Fig. 10. *Monthly top 10 source ASes at Samplepoint-F and -G. The bars visualize the amount of byte traffic from a specific AS type (NSP, ISP, education, business, content, hosting, or inactive).*

the traffic is mainly composed of TCP (roughly 91%–98%). UDP is the second largest protocol (1%–9%), followed by mostly ICMP, GRE, and ESP. The UDP traffic is mostly composed of QUIC. After Sep/19, the UDP traffic increases up to 0.18 TiB (before: 36 GiB), mostly due to the simultaneous increase of QUIC traffic. Compared to IPv4, the relative UDP share of Samplepoint-G is larger for IPv6; on average, the absolute UDP traffic share is 14.9 GiB (7.57%). The UDP traffic is mainly QUIC traffic with a negligible share 0.44% (865.31 MiB) of other applications. The absolute QUIC traffic volume has larger fluctuations in IPv6 than in IPv4. Further, the absolute share of UDP (without QUIC traffic) over IPv6 is similar for Samplepoint-G (865.31 MiB) and Samplepoint-F (1.07 GiB).

Figs. 9a and 9b show the application mix for Samplepoint-F, and Figs. 9c and 9d for Samplepoint-G. We see that Samplepoint-G has a roughly stable average HTTP-to-HTTPS ratio of 9-to-5. There is also a share of on average 1.5% (24.0 GiB) of `rsync` in the monthly application mix. The SSH share changes from 0.7% (16.5 GiB, Jan/20) over 6.4% (120.9 GiB , Mar/20) to 4.0% (59.9 GiB, Apr/20). This possibly indicates increased remote work due to COVID-19. In contrast, the HTTP-to-HTTPS ratio at Samplepoint-F is larger in favor of HTTPS with on average about 1-to-2. Moreover, the monthly QUIC share is stable after Sep/19 at Samplepoint-G on average 4.39% (84.52 GiB). Before and including Sep/19, it is twice as small on average 2.46% (28.28 GiB). On the other hand, after Sep/19, the QUIC share is between 0.02% (64.00 MiB) and 0.97% (1.39 GiB) at Samplepoint-F. Before Oct/19, the QUIC share is more than twice as large between 2.83%–7.65% (13.18–30.55 GiB). We suspect that the QUIC traffic shifts from Samplepoint-F to Samplepoint-G which leads to a vanishing QUIC share on Samplepoint-F and increasing QUIC share on Samplepoint-G. Over IPv6, the HTTP-to-HTTPS ratio at Samplepoint-G is slightly larger in favor of HTTPS, with on average 9-to-10. After Nov/19, the HTTPS relative share decreases slightly, while the relative and absolute HTTP share grows. The QUIC share is more than twice as large as over IPv4 and the relative `rsync` share is up to 4 times as large. In contrast, at Samplepoint-F, the QUIC share is less than 0.1% ($<$ 0.27 GiB). DNS and SSH have shares between 0.8% (142.10 MiB) and 5.8% (4.41 GiB). After Mar/20, the `rsync` share increases up to 82.78% (13.27 GiB), while the HTTPS share decreases rapidly after Feb/20. We assume that this is due to the Covid-19 pandemic.

Finally, we examine the monthly top 10 source and destination ASes. Figs. 10a and 10b show the traffic volume of the top 10 source ASes aggregated per month by types in GiB for Samplepoint-G and -F, respectively. The top 10 source ASes of Samplepoint-G are mainly of type NSP/ISP and education. The top ASes belong mainly to army and government networks, universities and research institutions, and data centers or hosting services appears only at place ten in Jun/19 and Feb/20. One reason might be that Google is an intermediate AS (transit) and not the end point. At Samplepoint-F, there is hardly any traffic from educational ASes and instead more traffic from content related ASes. Additionally, almost every month has a small percentage of traffic originating in ASes of type business. The monthly traffic volume destined to the top 10 destination ASes is for some months $>$600 GiB less than the traffic originating from the top 10 source ASes at Samplepoint-G. However, at Samplepoint-F, the monthly traffic is even slightly larger for the destination ASes. AS4713 (NTT) and AS4134 (CHINANET-BACKBONE) are always among the top three ASes. In fact, most destination ASes are assigned to telecommunication services or CDNs. The destination ASes at Samplepoint-F also belong mainly to NSP/ISPs but there is also around 14–92 GiB destined to educational ASes. Furthermore, there is unstable traffic share of business and content delivery traffic. The top ASes differ from month to month. However, most traffic is from telecommunication service providers or academic networks.

Takeaway: Peering with Google increases QUIC traffic over IPv4 and IPv6; the relative share over IPv6 is even larger than over IPv4. The average HTTP-to-HTTPS ratio is 9-to-5 over IPv4, while at Samplepoint-F it is about 1-to-2. The traffic mainly originates from ASes of type NSP/ISP and education. In contrast, at Samplepoint-F there is more traffic from content-related ASes.

VII. Conclusion

We analyzed and compared >6.6 TiB of traces of the MAWI dataset with the goal to understand the impact of various factors on the Internet traffic composition. From 2007 to 2019, there was a change in the use of IPv6: By 2019, the gap between the packet sizes in the range of 20 to 800 bytes of IPv4 and IPv6 increased significantly; 70% of the IPv6 flows had an average packet size between 100 and 500 bytes, while over IPv4, more than 90% of flows had an average packet size of less than 80 bytes. This shift might be due to an increased use of web over IPv6. In general, the increased total traffic volume in 2019 is mainly due to traffic to/from ports 443/UDP (QUIC), 443/TCP (HTTPS) and 80/TCP (HTTP). Moreover, the HTTP-to-HTTPS ratio changed in favor for HTTPS to 1-to-2; before, more than 70% was HTTP and less than 1% HTTPS. Since the Covid-19 outbreak in Jan/20, the traffic changed significantly, which led to different patterns. Due to closures of universities in Apr/20, traffic on weekdays was as low as on weekends. Daily traffic remained low even after the state of emergency was lifted, likely because the universities had switched to online classes for the remaining spring semester. The aggregated traffic between Jan/20 and Jul/20 decreased by roughly 43.8% compared to the traffic volume between Jan/19 and Jul/19. The main reason is the significantly decreased HTTPS share. Moreover, the data indicates a shift towards more remote and cloud access during the pandemic. Finally, we examined the impact of peering with Google. The average packet sizes per flow over IPv4 and IPv6 are similar and no trend towards larger IPv6 packets is visible. The HTTP share is slightly larger than HTTPS, having an average HTTP-to-HTTPS ratio of 9-to-5. The relative QUIC traffic share is larger over IPv6 than IPv4.

Acknowledgments

We would like to thank Kenjiro Cho (IIJ) for sharing the non-anonymized MAWI dataset with us. We also would like to thank Luca Schumann (TUM) for his exploratory work on this topic. This work was supported by the Volkswagenstiftung Niedersächsisches Vorab (Funding No. ZN3695).

References

[1] P. Dikshit *et al.*, "Recent Trends on Privacy-Preserving Technologies under Standardization at the IETF," *CoRR*, vol. abs/2301.01124, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.01124

[2] A. P. Felt *et al.*, "Measuring HTTPS Adoption on the Web," in *26th USENIX Security Symposium*, Aug. 2017. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/felt

[3] Z. Hu *et al.*, "Specification for DNS over Transport Layer Security (TLS)," *RFC*, vol. 7858, pp. 1–19, 2016. [Online]. Available: https://doi.org/10.17487/RFC7858

[4] T. V. Doan *et al.*, "Measuring DNS over TLS from the Edge: Adoption, Reliability, and Response Times," in *Passive and Active Measurement Conference*, 2021. [Online]. Available: https://doi.org/10.1007/978-3-030-72582-2_12

[5] P. Hoffman and P. McManus. (2018, Oct.) DNS Queries over HTTPS (DoH). [Online]. Available: https://tools.ietf.org/html/rfc8484

[6] T. Pauly (WWDC2020). Enable encrypted DNS. [Online]. Available: https://developer.apple.com/videos/play/wwdc2020/10047/

[7] T. V. Doan *et al.*, "Evaluating Public DNS Services in the Wake of Increasing Centralization of DNS," in *IFIP Networking Conference*. IEEE, 2021. [Online]. Available: https://doi.org/10.23919/IFIPNetworking52078.2021.9472831

[8] T. Shreedhar *et al.*, "Evaluating QUIC Performance Over Web, Cloud Storage, and Video Workloads," *IEEE Transactions on Network and Service Management*, vol. 19, no. 2, pp. 1366–1381, 2022. [Online]. Available: https://doi.org/10.1109/TNSM.2021.3134562

[9] M. Kosek *et al.*, "Beyond QUIC v1: A First Look at Recent Transport Layer IETF Standardization Efforts," *IEEE Communications Magazine*, vol. 59, no. 4, pp. 24–29, 2021. [Online]. Available: https://doi.org/10.1109/MCOM.001.2000877

[10] D. Schinazi, F. Yang, and I. Swett. (2020, Oct.) Chrome is deploying HTTP/3 and IETF QUIC. [Online]. Available: https://blog.chromium.org/2020/10/chrome-is-deploying-http3-and-ietf-quic.html

[11] M. Bishop, "HTTP/3," *RFC*, vol. 9114, pp. 1–57, 2022. [Online]. Available: https://doi.org/10.17487/RFC9114

[12] C. Huitema, S. Dickinson, and A. Mankin, "DNS over dedicated QUIC connections," *RFC*, vol. 9250, pp. 1–27, 2022. [Online]. Available: https://doi.org/10.17487/RFC9250

[13] M. Kosek, T. V. Doan, M. Granderath, and V. Bajpai, "One to Rule Them All? A First Look at DNS over QUIC," in *Passive and Active Measurement Conference*, 2022. [Online]. Available: https://doi.org/10.1007/978-3-030-98785-5_24

[14] M. Kosek, L. Schumann, R. Marx, T. V. Doan, and V. Bajpai, "DNS Privacy with Speed?: Evaluating DNS over QUIC and its Impact on Web Performance," in *ACM Internet Measurement Conference*, 2022. [Online]. Available: https://doi.org/10.1145/3517745.3561445

[15] C. Dietzel. (2020, Jun.) Why the Internet holds firm: Internet infrastructure in times of Covid-19. [Online]. Available: https://www.de-cix.net/en/resources/articles/why-the-internet-holds-firm-internet-infrastructure-in-times-of-covid-19

[16] DE-CIX. (2020, Mar.) Big upswing in Internet usage due to Covid-19 measures. [Online]. Available: https://www.de-cix.net/en/about-de-cix/news/big-upswing-in-internet-usage-due-to-covid-19-measures

[17] A. Feldmann *et al.*, "The Lockdown Effect: Implications of the COVID-19 Pandemic on Internet Traffic," in *ACM Internet Measurement Conference*, 2020. [Online]. Available: https://doi.org/10.1145/3419394.3423658

[18] L. Schumann *et al.*, "Impact of Evolving Protocols and COVID-19 on Internet Traffic Shares," *CoRR*, vol. abs/2201.00142, 2022. [Online]. Available: https://arxiv.org/abs/2201.00142

[19] PeeringDB. (2020) WIDE Project. [Online]. Available: https://www.peeringdb.com/net/831

[20] CERT NetSA Security Suite. Monitoring for Large-Scale Networks. [Online]. Available: https://tools.netsa.cert.org/index.html

[21] Team Cymru. IP to ASN Mapping Service. [Online]. Available: https://team-cymru.com/community-services/ip-asn-mapping/

[22] IPinfo. (2020) The Trusted Source for IP Address Data. [Online]. Available: https://ipinfo.io/

[23] V. Bajpai *et al.*, "A Longitudinal View of Dual-Stacked Websites - Failures, Latency and Happy Eyeballs," *IEEE/ACM Transactions on Networking*, vol. 27, no. 2, pp. 577–590, 2019. [Online]. Available: https://doi.org/10.1109/TNET.2019.2895165

[24] M. Zhang, M. Dusi, W. John, and C. Chen, "Analysis of UDP Traffic Usage on Internet Backbone Links," in *International Symposium on Applications and the Internet, SAINT 2009*, 2009. [Online]. Available: https://doi.org/10.1109/SAINT.2009.65

[25] M. Kosek *et al.*, "Measuring DNS over TCP in the Era of Increasing DNS Response Sizes: A View from the Edge," *ACM Computer Communications Reviews*, vol. 52, no. 2, pp. 44–55, 2022. [Online]. Available: https://doi.org/10.1145/3544912.3544918