A Cloud-Pattern based Network Traffic Analysis Platform for Passive Measurement

Hengbo Wang, Wei Ding, Zhen Xia
School of Computer Science and Engineering, Southeast University
Jiangsu Province Key Laboratory of Computer Networking Technology
Nanjing, Jiangsu, China
{hengbowang, wding, zhxia}@njnet.edu.cn

***Abstract*—The existing Internet traffic passive measurement solutions are mainly based on a technical route of downloading traffic dataset and analysis tool from the corresponding distribution site, and then carrying out a local place research based on off-line traffic analysis. However, the issues of massive traffic datasets acquisition and analysis as well as measurement achievement reuse and sharing still need a further consideration. The paper applies the ever upsurging cloud computing paradigm to network traffic passive measurement field in order to address the issues. On the basis of inducing the drawbacks of the conventional technical route, we proposed a novel cloud pattern of passive measurement work and designed an architecture of cloud-pattern based network traffic analysis platform. Furthermore, using the authentic traffic collected at a CERNET backbone (10 Gbps), we implemented a prototype system of the architecture, called IP Trace Analysis System, or IPTAS for short. Combined with IPTAS, the paper elaborates the critical implementations of the architecture and verifies its feasibility and flexibility through IPTAS application instances.**

Keywords-component; network traffic passive measurement; cloud pattern; traffic analysis; measurement achievement reuse and sharing; IPTAS

I. INTRODUCTION

Network traffic measurement is one of the most important means to establish accurate network models, validate new protocols and applications, diagnose network failures, and enhance network performance and quality of service [1], [2]. According to whether injecting the extra traffic into network during the measurement, network traffic measurement can be divided into active measurement and passive measurement. Passive measurement is usually launched through mirroring, monitoring and analyzing the traffic on a certain link. Compared with active measurement, it imposes no interference on the operation of network and the measurement result can reflect network behavior most realistically and accurately. When referred to a large-scale and high-speed network passive measurement, like Internet-wide, the emphasis is placed on massive network traffic collection, storage management and analytical processing. Unfortunately, most research groups or individual researchers generally lack the capacity of backbone traffic collection, as well as the traffic size-matched storage and computing power, which highly restricts them taking the network behavior research based on passive measurement.

Both domestic and international research institutions have carried out some projects against the common difficulties of backbone network passive measurement. The projects can roughly be classified into two categories. One stresses on traffic collecting and distribution, which is represented by CAIDA's [3] data distribution system and SNATT [4] traffic sharing platform of Tsinghua University. While the other focuses on tracking, describing and sharing the measurement associated resources accumulated from multiple vantage points, such as dataset, algorithm, experiment, etc., and provides the catalog and indexing service. The representatives of the second category include SIMR [5], IMDC [6] and MOME database [7], etc. The above projects provide researchers with a technical route of downloading traffic dataset and analysis tool from the corresponding distribution site and carrying out a local place research based on off-line traffic analysis. But using such a route, the researchers have to confront the following issues:

- Massive traffic datasets acquisition: Due to large traffic size, limited local storage, download bandwidth and cost constraint, invalid distribution site or other factors, researchers are often unable to download sufficient datasets to support the validation of the macro network behavior models and conclusions, especially download some international traffic.
- Massive traffic datasets analysis: When the size of traffic dataset to be analyzed comes up to TBytes level or higher, high-performance computing resource and high-speed I/O storage device are required to meet the efficiency requirement of data processing. The general research groups or individual researchers have to invest extra funds and time to build up and maintain their own computing environment or ask for computing service from the third party.
- Measurement achievement reuse and sharing: In the network behavior studies of research groups, there would accumulate some valuable analysis algorithms and general measurement results [1]. However, due to the lack of information sharing mechanism among research groups or even in a group, some work has to begin with designing and implementing the analysis algorithm, which causes a waste of time and energy on the repetitive work.

Cloud computing [8] is a new and emerging IT resource delivery and usage paradigm. It means the consumer obtains the required resources (hardware, platform, software, etc.) via a common network in an on-demand, scalable and pay-as-you-go manner. The paradigm is consistent with the utilization of public facilities and services like water and electronic. The so-called "cloud" refers to an IT resource pool containing huge storage and computing power and software resources. Through the abstraction and encapsulation of a virtualization layer, the cloud exposes to the consumer by means of XaaS (Everything as a Service), which means its internal implementation details are totally transparent to the outside consumer. Nowadays, the cloud computing systems or platforms are mainly concentrated on the general IT field [9], but such a macro application scenario of cloud computing is obviously applicable to the micro network passive measurement field.

This paper aims at using the ever upsurging cloud computing paradigm to address the above issues in passive measurement work. On the basis of inducing the drawbacks of the conventional technical route, we proposed a cloud pattern of passive measurement work and designed an architecture of cloud-pattern based network traffic analysis platform. Furthermore, using the authentic traffic collected at a CERNET backbone (10 Gbps), we implemented a prototype system of the architecture, called IP Trace Analysis System, or IPTAS for short. Combined with IPTAS, the paper elaborates the critical implementations of the architecture and verifies its feasibility and flexibility through IPTAS application instances.

The rest of the paper is organized as follows: related work is introduced in Section II. Section III discusses the cloud pattern of network passive measurement work; An architecture of cloud-pattern based passive measurement traffic analysis platform is proposed in Section IV; Section V elaborates the critical implementations of the architecture combined with IPTAS platform; The deployment and application of IPTAS platform are presented in Section VI; Section VII summarizes the paper and discusses our future work.

II. RELATED WORK

CAIDA (Cooperative Association Internet Data Analysis) is an international research association involved in Internet measurement and analysis. CAIDA began to distribute anonymized Internet passive traffic taken at an OC48 peering link as early as 2002. Since April 2008, CAIDA has distributed and updated the anonymized traffic dataset collected from two high-speed monitors on a commercial backbone link (equinix-chicago and equinix-sanjose). Those data can be used to research on the characteristics of Internet traffic, including application breakdown, security events, geographic and topological distribution, flow volume and duration.

Considering the fact that rapid growth of traffic collection will bring about the difficulty of management, some efforts have been made on Internet data tracking and describing. Mark Allman et al. detailed a Scalable Internet Measurement Repository (SIMR) in [5] for facilitating the data sharing within the research community. SIMR is centered around a database of measurements, tools, experiments, users and datasets. From the database, users can search for specified measurements, download the tools used to make and analyze those measurements, and quickly ascertain the relationships between various measurements. The MOME Project Consortium built a similar database for data indexing and distribution. The MOME database [7] contains packet, flow, application traces, as well as routing, HTTP and QoS data sources. Some data have analysis result including average traffic rate, package sizes, arrive rate and inter-arrival time. CAIDA began to build the Internet Measurement Data Catalog (IMDC) [6] in 2002. Compared with SIMR and MOME database, IMDC does not provide centralized sharing of measurement data, but rather tries to track and describe any available measurement data in the networking community.

III. CLOUD PATTERN OF PASSIVE MEASUREMENT WORK

The elements related to network passive measurement work contain traffic dataset, analysis algorithm, measurement result as well as storage and computing resource. Respectively, the participants of passive measurement work can be divided into four roles: traffic collector, algorithm developer, measurement worker and computing service provider.

The conventional pattern of passive measurement work can be characterized from the three issues in section I: in this pattern, traffic collector, algorithm developer and computing service provider were not effectively integrated, which mainly causes resources to become scattered and difficult to obtain and share. When planning to carry out measurement work, measurement worker must coordinate with the rest three roles in order to obtain the resources required, which calls into the issue of efficiency and cost. Meanwhile, the pattern lacks feedback process from measurement worker to the rest three roles, which means traffic collector and algorithm developer cannot gain benefits from the completed measurement work, whether in knowledge or economic payback. Although some measurement platform integrates the roles of traffic collector and algorithm developer, the algorithms provided by platform architect is far from enough compared with the algorithm resource scattered on various research groups.

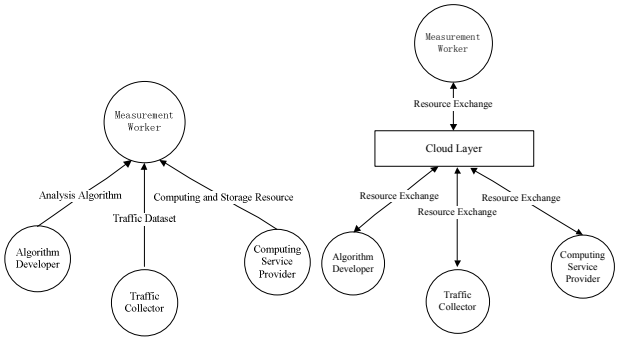

Fig. 1. Conventional (left panel) and cloud pattern (right panel) of passive measurement work

To overcome the drawbacks of the conventional pattern, we proposed a novel cloud pattern by deploying an intermediate cloud layer between measurement worker and the rest three roles (shown in Fig. 1). The cloud layer centralizes up-to-date

traffic datasets, reusable analysis algorithms, valuable measurement results and high-performance storage and computing infrastructure for management and exchange. The cloud pattern of passive measurement work can be characterized by the following new features:

A. *Efficient Access to Massive Traffic Datasets*

Massive traffic datasets transmission from the remote storage site to the local computing device is time-consuming and bandwidth-bound, especially concerning a centralized storage environment and multi-users concurrent downloading scenario. Naturally, an alternative thought is to replace data movement with computing task movement since the size of analysis algorithm is almost negligible compared with the traffic dataset. In this way, measurement worker doesn't need to endure a long time waiting for downloading massive traffic datasets. Alternatively, they could upload their private algorithms to the cloud and carry out the measurement work in a route of customizing traffic dataset, reusing shared or their private analysis algorithm, submitting analysis task and downloading measurement result. The new route provides measurement worker with an on-demand and efficient manner to use massive traffic datasets. Algorithm developer can use the same route to debug their new developed algorithms.

B. *High-Efficiency Analysis of Traffic Datasets*

There is no need for measurement worker to build up, configure and maintain their own computing environment or ask for computing service from the third party. The cloud layer provides high-performance computing and storage infrastructure and suitable processing pattern to ensure a high-efficiency and transparent execution of the formatted analysis task submitted by measurement worker. A well-known massive data parallel processing pattern in cloud computing is Google's MapReduce. However, it's still a challenging job for algorithm developer to abstract each specific analysis algorithm into a map and reduce function owing to the inherent correlation inside network traffic, especially those whose purpose is an overall analysis rather than a data-independent, output-mergeable metrics statistics, such as packet profiling, distribution and topology modeling, traffic recognition, etc. Therefore, we recommend the conventional processing pattern of independent task or workflow based on DAG (Directed Acyclic Graph) for traffic analysis. It could be improved with the future effort of algorithm developer.

C. *Reuse and Sharing of Measurement Achievement*

On the one hand, customizing traffic dataset as need and reusing the analysis algorithms with high utilization frequency uploaded by algorithm developer exert great help to relieve possible repetitive work of measurement worker. Conversely, on account of traffic analysis computation taking place inside the cloud layer, measurement result generated by measurement worker's analysis task can be shared with algorithm developer and traffic collector as a feedback. Algorithm developer can take advantage of the historical measurement results to validate the functional correctness of the new developed algorithms. Still through observing and analyzing existing measurement results, traffic collector is able to optimize the collection parameters and adjust sampling method [2] so as to make the traffic captured meet the study needs better. For example, the traffic collected by periodic sampling may cause measurement distortion when the network reveals a periodic behavior coincidently.

Besides the above notable advantages, another implicit but essential feature of the cloud pattern lies in that each role of measurement work is both a resource producer and consumer, which will keep the cloud resources always fresh and continuously adequate as the increasing measurements. Although it seems that computing service provider gains no benefit from the cloud, computing service provider is a natural candidate of the cloud layer architect since that storage and computing infrastructure is fundamental to measurement work. Consequently, computing service provider is able to obtain economic profits from the measurement activities through promoting resource pricing mechanism.

IV. ARCHITECTURE OF CLOUD-PATTERN BASED NETWORK TRAFFIC ANALYSIS PLATFORM

According to the above proposal, we designed a practical architecture of cloud-pattern based network traffic analysis platform to improve and facilitate the conventional pattern of measurement work. Above all, considering the conventional technical route is accustomed and convenient to a quantity of measurement workers, the architecture still supports the conventional pattern rather than simply replace it with the cloud pattern.

The overall architecture is divided into three layers: resource collection layer, resource management layer and open service layer from bottom to top (shown in Fig. 2). Function and composition of each layer is elaborated below.

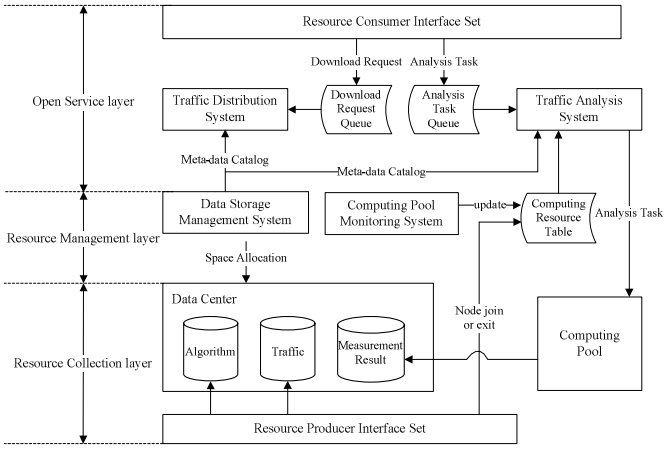

Fig. 2. Architecture of cloud-pattern based network traffic analysis platform

A. *Resource Collection Layer*

Resource collection layer provides traffic collector, algorithm developer and computing service provider corresponding formatted interfaces for uploading their own resource, called resource producer interface set. The uploaded traffic dataset, analysis algorithm and measurement result are stored in a data center whose component storage device has the

capacity of redundant protection and concurrent access based on network, no matter using centralized or distributed architecture. The "uploaded" computing nodes constitute a computing pool in logic to undertake traffic analysis computation and are registered in a computing resource table so that the upper management layer can perceiving the scaling of the computing pool. Because of the difference in the location of data storage and task execution, the routes between storage devices and each computing node must be network reachable.

B. Resource Management Layer

Resource management layer takes charge of computing pool monitoring and data storage management. Computing pool monitoring system initiates periodical heartbeat check and performance metrics collection for each node in the pool and then updates the computing resource table which assists the upper layer to make task assignment decision. The work can be partly resorted to the existing mature monitoring software, like Ganglia, to ease the complexity and difficulty of system development.

Data storage management system performs a centralized management in terms of data entry conduction and data description. The data entry conduction concentrates on appropriate space allocation for the each kind of data resources before entering the storage environment in order to enhance storage density and avoid possible space fragments. Particularly, an entry of large size traffic dataset may suddenly saturate the free storage space. Therefore, a historical traffic datasets elimination procedure should be activated ahead of time to vacate space for the upcoming traffic if necessary. The data description aims at giving measurement worker an intuitive cognition of traffic dataset, analysis algorithm and measurement result as well as supporting the open service layer's quick query and call. It's achieved by maintaining a metadata catalog that records the various attributes of each kind of data resources, which are provided by data-owner or evaluated by the system. Take traffic dataset for example, before a dataset being selected and analyzed for research, measurement worker would concern about its data size and source, collection date and duration, TCP traffic ratio, packet arrival distribution, etc., while the computing node engaged in analysis have to know its storage location. Thus, those attributes must be recorded in the metadata catalog.

C. Open Service Layer

The open service layer offers traffic distribution service and traffic analysis service. Through the well-formed resource consumer interface set, usually webpage, any measurement work participant can download traffic dataset, analysis algorithm and measurement result, as well as submit analysis task. User's traffic download request and traffic analysis task is inserted into download queue and analysis task queue respectively.

Traffic distribution system processes each download request in download queue at a certain concurrent degree. For each request, it locates the required traffic datasets by querying the metadata catalog, imposes an anonymization procedure on the traffic datasets according to the international convention and finally notifies the user to establish a transmission connection based TCP or UDP. The purpose of anonymization is to remove the privacy data (e.g., source IP address, payload) in each trace (i.e., useful fields set of an IP packet) of a traffic dataset for privacy protection.

Traffic analysis system is a centralized task scheduling and transferring system, with functions of scheduling each waiting task in the task queue and transferring it to a suitable online node in the computing pool based on certain strategies. When the target node completes the task, measurement result is sent to storage environment directly while task completion information is returned to the system for updating task queue and notifying the user to download the measurement result.

D. Summary of the Architecture

If taking business mode into consideration, payment and audit system can be integrated into the architecture, which charges for the usage of open services and determines the profit assignment among various resource producers according to the utilization frequency of traffic dataset, analysis algorithm, measurement result as well as storage and computing resource.

To implement such an architecture, some critical steps must be properly designed:

1) Consistency Constrains over Analysis Algorithm: Distinctive algorithms uploaded by various developers have different self-defined inputs and execution means. For the sake of reuse and execution in a consistent way, some specifications must be imposed on the uploaded algorithm.

2) Traffic Flowing into the Platform: Traffic collector could be an equipment or human data-owner. Thus, the process design of traffic flowing into the platform ought to support automation and manual intervention.

3) Historical Traffic Dataset Elimination: The elimination procedure should have a quantitative standard to make the lost value of eliminated datasets as little as possible.

4) Traffic Analysis Service Control: The service must have some control strategies to ensure the reasonable use of the platform resources and prevent resources being monopolized or inclining to one side caused by user's abuse and misuse behaviors.

5) Traffic Anomymization Procedure Design: Anonymization usually accompanies with the degradation of trace research value. Thus, the main concern of anonymization procedure is keeping the primitive features of trace to the maximal extent on the premise of privacy protection. What's more, computation efficiency of anomymization is the other concern in order to reduce user's waiting time before initiating downloading procedure.

V. IMPLEMENTATION OF THE ARCHITECTURE—IPTAS

The affiliation of the authors is a province-level network center of CERNET which is long-termed engaged in the operational management and access services of Internet. It possesses a high-speed network real-time packet collector WATCHER [10] which is deployed on the CERNET province border (10 Gbps), a border router Netflow forwarding system and an international traffic collection system GATHER. With

the help of the three traffic collectors, we can obtain the bi-directional IP Trace [11] and Netflow traversing the province border, as well as pcap format trace distributed by CAIDA as a useful complement to the province border traffic (shown in TABLE I). Moreover, most staffs in our lab are involved in network behavior research based on passive measurement, and considerable analysis algorithms and measurement results have accumulated in their previous studies. On the basis of these, we developed a network traffic analysis platform called IP Trace Analysis System, or IPTAS for short, which is consistent with the above architecture. The main purpose of IPTAS is to provide traffic distribution and analysis service for our daily studies and promote measurement achievements reuse and sharing among our research groups. Eventually, we hope IPTAS is able to open in a public network environment to facilitate Internet researcher's passive measurement work.

TABLE I. THE TRAFFIC SOURCES OF IPTAS

Traffic Collector	Traffic Format
Watcher	IP Trace
NetFlow Forwarding System	Cisco Netflow V5
Gather	pcap

IPTAS's resource collection layer and open service layer are both based on B/S (Browser/Server) model with webpage-styled user interfaces. Taking IPTAS as a prototype system, we elaborate the critical implementations of the architecture listed in the section IV.

A. Consistency Constrains over Analysis Algorithm

IPTAS platform imposes following consistent constrains over analysis algorithm for reuse and consistent execution.

- Rule 1: the execution parameters of analysis algorithm code in sequence must be parameter file path, dataset file path and output path. Parameter file contains the self-defined parameters. Dataset file contains the storage path of selected traffic datasets and is automatically generated by IPTAS. Output path is also specified by IPTAS for storing the measurement result. The developer takes charge of parsing each execution parameter in the algorithm.
- Rule 2: the algorithm code returns 0 when it normally exits, and returns a non-zero integer when any exception happens during the execution.

Through the above constrains, traffic analysis system of IPTAS is able to execute all algorithms in a consistent way and judge the termination state of each running algorithm code. Moreover, as the self-defined parameters are saved in a file rather than built in the code, other measurement workers can easily reuse the algorithm by modifying the parameter file and selecting new traffic dataset as needed.

B. Traffic Flowing into IPTAS

Among the three IPTAS traffic collectors, WATCHER and Netflow Forwarding system are both the real-time collection system. Therefore, IPTAS pulls the IP Trace and Netflow in a certain interval by means of assigning collection task artificially to derive WATCHER and Netflow forwarding system to work. While GATHER is an off-line collection system, the administrator of GATHER pushes the traffic dataset that has been saved in local place into IPTAS through a manual upload interface.

One advantage of such a pull and push combined flow-in pattern would be that both IPTAS and other traffic owners can select the flow-in traffic datasets artificially to maximize the value of the stored traffic datasets in view that storage space will inevitably approach to the up-bound. For instance, the traffic collected during the natural disasters will exert great help on large-scale network failure recovery research.

C. Historical Traffic Dataset Elimination

According to WATHER historical collections, suppose capturing the 44 bytes IP packet header from the province border using 1/4 sampling rate, the collection speed roughly reaches 50—60 GBytes per hour. It means the size of a 24 hours traffic dataset will approach to 1.2—1.4 TBytes. It's proved that historical traffic dataset elimination is totally necessary.

The data storage management system of IPTAS, called DSMS, picks multi-dimension attributes to quantify the importance of each traffic dataset by weight, including data size, collection date and duration, utilization frequency, stored time, etc. In addition, DSMS developed a spatial prediction model, which combines single exponential smoothing method with weighted moving average method, to estimate the traffic size generated in the next collection through N times historical collections. On the basis of above work, the elimination problem can be formalized as below:

A storage zone with free space V_{free} has stored n datasets, in which dataset D_i has size V_i and weight W_i. The size of upcoming traffic is donated as V_{new} ($V_{new} > V_{free}$). The target of the elimination is to pick up $D_i, \ldots, D_j$ ($1 <= i, j <= n$):

$$\text{minimize} \quad \sum_{k=i}^{j} W_k \tag{1}$$

$$\text{s.t.} \quad \sum_{k=i}^{j} V_k \geq V_{new} - V_{free} \tag{2}$$

DSMS converts the above elimination problem into 0-1 bin-pack problem and then uses dynamic programming to search for the datasets that can be eliminated. The optimization objective of elimination process is to minimize the total weight of eliminated datasets on the premise of vacate enough space for the upcoming traffic datasets.

D. Traffic Dataset Analysis Service Control

IPTAS regulates measurement worker's behavior through task entry strategy and task scheduling strategy.

Task entry strategy is used to ensure the fair sharing of storage resource among IPTAS users. Since the measurement result generated by an analysis task is associated with a certain IPTAS user, when the storage space allocated to the user is lower than the threshold, his follow-up tasks will be

automatically denied. Certainly, the user can manually delete some of his historical measurement results to regain the adequate free space and go on traffic analysis.

Task scheduling strategy is used to ensure the fair sharing of computing resource as well as processing efficiency. Currently, IPTAS applies independent task pattern for analysis service. Following factors are taken to make a comprehensive scheduling decision using weighted linear combination strategy.

1) Task Throughput: Shortest Job First (SJF) strategy is in favor of throughput in task scheduling. Traffic analysis is a data-intensive computation. As distinctive algorithms have slight difference in computational complexity but may highly vary in input size, the task with smaller traffic dataset size is prioritized to be scheduled.

2) User Fairness: Considering undifferentiated users sharing the computing resource, resource monopoly and inclination should be prohibited. IPTAS takes the total runnning time of historical tasks and the amount of present concurrent tasks as two indices to quantify the fairness degree of a certain user. The follow-up tasks belonging to those who earn a higher fairness degree are prioritized to be scheduled.

3) Task Hungry Degree: Impacted by the above two rules, some tasks could always encounter a disadvantage in competition with other tasks and endure a long time hungry. Thus, the long-time waitting task is supposed to be given a higher priority.

E. Traffic Anonymization Procedure Design

Currently, IPTAS's traffic distribution service is only limited to IP Trace. Reference [11] proposed a complete IP Trace anonymization procedure whose emphasis is an improved Crypto-PAn algorithm based on the distribution characteristic of IP Trace address. The improved algorithm accelerates the anonymization through precalculating and caching the anonymization results of the first 16 bits of all IP addresses. In addition, the algorithm keeps the address type identifier as it was and skips those non-privacy related addresses (e.g., private, multicast address) to preserve the micro features of IP Trace as possible.

VI. DEPLOYMENT AND APPLICATION OF IPTAS

IPTAS platform is deployed in the internal Gigabit Ethernet of the province-level network center (shown in Fig. 3). The storage environment of IPTAS is a DELL network disk array with a total capacity of 80 TBytes, which supports multi-machine concurrent access using iSCSI. IPTAS's computing environment is temporarily a high-performance blade server, which is divided into several virtual machines (VM) using XEN to reduce the coupling among the multi-task concurrent execution.

IPTAS platform's data storage and management system, traffic distribution system, traffic analysis system and the database of metadata are all deployed in the IPTAS server for the sake of sharing metadata among the three systems. For the security reason, user interfaces of IPTAS are independently deployed on a Web server with a firewall isolating it from the IPTAS server. The two servers carry out communications through XML-RPC mechanism.

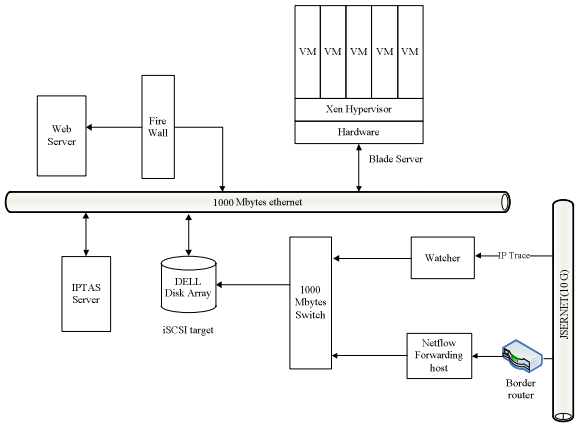

Fig. 3. Physical Depolyment of IPTAS

Here we demonstrate two application instances of IPTAS's open service layer to verify the feasibility and flexibility of the architecture.

A. Traffic Analysis Service

In network measurement, a set of IP packets complying with a certain flow specification and timeout limit is called data flow [12]. The flow measurement usually begins with the process of raw IP packets aggregation, which is known as "packet profiling". Here we display using IPTAS analysis service to complete IP Trace profiling in Fig. 4. Since packet profiling algorithm is a complicated but frequently used one, we added it into IPTAS algorithm repository for sharing. Thus, we don't need to develop such an algorithm. To initiate the analysis, the "packet profiling" algorithm iptrace2netflow and a dataset of 24 hours IP Trace on July 25th, 2012 are selected on the task submission webpage. Next, the self-defined parameters of iptrace2netflow are configured as needed on the page, for instance, we choose to profile the IP Trace falling into 00:00-01:00 and set sampling rate as 1/16. At last, submit the task and it will be scheduled and then run automatically. After completion, a set of data flow files (all flows in every 5 minutes are stored in a single file) consistent with Netflow V5 format can be downloaded for the further measurement.

Fig. 4. IPTAS Traffic Analysis Service

B. Traffic Distribution Service

Suppose a researcher would like to conduct a study about large-scale network failure recovery, some background traffic is indeed required. On the webpage of traffic distribution

system, a dataset of 24 hours IP Trace collected during Japan"3.11" earthquake in 2011 can be found in IPTAS distribution catalog (framed in red in Fig. 5). The first step to obtain the dataset is to fill out the download request, which contains information like user's email address, data duration, traffic direction, etc. (shown in the right panel of Fig. 5). After the request is submitted and the anonymization procedure is applied to the whole dataset, a download link will be sent to the user's mailbox. The user can download the anonymized IP Trace dataset using download tools like wget.

Fig. 5. IPTAS Traffic Distribution Service

VII. CONCLUSION AND FUTURE WORK

In this paper, we discussed the conventional and cloud pattern of passive measurement work, proposed an architecture of cloud-pattern based network traffic analysis platform and implemented a prototype system IPTAS. At present, IPTAS possesses various measurement-related resources with a total size of 40 TBytes, in which the amount of published IP Trace datasets is close to 140, while the amount of analysis algorithms for sharing is more than 30. Researchers can obtain IP Trace distribution service by accessing to iptas.edu.cn. But analysis service now is only available to the internal users. Our future work includes:

1) Extend the Traffic Category of IPTAS: At present, IPTAS mainly provides packet and flow level traffic dataset. But for supporting application level research, IPTAS should provide fine-grained traffic dataset (e.g., RRE, Session [13]). An available way to gain such dataset could be aggregating the existing data flow based on certain rules, or downloading from the possible distribution sites.

2) Open Traffic Analysis Service to the Public: To achieve this, two issues must be taken into consideration. One is the hierarchical management of user group to provide differentiated service. The other is the security inspection of uploaded algorithms. The potential malicious codes must be detected before execution according to the predefined rules.

REFERENCES

[1] Carey Williamson, "Internet traffic measurement," IEEE Internet Computing, vol. 5, 2001, pp. 70-74.

[2] Guang Cheng, Jian Gong, "A research on traffic measurement in a large-scale high-speed network," Computer Engineering and Apllications, vol. 38, 2002, pp. 17-19.

[3] Cooperative Association of Internet Data Analysis. http://www.caida.org.

[4] Feng-Hua Li, Xiao-Xin Shao, Shi-Jin Kong, et al. "SANTT: sharing anonymized network traffic traces among researchers," Journal of Dalian University of Technology, vol. 45, Suppl. October 2005, pp. 25-28.

[5] Mark Allman, Ethan Blanton, Wesley M Eddy, "A scalable system for sharing Internet measurements," In: Proceedings of the 2002 Passive and Active Measurement Workshop, Fort Collins, USA, March 2002.

[6] Colleen Shannon, David Moore, Ken Keys, et a1. "The Internet measurement data catalog," ACM SIGCOMM Computer Communication Review, vol. 35, October 2005, pp. 97-100.

[7] P. Aranda Gutierrez, A. Bulanza, M. Dabrowski, " MOME: an advanced measurement meta-repository," In: Proceedings of 3rd International Workshop on Internet Performance, Simulation, Monitoring and Measurement (IPS-MoMe), Warsaw, Poland, March 2005.

[8] Armbrust M, Fox A, Griffith R, et al. "Above the clouds: a Berkeley view of cloud computing," UCB Technical Report, 2009. http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-200928.pdf.

[9] Chen K, Zheng WM, "Cloud computing: system instances and current research," Journal of Software, vol.20, May 2009, pp.1137-1348.

[10] Ming-Zhong Zhou, Wei Ding, Ya-Dong Gao, "High speed Internet traffic analysis system—Watch1.0," Computer Era, vol.22, 2004, pp. 40-41.

[11] Bing Shi, Wei Ding, Ya-Dong Gao, Jian Gong. "IP Trace data based on a CERNET backbone," Joumal on Communications, vol. 27, November 2006, pp. 214-217.

[12] Ryu B, Cheney D, Braun HW, "Internet flow characterization: adaptive timeout strategy and statistical modeling," In: Proceedings of on Passive and Active Measurement Workshop 2001, Amsterdam, Netherlands, April 2001.

[13] J. Sommers, P. Barford, "Self-configuring network traffic generation," In: Proceedings of ACM Internet Measurement Conference, Taormina, Sicily, Italy, October 2004.