A Trace Measurement and Analysis System for the Multi-Links CERNET Backbone

Cheng Guang
College of Computer Science & Engineering
Southeast University, Nanjing, China
gcheng@njnet.edu.cn

Yongning Tang
Illinois State University
Normal IL, USA
ytang@ilstu.edu

Jiang Jiexin
College of Computer Science & Engineering
Southeast University, Nanjing, China

Ding Wei
College of Computer Science & Engineering
Southeast University, Nanjing, China

***Abstract*—Currently, researchers study the traffic difference of various Internet applications, and analyze the impacts on network performance and quality of service, mainly through network measuring, to explore the unknown behaviors performed by this huge complex nonlinear system. Passive measurement can get the measurement data which most truly reflect network behavior, so it is widely used in the network measurement. However, in the high-speed network, it is difficult to measure, process, storage and analyze the measured massive data. Under the background of passive measurement for the large-scale high-speed network, this paper focuses on measurement, collation, storage and analysis of the massive data, which was closely related to the measurement and behavior analysis. This paper will design and implement a Trace Measurement and Analysis System for the Multi-links CERNET Backbone (TMASM), and it would test TMASM using a JSERNET trace. In this paper, first, we design and implement TMASM, then discuss the features of the data captured by Watcher measurement subsystem. We study the strategy of processing the parallel links packet data collected from Jiangsu Province border multi-links of CERNET backbone. The Architecture of TMASM supports scalability in order to expand the analysis functions only through a simple approach. Finally, TMASM is tested, and a Trace collected from Jiangsu Province border channel of CERNET is analyzed using TMASM.**

Keywords- Trace, Packet Header, Measurement, Data Analysis Algorithm, CERNET

I. INTRODUCTION

With the rapid development of network technology, the amount of users, and the types of applications, has been growing almost exponentially. Driven by such an expansion, network scale and traffic grows, network structure and behavior becomes more complicated, and there are increasing technical problems in the running of the network. Currently, the researchers study the traffic difference of various applications, and analyze the impacts on network performance and quality of service, mainly through network measuring, to explore the unknown behaviors performed by this huge complex nonlinear system. Due to the operation of the network without interference, passive measurement can get the measurement data which most truly reflect network behavior, so it is widely used in the network measurement. However, in the high-speed network, massive data collection, collation, and analysis, are the difficult problems, which need to be resolved by passive measurement.

The Cooperative Association for Internet Data Analysis, CAIDA[1], measures and analyzes Internet traffic and performance, collects data for scientific analysis of network function. CAIDA collects several different data types at geographically and topologically diverse locations, and makes this data available to the research community. The Community Resource for Archiving Wireless Data At Dartmouth, CRAWDAD [2], is a wireless network data resource for the research community to store wireless trace data from many contributing locations, and staff to develop better tools for collecting, anonymizing, and analyzing the data. The Distributed Research & Academic Gigabits Open Network Lab, DRAGON-Lab [3], is a remote access networks lab that supports various network technique experiments, network products test, application system development and test of engineering plans. The lab can also be used as an online demonstration platform for various network and application products. Watcher [4] , located in the boundary of Jiangsu CERNET, is a network traffic measuring system for the high-speed network links, and it monitors three links between the CERNET backbone and the JSERNET backbone. In this paper, we will build a system based on the Watcher monitor to merge, analyze, and manage these different links data.

Under the background of passive measurement for large-scale high-speed network, this paper focused on the collation and analysis of the massive data, which was closely related to the measurement and behavior analysis. It would deal with the strategy of data collation, and the design and implementation of the Trace Measurement and Analysis System for the Multi-links CERNET Backbone (TMASM). And it would analyze a Trace via TMASM. The Remainder of the paper is structured as following. In Section 2, we describe related work. In Section 3, it designed and implemented TMASM. It analyzed the requirements of TMASM, and put forward that TMASM should have an architecture which supported scalability, in order to expand the analysis functions of TMASM only through a simple approach. Section 4 introduces the experimental and development environment of TMASM. In Section 5, TMASM is tested, and examples of some analysis

978-0-7695-3741-2/09 $25.00 © 2009 IEEE
DOI 10.1109/NAS.2009.63

tasks on a Trace collected from Jiangsu Province border channel of CERNET is executed using TMASM. Finally, Section 6 concludes our paper.

II. Related Work

Network measurement data are one basic research of the computer network technologies. But with the rapid growth of measuring data volume, the increasing of the corresponding statistical analysis tools (software) and the results of the analysis quantity, it is difficult to research how to encounter the effective management of measurement data. On the one hand, the measurement data stored in the storage medium will be likely to encounter the management difficulties because of hardware damage or update, as well as data migration. On the other hand, it is difficult to effectively track each measurement data set, that is, how to save and express every question and the analysis results.

A number of international Measurement agencies have been studied the measurement data management, mainly focusing on the measurement data tracking, feedback and improving availability. DatCat [5, 6], developed and run by CAIDA, is an Internet Measurement Data Catalog (IMDC), a searchable registry of information about network measurement datasets. It serves the global network research community by allowing anyone to find, annotate, and cite data contributed by others, and soon by allowing anyone to contribute new data. Information in DatCat is organized as Objects, each of which describes a real-world object or idea. DatCat is designed to work with any browser that supports standard HTML. The goals of the DatCat are to facilitate searching for and sharing of data among researchers, to enhance documentation of datasets via a public annotation system, to advance network science by promoting reproducible research.

Mark Allman [7] details a Scalable Internet Measurement Repository (SIMR) designed to distribute network data and the associated metadata, including details on user information, measurement tools, the dataset collection platform and location, dataset features, experiment information, and relationships between one dataset and others. From this set of databases users can search for particular measurements, download the tools used to make and analyze those measurements, and quickly ascertain the relationships between various measurements.

The MOME [8] was to co-ordinate activities in the field of IP monitoring and measurement by offering a platform for knowledge, tool and data exchange. The MOME database provides a knowledge exchange platform about measurement tools and measurement data. The data analysis tools can analyze packet-level traces stored in the following formats: pcap, erf, and tsh. The statistical data includes time-plots, average values of bit rate, and histograms of packet sizes, as well as split of captured traffic between different protocols and applications.

The above-mentioned three kinds of systems or methods provide centralized data sharing or distributed data catalog, which in essence are the measurement data regarding the management information systems

III. Network Measurement and Analysis Architecture

According to the design of data flow, network measurement and analysis architecture diagram is shown in Figure 1. In the figure, the system is consisted of data measure and storage subsystem, merging data subsystem, analyzing data subsystem, and measurement management subsystem. The measure and storage system with high-performance devices captures traffic packet trace from the different high-speed fiber links real-time. The merging data system requires a special strategy to merge a huge amount of data in a short time, and the data analysis system for trace-based analysis is the core of the system. The measurement management subsystem sets MIB to control the measurement and analysis in the system, and help the management and retrieval of research results. The following sub-section will introduce the four sub-systems.

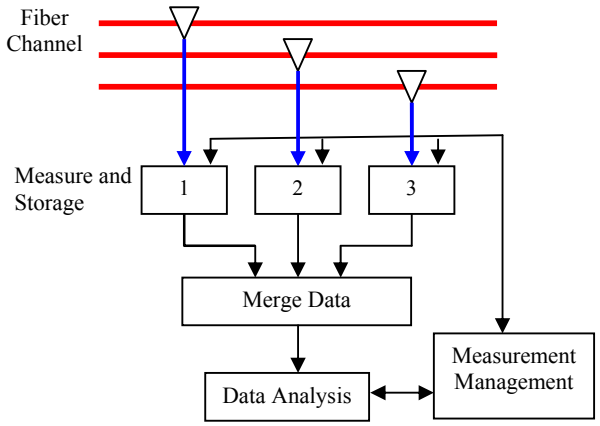

Figure1: Network Measurement and Analysis Architecture

A. *Measure and Storage Subsystem*

In this paper, the studied traffic traces are measured from the high-speed links between border routers and backbone routers of the CERNET network in Jiangsu Province (i.e. JSERNET) by the measure and storage subsystem (Watcher). Watcher designed by CERNET East China (North) Regional Network Center, can monitor the three channels in the Jiangsu CERNET network border to carry out real-time measurement. Moreover, if we continue to increase the fiber to expand the channel, as long as the installation of the corresponding fiber-optic splitter to increase the collector and memory, we can expand the system. Structure of the system is shown in Figure 2:

In the figure 2 structure, each bi-directional physical link is installed two fiber optical splitter and set up two servers to act as the collector and storage respectively. Collector captures all traffic from the optical interface directly, and intercepts each IP packet header, marks it with a timestamp, and then sends the packet header batch through the Gigabit link to storage, and last memory writes them to storage medium. All collectors of the Watcher system use NTP (Network Time Protocol) to keep

clock synchronization with a dedicated time server (synchronization interval: 64 seconds), which synchronizes with the standard time through GPS (Global Position System).

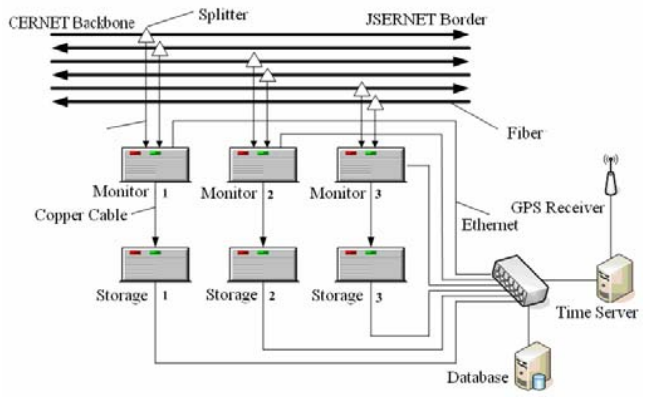

Figure 2 Watcher Architecture

Trace collected by the multi-links Watcher system exist two defects. First, because of the multi-link channel structure, the three pairs of parallel optical data are collected by the three different storages, so no storage contains all traces. Second, each monitor of the Watcher system is responsible for collecting a pair of optical fiber trace, and the two directions of packet timestamps will be possible out-of-order timestamps. Therefore, the storage needs to sort each link trace data, and then merge them in the different links according to their timestamp. The standardized and orderly traffic trace can be used for the future research work.

B. Merging Data Subsystem

If one communication between two machines is defined as a data stream, then the border router JSERNET with load balancing strategy may cause different packets in the same stream to choose a different fiber channel to enter the CERNET backbone, so the collected packet trace in the same packet stream may be distributed in more than storages. Before the traces are used, the merging data subsystem will sort and merge these scattered traces into a standardized message flow in an orderly timestamp manner. The aim of the merging data subsystem is to compare the timestamp of packet traces, sort them, and then merge different traces into one trace. In order to improve the efficiency of merging data, we use the high-performance computer cluster of Computer School Southeast University to merge these traces. First, each storage data is divided into sections by time period. Then for each time period, the data merging system in the high-performance computer cluster opens a merging process, which will merge the different storage sections into one section. A number of processes run in parallel, shortening the time to collate the data merge.

C. Data Analysis Subsystem

The data analysis subsystem to provide services directly to customers, the completion of trace-based analytical task, is the core of the system. Because of different needs, and the ever-changing and increasing tasks, the data analysis subsystem can not satisfy all needs, so the designed aim of the system is to provide users with general and basic data analysis services. For other more advanced personalization data analysis task, the system provides them with the analysis of secondary source data.

A well-designed architecture of the system is the key. The analysis subsystem at the time of development can only take into account the part need, so the system structure should have good scalability, so that the system will facilitate the expansion of functions in the future development. We consider two kinds of way to expand functional modules. One of which is to write the new module's source code, then the procedure in the original system adds a corresponding statement to call the new module and re-connect the compiler to generate the new implement code of the system. In this way, the system can be better integrated, but its shortcomings are too closely coupled, and only the development staff familiar with the systems can expand its functions. The approach is only applied the system development phase, and does not reflect scalability of the system.

The second expansion is the new function module itself which is an independently running program. The executable file of the expansion module is added a specified location, and then registered in the system file to add the module information. According to the user's configuration, the system can load the corresponding function module automatically. In this way the system assumes the management function, which registers the new module into the system, so the expansion function of the system can be realized. The advantage of the development approach is that the system code does not need to be modified to expand function.

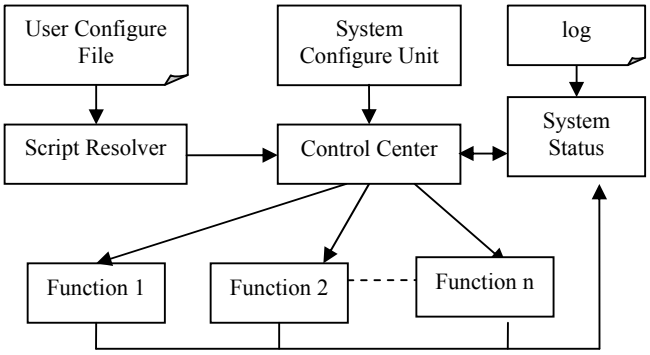

Figure 3(a) Control Flow Diagram

When system architecture is designed, the performance of the system is an important factor to be taken into consideration. The different modules in the system are designed to run in parallel to improve the performance of the system through multi-threaded way, and multiple users simultaneously use the data analysis subsystem. The user's configuration information and related data formats can be described by a script language, and a designed script resolver in the system resolves these scripts. Based on the above designed rules, the figure 3(a) shows the control flow diagram of the system architecture, and figure 3(b) gives its data flow diagram.

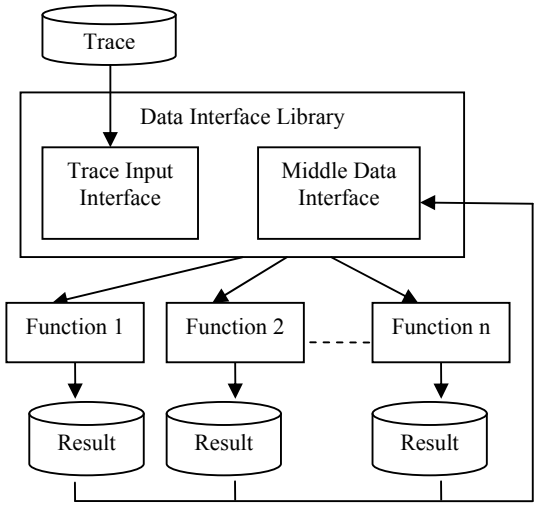

Figure 3(b) Data Flow diagram

The aim of the data analysis subsystem is to design and realize the data interface library, in order to read the trace and data results. The subsystem is consisted of four basic types of algorithms, which are respectively flow processing, packet sampling, and packet and flow metrics calculation. The packet sampling includes random sampling, systematic sampling and stratified sampling. The packet and flow metrics calculation is to obtain some statistics. Because of the deferent of the flow definition, the flow processing algorithm is more complicated. Ryu [9] defined that flow is a collection of a series of data packets, which meet specific flow specification and timeout condition. Different flow specifications and timeout conditions lead to different flow definition. The current system mainly uses the flow five-tuple (source IP, destination IP, source port, destination port, protocol type) and timeout to define flow.

D. Measurement Management Subsystem

The measurement management subsystem is a support facility of the measure platform. Its goal is to manage and store the original data, studied result, experimental code, experimental data, experimental results, as well as various effective management documents, provide convenient query methods, and supervise user behavior. The management subsystem can help researchers to reduce duplication of effort, improve research efficiency.

The supported function of the management subsystem is mainly reflected in four areas: (1) it has a reliable security management and user management strategy, and can oversight and audit user behavior; (2) it manages the original data packet trace, and trace-related research task; (3) it has a user-friendly man-machine interface to publish and retrieve research documents and data; (4) based on user demand, it generates suitable form for users to submit data, as well as intuitive diagram.

IV. THE TMASM IMPLEMENTATION

Consider the operating system, software development environment, such as copyright issues, the system will all make use of open source software, including open-source operating system, compilers, database management systems, etc. to achieve development. The system uses a Linux + Apache + PHP + PostgreSQL to realize the program, referred to as LAPP program. Its management subsystem is built through the web site structure to provide users with services, and the implementation of the subsystem runs directly as a background process on the Linux operating system. Table 1 shows the detailed hardware and software environment of this system.

TABLE I. HARDWARE AND SOFTWARE ENVIRONMENT

CPU	Intel(R) Xeon(TM) CPU 2.80GHz×2
Memory	DDRAM 2GB
Operation System	Red Hat Linux release 8.0 (Kernel v2.4.18)
Web Sever	Apache v2.2.3
Database	PostgreSQL v8.1.4
Compiler Environment	Gcc v3.2 PHP v5.1.4

The system uses PostgreSQL as database. PostgreSQL is an open source object-relational database management system. First, TMASMDB database is created in the PostgreSQL database management system, then a variety of data tables in the system are defined according to the actual needs, trace information table, analysis functional categories table, analysis algorithms information table, task configuration table, task result table, user information table, user behavior log table, and so on.

Some tables use to exchange external data, and some others exchange internal data. External interaction is to exchange data between user and management subsystem, and its tables involve user information table, user behavior log table, trace information table, analysis functional categories table, analysis algorithms information table, task configuration table, task result table, and so on.

V. TESTING TMASM

This section tests and verifies its functionality. It also uses some specific task data to analyze and test system stability. This paper will give some examples to illustrate its functionality and data analysis.

A. Testing Functionality

The test uses one hour traces from 20:00:00-21:00:00, named JS_20051110_20_21, and the trace is inputted into the TMASM. Its registered information in trace information table in the management information database shows in Table 2.

The testing steps include as following. According to the record format of the JS_20051110_20_21 trace, and the consistency constraints of the analysis algorithm, the first step one is to write a PPS (Packets per Second) statistical algorithm, which is named PPS_Stat, the corresponding parameter files named PPS_Stat.para, and documentation named PPS_Stat.summ. The second step is to upload executable files, parameter files and documentation to the TMASM upload area, and submit the registration information. After the Administrator audits these analysis algorithms, they are

removed into TMASM algorithm library. The third step submits the task to the TMASM, and the TMASM automatically distributes a code No. 23 for the task. According to the no. 23 TMASM automatically generates the data storage path. Based on the trace name of the user selected, the system automatically gives the data source storage path, and so on. Finally, researcher receives notification issued by e-mail after the completion of task, and the task information can be viewed, such as shown in table 3.

TABLE II. JS_20051110_20_21 TRACE REGISTERED INFORMATION

Trace	JS_20051110_20_21
Trace Origin	The Border of CERNET Jiangsu Province Network
Measuring Time	2005-11-10 20:00:00-21:00:00
Format	XML description
Upload User	gcheng
Status	Y (have been verified)

TABLE III. COMPLETION INFORMATION OF THE NO. 23 TASK

Task No.	23
Begin time	2007-4-20 15:08:05
End	2007-4-20 16:33:57
Existence of data	Yes
Data Storage Path	/home/gcheng/TMASM/Result/23
Data Size(MB)	1

B. Trace-based Data Analysis and Test

The example revises the PPS_Stat algorithm to count two way PPS in the JS_20051110_20_21 trace, and the improved algorithm is named PPS_Bidirect. The example also contains a testing of the expansion functionality.

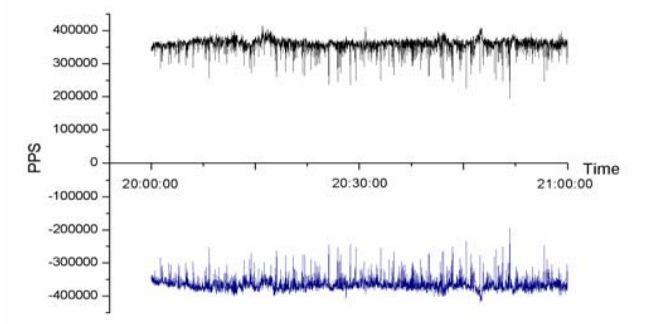

Figure 4 Bi-direction PPS between CERNET and JSERNET

The testing steps include as following. The first step is to extend PPS_Bidirect into the TMASM. The second step submits the take request to the TMASM, and the system automatically assigns the task no. 24. The third step issues a notification email after the task completion through the management subsystem. The analysis result is shown in the figure 4, where X axis is time and Y axis is PPS. The curve above the X axis is the direction PPS from JSERNET to CERNET, and the second curve below the X axis is the direction PPS from CERNET to JSERNET.

Based on the five-tuple of the packet characteristics, the sampling algorithm 5Tuple_Samp is given. The algorithm can randomly select packet to extract the five-tuple, and take all following packets with the same five tuple of the sampled packet. The number of five-tuple is a parameter of the user profile. The continuous packet time intervals in a flow are recorded. One flow with 80303 packets is sampled from the measured five-tuple randomly, so 80302 time intervals in the flow are recorded. Figure 5(a) is the measurement time of each time interval, and figure 5(b) is its cumulative distribution.

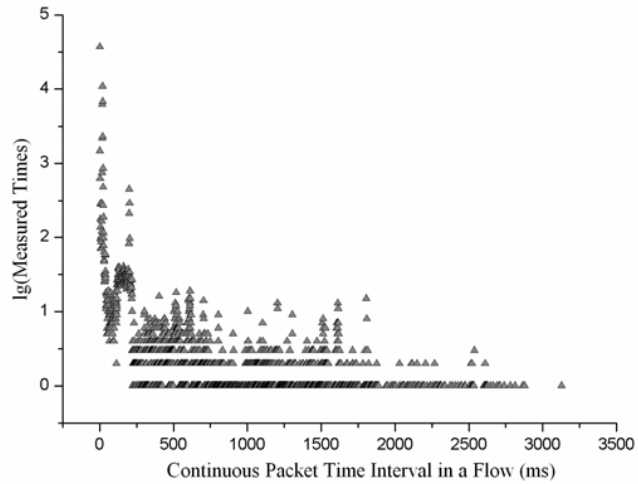

Figure 5(a) Measured Time of Continuous Packet Time Interval in a Flow

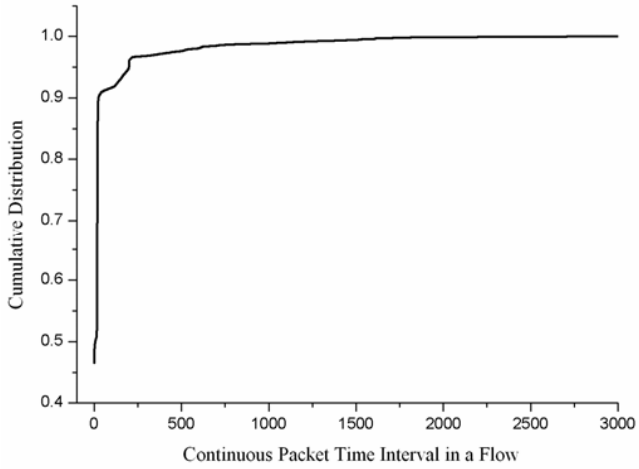

Figure 5(b) Cumulative Distribution of Continuous Packet Time Interval in a Flow

VI. CONCLUSION

The TMASM has been put into use. Practical operation shows that the Watcher measured and storage subsystem can be used to monitor the high-speed multi-links, and control packet loss rate at a very low value. The merging data subsystem in the high-performance computer cluster is able to merge the measured traces from the different links efficiently. Finally, the

TMASM testing shows that the data analysis algorithms, which obey the consistency constraints, can be expanded successfully into the system under the framework. The TMASM is able to use the analysis algorithms under the framework to process related tasks, and the trace analysis can achieve the desired results.

ACKNOWLEDGMENT

This work has been supported by the National Grand Fundamental Research 973 program of China under Grant No. 2009CB320505, the 2008 Natural Science Fundamental Program of Jiangsu Province under Grant No. BK2008288, the Excellent Youth Teacher of Southeast University Program under Grant No. 4009001018, and the Free Research Program of Key Lab of Computer Network in Guangdong Province under Grant No. CCNL 200706.

REFERENCES

[1] CAIDA, the Cooperative Association for Internet Data Analysis, http://www.caida.org

[2] CRAWDAD, A Community Resource for Archiving Wireless Data At Dartmouth, http://crawdad.cs.dartmouth.edu/

[3] DRAGON-Lab, Distributed Research & Academic Gigabits Open Network Lab, http://www.dragonlab.org/index.php

[4] WATCHER, IP Trace in the Jiangsu CERNET backbone, http://ntds.njnet.edu.cn/data/index.php

[5] DatCat, Internet Measurement Data Catalog, http://imdc.datcat.org/Home

[6] C Shannon, D Moore, K Keys, M Fomenkov, B Huffaker - ACM SIGCOMM Computer Communication Review, Volume 35, Number 5, October 2005, page: 97-100

[7] Mark Allman, Ethan Blanton, and 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

[8] The MOME Project Consortium. Information Technologies Society - Cluster of European Projects aimed at MOnitoring and MEasurement. http://www.ist-mome.org/.

[9] Ryu B, Cheney D, Braun H.W. Internet Flow Characterization: Adaptive Timeout Strategy and Statistical Modeling[J]. In Workshop on Passive and Active Measurement(PAM), Apr, 2001.