




Look at my Network: An insight into the ISP Backbone Traffic

Tomas Benes , Jaroslav Pesek 
Czech Technical University in Prague & CESNET a.l.e.
 Prague, Czech Republic
 {benesto3, jaroslav.pesek}@fit.cvut.cz

Tomas Cejka 
CESNET a.l.e.
 Prague, Czech Republic
 cejkat@cesnet.cz

Abstract—High-speed ISP networks provide several challenges that prevent the creation of long-term datasets for giving insight into the traffic. Currently, there are no publicly available long-term datasets capturing the entirety of high-speed ISP networks. Such networks are traditionally monitored using IP Flows, which provide enough high-level information about the situation in the network and support various use cases, such as the detection of outages or security threats. Even with this type of aggregation long-term datasets are very unpractical due to their size. The other problem is that flow monitoring comes with significant aggregation and common traffic statistics are brief and lack useful details and require further processing. This paper addresses these problems and presents a new long-term aggregated dataset, a detailed analysis of public network traffic measured on the ISP backbone, and a monitoring architecture composed of open-source tools capable of using an existing flow exporter infrastructure. Such insight into traffic helps to design and develop hardware optimizations, tuning the performance of monitoring systems, and adapting security detection algorithms.

Index Terms—traffic monitoring; IP flows; traffic statistics; heavy-tailed distribution; ISP network

I. INTRODUCTION

Network technologies are evolving, and the complexity of network infrastructures is increasing. Increasing the speed of network lines requires a sophisticated and fast approach to packet processing and sufficient memory to track connections. Since hardware resources and computing power are limited, it is necessary to consider optimizations of hardware-software applications. However, such optimization requires a deeper understanding of the traffic and its distribution.

In addition, networks are increasingly becoming larger, more complex, and more distributed, making them more challenging to understand and manage. Monitoring systems help maintain situational awareness and provide continuous streams of information about data transfers and activities of network devices. This functionality is essential, e.g., for network security, especially enforcing security policies and detecting security threats. Therefore, monitoring probes must be precise and powerful enough to analyze all packets and compute statistics without loss.

Understanding the complexity of a large network allows proactive management and planning and maintains network performance, reliability, and security. It can help identify bottlenecks, recognize patterns, predict future capacity needs, and detect security threats. In an ISP context, understanding peering traffic can form peering decisions, optimize traffic

routing, and provide adequate service quality to end users. Additionally, a deep understanding and monitoring of such a network are necessary to enhance the development further and reduce the cost of connection tracking tables and flow caches designed for use in hardware-based solutions. Therefore, deep insights into large networks are beneficial and necessary in our increasingly digital and connected world.

The consecutive survey of the latest resources and related work revealed that it is very challenging to find or create real traffic datasets that could be used to study the characteristics and distribution of high-speed network traffic. The main problem is the sheer amount of data transferred throughout the networks that cannot be simply captured and stored. To our knowledge, there is no publicly available capture of high-speed that would depict the entirety of traffic over a long period of time.

Therefore, we present a long-term aggregated dataset addressing the impracticality of dealing with and sharing a tremendous amount of data. This paper also presents a comprehensive analysis, traffic properties, and detailed statistics of the presented dataset. The dataset has been created from a large-scale national research and education network using our monitoring architecture. To provide a better understanding of the dataset we briefly present the monitoring architecture, which has been used to capture the dataset.

II. RELATED WORKS

There are many existing datasets — Table I summarizes some publicly available datasets of observed network traffic. Most are outdated or deal with only a limited traffic subset, or the measurement is not continuous. This exact problem is also mentioned in [14].

The most relevant group that provides an ISP dataset is the MAWI Working Group [1, 15]; the most recent available dataset is from samplepoint-F dated early 2023 within network speed of 1 Gbit/s. Even some of the recent publications [16] refer to these datasets, which may provide insufficient insight into the performance of high-speed networks. The main problem with providing high-speed network datasets is the amount of data required to store such datasets. The most recent MAWI dataset, less than 500 GB in size, would occupy between 12.5 TB and 50 TB of space if it were at a traffic speed of 100 Gbit/s.

When it comes to network monitoring approaches operating on large scale network the preferred approach is the flow-based

Table I: Network Traffic Datasets

Dataset	Description
MAWI's Packet traces from WIDE backbone [1]	Recognized and up-to-date dataset; however, it is captured on a relatively slow line, not reaching modern standards in tens of Gbit/s
CESNET-TLS22 [2]	Two-week long dataset containing only a TLS protocol from CESNET network
CESNET-QUIC22 [3]	One-month long dataset containing only a QUIC protocol from CESNET network
The CAIDA UCSD Anonymized Internet Traces 2008-2019 [4]	Anonymized packet headers in PCAP format, traffic is sampled and non-continuous, captured on the relatively slow network
UMassTrace Repository [5]	Outdated generic captures and more recently capture targeting specific traffic
10 Days DNS Network Traffic from April-May [6]	Only DNS traffic and only ten days of 2 months. It does not describe traffic on a full scale; however, it could be useful for specific tasks.
University of Oregon Route Views Project [7]	It provides AS path visualization, topological mapping, host geolocation, etc. BGP perspective.
UNIBS-2009 [8]	Outdated, it describes a completely different reality than today's; moreover, it contains relatively little data (approximately 79k flows only).
Internet Traffic Archive [9]	Wide variety of outdated captures from wide area network and dedicated servers between 2000–2008.
Labeled Network Traffic flows [10]	Small amount of raw flows, but all flows are labeled, so could be used as a dataset for machine learning
UNSW-NB15 [11, 12]	Small generated flow-dataset with the primary goal of benchmarking intrusion detection systems.
TON_IoT [13]	Small flow and capture dataset focused on IoT devices for cybersecurity

oppose to high resource intensive packet-based usually called deep packet inspection (DPI).

For the flow-based approach (surveyed in [17] and further elaborated, e.g., in [18]), the monitoring infrastructure extracts and computes metadata and statistical data for the observed traffic. This aggregation dramatically reduces the amount of information that must be stored or analyzed.

There is an existing category of network monitoring tools using an information aggregation [19, 20]; these usually deal with similar problems as flow exporters for maintaining per-flow information. Subsequently, they perform aggregation and filtering functions with stored flows. However, they typically perform very specific measurement tasks, for example, *Heavy hitter detection*, and etc. The main point of these tools is to provide accurate and real-time feedback to network operators to evaluate the situation on the distributed network, even under DDoS conditions. They usually require a dedicated deployment of measurement nodes for the measurement technique, which is impractical.

The most recent flow-based measurement/analysis tool by Saidy et al. [21] focuses on query-based architecture. The system *Flowyager* stores the flows in memory to perform queries on them. The authors constructed flow datasets IXP and ISP; however, they have not been published.

Trevisan et al. [22] have looked inside the Italian national ISP for more than five years and observed the trend of increasing downloaded data for the average network user. However, the author primarily focused on providing insight into services within the application layer (L7 in the ISO/OSI model). The size of their dataset is 31 TB over five years within the lines of 4 Mbps-20 Mbps for ADSL users and 100 Mbps for fiber-to-the-home (FTTH) users, while providing no information about their backbone network.

Benson et al. [23] is also an outdated article from 2010. They studied 10 data centres of different kinds (university campuses, private data centres, and cloud data centres) and modelled distributions of the interarrival times of packets.

Gebert et al. [24] analyzed a network for 14 days with 600 users, but their observations are rather outdated since the measurement is more than 10 years old.

There are flow-based datasets [11–13, 25, 26], which are primarily focusing on use in machine learning. However, these types of data sets do not represent the actual behaviour on the network; they try to cover as many labelled traffic cases as possible.

To our knowledge, no recent related work focuses on the extensive detail of long-term analysis of real high-speed network traffic (100 Gbit/s and potentially more) using flow-based online monitoring tools that collect advanced statistics regarding the distribution of data streams over time. However, we have drawn from previous publications to include commonly used statistics and metrics the community monitors.

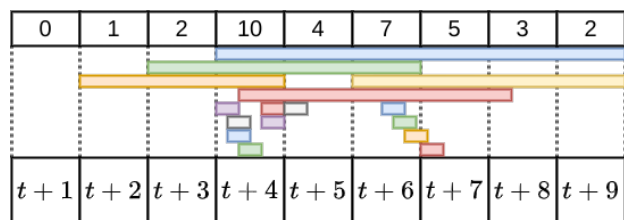
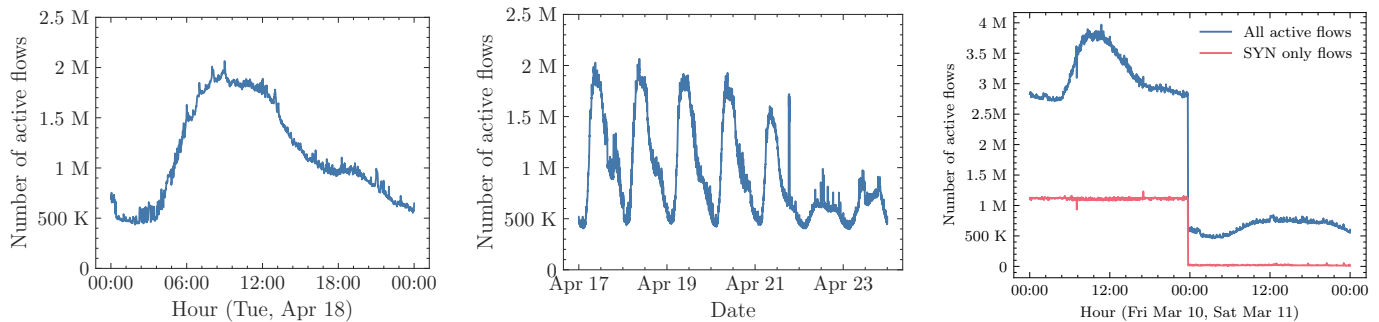


Figure 2: Flow intervals visualized across aggregation windows. The top numbers visualize the number of active flows inside the given window.

III. METHOD OF INTERPRETATION OF FLOW INFORMATION IN A TIME-BASED DOMAIN

A, Flows: The flow-based monitoring approach was explained in detail, e.g., by Hofstede et al. in [27]. The commonly used representation of flow data is currently Net-flow [28] or IPFIX format [29]. This section emphasizes how packets are associated with a given flow using flow keys because it is the most crucial concept for fully understanding our aggregation architecture.



(a) The number of active flows during one day. (b) The number of active flows during one week. (c) Unusual traffic containing possible scanning or an SYN flood.

Figure 1: Number of active flows from different time points of the measurement.

B, Flow data in time windows: Bidirectional flow [30] information is better for the interpretation and understanding of traffic than traditional unidirectional ones. This leaves us with aggregated information, which is contained in a time interval defined by the first packet and the last packet inside the flow. Unfortunately, even this aggregation creates too large an amount of data in 100G networks to create any practical dataset.

Therefore, it is useful to aggregate the data on the fly and compute statistics that can explain the distribution of the traffic within each time interval of the observed long-term period. We have chosen a time-based aggregation using a time window of one minute. A predefined set of statistics is computed using a set of overlapping flows for a given time window. This example situation is shown in Figure 2, where the flows are displayed as bars spreading across single or multiple time windows $t+n$. Additionally, on the top, we can always see the current number of active flows for a given time window. Note that the aggregation intervals are in the order of minutes, which keeps the number of splittings of bursts, which are common in cloud infrastructures, into separate windows relatively low compared to the total volume of traffic.

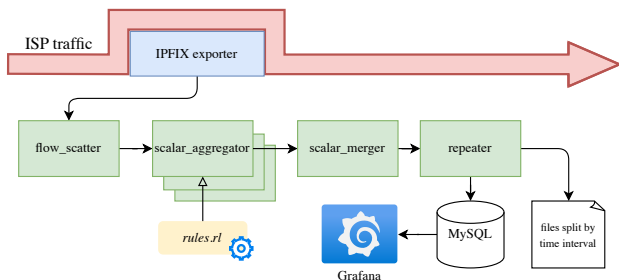


Figure 3: The schema shows flow-based infrastructure to capture, process, and visualize information about the network traffic of an ISP peering line.

Unfortunately, the traditional flow information does not contain any items regarding the distribution of the packets over its time interval. For statistics such as the “number of bytes,” “number of packets,” etc. We chose to proportionally distribute the information across all overlapped time windows

Table II: Flow features and their descriptions

IP Flow Feature	Description
BYTES	Total number of bytes in the flow
PACKETS	Total number of packets in the flow
SRC_IP	The IP address from which the flow originated
DST_IP	The IP address to which the flow is destined
SRC_PORT	The port of the originating device
DST_PORT	The port of the destination device
PROTOCOL	The L4 protocol used for the flow (e.g., TCP, UDP)
TIME_FIRST	Timestamp of the first packet in the flow.
TIME_LAST	Timestamp of the last packet in the flow.
FLOW_KEY	Flow key for unidirectional flow
TCP_FLAGS	TCP flags if TCP protocol is used

to avoid overestimating the characteristics in long flows, as some values would then be aggregated multiple times across multiple windows or one time into a single window.

Other more simple information, such as “how many active flows” or “how many source ports”, simply sums the number of flows overlapping on top of the current time window t . These are not affected by inaccuracies caused by the aggregation of flow data.

IV. MONITORING ARCHITECTURE

In the following section, we describe the components of our scalable monitoring infrastructure composed of open-source tools, as shown in Fig. 3. These days flow exporters are commonly used for gaining basic insight into the networks. We are trying to extend the insight by using the pre-existing infrastructure of these exporters. The measurement is compatible with any flow exporter as long as it provides features shown in Table II and is using a widely used IPFIX data protocol. We take the input IPFIX stream and separate it into N parts to enable parallel processing using our aggregation method. The aggregation of computed statistics is fully configurable and controlled using a rules file for the scalar aggregator. Any additional fields provided by flow-exporters are passed through the infrastructure architecture to the aggregators and can be used by the rule set. After the aggregation, we merge the results and save them into permanent storage.

A, Flow exporter: As a flow data source, we use *ipfix-probe* [31], a robust open-source exporter offering a comprehensive suite of features for inspecting network traffic. It

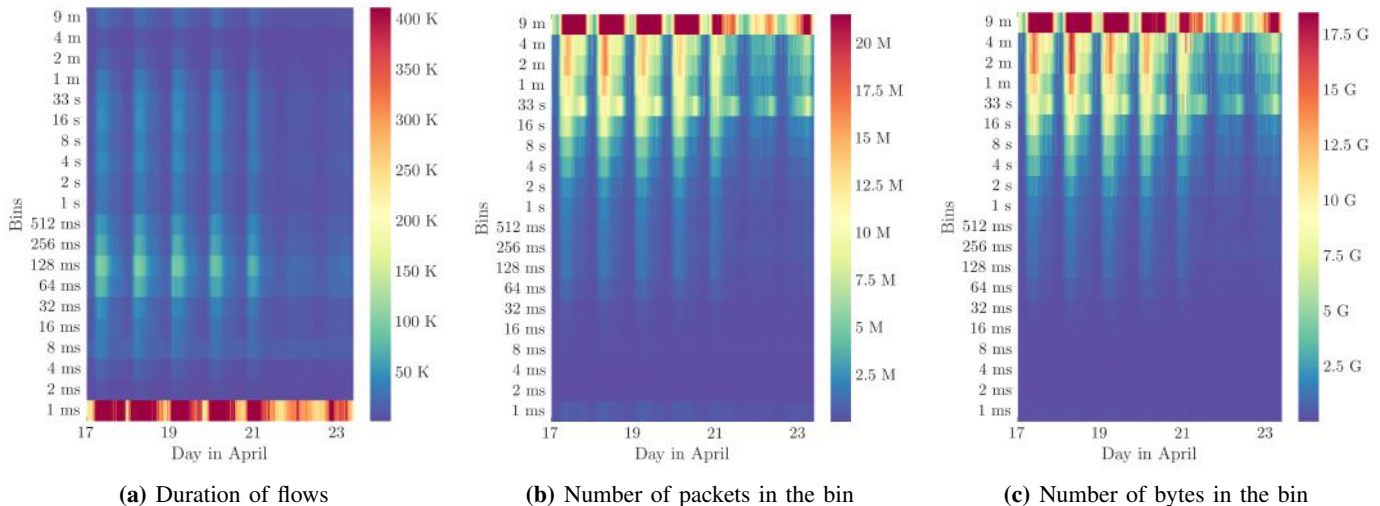


Figure 4: Time-oriented histograms of duration. The vertical axis uses a logarithmic scale, and each histogram bin aggregates flows with similar duration. The horizontal axis corresponds to the 1-minute time windows.

supports hardware-accelerated network cards to monitor even 100 Gbit/s lines and is faster using Network Development Kit (NDK) or Data Plane Development Kit (DPDK).

B, Scalability: The architecture can deploy any number of aggregation nodes to process high volumes of data efficiently. This is managed by the flow scatter node shown in Fig. 3. It splits the input flow stream into N parts using a hashing method based on the flow key. We can then scale the number of nodes to provide the necessary processing power for a given network.

C, Aggregation of flows: In our study, we employ a scalar aggregation (included in the NEMEA system [18]) to analyze the properties and behaviour of flows. These scalar aggregator nodes are shown in Fig. 3. It is a special case of an aggregation process over time, where the output is a single vector of statistics per fixed time window.

The main power of scalar aggregation is the ability to configure a set of computed statistics and the time window size. The set of statistics is specified using a series of rules. The rules always consist of a *label* (name of the new computed statistic), *operation*, and a *filter*. Operation defines a statistic calculation consisting of various functions, which always operate on the fields exported by the flow exporter listed in Table II. The scalar aggregator then uses the set of rules as input depicted in Figure 3. Thanks to this the aggregation process can be simply modified/configured to accommodate most of the specific monitoring needs.

D, Scalability merging: At the end of the multiple aggregations, the results need to be aggregated into a single one to represent the entire input flow stream. This is done by the scalar merger node shown in Fig. 3. The scalar merger uses the same rules set as the aggregation and carefully applies appropriate functions to merge the partial results into a single one.

Table III: Basic descriptive statistics for a 1-minute window calculated across the whole measurement period.

Variable	Mean (SD)	Min	Q1	Q3	Max
Flow average duration [s]	12.04 (3.06)	2.71	10.20	14.46	18.07
Data transmitted [GB]	45.59 (27.67)	5.38	21.84	65.66	215.66
Data TCP [GB]	28.57 (17.33)	3.24	14.02	41.11	159.99
Data UDP [GB]	16.12 (10.63)	0.90	7.20	23.69	54.85
Number of flows [MF]	1.10 (0.78)	0.38	0.58	1.44	5.83
Packets [MP]	53.85 (32.25)	7.54	26.02	77.85	243.92
Average packets per flow	53.14 (18.33)	5.49	38.56	64.62	172.26

E, Output: The computed statistics are efficiently stored in two primary ways by the repeater node shown in Fig. 3. Firstly, the results are stored in a MySQL database for real-time visualization using the open-source platform Grafana. Secondly, they are stored in a file system with periodic backups.

V. MEASUREMENT AND ANALYSIS

This section presents traditional metrics, such as the number of flows or packets on the network, etc. We demonstrate the differences in traffic during weekdays, weekends, days, and nights and point out some of our observations from our extensive traffic data analysis. Lastly, we focus on showing the heavy-tail nature and shape of the traffic.

A, The environment of the experiment: The measurement was conducted in CESNET2, the national research and education network in the Czech Republic. It interconnects many academic institutions, research organizations, governmental offices, and others. It represents around 500 K users. There are multiple 100 Gbit/s peering and transit links; six monitoring probes monitor them at the infrastructure perimeter. The probes are equipped with custom hardware cards with FPGA to accelerate packet processing created by Liberouter project [32]. The flow collector receives an average rate of about 150 K bidirectional flow records per second from 8 peering lines. Measurement for this paper was performed from February 25th, 2023, to May 3rd, 2023, using one monitoring

Table IV: Description of network traffic volumes (flows, bytes, packets) for different periods (day/night, weekday/weekend).

		Flows [MF]					Bytes [GB]					Packets [KP]				
		Mean (SD)	Min	Q ₁	Q ₃	Max	Mean (SD)	Min	Q ₁	Q ₃	Max	Mean (SD)	Min	Q ₁	Q ₃	Max
Day	Weekday	82.27 (25.65)	13.87	66.88	101.57	243.92	69.97 (22.40)	10.73	56.39	86.54	215.66	1.56 (0.79)	0.48	1.00	1.82	5.83
	Weekend	38.30 (12.13)	10.68	30.17	44.64	88.44	32.52 (10.02)	7.76	26.06	37.69	88.40	0.70 (0.79)	0.44	0.59	0.74	3.15
Night	Weekday	34.21 (21.36)	7.54	17.07	46.63	108.10	28.50 (17.92)	5.37	14.09	39.64	98.32	0.89 (0.72)	0.38	0.51	0.84	3.67
	Weekend	27.69 (17.73)	8.38	14.43	37.01	93.43	23.28 (14.76)	5.54	12.07	31.59	98.58	0.59 (0.36)	0.38	0.47	0.59	3.27

probe that observes one of the 100Gbit/s lines to the Czech internet exchange point (NIX.CZ), which carries the majority of public traffic of CESNET2.

B, Measured statistics: The day is defined as the time period of from 6 to 20 hours of local time; otherwise, there is a night. The weekday is defined as commonly understood Monday till Friday; otherwise weekend. These heavily depend on the country’s culture where the measurement is conducted. We primarily defined these properties to separate the results and make them more presentable. Students and workers mainly use the network during their work time period. Due to the specific nature of our network, we have chosen this view to separate weekends and weekdays due to their significant traffic difference. Additionally, we wanted to highlight the rush hour, which is on our network during the middle of the weekdays. Traditionally is the rush hour, on home-connected ISP networks, around 6 pm-8 pm when people are coming back from home.

Table V: Top prevalent ports (source and destination) from one-month perspective.

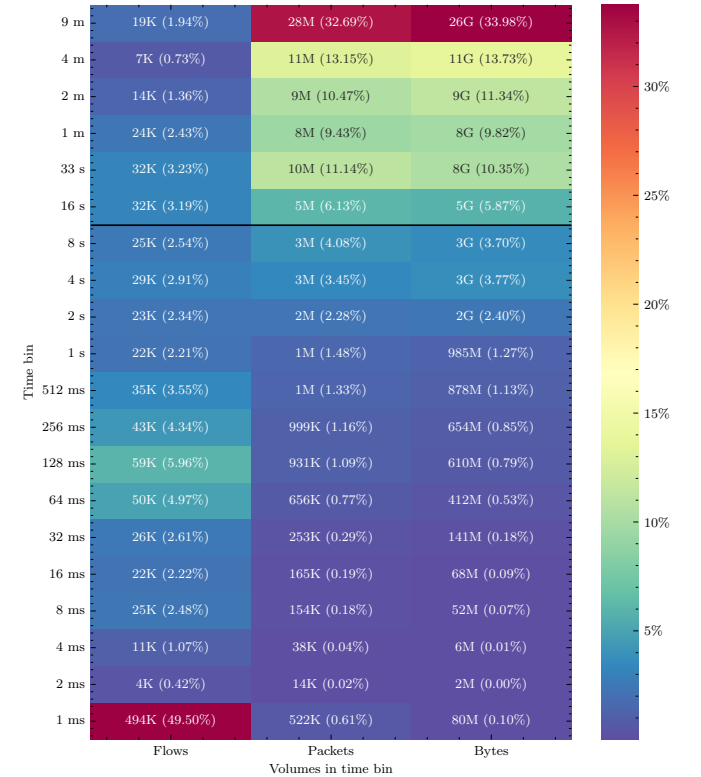
Top ports by flows			Top ports by packets		
Port	Portion [%]	Service	Port	Portion [%]	Service
443	55.10	https	443	76.74	https
53	31.73	domain	80	7.62	http
123	4.24	ntp	1095	2.04	nicelink
0	2.12	icmp	0	1.88	icmp
80	1.67	http	873	1.74	rsync
5222	1.10	jabber-client	53	0.54	domain
993	0.23	imaps	1194	0.33	openvpn
25	0.14	smtp	993	0.25	imaps
445	0.13	microsoft-ds	5222	0.18	jabber-client
22	0.10	ssh	554	0.17	Rtsp

When we mention the number of flows, we always refer to the number of active flows, ergo the number of overlapping flows in a given time window. Such a number is related to the number of flows kept inside a flow cache of flow exporters, which is controlled (and prolonged for short flows) by inactive timeouts. Typically, UDP flows do not have the connection-ending condition; thus, flow exporters have an inactive timeout to wait for possible following packets within the same stream. Due to this property of the flow exporters, the number of flows always represents the worst-case scenario on the network. It is essential to understand this metric to interpret the results accordingly. This metric should be more accurate for network applications than commonly used flows per second, obscuring the network traffic’s reality in certain situations.

C, Volumetry: This section presents an empirical analysis of an approximately four-month data collection period, scru-

Table VI: Table of throughput (byte-wise and packet-wise) across all one-minute windows.

Estimated throughput [Gbps]					Packet rate [MPps]				
Mean (SD)	Min	Q1	Q3	Max	Mean (SD)	Min	Q1	Q3	Max
6.08 (3.69)	0.72	2.91	8.76	28.75	53.82 (32.25)	7.54	26.02	77.85	243.92

**Figure 5:** Histogram of detailed flow, packet, and size distribution in 1-minute windows during the middle of the day. The black horizontal line denotes the duration mean of flows.

tinizing multiple components — flows, packets, and bytes transmitted. We show off the different views of the network traffic to provide insight into any network application dealing with a specific network layer and emphasize the difference between flow-based and packet-based network applications. Table IV shows basic volumetrics (flow-wise, packet-wise, byte-wise). Here we can see the total volumes of flows, packets and bytes being transferred throughout the network in different time periods. The most notable difference is between a weekday and weekends across all of the domains. The traffic during the night is very similar to the traffic during the weekend, which is caused by most of the users being inactive during this period of time.

Figure 1 shows a progression of the number of active flows throughout the day (1a) and the week (1b). It is worth noting,

Table VII: Proportion (fraction of all active flows) of L3 (IPv4/IPv6) and L4 (TCP/UDP) protocols across time windows.

	Flows [%]			Bytes [%]				Packets [%]				
	Mean (SD)	Min	Max	Sum [GF]	Mean (SD)	Min	Max	Sum [PB]	Mean (SD)	Min	Max	Sum [TP]
IPv4	72.14 (7.24)	14.73	95.58	80.62	80.93 (5.97)	9.19	95.73	3.53	83.15 (4.40)	17.75	94.41	4.31
IPv6	27.86 (7.24)	4.42	85.27	25.54	19.07 (5.97)	4.27	90.81	0.89	16.85 (4.40)	5.59	82.25	0.90
TCP	54.19 (7.67)	12.14	81.39	59.98	63.69 (6.61)	40.95	96.76	2.77	65.14 (5.63)	47.89	94.72	3.36
UDP	43.73 (7.20)	17.88	87.19	44.52	33.37 (7.24)	2.26	52.06	1.56	32.28 (6.41)	3.95	49.05	1.77

contrary to expectations for public ISP networks, that the peak traffic in our monitored network does not occur in the evening. Instead, as this is predominantly a working network, the rush hour appears before noon, indicating a specific usage pattern and showcasing the characteristic behaviour of this type of network.

Figure 1c illustrates a time period where we noticed an unusual traffic of around 2 million flows, half of which were TCP flows containing only SYN packets. While this pattern could suggest a potential scanning or SYN-flood attack treating this observation as a hypothesis requiring further investigation is essential.

Figure 6 visualizes all the days included in our measurement grouped by their weekday. The main curve represents the mean value of all respective weekdays, and the gradient represents the overlap of all the respective weekdays. Here we can see the other specific factor of your network, which is that the weekend traffic is absolutely different from the weekday traffic. This makes sense because students and workers mainly use our network during their work time period.

Table III summarizes the volumetric and overall information across all the measurement windows. It represents information transmitted every 1-minute window over the network. This information can be used to have an approximate load for any network application running on such 100 G link.

Table VI shows the byte and packet rates on the measurement probe. This metric is an approximation for these rates from the flow interval information. Due to the nature of the measurement, we do not have the exact values for each time window on the flow-based aggregator. The number of packets/bytes is always equally distributed to all the time windows the flows cover, decreasing the actual network's peaks. Thus it should be taken as an orientation value instead of an exact view of the traffic.

D, Protocols: Table VII shows the portions of traffic transmitted throughout the network by L3 and L4 protocols. The percentage is stated for each respected domain to demonstrate the difference in traffic nature. We can see that IPV4 is still the dominant protocol over IPV6 in all of the cases. Additionally, we can unexpectedly see that the TCP is dominant over UDP in the number of flows, which is rather unexpected due to the overall number of short flows inside the traffic. Moreover, the average of transmitted bytes and packets by the UDP flows is also unexpectedly very close to the number of bytes and packets transmitted by the TCP flows. The other most notable information shown in Table VII is the total number of flows observed by our measurement which reaches the order of

Giga Flows, the number of packets reaches Tera Packets, and the number of processed reaches the order of Peta Bytes.

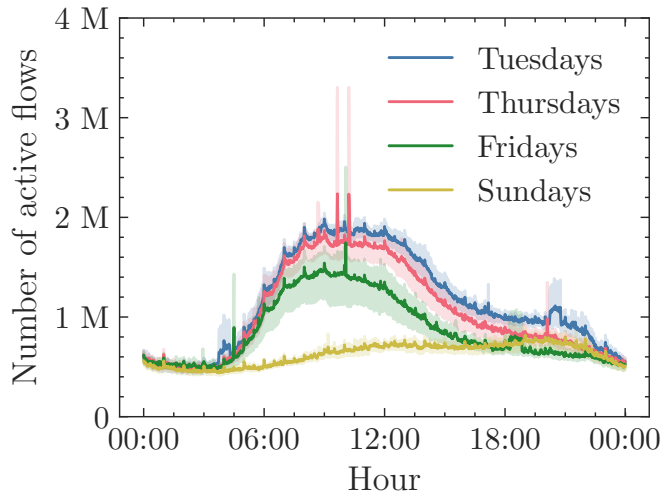
**Figure 6:** Number of active flows throughout selected days of the week.

Table V summarizes our network's most frequently observed protocols/ports from both packet and flow perspectives, based on the top twenty source and destination ports. We currently estimate the L7 protocol based on the lower of the source and destination ports. In our list there are some unexpectedly high-range ports, we surmise this might be due to the use of common protocols like HTTPS or, more likely lesser-known services' usage of these ports. Our dataset does not yet contain the port aggregation part of the traffic that will be added in the future. We acquired this statistic using our internal flow monitoring tools mainly to complete the gathered statistics as other related work is doing the same [1].

E, Traffic distribution: Traffic on ISP networks tends to be heavy-tail in nature. This is shown by the histograms in Figure 5, showing the distribution of flows according to their time lengths in the chosen time windows. Each time interval has a number of flows, packets, and bytes that belong to a group of flows with the given length. More than 70% of the flows (short flows) contain less than 3% of packets and 2% of bytes. More than 70% of packets and bytes are contained in less than 10% of flows (long flows), commonly called heavy flows. All of our bin measurements end with the last bin (9–18 min) due to the configuration of the active timeout of the flow exporter. It is set to 10 min during all of our measurements. This causes all flows longer than the active timeout to be separated into two flow records. Our measurement does not contain any collector,

which would stitch these flows together, so our view is limited to a maximum of 10 minutes.

Looking at the flows' min, max, and mean values can be very misleading when inferring information about the network's traffic. This obstruction usually causes most visualization attempts to depict the situation on the network incorrectly. This can be seen in Figure 4, a histogram that tries to visualize the progression of the number of flows according to their lengths during a week. The first figure shows that most flows are shorter than 1 ms. To display the proper magnitude of the actual traffic of the flows, we have to show the number of packets and bytes transmitted instead of the flow count. The second figure shows the sum of bytes transferred inside the bin, and the third is the sum of packets transferred inside the bin.

VI. CONCLUSION AND FUTURE WORK

We published all the information gathered for our CESNET ISP network as a dataset in the form of a CSV file with the statistics gathered available at [33] for other researchers. We have given researchers access to detailed information about an ISP network, a complex field that lacks comprehensive data. We processed and extracted the most interesting metrics, in our opinion, to allow researchers to optimize future research for ISP networks. We presented novel views on the data gathered, such as packet, byte, and flow domains, to highlight the effect of heavy-tail traffic on the ISP network. Lastly, provide information on the difference between traffic during different periods of time according to the system's location to allow the service to optimize its computational resources and reduce maintenance costs.

REFERENCES

- [1] "MAWI Working Group Traffic Archive," [Online]. Available: <https://mawi.wide.ad.jp/mawi/>.
- [2] J. Luxemburk *et al.*, "Fine-grained TLS services classification with reject option," *Computer Networks*, vol. 220, p. 109467, Jan. 1, 2023, DOI: 10.1016/j.comnet.2022.109467.
- [3] J. Luxemburk *et al.*, "CESNET-QUIC22: A large one-month QUIC network traffic dataset from backbone lines," *Data in Brief*, vol. 46, p. 108888, Feb. 2023, DOI: 10.1016/j.dib.2023.108888.
- [4] "The CAIDA Anonymized Internet Traces Data Access," CAIDA. (Mar. 21, 2019), [Online]. Available: https://www.caida.org/catalog/datasets/passive_dataset_download/ (visited on 06/18/2023).
- [5] "UMass Trace Repository," [Online]. Available: <http://traces.cs.umass.edu/index.php/Network/Network>.
- [6] M. Singh *et al.*, "10 days DNS network traffic from april-may, 2016," vol. 2, May 7, 2019, DOI: 10.17632/zh3wnddzxy.2.
- [7] "University of Oregon Route Views Project," [Online]. Available: <http://www.routeviews.org/routeviews/>.
- [8] "The Telecommunication Networks Group. (UNIBS)," [Online]. Available: <http://netweb.ing.unibs.it/~ntw/tools/traces/>.
- [9] "Traces In The Internet Traffic Archive," [Online]. Available: <https://ita.ee.lbl.gov/html/traces.html> (visited on 06/18/2023).
- [10] "Labeled Network Traffic flows - 141 Applications," [Online]. Available: <https://www.kaggle.com/datasets/jsrojas/labeled-network-traffic-flows-114-applications> (visited on 06/18/2023).
- [11] N. Moustafa *et al.*, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Mil. Commun. Inf. Syst. Conf. MilCIS*, Nov. 2015, pp. 1–6, DOI: 10.1109/MilCIS.2015.7348942.
- [12] M. Sarhan *et al.*, "NetFlow Datasets for Machine Learning-based Network Intrusion Detection Systems," version 1, 2020, DOI: 10.48550/ARXIV.2011.09144.
- [13] N. Moustafa, *ToN_IoT datasets*, IEEE DataPort, Oct. 16, 2019, DOI: 10.21227/FESZ-DM97.
- [14] A. D'Alconzo *et al.*, "A Survey on Big Data for Network Traffic Monitoring and Analysis," *IEEE Trans. Netw. Serv. Manag.*, vol. 16, no. 3, pp. 800–813, Sep. 2019, DOI: 10.1109/TNSM.2019.2933358.
- [15] K. Cho *et al.*, "Traffic data repository at the WIDE project," in *2000 USENIX Annu. Tech. Conf. USENIX ATC 00*, San Diego, CA: USENIX Association, Jun. 2000, [Online]. Available: <https://www.usenix.org/conference/2000-usenix-annual-technical-conference/traffic-data-repository-wide-project>.
- [16] G. Vormayr *et al.*, "Why are My Flows Different? A Tutorial on Flow Exporters," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 3, pp. 2064–2103, 2020, DOI: 10.1109/COMST.2020.2989695.
- [17] B. Li *et al.*, "A survey of network flow applications," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 567–581, Mar. 1, 2013, DOI: 10.1016/j.jnca.2012.12.020.
- [18] T. Čejka *et al.*, "NEMEA: A framework for network traffic analysis," in *2016 12th International Conference on Network and Service Management (CNSM)*, Montreal, QC, Canada: IEEE, Oct. 2016, pp. 195–201, DOI: 10.1109/CNSM.2016.7818417.
- [19] Y. Li *et al.*, "FlowRadar: A better NetFlow for data centers," in *13th USENIX Symp. Networked Syst. Des. Implement. NSDI 16*, Santa Clara, CA: USENIX Association, Mar. 2016, pp. 311–324, [Online]. Available: <https://www.usenix.org/conference/nsdi16/technical-sessions/presentation/li-yuliang>.
- [20] D. Yu *et al.*, "dShark: A general, easy to program and scalable framework for analyzing in-network packet traces," in *16th USENIX Symp. Networked Syst. Des. Implement. NSDI 19*, Boston, MA: USENIX Association, Feb. 2019, pp. 207–220, [Online]. Available: <https://www.usenix.org/conference/nsdi19/presentation/you>.
- [21] S. J. Saidi *et al.*, "Exploring Network-Wide Flow Data With Flowyager," *IEEE Trans. Netw. Serv. Manag.*, vol. 17, no. 4, pp. 1988–2006, Dec. 2020, DOI: 10.1109/TNSM.2020.3034278.
- [22] M. Trevisan *et al.*, "Five years at the edge: Watching internet from the ISP network," in *Proc. 14th Int. Conf. Emerg. Netw. Exp. Technol.*, Heraklion Greece: ACM, Dec. 4, 2018, pp. 1–12, DOI: 10.1145/3281411.3281433.
- [23] T. Benson *et al.*, "Network traffic characteristics of data centers in the wild," in *Proc. 10th Annu. Conf. Internet Meas. - IMC 10*, Melbourne, Australia: ACM Press, 2010, p. 267, DOI: 10.1145/1879141.1879175.
- [24] S. Gebert *et al.*, "Internet Access Traffic Measurement and Analysis," in *Traffic Monitoring and Analysis*, A. Pescapè *et al.*, Eds., vol. 7189, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 29–42, ISBN: 978-3-642-28533-2 978-3-642-28534-9, DOI: 10.1007/978-3-642-28534-9_3.
- [25] I. Sharafaldin *et al.*, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Priv.*, Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018, pp. 108–116, DOI: 10.5220/0006639801080116.
- [26] N. Koroniotis *et al.*, "Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset," version 1, 2018, DOI: 10.48550/ARXIV.1811.00701.
- [27] R. Hofstede *et al.*, "Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX," *IEEE Commun. Surv. Tutor.*, vol. 16, no. 4, pp. 2037–2064, 2014, DOI: 10.1109/COMST.2014.2321898.
- [28] B. Claise, *Cisco systems NetFlow services export version 9*, RFC 3954, RFC Editor, Oct. 2004, DOI: 10.17487/RFC3954.
- [29] P. Aitken *et al.*, *Specification of the IP flow information export (IPFIX) protocol for the exchange of flow information*, RFC 7011, RFC Editor, Sep. 2013, DOI: 10.17487/RFC7011.
- [30] B. Trammell *et al.*, *Bidirectional flow export using IP flow information export (IPFIX)*, RFC 5103, RFC Editor, Jan. 2008, DOI: 10.17487/RFC5103.
- [31] *Ipfixprobe - IPFIX flow exporter*, CESNET, Jun. 8, 2023, [Online]. Available: <https://github.com/CESNET/ipfixprobe> (visited on 06/26/2023).
- [32] "Publications — Liberouter / Cesnet TMC group," [Online]. Available: <https://www.liberouter.org/publications/> (visited on 09/07/2023).
- [33] T. Benes *et al.*, *CESNET-AGG23*, Zenodo, Jun. 18, 2023, DOI: 10.5281/ZENODO.8053021.