A Traffic Visualization Framework for Monitoring Large-scale Inter- DataCenter Network

Meryem Elbaham
École de Technologie supérieure
Montréal, Canada
Meryem.el-baham.1@etsmtl.net

Kim Khoa Nguyen
École de Technologie supérieure
Montréal, Canada

Mohammed Cheriet
École de Technologie supérieure
Montréal, Canada

***Abstract*— Diversity, dynamicity, and the huge volume of traffic in the network between datacenters has risen network administrators concerns on how to efficiently visualize their system in real-time. To deal with these challenges, we present in this paper a visualization framework based on advanced machine learning, traffic characterization, sampling, and graphical visualization algorithms, which aims to efficiently support inter-datacenter network monitoring. Experimental results show the framework is able to process real-time big flows and provides human-friendly interactive graphical representations.**

Keywords—Machine learning; Network monitoring; traffic classification; traffic sampling; visualization.

I. INTRODUCTION

Today, Internet has become an essential and a universal media for transmitting any kind of information between remote network entities. As a result, the amount of data the network administrators have to analyze is continuously increasing. Thus, network monitoring is more challenging, and traditional tools are no longer efficient to provide an overview of traffic patterns. So, other methods which are more adaptive to characterize traffic and describe network state efficiently in real time are required.

Visualization system is a promising solution to simplifying network monitoring in an efficient way. Many visualization tools have been developed. However, most of them are customized and used only for some specific use cases, especially in security field [1]. For this reason the network administrator has to use multiple tools together to have a complete view of the network state. As well, some issues are still not well addressed, for example regrouping hosts on external and internal groups where the user may focus on the external events. However, serious problems or attacks may come from internals hosts. In addition, many visualization systems are not suitable for real-time visualization as they process all the traffic [1].

Our solution to address these issues is to use a behavioral learning approach combined with visual analytics to characterize, identify and classify IP traffic. The contributions of our work are as follows:

1. We propose a solution combining behavioral and statistical approaches to characterize and classify traffic. Thus, different applications can be identified using a graphical dispersion of flows.
2. A new statistical based approach to classify dynamic ports and encrypted applications in real time.
3. A method providing real-time traffic analysis which helps to detect and prevent external and internal malicious events through a graph of ports activities.
4. Employing sampling technique to collect traffic and to reduce the amount of processes data.

The remainder of this paper is organized as follows. Related work in the areas of traffic classification, traffic visualization and sampling techniques is presented in Section II. Section III describes our proposed real-time traffic visualization solution. Section IV discusses examples of monitoring use cases using our framework. Finally, conclusions and future work are drawn.

II. RELATED WORK AND BACKGROUND

Regarding the complexity of inter-datacenter network, as well as its huge volume of traffic, traffic visualization is an inseparable part of network monitoring. This section reviews existing traffic visualization tools as well as previous works done in traffic classification and traffic sampling areas.

A. *Traffic visualization system*

Visualization systems have been developed to support and facilitate traffic monitoring. As traffic volume is continuously increasing, they exploit human's visual system by creating insightful graphical representations to explore these huge amounts of data quickly and efficiently.

VISUAL is a visualization system displaying communications patterns between internal network and external networks [2]. In this system, internal network is represented by a grid-based graph where each cell corresponds to a local host. Hosts in external network are represented as squares whose sizes indicate activity levels. Connections between internals and externals hosts are represented by a line. A filtering system can be used to display details for a specific internal or external host. TNV is designed to prevent the loss of sight of the general context of the network while investigating low-level of details of attacks by integrating packet level details in big picture of the network [3]. NvisionIP is a

978-3-901882-85-2 (c) 2016 IFIP

visualization system focusing on host activities in a class B network [4]. All subnets of the network are projected along the horizontal axis of the galaxy view while the hosts of each subnet are listed down the vertical list. RTA visualization system displays communications distribution of a particular host in a radial graph [5]. As it uses port number to identify applications, it may not be accurate when an application uses dynamic ports. A system to visualize port activities through a stacked histogram graph is presented in [6], to assist the network analyst in detecting malicious activities like zero-day exploits which cannot be detected by the conventional methods. Other works have emphasized on a particular kind of traffic like the software in [7], which visualizes SNMP traffic.

B. Network traffic classification techniques

Traffic identification and classification is an integral part of supervision and networks management. Works done in this area can be categorized into three groups based on their traffic classification approaches. The first one uses port number based techniques [8], which are accurate only for legacy applications, and are not appropriate for applications using non-standard ports such as P2P [9,10] . To overcome the port-based classification limitations, the second group uses deep-packet inspection approach which inspects the payload of packets looking for signatures at the application level [10]. Although deep-packet inspection resolves the problem of port-based approach it is still inaccurate when dealing with encrypted applications. As the previous techniques are becoming unreliable for an accurate classification of applications, statistical approaches were used by the third group to accurately classify network traffic using machine learning techniques. A simple approach based on Naive Bayesian technique is proposed in [11]. However, this work is limited to some legacy applications. A traffic classifier using Support Vector Machine is developed in [12]. A novel approach based on host behavior was introduced as a different way to classify traffic [13]. It uses Traffic Dispersion Graph (TDG) and a series of different metrics to quantify and evaluate the TDG such as InO metric which has a good correlation with P2P applications.

C. Traffic sampling and Sflow

Real-time traffic visualization is facing exponentially growing traffic. For this reason data reduction methods are required. This section presents a brief description of sampling methods in particular Sflow[14]. The sampling consists in selecting a packet from a stream of packets with a probability q. Sflow uses a random sampling algorithm, in particular 1-out-of N according to the sampling rate to prevent synchronization for periodic pattern in traffic.

III. TRAFFIC VISUALIZATION FRAMEWORK

The proposed visualization framework, as shown in Fig.1, provides different traffic analytics in real-time. In particular, it deals with temporal characteristics, flow-based analysis and application classification in addition to port activity based monitoring. Thus, the user can navigate over different views to get more details and clear idea on the network state.

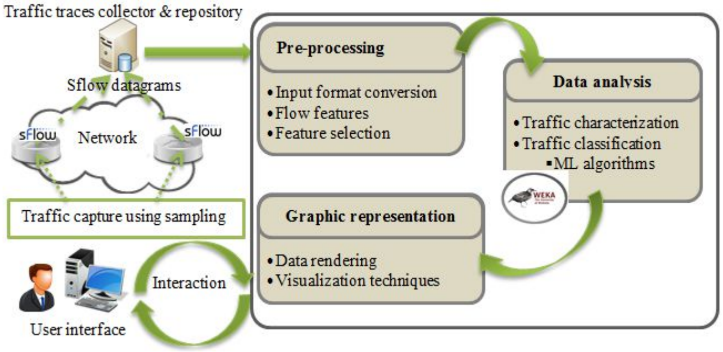


Fig. 1. Visualization mechanism architecture

A. Traffic capturing and sampling

Due to the big size of traffic flows, capturing all packets transiting across the network is no longer appropriate because of high resource consumption and computational time requirements. To address this issue, a sampling technique is mandatory to reduce the volume of data to be processed, thus real-time visualization can be achieved. In this work, Sflow is used for packet sampling because of its flexibility and ability to be used in SDN context and high-speed networks regarding its sampling strategy [15].

Although Sflow is useful to monitor high-speed networks, its deployment can influence on the sampling performance as the traffic volume fluctuates continually. To improve this limitation we propose a dynamic sampling rate assignment based on throughput. In contrast of the conventional Sflow implementation which associates to an interface a static sampling rate, the dynamic one allows to define sampling rate according to throughput as described by equation (1).

$$T_e(\mathrm{t}) = \left(\frac{p_{nb}}{D_t \times d} \right) \times 100 \qquad \textit{(1)}$$

Where $T_e(t)$ is the sampling rate and D_t is throughput (number of packets/s) at time t, P_{nb} is the needed number of samples for the measurement in the period d.

B. Data preprocessing

This step aims to prepare data traffic by selecting information, performing flow statistics from the raw packets and store the relevant information in a compatible format. A set of variables is considered depending on the analysis purpose. The information relative to a particular interface of a monitored network node is first extracted. At this level, the counters interface statistics are processed and exported to data analysis process. In flow-based analysis level, unidirectional flow is considered which is defined as sequence of packet with the same: source IP address, source port, destination IP address, destination port and protocol. Also, the preprocessing step consists in flow features construction and feature selection tasks. The first process computes the flow features, while the second one selects a subset of the entire set of flow features in order to reduce data volume and keep just the significant features. In this work the greedy Stepwise regression algorithm is used for selecting features [16].

C. Data analysis and Classification

The data, which has been preprocessed, will then be analyzed in three complementary levels.

The first level aims to give a general view on network state, in particular for each interface on a specific network node. It provides interface statistics independently to the types of traffic passing through it . A JSON data model is used to interact with Sflow collector and periodically read interfaces statistics and compute their accumulative values over time.

The second level, flow analysis level is intended to provide relevant information on the flows exchanged between the local network and the Internet. At this level, network traffic is visualized in such a way to reveal the repartition of flow and their behaviors, and highlight the relationship between hosts within local network, and between internal and external hosts. Also port activities analysis in implemented.

The third level provides an overview of network at application level. A network traffic classifier based on machine learning is implemented. It deals with flow characteristics selected by the feature selection process. In order to perform online classification, multiple sub-flow approach [17], is adopted in training phase. In the classification phase, only the first sub-flow is used to assign a flow to a class instead to wait until the end of flow to classify it. This approach offers an online classification. A tree decision classifier has been implemented in this context, using the C4.5 algorithm [18]. Decision tree algorithms present several advantages like its comprehensible nature, robustness and low computational cost for generating the models [18].

D. Visualization

The visualization process focuses on visualization techniques and the preparation of data to be in a displayable format by mapping and rendering data processes. This process is carried out according to the type of information and the desired detailed level to display and communicate in a graph. Thus, the graphic representation must be simple while restoring maximum relevant information in order to help the analyst in decision-making process. Three different visualization techniques mainly multiple time-series, bar chart, parallel coordinates and network graphs described in [19] are used. The flow graph, displays flow distribution, while multiple time series and bar chart graph display general characteristics and packets size distribution. The user can get details on demand by drill down and filtering functionalities.

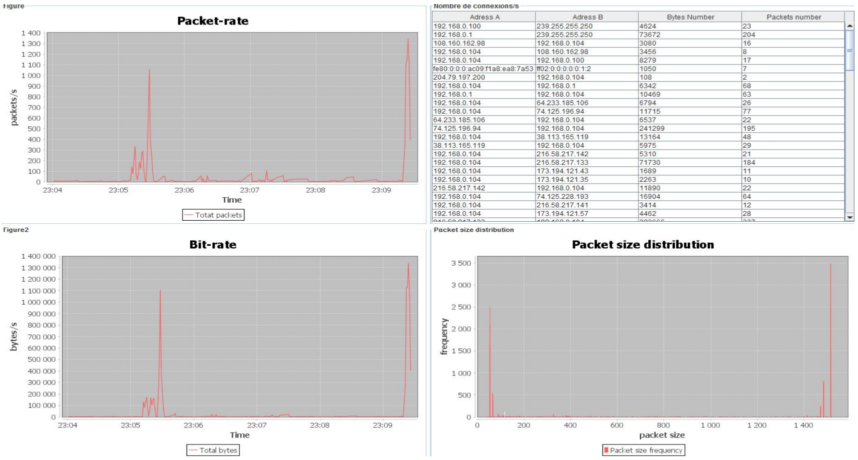
IV. Experimental results

The proposed framework has been developed in Java programming language. A set of libraries and tools has been used in this context. Sflow and sflow-rt [20] are used to sample and collect traffic from network, in addition of Weka Java library and NetMate[21] for traffic classification. Then, the processed data is visualized using Piccolo2D [22].

As shown in Fig.2, a set of general information on a given interface is visualized on three time series based graphs and a table. Two time series graphs on the left show respectively traffic volume (number of packets and bytes) over time. These two charts give a good indicator if there is a burst and if it is a temporary or a persistent burst and so on. The table shows different sessions, and the bar chart at bottom-right draws packets size distribution. So, by combining all information the user can have an overview of a given interface.

At the flow level, the graph displays flows distribution in such a way that the user can distinguish locals and externals nodes. So, a network graph is implemented, where local nodes are represented by the larger circles and the remote nodes are represented by smaller ones connected by links representing communications between them. So, nodes having a potential abnormal behavior can be analyzed. Fig.3 shows traffic flow distribution. The selected local node (the green node) has a lot of activities, which is a good sign on traffic type and nodes states.

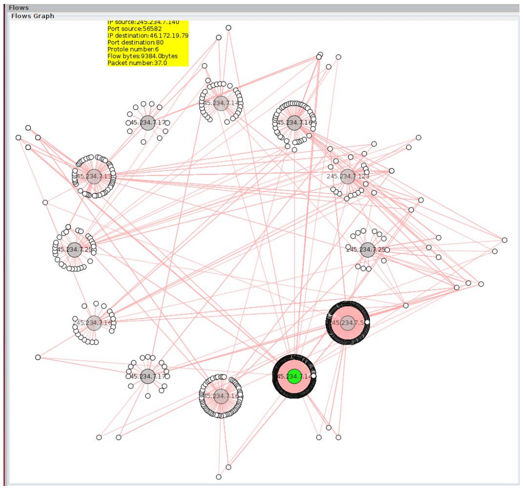
As mentioned earlier, previous works have emphasized on internal/external network behavior like in [4, 5], and they have not paid sufficient attention to the analysis of internal traffic. This approach can have an impact on network performance. Those methods cannot be helpful in the case when a problem or attack comes from one or several internals nodes.



Nombre de connexions/s

Adress A	Adress B	Bytes Number	Packets number
192.168.0.100	239.255.255.250	4624	23
192.168.0.1	239.255.255.250	73672	204
108.160.162.98	192.168.0.104	3080	16
192.168.0.104	108.160.162.98	3456	8
192.168.0.104	192.168.0.100	8279	17
fe80:0:0:0:ac09:f1a8:ea8:7a53	ff02:0:0:0:0:0:1:2	1050	7
204.79.197.200	192.168.0.104	108	2
192.168.0.104	192.168.0.1	6342	68
192.168.0.1	192.168.0.104	10469	63
192.168.0.104	64.233.185.106	6794	26
192.168.0.104	74.125.196.94	11715	77
64.233.185.106	192.168.0.104	6537	22
74.125.196.94	192.168.0.104	241299	195
192.168.0.104	38.113.165.119	13164	48
38.113.165.119	192.168.0.104	5975	29
192.168.0.104	216.58.217.142	5310	21
192.168.0.104	216.58.217.133	71730	184
192.168.0.104	173.194.121.43	1689	11
192.168.0.104	173.194.121.35	2263	10
216.58.217.142	192.168.0.104	11890	22
192.168.0.104	74.125.228.193	16904	64
192.168.0.104	216.58.217.141	3414	12
192.168.0.104	173.194.121.57	4462	28

Fig. 2. General network information



IP source:245.234.7.140
Port source:56582
IP destination:46.172.19.79
Port destination:80
Protole number:6
Flow bytes:9384.0bytes
Packet number:37.0

Fig. 3. flow distribution graph

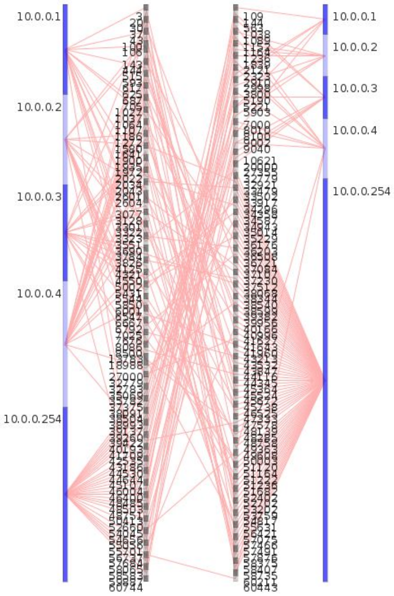


Fig. 4. ports activities graph

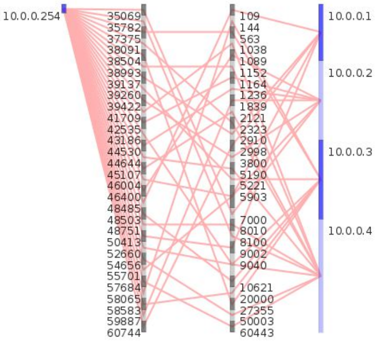


Fig. 5: host-based port activities

To overcome such problem new design to visualize port activities is developed based on parallel coordinate graph with four axes .The first and last axes represent respectively IP source and IP destination, each of them is divided into two segments where the upper one represents local network addresses while the lower one is for remote addresses. Source ports are listed among the second axe and destination ports are listed among third axe from the left. Fig.4 shows a network port scan coming from an internal node while the external activities were normal. To further facilitate the analysis and prevent a cluttered graph the user can select a server having a potentially abnormal behavior and display its information in a separate subgraph as in Fig.5.

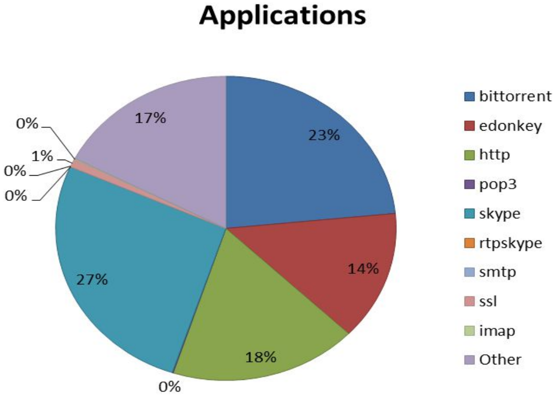


Fig. 6. Traffic classification chart

In order to train and test the classifier, we used three traffic traces collected on the edge router of the campus network of the University of Brescia [23]. The sub-flows characteristics are computed by NetMate. By using the sub-flow a nearly real-time flow classification can be performed, as the sub-flow characteristics are calculated immediately without having to wait until the end of a flow. The implemented C4.5 classifier achieves an accuracy of 99.3%. Fig.6 illustrates the integration of the traffic classification results as a pie chart.

V. Conclusion

The visualization framework described in this paper attempts to address the combined problem of traffic collection, sampling, processing, classification, and data visualization using several tools together in order to understand large-scale network in real-time. A new sampling technique has been proposed to reduce the cost of traffic collection, and a number of new concepts such as flow graph distribution, and traffic and port classification have been used to achieve real-time traffic analysis.

Although the results are promising , noticed that the flows may change their behavior during a transmission; we will investigate in a future work the impact of the sampling and sub-flow utilization on classification accuracy, in such cases.

Acknowledgment

The authors thank NSERC and Ciena for funding the project CRDPJ 461084. This research also receives support from the Canada Research Chair, Tier 1, hold by Mohamed Cheriet.

REFERENCES

[1] V. T. Guimarães, C. M. D. S. Freitas, R. Sadre, L. M. R. Tarouco and L. Z. Granville, "A Survey on Information Visualization for Network and Service Management," in IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 285-323, Firstquarter 2016.

[2] Ball, R., G.A. Fink, and C. North. Home-centric visualization of network traffic for security administration. in Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security. 2004. ACM.

[3] Goodall, J.R., et al. Preserving the big picture: Visual network traffic analysis with tnv. in Visualization for Computer Security, 2005.(VizSEC 05). IEEE Workshop on. 2005. IEEE.

[4] Lakkaraju, K., W. Yurcik, and A.J. Lee. NVisionIP: netflow visualizations of system state for security situational awareness. in Proceedings of the 2004 ACM workshop.

[5] Keim, D.A., et al. Monitoring network traffic with radial traffic analyzer. in Visual Analytics Science And Technology, 2006 IEEE Symposium On. 2006. IEEE.

[6] Abdullah, K., et al. Visualizing network data for intrusion detection. in Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC. 2005. IEEE.

[7] E. M. Salvador and L. Z. Granville, "Using visualization techniques for SNMP traffic analyses," Computers and Communications, 2008. ISCC 2008. IEEE Symposium on, Marrakech, 2008, pp. 806-811.

[8] Schneider, P., *Tcp/ip traffic classification based on port numbers.* Division Of Applied Sciences, Cambridge, MA, 1996. **2138**.

[9] Nguyen, T.T. and G. Armitage, *A survey of techniques for internet traffic classification using machine learning.* Communications Surveys & Tutorials, IEEE, 2008. **10**(4): p. 56-76.

[10] Sen, S., O. Spatscheck, and D. Wang. Accurate, scalable in-network identification of p2p traffic using application signatures. in Proceedings of the 13th international conference on World Wide Web. 2004. ACM.

[11] Gu, R., H. Wang, and Y. Ji. Early traffic identification using Bayesian networks. in Network Infrastructure and Digital Content, 2010 2nd IEEE International Conference on. 2010. IEEE.

[12] Gu, C., S. Zhang, and H. Huang, *Online internet traffic classification based on proximal SVM.* Journal of Computational Information Systems, 2011. **7**(6): p. 2078-2086.

[13] Iliofotou, M., et al. Graph-based p2p traffic classification at the internet backbone. in INFOCOM Workshops 2009, IEEE. 2009. IEEE.

[14] Phaal, P., S. Panchen, and N. McKee, InMon corporation's sFlow: A method for monitoring traffic in switched and routed networks. 2001.

[15] http://www.sflow.org/packetSamplingBasics/

[16] Rocha, Guilherme V., and Bin Yu. "Greedy and relaxed approximations to model selection: a simulation study." Festschrift in honor of Jorma Rissannen on the occasion of his 75th birthday(2008): 63-80.

[17] Nguyen, T. T., & Armitage, G. (2006, November). Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks. InLocal Computer Networks, Proceedings 2006 31st IEEE Conference on (pp. 369-376). IEEE.

[18] Barros, Rodrigo C., André CPLF de Carvalho, and Alex Alves Freitas.Automatic design of decision-tree induction algorithms. Springer, 2015.

[19] Aigner, Wolfgang, et al. Visualization of time-oriented data. Springer Science & Business Media, 2011.

[20] http://www.inmon.com/products/sFlow-RT.php.

[21] DUPAY, A., SENGUPTA, Soumitrn, WOLFSON, Ouri, et al. NETMATE: A network management environment. Network, IEEE, 1991, vol. 5, no 2, p. 35-40.

[22] http://www.cs.umd.edu/hcil/piccolo/

[23] http://netweb.ing.unibs.it/~ntw/tools/traces/.