A Study of Recent Trends in Network Traffic Analysis on Large Scale

Hasmukh B. Domadiya
Assistant Professor, National Computer College, Jamnagar, Gujarat, India

Dr. Girish C. Bhimani
Head of Department, Department of Statistics, Saurashtra University, Rajkot, Gujarat, India

Abstract – A computer network is a set of devices connected with each other for the purpose of communication. For the global Internet to the local Intranet, from a satellite network to a sensor network, every communication needs to go through a scrutiny to analyze and ensure security of the system for which a particular network is being used. This paper discusses some of the well known ways of doing such network analysis and their limitations. With the growing increase of digitalization, the usage of the computers and subsequently the computer networks is increased drastically. The same is true for the usage of the Internet too. This paper puts insight into some of the recent and advanced ways of network analysis.

Index Terms – Network, Traffic, Analysis, Snifters, Firewall.

1. INTRODUCTION

A computer network is a collection of devices which are connected with each other. These devices are of two types, end devices and intermediate devices. End devices are used by the users to access the network. Intermediate devices are used by the network itself to make communication possible. With the growing usage of the digitalization, more and more people have started using computers for their business, education and daily routines. The Internet is one of the widest and fastest growing world wide computer networks.

Every network needs to perform many communications to enable its users to provide desired services like file transfers, mail service, remote access, resource sharing etc. At the same time various ISPs need to perform many activities to keep track of utilization. The administrator of the network has to look into what kind of communications is being done by their users. Bandwidth utilization related analysis is required to calculate billing information for various users, to limit speed after a certain amount of data (in MBs, in GBs) have done, restrict usage if bills are due even after due date. Suspicious activities related analysis is required to detect pirated downloads, prohibited content and sites visit, malicious activities to ensure computer, network and information security.

This paper explains various existing approaches which are used by network administrators to analyse networks. Section 2 explains some of the well known tools and section 3 explains recent and future trends to perform analysis of network. Section 4 concludes with comparison and section 5 shows future directions from the research perspectives.

2. NETWORK ANALYSIS TOOLS

This section discusses some of the well known tools to perform network analysis by network administrators.

2.1 Firewall

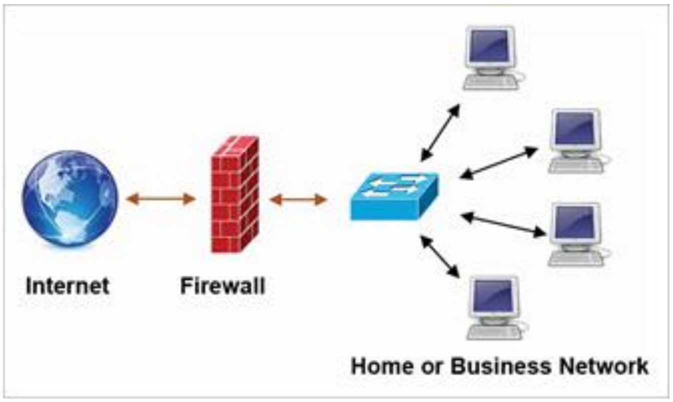


Figure 1. – Firewall

Figure 2. – Cyberoam administration

A firewall is used to perform user management, traffic management and policy management for a network. Mostly a firewall is placed to connect the Internet with the network of an organization. Firewall makes sure that only genuine and authentic traffic flows between the Internet and the users accessing the Internet from the network. One such scenario is shown in Figure 1. Figure 2 shows a screen of cyberoam firewall configuration [1].

Firewall can set various policies to perform limited access facility, limited bandwidth facility to the users. It also supports antivirus facility so that every activity can be scanned before permitted. Various companies like cyberoam, fortinet provide hardware units as a part of firewall product. Every firewall has a built in operating system to perform its activities too. So we can say that a firewall is a combination of hardware and software. The list of facilities provided by a firewall is given below [1].

- User and Group Management
- Bandwidth Management
- Usage and Traffic Policies Management
- Antivirus, Filtering, Intrusion Detection and other Security Management
- IP Addressing Management
- Log Management

2.2 Sniffers

A packet sniffer is software which can be used to analyse various packets. Here analysis is from higher level to the lower level. At application layer to the mac layer, every packet can be analysed in detail. The purpose of a packet sniffer is to look into the header and related information of various packets, handled at various levels of network based communication. A sniffer supports finding particular piece of information based on specific protocol too. A sniffer can be used to analyse activities of a single computer – that is called user level analysis using a user sniffer and at the network level – that is called network level analysis using a network sniffer. One such scenario is shown in Figure 3. Figure 4 shows a screenshot of wireshark packet sniffer which is capturing packets [2].

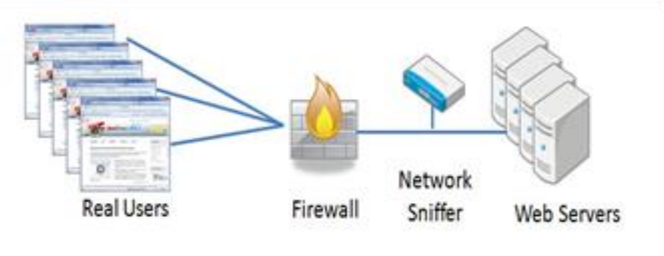
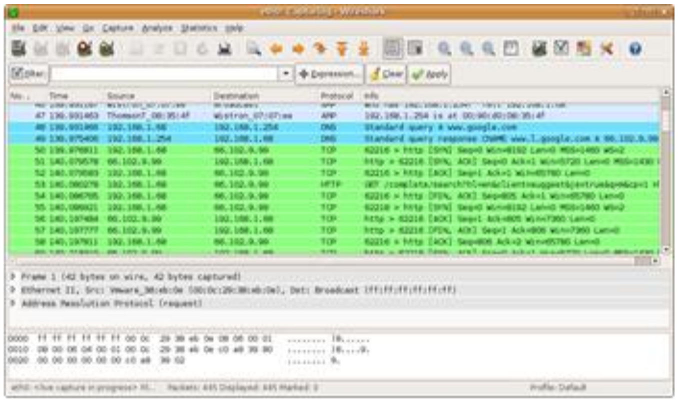


Figure 3 – Sniffer

The list of facilities provided by a packet sniffer is given below [2].

- Application Layer Protocol based Analysis – HTTP, DNS, DHCP, FTP etc.
- TCP/IP based Analysis
- Header Analysis of various protocols.
- Graphical analysis of various communications.



```
Frame 1 (42 bytes on wire, 42 bytes captured)
Ethernet II, Src: Vmware_38:eb:0e (00:0c:29:38:eb:0e), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
Address Resolution Protocol (request)
```

Figure 4 – Wireshark

3. RECENT TRENDS

This section discusses some of the well known, recent and very advance tools and techniques to perform network analysis. The purpose of introduction of such tools is to analyse the network more precisely and accurately. These techniques were not present in traditional firewalls or packet sniffers.

3.1 Firewall and related Devices

With the growing usage of the Internet and with the growing scaling of companies, to perform network analysis across all braches – geographically separated locations, a single firewall can not be used. Several companies have introduces aligned devices which are used along with the firewall to perform network analysis at very large scale. One such scenario is shown in Figure 5 [3][4].

Fortinet is a company which manufactures and design various devices related with network, Internet and Information security. The devices are named with Forti followed by a word suggesting the purpose of using them. Here is a list of devices and their primary usage [4].

- FortiGate : Firewall
- FortiOS : Operating System for Firewall
- FortiAnalyzer : A centralized and separate system to store log and perform analysis.
- FortiManager : A centralized and separate system to manage a group of firewalls.

- FortiCloud :A cloud based services to manage log and perform analysis.

Figure 5 shows a scenario where a company has a head quarter (main office) and two branches. All these three offices are separated from each other geographically and connected with each other through the Internet. At stage 1, every office is equipped with a firewall (Here it is FortiGate). These firewalls are responsible to manage traffic at office level. At the same time, to apply global policies across all the three offices and subsequently all the users of all the three offices, headquarter has a FortiManager which acts as a centralized device for policy management for all the three firewalls. Every firewall has a limited capacity of log management, to extend the log management facility, the headquarter has a FortiAnalyzer which acts a centralized device for logging and reporting. As we can see that FortiManager controls FortiGates as well as FortiAnalyzer. The FortiAnalyzer collects logs from all the firewalls either through local network or through the Internet [4].

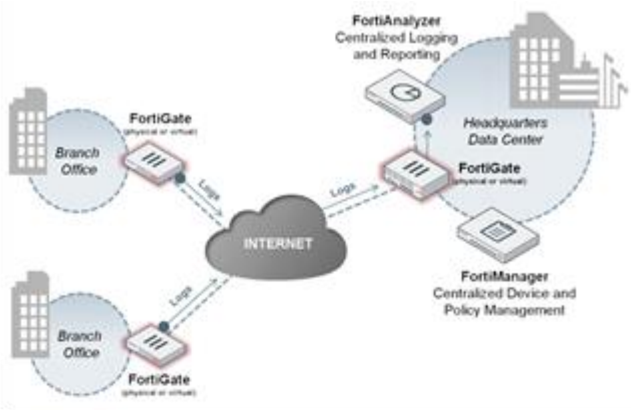

FortiAnalyzer
Centralized Logging and Reporting
Headquarters Data Center
FortiGate (physical or virtual)
Branch Office
FortiGate (physical or virtual)
Logs
Logs
INTERNET
Logs
Branch Office
FortiGate (physical or virtual)
FortiManager
Centralized Device and Policy Management

Figure 5 – Fortinet Devices

3.2 Hadoop Based Solutions

As the Internet usage is drastically increasing day by day, corresponding traffic logs are also requiring more and more space to be stored. Even a small scale organization has to maintain logs in GBs while a large organization in TBs. The volume required to store all logs increases more if we want logs to be maintain for longer period as every second, new logs are being added. The more challenging task here is to provide space for such a huge data and after that to provide computation resources to process such a huge data [5].

As discussed, a single centralized system needs to be having a large disk and large amount of processing units to handle large logs. A super computer could be of a solution with extreme parallelism and high speed computation. Neither there is a possibility that an organization affords firewall based other devices nor has a super computer but still needs to maintain and analyze large number of logs at a reasonable cost. The solution which has gained immense popularity is the usage of Hadoop. A Hadoop based solution could be considered as a distributed solution having a set of sub solutions. Each of the sub solutions can be performed at remote devices – with sufficient but not extra ordinary, architecture, organization or computational resources. Later on the results of all these sub solutions could be combined to find the final solution. At the same time Hadoop based file system provides easy way to access extremely large files as compared to conventional file systems like FAT and NTFS. Such scenario is called a multi node Hadoop cluster shown in Figure 6 [6][7].

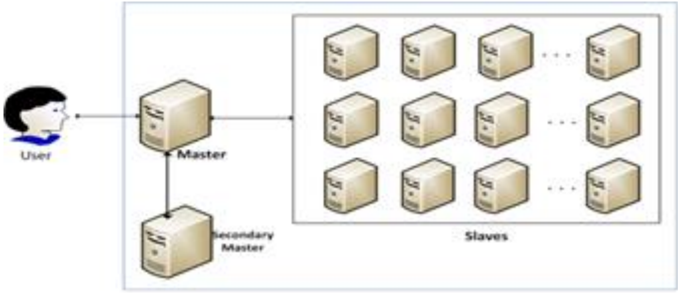


Figure 6 – Hadoop Cluster

Figure 6 shows that a user incapable of having a single system with strong resources (large memory, disk, extreme parallelism) can setup a Hadoop based cluster. A few master nodes coordinate tasks and collect results from the set of slave nodes. Slave nodes perform data analysis.

Hadoop based solution can be achieved using the concept of map-reduce. The functionalities could be achieved with implementation of Mapper and Reduce interfaces. A Map transforms input records into intermediate records. A record is in the form of key/value pair. Reduce is used to reduce a set of intermediate records to a small set of output records. Figure 7 shows how a Mapper works. Figure 8 shows how a Reduce works.

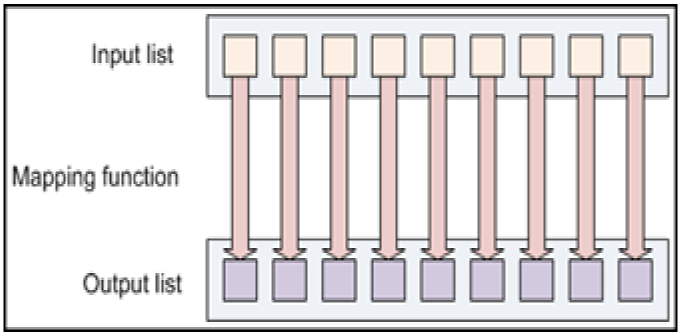


Figure 7 – Mapper

One such example to count words using map-reduce is shown in Figure 9. The same functionality could be used by while processing network traffic for pattern recognition.

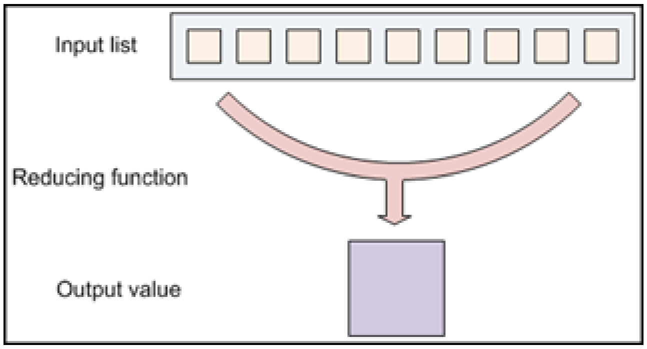


Figure 8 – Reduce

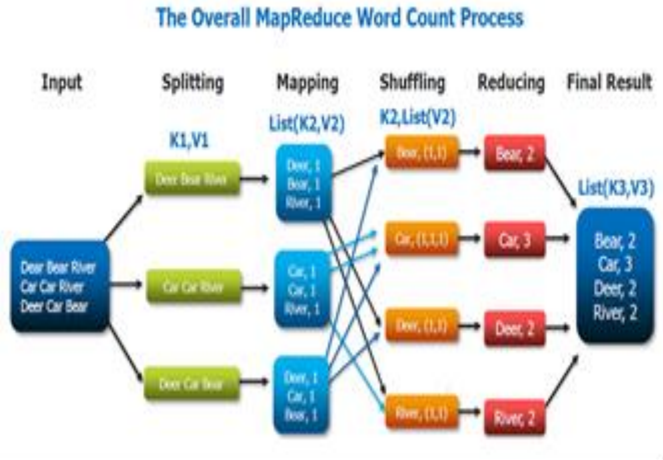


Figure 9 – MapReduce

In this example, we can see that the actual data is input to a set of nodes involved in analysis. Here it is 3.

Every node gets a part of the actual data to process. Initially there is no value and so each of the input data field is considered as a key. Later on value will refer to the number of occurrences of a specific data field. This process is called splitting of input data among the nodes. Mapping is a process of calculating intermediate records. Here an intermediate record means the word count of on a specific node. A set of values refer to the counting at every node in a small subset of dataset assigned to it. Shuffling is a process of distributed partial intermediate records with other nodes so that final computation could be more easier and less resource consuming. After sharing of similar intermediate records with each other, reduce operation performs the final calculation which will be merged at a single node to produce a set of output records.

4. CONCLUSION

This work discusses the existing approaches which are popular for network traffic analysis and what are the recent and advance trends towards improvement of it. We have discussed the purpose of network traffic analysis. Two most popular approaches firewall and sniffers are discussed along with purposes of each of them. From home users to small organization, firewall and sniffers are applicable. Many manufacturers provided firewalls based on user need from low cost to high cost. But these conventional methods are not suitable always in every case. Those large organizations who want to maintain and analyse network beyond the way a single firewall does have to select advance – expensive methods. Two of the most widely used methods are explained. One is based on developing a logical network of security devices using firewall and related devices. One even recent approach is using Hadoop based distributed structure to process extreme large volume of data at reasonable cost.

5. FUTURE WORK

With the growing usage of the Internet, it is indeed necessary to perform network analysis to identify threats and vulnerabilities. Apart from the techniques discussed in this study, more research work could be done towards improving the performance. Hadoop based solution is still an open problem. A few researchers [8][9] have shown some solutions but it is yet to be analyzing in-depth and yet to be implemented world wide at large scale level. More efficient Hadoop based systems could be designed. An interesting way is to use cloud based resource access with the Hadoop. So that the cost can be controlled too.

REFERENCES

[1] Law K. SE 4C03 Winter 2005 An Introduction of Firewall Architectures and Functions. 2005.

[2] Wireshark, http://www.wireshark.org.

[3] CAIDA CoralReef Software Suite ,http://www.caida.org/tools/measurement/coralreef.

[4] Fortinet CookBook

[5] Hadoop, http://hadoop.apache.org/.

[6] Shvachko, Konstantin, et al. "The hadoop distributed file system." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. IEEE, 2010

[7] J. Dean and S. Ghemawat, MapReduce: Simplified Data Processing on Large Cluster, OSDI, 2004

[8] Lee, Yeonhee, and Youngseok Lee. "Toward scalable internet traffic measurement and analysis with hadoop." ACM SIGCOMM Computer Communication Review 43.1 (2013): 5-13.

[9] Lee, Youngseok, Wonchul Kang, and Hyeongu Son. "An internet traffic analysis method with mapreduce." Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP. IEEE, 2010.