LABORATORY MANUAL

SC2008/CZ3006/CE3005 Computer Network

No.4: Analyzing Network traffic log data using python

ANALYZING NETWORK DATA LOG

1. OBJECTIVE

To understand and analyze network data log file

2. LABORATORY

For the lab location, please check
https://wish.wis.ntu.edu.sg/webexe/owa/aus_schedule.main

3. EQUIPMENT

PC and access to the Internet.
Access is provided at Hardware Lab/Software Lab at SCSE (Please see the schedule for more information)

4. DURATION

2 hours.

5. INTRODUCTION TO NETWORK TRAFFIC LOG

Network is an essential part of the infrastructure. Thus, it is essential to ensure that it is functioning efficient and effectively. To ensure its operation effectiveness, network devices status need to be continuously monitored. A number of protocol has been standardize and widely deployed by network operators. Simple Network Protocol (SNMP)[xxx] is widely used to monitor the status of network devices. It can provide live view of the traffic utilization, as well as status(UP/DOWN) of switch/router ports, etc.. Vendor has their own SNMP visualization tools, which are expandable to read other vendor SNMP enabled devices. A popular open source tool that is used to plot the traffic on individual port is MRTG, as shown in Figure 1.

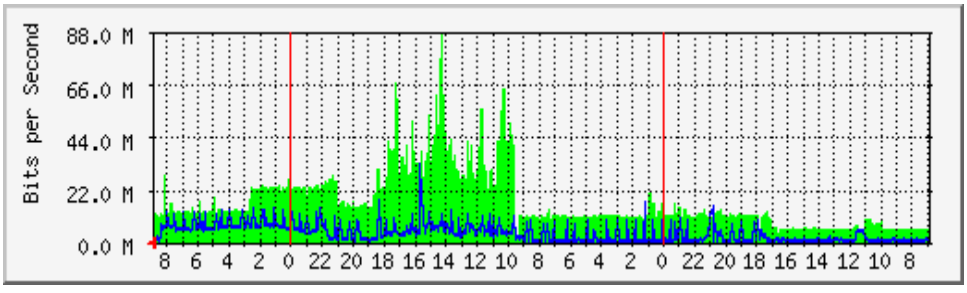


Figure 1: MRTG plot

However, it does not provide the packet details. Thus, NetFlow and SFlows were developed so that packet information can be archieved and analysed. Network operators reports would contained information about packets flowing through the network devices, eg. TOP 10 Talker(traffic generator), TOP 10 application, etc. The flow analysis is used for network planning purposes and network cybersecurity. A large number of tools have been developed to analised the NetFlow and SFlow data., eg. SolarWinds[https://www.solarwinds.com/network-management-software]. Solarwinds provides a large number of ways to view the data, eg. Most popular, Most Talkers, communication pair, % of IP application, IP protocol, bandwidth, etc…. An online

website that shows the traffic flows across continents among research and education entities is NetSage (https://portal.netsage.global/grafana/d/000000003/bandwidth-dashboard?refresh=1d&orgId=2)

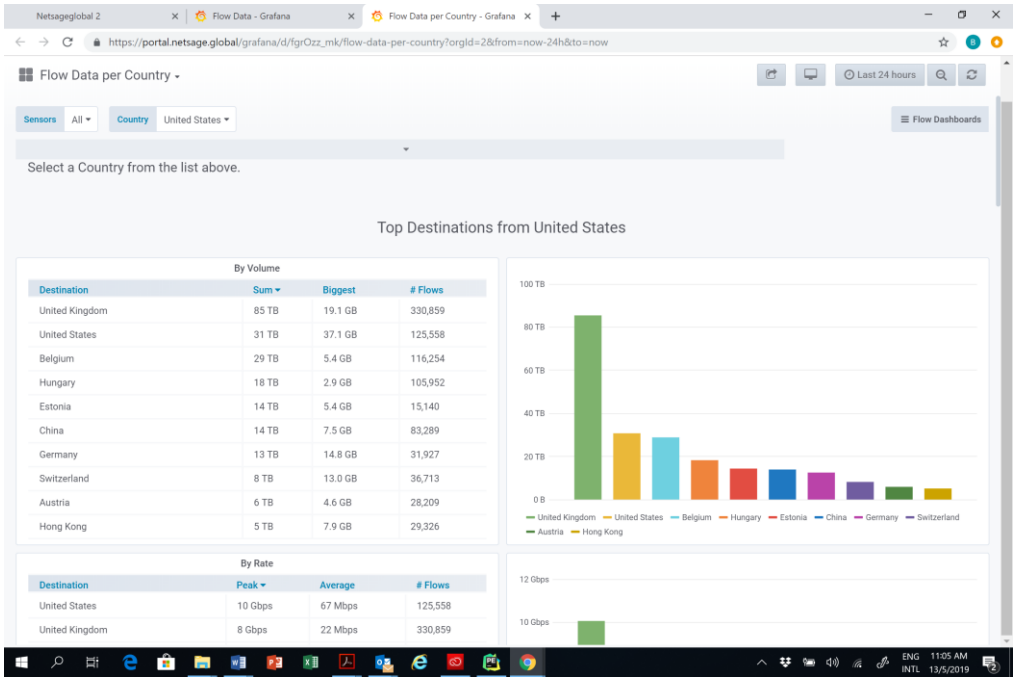


Figure 2; NetSage view for data traffic

5.1 Understanding traffic data collection using Netflow and SFlow

Netflow and SFlow are created to monitor and collect IP traffic. They capture packet informations, eg. Source-destination information which are found in the packet heads. By monitoring a flow, defined as a sequence of packets with the same source and destination, we are able to identify causes of congestion and detect any form of abnormal traffic.

Figure 3 shows and overview of the SFlow deployment, likewise similar for NetFlow. The Netflow/SFlow agent software need to be first installed on routers and switches. Whenever packets enter the router, an entry is made in the Flow Cache in the router consisting of the IP packet header information and the incoming port and outgoing port. The packet is then routed out of the destination interface of the router. The Flow cache table information is then exported out to the Collector at specified regular interval where it is analyzed.

There is a slight difference between NetFlow and SFlow, which is basically NetFlow captures all the packets while SFlow only captures a sample of the data. Based on a chosen sampling rate, an average of 1 out of n packets is randomly captured and sent to the collector to be analyzed. Although this method does not reflect a 100% accurate result, it has been shown to be sufficiently accurate for overall analysis. Furthermore, it requires less compute, storage and network requirement. Typical sampling rate could be as high as 1 in 2048 packets.

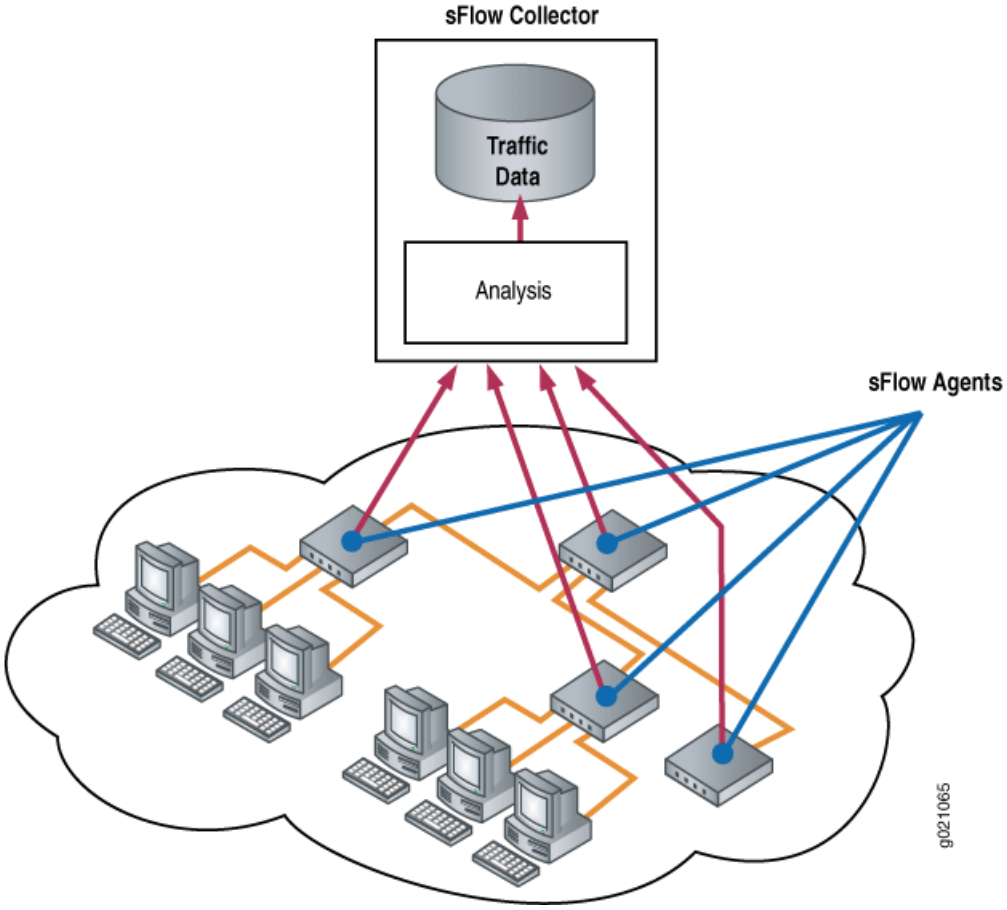


Figure 3: SFlow Architecture Diagram
[https://www.juniper.net/documentation/en_US/junos/topics/example/sflow-configuring-ex-series.html]

5.2 Traffic log data

The SFlow traffic data is obtained from a Network exchange point router. It has peering with multiple sites both local and international. The router support 802.1Q thus, supporting VLAN tagging, creating VLANs for special peering between different ASs(organization).

The format of the traffic log data is shown in Table 1. It basically captures the traffic packet header information of each packet going through the router.

	Field	Description	Example
0	Type	FLOW	FLOW
1	sflow_agent_address	IP address of the agent	aa.aa.aa.aa
2	inputPort	Router/switch port number receiving the packet	137
3	outputPort	Router/switch port number through which the packet is send out	200
4	src_MAC	MAC address of the transmitting host	d404ff55fd4d
5	dst_MAC	MAC address of the receiving host	80711fc76001

6	ethernet_type	802.3/Ethernet	0x0800 refer to Ethernet packet
7	in_vlan	VLAN on which the packet is received	919
8	out_vlan	VLAN on which the packet is send out.	32
9	src_IP	IP address of the sundering host of the packet	130.246.176.22
10	dst_IP	IP address of the receiver host of the packet	140.115.32.81
11	IP_protocol	IP protocol type	TCP = 6 UDP = 17
12	ip_tos	Type of service	0
13	ip_ttl	Value of the Time to Live attribute of the packet.	
14	udp_src_port/tcp_src_port/icmp_type	Source port address at the transport level	
15	udp_dst_port/tcp_dst_port/icmp_code	Destination port address which define the application service requested. https://www.webopedia.com/quick_ref/portnumbers.asp	http = 80
16	tcp_flags	TCP flag attribute specifying type of packet, SYN, etc..	
17	packet_size	Packet size including MAC headers	
18	IP_size	IP packet size	
19	sampling_rate	Sampling rate of the packets collected.	2048

Table 1: SFlow format (https://github.com/sflow/sflowtool#line-by-line-csv-output)

6. TASKS

The objective of this laboratory session is to have a first hand experience in doing basic analysis of data log. The student is required to write simple python codes to decipher the network traffic data captured and stored as Microsoft Excel file (.csv), download from the NTULearn. The program shown be able to generate the following information.

Top 5 Talkers. (ie sender nodes)
Top 5 Listeners (ie receiving node)
Top 5 applications
Total traffic
Proportion of TCP and UDP packets
Top 5 communication pair
Visualizing the communication between different IP hosts.

Students are required to do the following at the end of the laboratory session.

- Submit the answer template to the NTULearn at the end of the laboratory session along with the code in appendix.

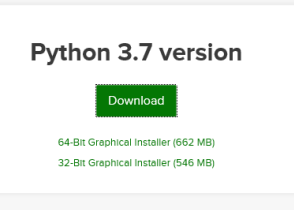
- Demonstrate their software to the Teaching assistance/Lecturer. (Optional)

7. Conclusions

The laboratory session , should allow the students to understand some of the basic analysis of network traffic log file. Students are to submit the answer template together with their source code.

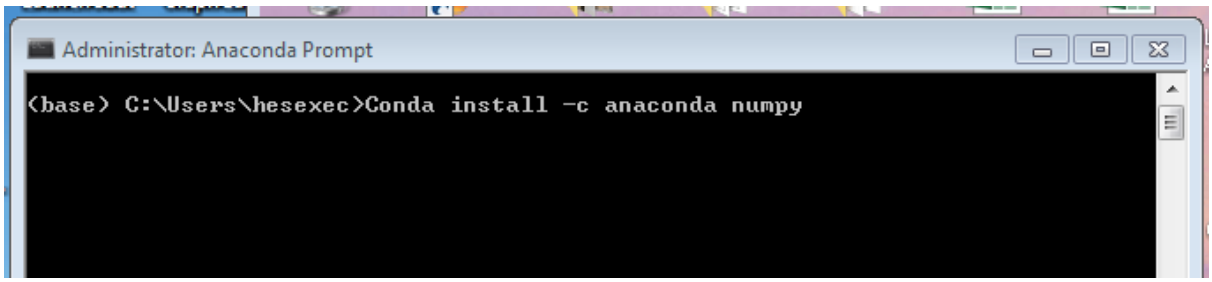
Appendix 1: Instructions to install Environment for Python for CE3005/CZ3006

1) Install Anaconda Environment for Python (Python 3.7 or above)
https://www.anaconda.com/download/



2) Install the library you need using conda install
   a. https://docs.anaconda.com/anaconda-cloud/user-guide/howto#use-packages
      e.g https://anaconda.org/anaconda/numpy

   Goto Start All Program-> Anaconda3 (64bit)-> Anaconda Prompt-> right click run as Administrator



   1) Conda –c anaconda numpy

3) Install PyCharm https://www.jetbrains.com/pycharm/

   Install the pycharm-community (latest one)

Some Help:
1) How to organize python code into packages:
   a. http://intermediate-and-advanced-software-carpentry.readthedocs.io/en/latest/structuring-python.html