International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-5)

Research Article | **May 2015**

Hadoop Based Network Traffic Measurement with Scalable Data Processing In Fine Grained Networks

D. Venkatesh
Associate Professor in CSE
GATTES Institute of Tech,
Gooty, Anantapur(D) Inida

M. Prasanna
Department of CSE
GATES Institute of Tech,
Gooty, Anantapur(D) India

N. Madhava Reddy
M.Tech Student, 139B1D5809
S.C.V.R.I.T.S, Tadipatri
Anantapur(D) India

Abstract:

Perfect traffic measurement and monitoring is important in a broad range of network applications such as security attacks, discovery of anomalies and traffic engineering. A numeral important network management decisions like jamming traffic to a victim destination, detection of anomalies, or re-routing of traffic need mining and investigation of real-time spatiotemporal patterns in network traffic. A high class network measurement tool is the key for mining such patterns of significance and making learned decisions to make certain proper network operation. The main aim of this paper is to utilize the Hadoop-based traffic observing system which performs HTTP, IP, TCP and NetFlow analysis on enormous Internet traffic in a scalable manner. Hadoop based traffic observing provides scalable data processing with huge storage services based on a distributed computing system which can be best utilized for traffic measurements.

Keywords: Classification algorithms, Computer network management, intrusion detection, Multi Resolution, Traffic

I. INTRODUCTION

Network analysis is the procedure for taking network targeted visitors along with examining the idea closely to find out what exactly is transpiring on the network". Two Overseeing Approaches are mentioned in the using parts: Router Centered along with Non-Router Centered. Neither overseeing functionality which have been built-into your routers Independently, nor need added installing electronics or software package is termed as Router Centered approaches. Non-Router based approaches need added electronics along with software package to be mounted and gives larger flexibility. Equally approaches are additional mentioned in the following paragraphs.

1. 1 Router Centered Overseeing Approaches:

Router Centered Overseeing Approaches are hard-coded into your routers and so deliver minor flexibility. A quick justification of the extremely commonly used keeping track of approaches is offered underneath. Just about every method possesses underwent several years connected with improvement to become a consistent product.

1. 1. 1 Simple Circle Overseeing Method (SNMP):

SNMP [Cisco5606] is surely an application covering method which is section of the TCP/IP method selection. That will allow Managers to control network performance, come across along with solve network issues, along with policy for network increase. That collects targeted visitors data by way of passive receptors which have been implemented via router to end sponsor. Whilst a couple types really exist, SNMPv1 along with SNMPv2, this specific section refers to SNMPv1. SNMPv2 develops upon SNMPv1 and will be offering improvements, for example added method procedures.

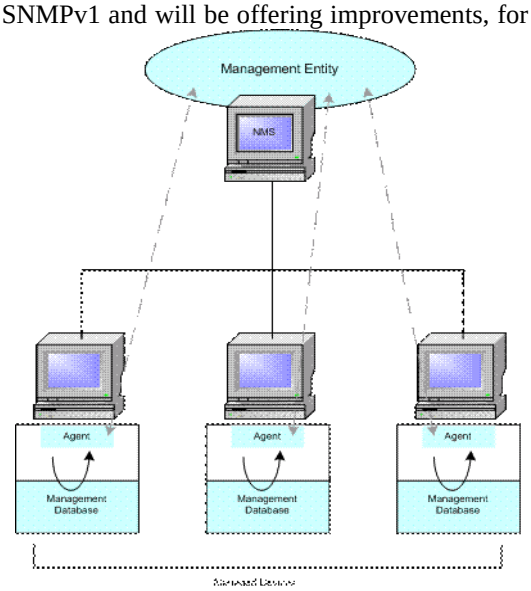

Figure 1: SNMP Components

Standardization connected with another version connected with SNMP. SNMP Variation 3 : (SNMPv3) is impending. You will find 3 important parts in order to SNMP: Handled Gadgets, Agencies, along with Circle Operations Methods (NMSs). They're proven inside Determine 1 underneath.
The Managed Devices contain the SNMP Adviser and may include routers, switches, hubs, computers, units, in addition to objects like most of these. That they have the effect of amassing facts in addition to rendering it open to your NMSs. The Providers incorporate application in which have knowledge of operations facts in addition to results this data right into a form works with SNMP. They may be situated on some sort of been able system.
The NMSs implement programs in which keep an eye on in addition to manage your been able units. Control in addition to memory space assets which have been required for circle operations are provided with the NMSs. A minimum of one NMS must occur upon any been able circle. SNMP could act entirely as a NMS as well as a realtor, as well as is capable of doing your obligations regarding each. You will discover 4 basic commands utilised by SNMP NMS to be able to keep an eye on in addition to manage your been able units: go through, create, lure, in addition to traversal functions. The go through command has a look at your parameters which have been kept with the been able units. The create command changes your beliefs from the parameters stashed with the been able units. Traversal functions search to find out just what parameters been able units facilitates in addition to collects facts from your supported varying furniture. The lure command is used with the been able units to be able to survey your happening regarding a number of functions on the NMS.
SNMP uses 4 method functions in order to function: Receive, GetNext, Collection, in addition to Capture. The Receive command is used in the event the NMS problems some sort of request facts to be able to been able units. The SNMPv1 information (request) that's delivered consists of a information header and a Protocol Data Unit (PDU). The PDU from the information provides the facts in which is necessary to properly comprehensive some sort of request that can either retrieve facts from your broker as well as established some sort of benefit from the broker. The been able units use the SNMP agents situated on these phones retrieve your desired facts, and interact to your NMS with an reply to your request. If your broker does not have any facts regarding the request, very easy return anything. The GetNext command will likely then retrieve on-line from the future object example. It's also easy for your NMS to be able to deliver some sort of request (Set operation) in which models your beliefs regarding objects from the agents. As soon as a realtor would need to enlighten your NMS of an event, it will use the Capture operations.

As discussed, SNMP is usually an Software Coating method in which uses passive detectors to aid directors keep an eye on circle targeted traffic in addition to performance. Despite the fact that SNMP could be a useful device to be able to System Managers it will develop a susceptability to be able to safety threats since it is lacking in any authentication capabilities. It truly is as opposed to Remote Overseeing (RMON) that's discussed inside the subsequent area in this RMON monitors with the System Coating in addition to below, instead of with the Software Coating.

1. 2. only two Remote Overseeing (RMON):

RMON [Cisco5506] enables numerous circle monitors in addition to console techniques to change network-monitoring facts. It truly is a file format from the SNMP Administration Information Data bank (MIB). As opposed to SNMP that has got to transmit some sort of request facts, RMON can established sensors that can keep an eye on your circle based on a number of criteria. RMON permits Managers to manage local communities in addition to out of the way web-sites from one middle area. The idea monitors with the System Coating in addition to below. RMON has only two variations RMON in addition to RMON2 this particular cardstock merely relates to RMON. RMON2 makes for keeping track of regarding packets upon most circle tiers. The idea focuses on IP targeted traffic in addition to app degree targeted traffic.

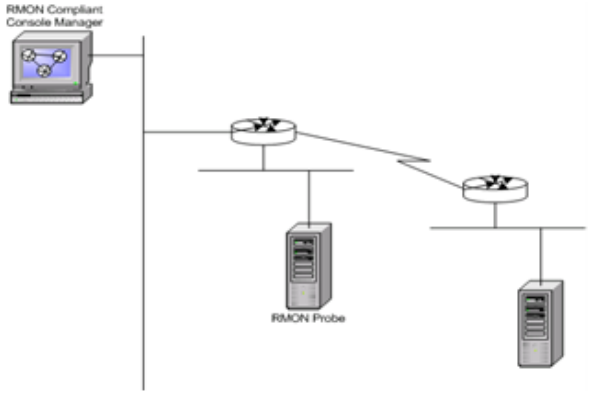

Figure 2: RMON Components [Cisco5506]

Even though you'll find 3 key ingredients on the SNMP keeping track of surroundings you'll find merely only two inside the RMON surroundings. They may be demonstrated in Physique only two below.
The two the different parts of RMON include the probe often known as the actual real estate agent as well as check, as well as the purchaser often known as the actual managing stop. Not necessarily not like SNMP the actual RMON probe as well as real estate agent collects and also stores the actual system info. Your probe will be set software program

around the system equipment, such as routers and also turns. Your probe can also run using a personal computer. Your probes have to be put on every various LAN as well as WAN section because they only are able to discover targeted traffic that passes as a result of only their particular hyperlink, and also don't know outside the house inbound links. The client is a managing stop that communicates with all the probe making use of SNMP to get and also correlate the actual RMON Files.

RMON [RMON] uses 9 various checking organizations to get details about the actual system.

- Statistics : numbers calculated with the probe for each checked screen with this unit
- History : data intermittent record trials coming from a system and also shop regarding access
- Alarm : routinely takes fact trials and also even comes close them using a collection of thresholds regarding celebration era
- Host : consists of data linked to every coordinator identified around the system
- HostTopN : prepares platforms that summarize prime serves
- Filters : allow packets to be matched by way of filter equation regarding acquiring occasions
- Package catch : catches packets after they circulation with the funnel
- Events : controls era and also notification involving occasions coming from a unit
- Token band : supports token band

Because stated preceding RMON, develops on the actual SNMP method. While targeted traffic checking can be carried out with your methods, evaluation in the info furnished by SNMP and also RMON has a small further do the job.

II. RELATED WORK

Traditional measurement schemes typically maintain unique "per-flow"-based statistics. The collected information is post-processed offline for answering higher-level user queries [6] such as detecting an anomalous behavior. The per-flow schemes, however, require storing information about potentially huge number of flows, straining the limited SRAM budgets of measurement hardware. The scalability issues of the per-flow scheme have traditionally been addressed using packet- or flow-based sampling approaches [7]. Studies have shown, however, that sampling leads to inaccuracies in answering the user queries [1]. Recently, smart sampling approaches, such as cSamp [8] and FlexSample [2], are proposed to balance the monitoring goals with resource constraints through smarter provisioning of resources based on application requirements. However, these schemes require measurement goals to be defined a priori, which can be challenging with highly dynamic network or traffic conditions.

More recently, iterative measurement schemes were proposed to address the challenges mentioned above [3], [4]. The key idea is to perform top-down and goal-oriented measurements that directly reflect the requirements of high-level user queries by taking in account dynamic traffic variations. In this context, the MRT algorithm [5] was proposed to answer the user query in an iterative manner through a progression of finite set of inter mediate measurements, also referred to as rules. A rule can be viewed as an intermediate question in pursuit of the user query that, if answered, can help lead the search in a more intelligent manner.

MRT^2 [5] is a recursive top-down heuristic that relies on a simple but powerful observation that if a flowset does not contain an anomaly, then no flow in that flowset can be anomalous. For instance, in the case of elephant flows, if a flowset does not consume fraction of the entire network bandwidth, then no flows within that flowset may be an elephant flow. In terms of CIDR notation, the algorithm states that if a prefix is not an elephant, then all its constituent prefixes of larger number of bits (finer granularities) can be discarded from further consideration. MRT iteration for two-dimensional tuple space {Source, Destination} involving a zoom ratio (ZR), or expansion ratio, of four is illustrated in Fig. 3. The ZR defines the rate of exploration within a subregion. For instance, the ZR of four in the figure dictates that a given tuple space is partitioned into four subregions at a time, implying exploring additional 2 bits in the tuple space per iteration. Data are generally obtained intended for particular person subregions for a offered measurement time period. It is accomplished utilizing policies that will partition the actual site visitors in to the a number of subregions.

Thus, a rule in the context of the MRT can be viewed as a Boolean bit-mask on specific header bits that helps qualify the incoming packets. In the case of HH, statistics corresponding to the traffic mapped to the subregions will be collected. The subregions that exceed the threshold θ , marked with a cross in the figure, are then selected for further zooming in, or expansion, in the next iteration.

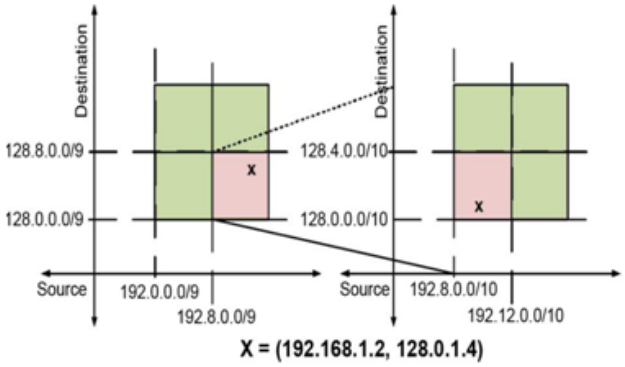

Figure 3: MRT with zoom ratio of four.

A zooming-in corresponding to a selected subregion means partitioning it further into subregions as per the ZR. Thus, each iteration in the given example results in resolution of 2 bits. MRT thereafter continues iterating between partitioning, statistics collection, and expansion phases until the anomalous flow is isolated or the all the header bits are resolved. Thus, in the given 2-tuple-space example, a flow will get isolated in a maximum of 64 iterations.
The MRT algorithm is too responsive to exact traffic patterns and is blind to constraints or opportunities which arise from measurement platforms. So, an advanced and easy way of network measurement algorithm is needed where every aspect of the network analysis like bandwidth and packet size can be measured and also adjusts the congestion problem. To do so this paper proposed a measurement tool which uses Hadoop network analysis tools to measure the bandwidth and the time delay to transfer any particular packet to destination.

II. HADOOP BASED NETWORK ANALYSIS TOOLS

This section presents the proposed Hadoop Network analysis that aim to address the challenges associated with application-aware rule-based online traffic measurements, with the goal of providing accurate response despite highly volatile traffic. As discussed earlier, a key design challenge is in maintaining computation and storage scalabilities while offering high accuracies and minimum latencies in answering the user query.

3.1 Overview:

The Key components for the Flow Analysis using Hadoop consist of three main layers which include Data Exchange layer, Analysis layer and User Interface layer [6]. Fig 1 shows the key components required for Flow Analysis. The functions of the above 3 layers are described below:

- ***Data Exchange layer*:** This layer implements HDFS (Hadoop Distributed File System) to store the information related to the Internet traffic. This layer is mainly concerned about the storage and it provides support to the other layers. In this layer preprocessing of the local file system is done. Here the network information and the traffic information are extracted from the packets which are got from the network.
- ***Analysis layer:*** This layer focuses of the internet traffic analysis and its management. In this layer multiple types of analysis are done. In this layer network analysis, node analysis, link analysis and flow analysis are done. Analysis layer also implements various algorithms needed for the flow analysis.
- ***User Interface layer:*** In this the user can interact with the system. The system will display graphical images to the user so the user can better understand the flow analysis. This layer implements few API and GUI tools for the better communication purpose.

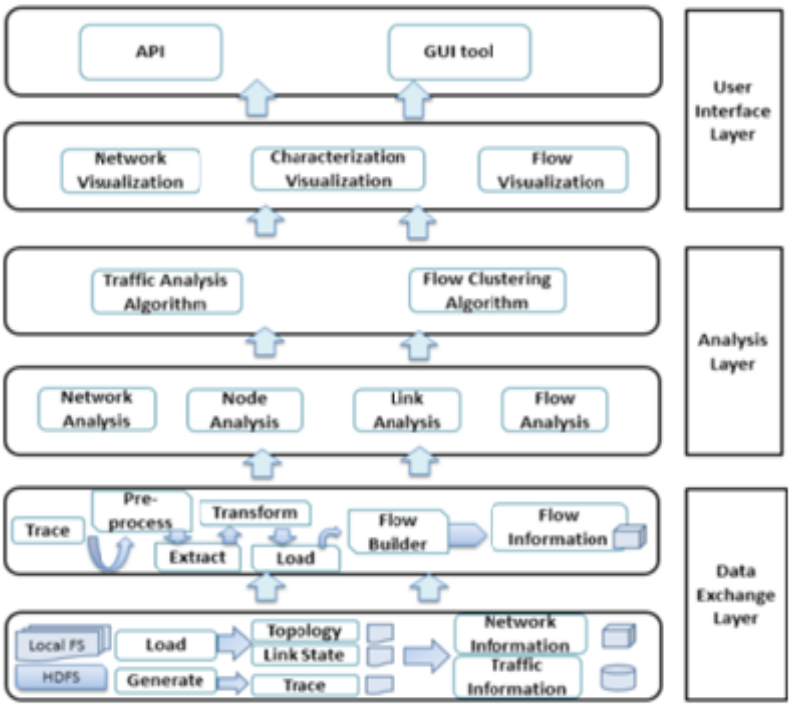

Figure 4: components for the Flow Analysis using Hadoop.

3.2 Flow Analysis with Hadoop :

Hadoop is a framework of tools which is used to address the challenges faced by Big data. Hadoop consist of Distributed File system and MapReduce engine. The Distributed File system divides the large data blocks into many smaller units and MapReduce engine will process and implement each and every data blocks independently.

In this system, the input is Bigdata which consist of huge amount of packets flowing from different network. The system will accept this large input of trace file from traffic measurement tool named Wireshark. This tool identifies the traffic flows running on the network. Once the input is stored in Hbase(Hadoop database) the next step is to analyse the input. Analysing the input is one of the difficult jobs. Analysis of the input is done based on the source IP address, source port address, destination IP address, destination port address, type of the packet and size of the packet. Fig 4 shows architecture of this system. To maintain the flow of the packet transfer this system uses Flow Momentum algorithm.

3.2.1 Stream energy:

The particular Flow Energy (FM) formula details the issue connected with MRT resets by means of using the average touch charge which is encountered more than a hierarchical way accomplishing a new flowset. This really is in contrast to an original MRT, in which the expansion/rejection decision can be just good flowset's recent size. The particular FM formula thus effectively provides the leaf nodes a new elegance time inside lively rule-set to face your temporary different versions inside anomaly. The particular stays on the elegance time are usually proportional towards high intensity, or even impetus, on the anomalous flows that advised your dimensions toward your leaf node to begin with. Thus, in the case of the heavy-hitter (HH), a leaf node may be more active if the anomalous flow has larger volume.

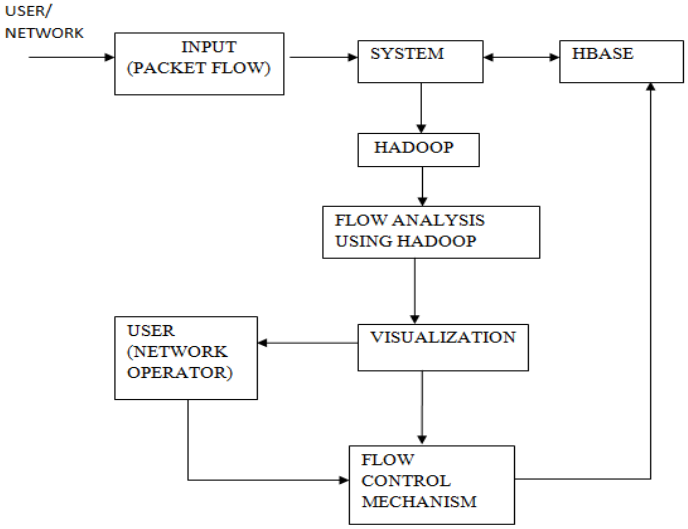

Figure 5: Architecture of the flow analysis system

In the figure above we can see the flow of the packets into the system. Initially the user or the network will give the input to the system. The input is the very large amount of packets flowing from different networks. All the packet information will be stored in the Hbase. Then the system will perform HDFS and MapReduce functions on these huge amount of packets. Once the packets are sorted accordingly various flow control mechanism on the packets are done. The user can view the flow of packets statically about the flow of packets.

For portability across different platform like Windows, Linux, Mac OS/X, FreeBSD, components are written in Java and only require commodity hardware. Initially we begin parsing the input given from the network. The input need to be in specific format only then parsing can be done. If the input is not in format as required then sorting of the input file is necessary. Once the input file is sorted the next step is parsing each and every input line. Parsing is done on the basis of source IP address, destination IP address, source destination port address, type of the packet. The input from the same source and destination port address will be got together and input from same source IP address and destination IP will be clustered and stored in the database as unstructured data.

MapReduce is a function which allows the programmers to write programs to parse a huge amount of unstructured data in parallel over distributed clusters of stand-alone computers. MapReduce function will take the unstructured input from the database and parse them. This function will calculate the sum of the bytes of the data from specific port address to specific port address or from specific IP address to specific IP address. Once this is parsing is done, the next step is visualization.

For portability across different platform like Windows, Linux, Mac OS/X, FreeBSD, components are written in Java and only require commodity hardware. The total process is done between two phases, one is in the

- client phase and
- Other at the side of receiver.

The below figure depicts the client side user interface with the following functionalities:

- Selecting the file to be transferred
- Splitting the file for transmitting through three routers A,B,C
- A key assignment paradigm for three routers.
- And the file transferring delay in milli seconds.

All the functions and the appropriate functionalities are defined on the client side.

Figure 6: Client Side User Interface

Coming to the receiver or the server side the below figure depicts the same as well.

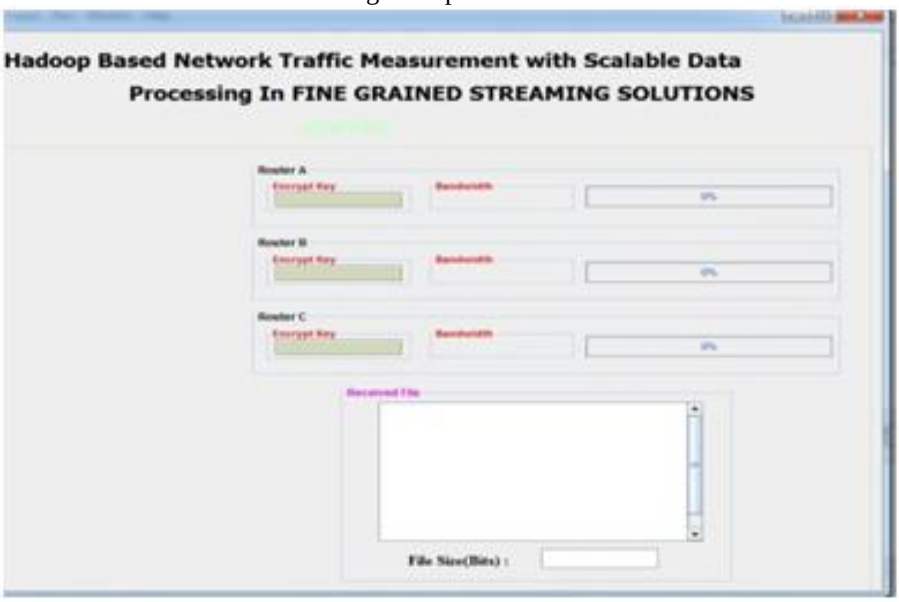

Figure 7: Server Side User Interface

The server side interface has pretty simple functionalities which includes the following:

- Router A, B, C Encrypted keys.
- Bandwidth from which the files are being transferred.
- The transferring progress.
- The received file and the size of the received file.
  The complete process is done by first starting from the client side
- The selected file is divided into appropriate number of small files and
- The keys are assigned to each router, note that if the keys are matched then only the transfer procedure is started else shows and error message.
- Next at the server side the transfer progress along with the bandwidth is calculated and displayed in the appropriate sections, and simultaneously the time delay is showed on the client side.
- The routers A, B, C just acts as the intermediate nodes to transfer the file. These routers are used to avoid congestion caused due to heavy traffic.
- Finally the file will received will be shown in the server side with its size and the transferred bandwidth along with the time delay.

IV. PERFORMANCE ANALYSIS:

The proposed technique is tested under different conditions and different systems with different bandwidth. The system is tested for different aspects including the through put improvements.

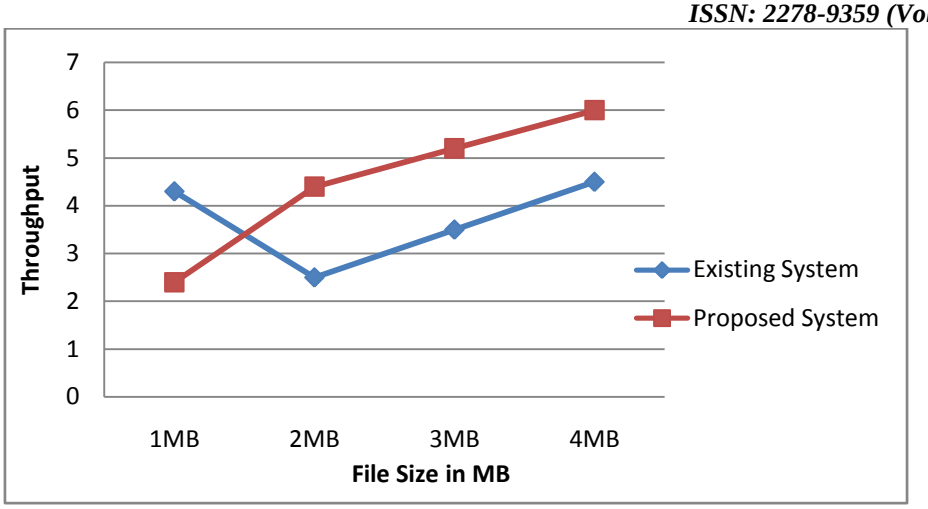

Figure 8: Throughput Comparison

The above figure depicts that the throughput of the proposed systems keeps on increasing while compared to the existing system as the bandwidth utilized is less. This is due to parallel transmission of data using the three routers rather than single one. The next aspect is to calculate the time delay to transfer a single file from source to destination. The comparison is made with the system of [1] where each packet is monitored effectively.

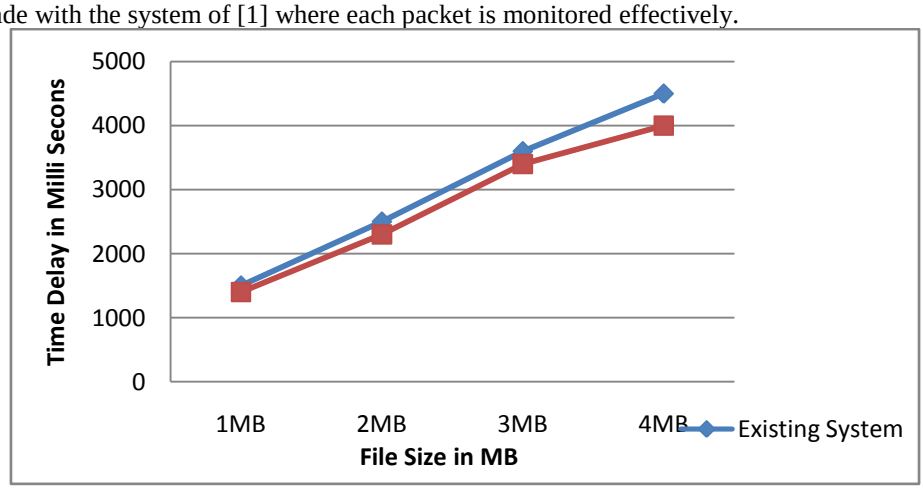

Figure 9: Time Delay Comparison

As the figure clearly depicts that the time taken to transfer the file is comparatively low compared to the existing system as the proposed technique uses parallel data transfer technique the time delay will be reduced to greater extent.

V. CONCLUSION

This paper presented a scalable Internet traffic measurement and analysis scheme with Hadoop that can process parallel file transferring capabilities. Based on the efficient computing platform, Hadoop, we have devised IP, TCP, and HTTP traffic analysis algorithms with a new input format capable of manipulating large file by splitting them into small files and transferring them in parallel to the destination. The paper also devised a key exchange algorithm which provides efficient data security to the packets before transferring them into the network. The traffic measuring is also kept quiet simple in order to reduce the burden on the measuring technique. The congestion control is handled well using multiple routers compared to the single router data transfers and finally the system is tested for better throughput and lower file transferring delays and the same is depicted in the analysis section. This is a basic step ahead for measuring network traffic lots of work has to be done to accomplish greater results.

REFERENCES

[1] Faisal Khan, Nicholas Hosein, Soheil Ghiasi," Streaming Solutions for Fine-Grained Network Traffic Measurements and Analysis", IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 22, NO. 2, APRIL 2014, pp.no:377-390.

[2] J. Mai, C.-N. Chuah, A. Sridharan, T. Ye, and H. Zang, "Is sampled data sufficient for anomaly detection?," in *Proc. 6th ACM SIGCOMM IMC*, 2006, pp. 165–176.

[3] A. Ramachandran, S. Seetharaman, and N. Feamster, "Fast monitoring for traffic subpopulations," in *Proc. 8th ACM SIGCOMM IMC*, 2008, pp. 257–270.

[4] C. Estan, S. Savage, and G. Varghese, "Automatically inferring patterns of resource consumption in network traffic," in *Proc. SIGCOMM*, 2003, pp. 137–148.

[5] L. Jose, M. Yu, and J. Rexford, "Online measurement of large traffic aggregates on commodity switches," in *Proc. , Hot-ICE*, 2011, p. 13.

[6] L. Yuan, C.-N. Chuah, and P. Mohapatra, "ProgME: towards programmable network measurement," in *Proc. SIGCOMM*, 2007, pp.97–108.

[7] [11] N. Brownlee, C. Mills, and G. Ruth, "Traffic flow measurement: Architecture," RFC 2722, 1999 [Online]. Available: http://www.ietf.org/rfc/rfc2722.txt

[8] [12] N. G. Duffield, "Sampling for passive Internet measurement: A review," *Statist. Sci.*, vol. 19, no. 3, pp. 472–498, 2004.

[9] [13] V. Sekar, M. K. Reiter, W. Willinger, H. Zhang, R. R. Kompella, and D. G. Andersen, "CSAMP: A system for network-wide flow monitoring," in *Proc. 5th USENIX NSDI*, San Francisco, CA, Apr. 2008, pp. 233–246.

