Network-monitoring Method based on Self-learning and Multi-dimensional Analysis

Isao Shimokawa and Toshiaki Tarui
Network Systems Research Department, Hitachi, Ltd.
292 Yoshida, Totsuka, Yokohama, Kanagawa 244-0817, Japan
{isao.shimokawa.sd, toshiaki.tarui.my}@hitachi.com

***Abstract*—A novel network-monitoring system for detecting abnormal network conditions (such as hidden network congestion) is proposed. The proposed monitoring system is based on self-learning and multi-dimensional analysis. It analyzes multiple parameters such as consumed bandwidth, packet size, and arrival interval of network packets simultaneously. By executing high-quality network monitoring it thereby achieves multi-dimensional analysis by use of Mahalanobis distance. A prototype monitoring system was constructed and evaluated. The evaluation results indicate that the monitoring system can accurately detect a hidden change of network-traffic conditions and reduce the number of unnecessary alerts for monitoring excess bandwidth according to a set threshold.**

Keyword-*Monitor; Network Fault; Mahalanobis distance.*

I. Introduction

To reduce environment load from the viewpoint of energy efficiency, lowering the power consumption of cloud-service systems is attracting much interest. In a current cloud-service system, to reduce power consumption of the system, a management server triggers migration of a virtual machine (VM), aggregates virtual servers from one server to another, and switches off unused physical servers. A power-saving information and communication technology (ICT) platform is previously proposed. [1]

The ICT platform has to guarantee network bandwidth for cloud-service systems. If a system-management server executes VM migration without considering network-link capacity, volume of network traffic may surpass network-link capacity (because network flows connected to the VM are also moved from one network to another). As a result, unexpected network congestion may occur. Moreover, quality of service (QoS) such as bandwidth guarantee may not be maintained. It is therefore important to rapidly and accurately monitor the network and to execute VM migration according to the monitored network conditions.

On the contrary, if the management server switches off a preliminary server used for redundancy in order to lower the power consumption of the cloud-service system, a "cloud-service fault" may occur because the redundant server is switched off. Accordingly, to run a cloud-service system 24 hours a day all year and maintain QoS, network faults must be rapidly detected.

To address the above-mentioned issues, so-called "feedback control" [2] by monitoring a system is expected to provide stable and high-quality cloud services. For finding network faults, it is especially critical that network monitoring rapidly detects abnormal increases or decreases of network traffic. In the present work, to meet that need, a novel network-monitoring method for rapidly detecting abnormal changes of network-traffic conditions was devised.

The rest of this paper is organized as follows. Section II summarizes issues about network operation and management. Section III outlines a proposed method. Sections IV describe a prototype network-monitoring system. Section V evaluates the prototype network-monitoring system. Section VI concludes this paper.

II. Issues concerning network operations and management

Network-traffic conditions are typically monitored by a method for network operations and management such as simple network management protocol (SNMP). In addition, the monitored data is analyzed according to a one-dimensional threshold (such as consumed network bandwidth) without distinguishing different network flows. If network traffic fluctuates around a predefined alarm threshold such as network bandwidth, however, an alarm may occur frequently. In that case, the administrator of the network will receive many alarms even if no fault or problem has occurred in the network. In other words, applying a one-dimensional judgment such as bandwidth threshold for detecting abnormal network conditions may raise too many alarms. As a result, it is difficult to accurately monitor network conditions. Consequently, it is necessary to establish a monitoring system that detects network congestion or faults without generating too many alarms.

III. Proposed method for monitoring and analyzing network traffic

To address the issue described in the previous section, a novel network monitoring system—based on self-learning and multi-dimensional analysis (SLMDA) [3]—is proposed here. The system monitors all network flows in real time from dimensions such as bandwidth, packet size, and packet interval. It is composed of an analyzing part and a monitoring part equipped with a node called "aggregated flow mining" (AFM) [4] implemented in each node. With regard to the analyzing part, a new evaluation scheme based on the multi-dimensional Mahalanobis distance [5] is applied.

ISBN: 978-1-61208-227-1

A. *Aggregated flow mining (AFM)*

When an administrator of a network monitors network-traffic conditions, it is necessary to distinguish many network flows and analyze them in detail. For that purpose, AFM (which distinguishes many kinds of network flows and provides statistical information about those flows) is used. The network administrator finds anomalous flows or conditions (or both) by analyzing of statistical information provided by AFM. A flow is conventionally defined as a collection of packets with five tuples (source IP address, destination IP address, source port, destination port, and protocol). In regard to AFM, the concept of the flow is extended, and an “aggregated flow” is defined by an arbitrary combination of each tuple. For example, one aggregated flow is defined by only the destination IP address irrespective the other tuples. Statistical information (such as number of packets and bytes) about flows that have the same destination IP address is therefore produced as statistics about one aggregated flow. As described above, if the concept of aggregated flow is introduced, flows that travel between one host and multiple servers are regarded as one aggregated flow. It is therefore possible to analyze network traffic or flows.

B. *Algorithm for self-learning multi-dimensional analysis*

Hereafter, the node at which “integrated mining of flow” is performed is simply referred to as “the IMF”. The proposed system for monitoring network-traffic flow is shown in Figure. 1 , where several users are connected to a data center. The IMF collects statistical information from multiple AFMs for analyzing network traffic and send alarms to a network-management server when it detects abnormal network conditions.

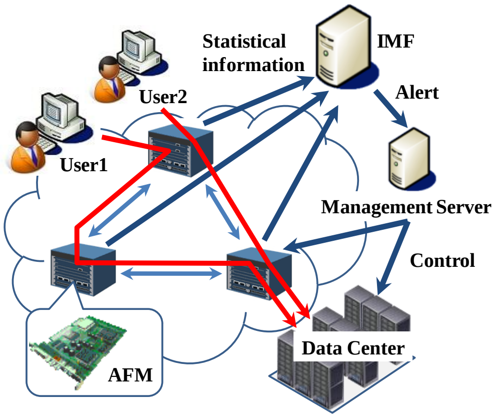

Figure. 1 Proposed monitoring system

In this system, network traffic is analyzed by using statistical information gathered from all AFMs in the network. Specifically, SLMDA using the Mahalanobis distance is used to analyze network conditions in detail. Although the system uses the same analysis parameters originally implemented in the AFM, the proposed SLMDA method is applicable to various analysis parameters. This method follows the procedure described below:

1. The standard distribution for each analysis parameter is defined.
2. A Statistical distribution for each analysis parameter is measured by an AFM in real time.
3. The Mahalanobis distance is calculated by comparing the distance between the defined standard distribution and the measured statistical distribution of each analysis parameter in real time, and the occurrence of abnormal network traffic conditions is judged.
4. The standard distribution for each analysis parameter is updated by using the measured statistical distribution (step 2) in real time.
5. If a rapid change of network traffic condition is detected, its cause is identified by analyzing the conditions in detail.
6. Return to step 2.

C. *One-dimensional judgment based on Mahalanobis distance*

To detect an abnormal network condition, it is necessary to analyze the changes of network traffic (such as bandwidth) in detail. Accordingly, a method for determining whether the network condition changes is proposed here. This one-dimensional judgment method based on the Mahalanobis distance is explained in Figure. 2 .

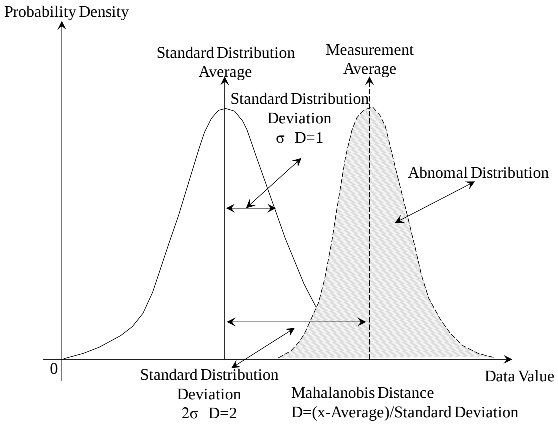

Figure. 2 One-dimensional judgment based on Mahalanobis distance

To analyse a rapid change of network-traffic distribution, it is necessary to define the standard distribution as a normal condition and compare that standard distribution and a statistical distribution measured by AFM. The comparison procedure is explained as follows. An initial value is set to define the standard distribution as a target for comparison with a measured distribution. As the parameters for comparison, average and standard deviation of the standard distribution are used. These parameters are set according to the administrator's experience or knowledge. To analyse a rapid change of network-traffic conditions, statistical information (such as data throughput) from the AFM of each router is measured in real time. Mahalanobis distance of the standard distribution is then calculated by using the throughput distribution measured by the AFM in real time. The Mahalanobis distance is defined as

D=(x-average)/standard deviation [a.u.: arbitrary unit] (1)

If the Mahalanobis distance is very large, it is considered that an abnormal traffic condition exists. For example, if the calculated Mahalanobis distance is larger than 2 and the measured distribution follows a normal distribution, it is judged that the throughput distribution (namely, data rate/bandwidth of a traffic flow) is not significant according to a 5% significance level. As a result, it can be regarded as an unusual distribution that occurs at a probability of 5%.

D. Multi-dimensional judgment method based on Mahalanobias distance

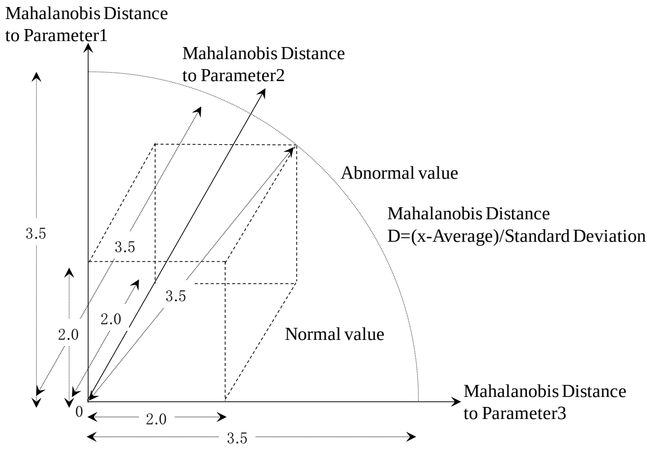

Figure 3. Three-dimensional Mahalanobis distance

If a one-dimensional judgement method is used to judge whether the measured network condition is normal or abnormal, an erroneous decision might occur frequently owing to the limited amount of information available for analyzing traffic flow. In the case that erroneous decisions occur, a management server may receive too many alerts, which might cause control errors. Accordingly, to improve the accuracy of the judgment, a multi-dimensional judgment method (as shown in Figure 3) is proposed here. This method involves several steps. First, each analysis parameter is assigned to each axis in the figure as a dimension for analyzing 3D Mahalanobis distance. On each axis, Mahalanobis distance is calculated. Multi-dimensional Mahalanobis distances are then calculated on the basis of multiple one-dimensional distances as follows.

Mahalanobis distance with three dimensions=

$$\mathrm{sqrt}(\alpha * x^2+\beta * y^2+\gamma * z^2) \qquad (2a)$$

$$\alpha+\beta+\gamma=3 \qquad (2b)$$

α,β,γ is not unique because reasons of network fault are not always same. So It is necessary to investigate α,β,γ from past data and system condition. In this report each of α,β,γ values is 1.

E. Updating standard distribution on the basis of feedback.

To analyze a measured network-traffic condition, a standard distribution as a target for comparison should be defined correctly. However, it is not easy to define a normal network traffic condition that changes day by day. A new self-learning method, by which a normal standard distribution is dynamically updated by using feedback data, is therefore proposed here. Basically, the network-traffic condition is monitored, and its changes are analyzed in real time. The standard distribution is then updated according to the change of the average data value of a measured distribution.

The example of a standard distribution updating method is shown in Figure 4. With the proposed method, average and standard deviation of the standard distribution are updated dynamically by calculating a moving average according to the following formula:

Moving average of average on standard distribution= (average on standard distribution + average on measured distribution)/2 (3a)

Moving average of deviation=
(deviation on standard distribution + deviation on measured distribution)/2 (3b)

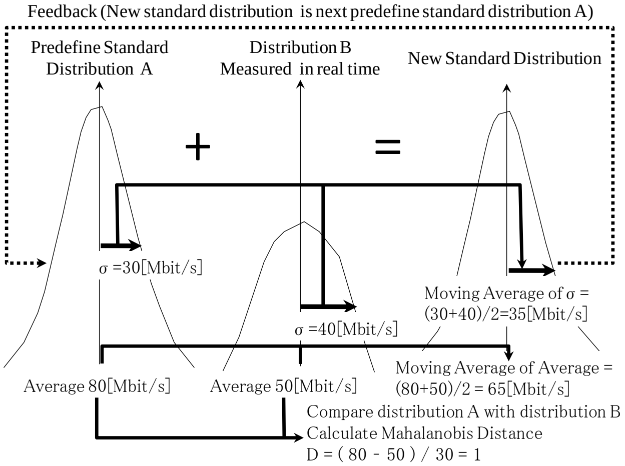

Figure 4. Example of standard distribution updating method.

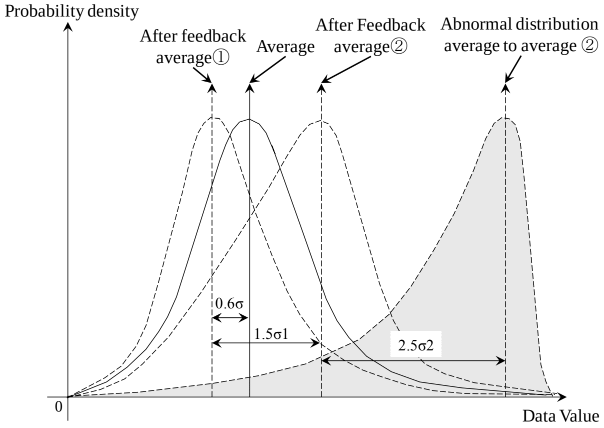

Figure 5. Abnormal distribution to standard distribution

ISBN: 978-1-61208-227-1

F. Detection of factors for changing network condition

Detection of factors that change a network condition is explained in Figure 6. The criterion for detecting an abnormal condition is as follows:

average of standard distribution + two standard deviations < measured throughput of network flow (4)

When a rapid change of a network-traffic condition is detected, a flow that is further than two sigmas from the average value of the standard distribution is considered as a peculiar flow. Although two standard deviations is set as the threshold for detecting a peculiar flow, the threshold is set by the administrator of a network. If the threshold is two standard deviations and the measured data distribution follows a normal distribution, this condition is equivalent to a significance level of 5%, and it only occurs at a probability of 5%. In addition, flows that cause such a condition are judged as peculiar flows.

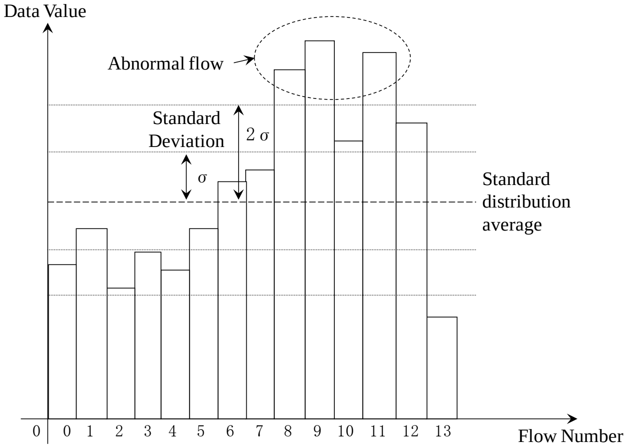

Figure 6. Detection of flows that cause abnormal condition

IV. IMPLEMENTATION OF PROTOTYPE TRAFFIC-MONITORING SYSTEM

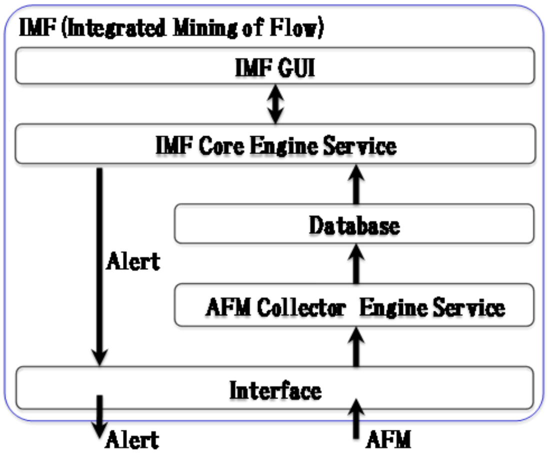

Figure 7. Block diagram of IMF

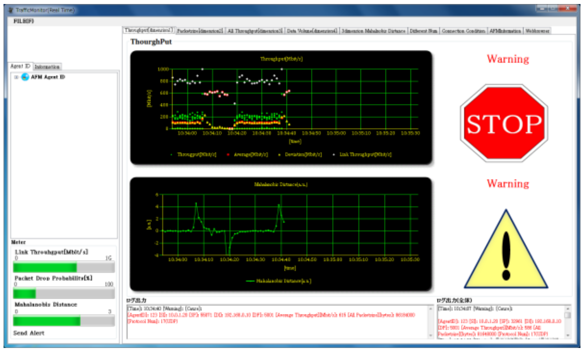

Figure 8. Example of IMF GUI

A block diagram of the IMF is shown in Figure 7. The IMF provides two functions (IMF core-engine service and AFM collector-engine service) for analyzing monitored data. The AFM collector engine service collects statistical information through a interface (such as Ethernet) and stores it in a database. The IMF core engine service then reads the statistical information from the database and analyzes it. If necessary, it sends an alert message to a network management server.

The GUI of the IMF is shown in Figure 8. The IMF analyzes statistical data from the AFM and shows real-time conditions on the GUI. The network administrator can check the flow that is presumed to be the factor causing an abnormal condition and the time of occurrence, when the change of condition is detected on GUI.

V. EVALUATION OF PROTOTYPE MONITORING SYSTEM

A. Verification of proposed method

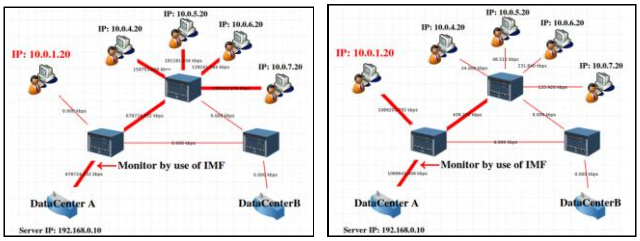

Figure 9. GUI screen of CORE

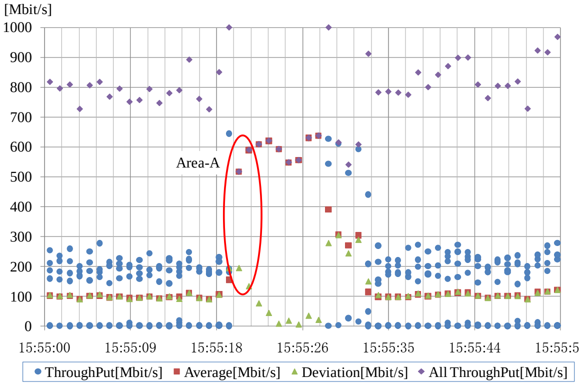

Figure 10. Throughput per network flow

ISBN: 978-1-61208-227-1

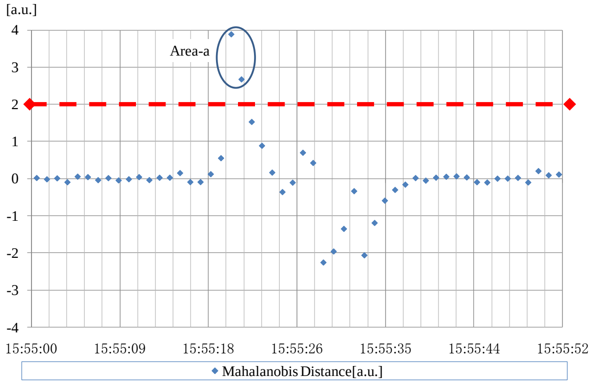

Figure 11. Mahalanobis distance with throughput per network flow

To evaluate whether detect a rapid change of network traffic by an SLMDA-based function, network congestion (as a change of network condition) is focused on, and the SLMDA-based function for detecting it is verified. Specifically, to produce network congestion, user datagram protocol (UDP) traffic is intentionally generated under the condition that only transmission control protocol (TCP) traffic is present. A rapid change of the network-traffic condition is then generated since transmission rate of TCP is rapidly decreased when TCP detects network congestion. The Common Open Research Emulator (CORE) [6] (which enables network emulation) was used to produce the above-described condition. The GUI screen of CORE, representing both conditions (before and after inserting UDP traffic), is shown in Figure 9.

An experiment to investigate traffic condition was performed by means of AFM. In the experiment, a rapid change of network condition is intentionally produced by inserting UDP traffic under the condition that only TCP traffic is flowing. As shown in Figure 9, multiple users connect to a data center by TCP communications. Network traffic via the router connected to data center A is monitored by the AFM. A user (IP:10.0.1.20) connects to data center A for a certain period by UDP communication. TCP controls the window size for data transmission when it detects congestion. On the other hand, UDP does not control transmission rate. Therefore, when UDP traffic is inserted into the TCP traffic, UDP traffic occupies most of the network link. As shown on the right of Figure 9, when UDP traffic is generated, the traffic (IP:10.0.1.20) occupies most of the network link. The throughput measured by AFM is shown in Figure 10. As shown in area A in the figure, a rapid change of network condition occurred at 15:55:20. Average throughput is not high until 15:55:20, since only multiple TCP communications share the network link. However, the average throughput increases rapidly, since a UDP communication occupies the network link from 15:55:20. As shown in Figure 11, a significant change of network-traffic condition is detected at 15:55:20, since the Mahalanobis distance is over 2 at 15:55:20. It is thus possible to detect abnormal traffic conditions that the network administrator cannot recognize by a conventional method. Moreover, the IMF could detect the UDP flow (IP:10.0.1.20) as a potential factor for causing a significant change of network condition. The network administrator can therefore easily find the factors causing significant changes in network condition and take appropriate measures for handling them by using the following information produced by IMF.

Detecting Flow by IMF: [Time] 15:55:20 [Source IP] 10.0.1.20 [Source Port] 56165 [Destination IP] 192.168.0.10 [Destination Port] 5001[Average ThroughPut][Mbit/s] 517 [AllPacketSize][bytes] 80136000 [Protocol Num] 17(UDP)

B. *Experiment on intranet*

The purpose of this experiment (see concept shown in Figure 12) is to verify whether the proposed method can detect a significant change of network-traffic condition on a real intranet. In the experiment, real intranet traffic was measured for one day by AFM. The bandwidth of the link used for the experiment was 100 Mbit/s. Four parameters (throughput, average packet size, link throughput, and data volume) were used as parameters in this evaluation. The results of the evaluation from AM0:00 to AM9:00 are shown in Figure 13 to Figure 21.

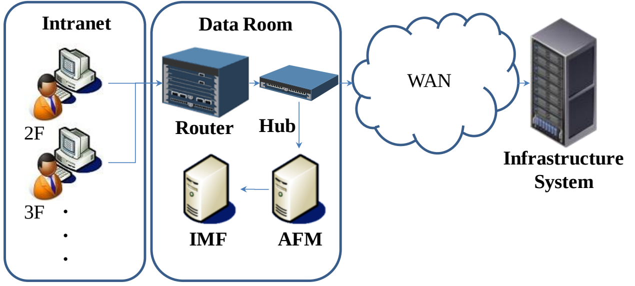

Figure 12.Concept of experiment on intranet

As shown in Figure 13, many flows have high throughput. It is therefore difficult to pinpoint in the figure whether a significant change of network-traffic condition was generated or not. On the contrary, in Figure 14, with the proposed method, significant changes of network-traffic condition are judged to occur two or more times. In the time zone when the Mahalanobis distance is over 2, namely, from area A to area C in the figure, traffic rapidly decreases and stays low for a short time. After that time, the traffic rapidly increases again. It is concluded that the proposed scheme could detect such significant changes of network-traffic condition.

The average packet size is shown in Figure 15, and calculated Mahalanobis distance that corresponds to average packet size is shown in Figure 16. As shown in Figure 15, average packet size does significantly not deviate from an average of 500 bytes. However, in Figure 16, two points (shown in area D and area E in the figure) are detected as significant changes of average packet size. The significant change of network-traffic condition is therefore detected according to average packet size. The times at which packet-size changes are detected are equivalent to the times at which throughput are detected in Figure 14. (Area-A,Area-B)

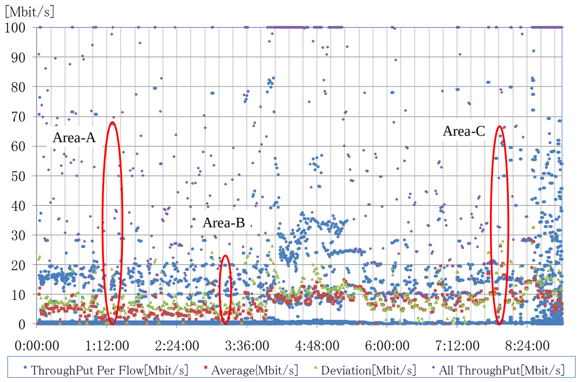

Figure 13. Throughput per flow

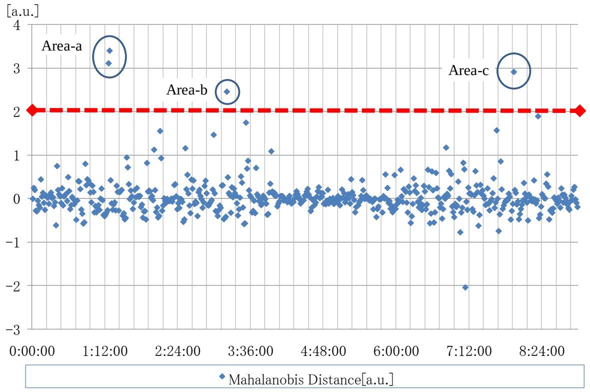

Figure 14. Mahalanobis distance with throughput per flow

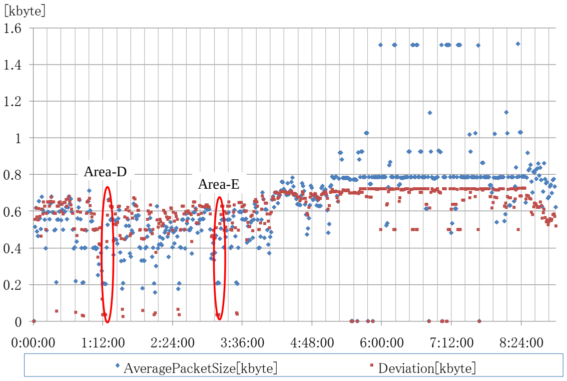

Figure 15. Packet size

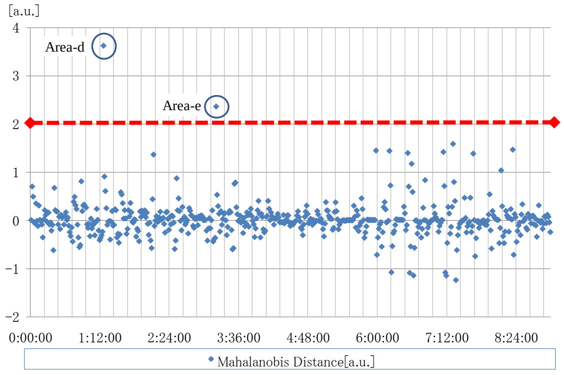

Figure 16. Mahalanobis distance with packet size

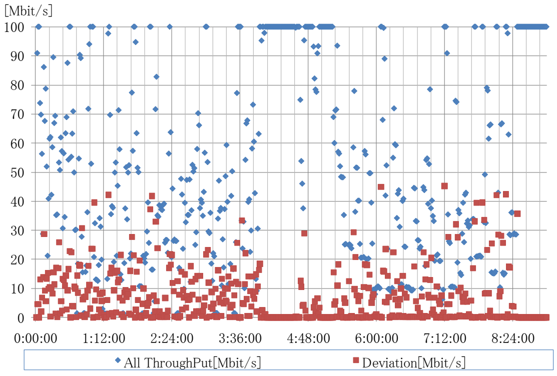

Figure 17. Link throughput

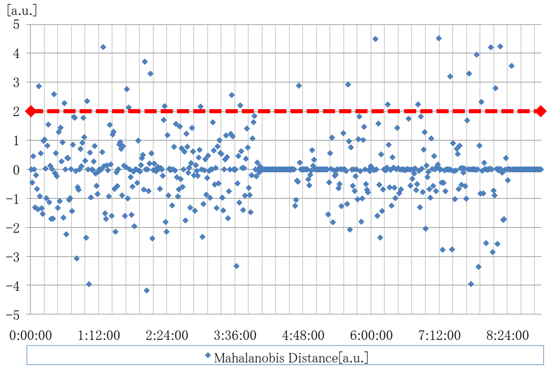

Figure 18. Mahalanobis distance with link throughput

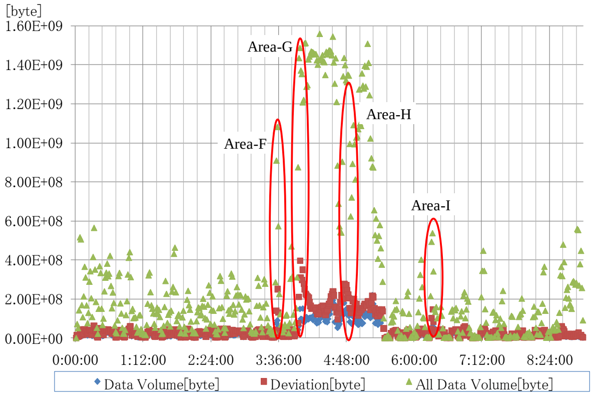

Figure 19. Data volume

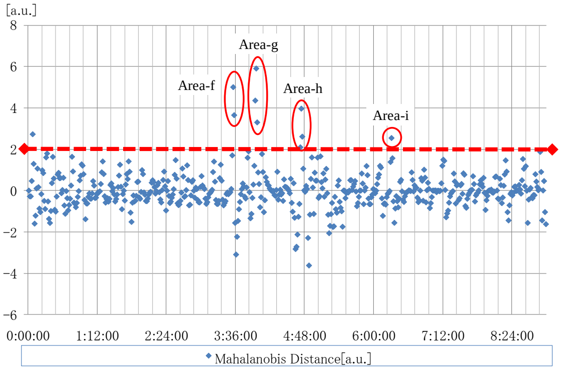

Figure 20. Mahalanobis distance with data volume

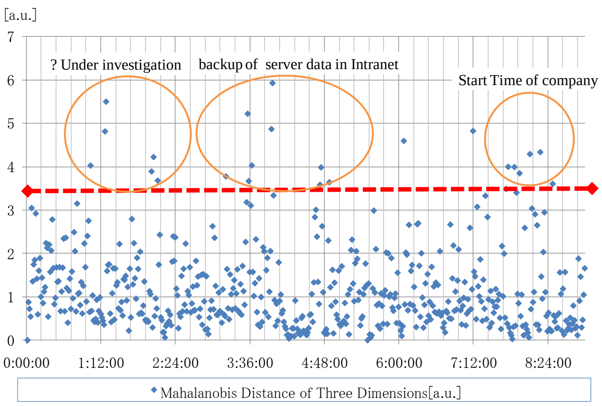

Figure 21. Mahalanobis distance with three dimensions

The measured link throughput is shown in Figure 17, and the calculated Mahalanobis distance with the link throughput is shown in Figure 18. It is difficult to pinpoint from Figure 17 the points at which a significant change of network-traffic condition was generated. On the other hand, significant changes are detected two or more times in Figure 18. However, there is no commonality between the detected changes in Figure 17 and the previously detected changes in Figures 14 and 16.

The measured data volume is shown in Figure 19. As shown from area F to area I in the figure, data volume rises rapidly in a certain time zone. Moreover, significant changes of network-traffic condition (i.e., data volume) were detected two or more times. The times of the change in data volume shown in Figure 19 is almost equivalent to those of the change in Mahalanobis distance shown in Figure 20. It is concluded that server data on the intranet should be backed-up at these times.

The calculated Mahalanobis distance converted into three dimensions is shown in Figure 21. In the calculation, three-dimensional data are selected from four types of monitored data (Throughput per flow, Packet size, Link throughput, data volume). This result bases on formula (2a) and (2b). Each of α,β,γ value used in this experiment is 1. α,β,γ can't be decided from theory because reasons of network fault are not always same and abnormal condition depends on system condition. So α,β,γ must be decided from past data and system condition, administrator experience. Currently how to decide α,β,γ is an issue in the future. Self-learning method could be used to decide those parameters.

The following focuses on from AM2:24 to AM4:48. As a result of setting the threshold to 80 Mbit/s and analyzing intranet traffic, the number of alarms exceeding the threshold value was detected 54 times in Figure 13. On the other hand, as shown in Figure 14, by proposed method, the number of alarms detected the significant changes of network- traffic condition is four times. Consequently, it is possible to reduce the number of alerts by 92%. Moreover, the one-dimensional judgment method was used to the four types of monitored data and the number of alarms exceeding the threshold value with Mahalanobis distance greater than the threshold of two is 14 times as sum of alarms with the four types of monitored data. When my proposed method was judged by the three-dimensional Mahalanobis distance, it was seven times, and the alarm decreases by half compared to one-dimensional judgment method. As a whole, the proposed method can reduce the number of alarms by 96% compared to one-dimensional judgment method.

VI. Concluding Remarks

A new network-monitoring system based on self-learning and multi-dimensional analysis (SLMDA) using the Mahalanobis distance was proposed. This system detects a significant change of network traffic. It uses Mahalanobis distance converted to multiple dimensions between standard distribution and measured distribution in real time. If the distance is larger than two standard deviations of standard distribution, the system judges that the monitored condition is abnormal. The system can therefore detect a rapid change of network traffic condition. A prototype network-monitoring system was developed and evaluated, When user datagram protocol (UDP) traffic is intentionally generated under the condition that only transmission control protocol (TCP) traffic is present even if the whole consumed bandwidth is not changed. As a result, the system could detect a significant change of network traffic. In addition to using real traffic, the system can reduce the number of unnecessary alerts by about 96% when throughput is fluctuating at normal rate of bandwidth consumption near a predefined threshold. As for future work, the proposed monitoring system will be extended to large-scale networks, and its performance will be evaluated.

Acknowledgment

Part of this research was supported by MIC (Ministry of Internal Affairs and Communications) as part of the "Research and development on network control technologies supporting cloud services for high availability and power saving" project

References

[1] T. Suzuki et al., "Power-Saving ICT Platform That Guarantees Network Bandwidth for Cloud-Service Systems," World Telecommunications Congress (WTC), 2012

[2] Xiao Wei et al., "A Network Monitor System Model with Performance Feedback Function" E-Business and Information System Security, 2009. EBISS '09. International Conference on Digital Object Identifier: 10.1109/EBISS.2009.5137879 Publication Year: 2009 , Page(s): 1 - 5

[3] I. Shimokawa et al., "Examination of network fault detection method by use of AFM," IEICE CPSY, computer system 110(473), 31-38, 2011-03-11.

[4] Y. Shomura et al., "Analyzing the Number of Varieties in Frequently Found Flow," IEICE Trans. Commun., vol. E91-B, no. 6, pp. 1896-1905, Jun. 2008.

[5] Mahalanobis, Prasanta Chandra (1936), "On the generalised distance in statistics," Proceedings of the National Institute of Sciences of India **2** (1): 49–55.

[6] http://code.google.com/p/coreemu