社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－DPS－122 (47)
2005－CSEC－ 28 (47)
2005／3／23

ディジタルフィルタの概念を用いたネットワークイベントの検出

張 鼎暉† 小出 和秀† 北形 元† Gleen Mansfield Keeni‡ 白鳥 則郎†

†東北大学大学院情報科学研究科/電気通信研究所
980-8577 宮城県仙台市青葉区片平 2-1-1

‡（株）サイバー・ソリューションズ
989-3204 宮城県仙台市青葉区南吉成 6-6-3 ICR ビル 3F

ネットワークトラフィックにおけるイベント検出は，ネットワークの効率的な運用管理のために重要である．本稿では，ディジタル信号処理の概念に基いたイベント検出手法を提案する．ディジタル信号処理技術を用いてネットワークイベントを検出する既存手法として，Deviation Scoreに基づく手法に注目する．この手法は他のものと比較し，イベント検出に要する時間コストが小さいという特徴がある．しかしながらこの手法は，3つの問題点，すなわち，(1) 検出の正確性，(2) 計算量，(3) 閾値が手動設定である点，を持つ．本研究はこれらの問題点を改善し，高い正確性，低い計算量，および閾値の自動設定を実現するイベント検出手法を提案する．本手法を利用することによって，ネットワークの効率的な管理が可能となることが期待される．本稿では，実トラフィックを用いた実験を通じ，提案手法の有効性を示す．

Detection of Network Events based on Digital Filtering

Ding Hui Zhang† Kazuhide Koide† Gen Kitagata†
Glenn Mansfield Keeni‡ Norio Shiratori†

†Graduate School of Information Sciences
Research Institute of Electrical Communication, Tohoku University
2-1-1 Katahira Aoba-ku, Sendai city, Miyagi, Japan ZIP code:980-8577

‡ ICE Bld. 3F 6-6-3 Minamiyoshinari Aoba-ku
Sendai city, Miyagi, Japan ZIP code:989-3204

Network traffic analysis and detection of Event are essential for efficient network administration. The main goal of this research is to provide useful and meaningful information to the network administrator. After a thorough investigation in the field of detecting network Event by digital signal technique, we focus on a method called Deviation Score. Deviation Score takes less time on anomaly detection than any other related methods, such as Statistical method, Markov Models and Neural Networks method. However, there are three problems with Deviation Score, which are (1) inexact detection result, (2) high computational complexity, (3) requirement of manually set threshold. In this research, we propose a new method of Event detection of network traffic, which is exact, real-time and automatic, i.e. no manual intervention is needed. Experiment results show that our proposed method can detect Event with more efficiently than Deviation Score, especially pertaining to the detection exactness point.

1 Introduction

In recent years, network system is more and more important to people and organization. However, network anomaly Event may cause heavy damage to network system. For an uninterrupted smooth use of network system, it is important to detect the network anomaly Event as quickly as possible. A network event can be a network illegal access, computer viruses, DDoS attack, network failures. This has prompt a wide range of of in-depth research on anomaly event detection. Until now, the most commonly known solutions are IDS (Intrusion Detection System). In IDS, there two variations. One is Misuse Detection, the other is Anomaly Detection. Misuse Detection can identify intrusions, based on a known pattern for the malicious activity. These known patterns are referred to as signatures. As Misuse Detection can't detect the new pattern, the chance of false negatives is high. Anomaly Detection can identify malicious traffic based on deviations from established normal network traffic patterns. Anomaly Detection technique takes time to adjust with the new pattern. In a situation where traffic pattern changes frequently, Anomaly Detection considers this apparently normal traffic as a deviation from the original and thus the number of false positive rises. Because Anomaly Detection can detect

network anomaly Event even the traffic pattern changes. we decide to focus on Anomaly Detection methods.

It is known that there are three phases in network administrators' tasks. The first phase is monitoring the network traffic, then detecting the Events from the traffic. The second phase is analyzing the traffic which has the Events information and checking whether there are really Events or not. The third phase is taking effective action to overcome the problems. In the first phase, there are two tasks. One is to collect the traffic data correctly and in real time, the other one is to isolate anomalous traffic from the original traffic data. For the first task, there are some effective solutions like SNMP and IP flow monitor. The second task is to detect the network anomaly Event as quickly as possible. But since network administration is a human controlled job, it is impossible for network administrators to spend all the time on monitoring network systems. They also have no time to plow through tones of network system log files. The goal of our research is to help network administrators to detect the network system anomalies and manage the network efficiently, and propose a method which can detect the network system anomaly Event with exactness, real-time, automation and with less computational effort.

This paper is organized as follows. In Section 2 we introduce some related methods especially the Deviation Score method. After explaining the problems of the Deviation Score in Section 2.2, we will explain our proposed scheme in Section 3. In Section 4 we present the results experiment and evaluation. We conclude this work in Section 5 and discuss about our future works in Section 6.

2 Related Methods

There are various methods to recognize and identify the anomalies, such as statistical method [1] clustering methods [2], neural networks [3] and Markov models [4] to recognize intrusions. In clustering method, they define the normal cluster by pattern recognition. If the unmoral pattern is detected, the anomalies will be detected. Neural network define the normal situations by AI (Artificial Intelligence). In any case, it is important to detect the Events as early as possible even the network normal situation changed extensively. But those methods consume too much time to decide their parameters once the network normal situations changed, they are impractical to serve the purpose.

Other than these methods mentioned above, we focus our attention on a method called Deviation Score [5]. We focus on Deviation Score is because it takes less time on anomaly detection than any other related methods, such as Statistical method, Markov Models and Neural Networks method. However, there are three problems with Deviation Score, which are (1) inexact detection result, (2) high computational complexity, (3) requirement of manually set threshold. In this research, we propose a new method of Event detection of network traffic, which is exact, real-time and automatic, i.e. no manual intervention is needed. The basic idea is to use the digital filter to isolate the Event data from the original traffic data and output the Events information, e.g. Event count and occur time. The most difficult point is how to establish a standards to define whether the traffic data is an Event or not. After all, there is no one have defined what the Event is. Maybe we should consider any changed in the traffic data to be an Event.

2.1 Introduction of Deviation Score

Deviation Score uses the Digital Signal Processing called wavelet system to analysis the traffic data. They use wavelet system because wavelets provide a powerful means for isolating characteristics of signals via a combined time-frequency representation. They analyzed a variety of traffic data by applying a lot of general wavelet filters. In their chosen system, there is one low-pass filter L and three high-pass filters H1, H2, H3. In their algorithm, to detect the Events from the collected traffic, they defined a simple nomenclature which was used to label Events as follows: Network anomalies, Attack anomalies, Flash anomalies and Measurement anomalies. Network anomalies: A network failure Event or temporary misconfiguration resulting in a problem or outage. Attack anomalies: Typically a Denial-of Service Event, usually flood-based. Flash anomalies: A flash crowd [6] Event. Measurement anomalies: An anomaly that they determined not to be due to network infrastructure problems nor abusive network usage.

From the four categories of traffic anomalies, we can see that network anomalies, attack anomalies and measurement anomalies are all short-term characteristics while flash crowd anomalies is long-lived characteristic. They want to analyze the traffic via their original algorithm, grouping the observed anomalies into four categories and outputting the Events classification information. Before applying the source data to their algorithm which they call a Deviation Score, they use the wavelet system to organize the source data into three stratas as follows: The Low frequency-part of the signal, the Mid frequency-part of the signal, the High frequency-part. At each part of the signal, they apply the output signal to the Deviation Score algorithm to detect the Events which grouped into 4 categories.

2.2 Problems of Deviation Score

Since the both methods can detect Events in real-time, we don't use the real-time as the evaluation standards. But we think the computational time will be a good point. In addition, from the result they provide, we confirm that their method works well on detecting the Events from the original traffic and cataloging the Events into 4 groups. But we have to manifest that it is very difficult to get a good detecting result unless we sniff out a suitable threshold. For network administrators, it is also important to detect the Event easy. So the problems are as follows:

(1)Inexact detection result

Because Deviation Score use the wavelet system to transform the traffic data, and use a moving-window

which calculate the variance of traffic to analyze the traffic, there must be a statistical error. To network administrators, it is important to know where, when, what happened in their network. So the exactness of accrual Events will be very important. When we use Deviation Score to analyze the traffic, we found there are some false positive, false negative and time misaligned problems in the detection results. We will interpret the results in Section 4.

(2) Huge computational Complexity

In Deviation Score, they use the wavelet system to extract the traffic into 3 binds: Low Frequency Bind, Mid Frequency Bind and High Frequency Band. The wavelet system's Complexity is order of n. That means the Deviation Score always requires n times volume if the input traffic multiplies one time. If the speed of input of traffic data is too quick not to processed in time such as in network backbone, the delay problem will be visible. Comparing with Deviation Score, our scheme's order is just 1.

(3) Threshold Configuration

In Deviation Score, the most difficult to decide is the detecting threshold. To find a suitable detecting threshold, we must do many experiments. That means the Deviation Score is not human-friendly which is manifested in the last of their thesis. A new process should be proposed which can decide the threshold or parameters automatically.

3 Proposed Scheme

We have demonstrated clearly the three problems in the Deviation Score at the end of Section 2. To overcome these problems, we propose a new method called Difference Calculus which can analyze the traffic data and detect the Events accurately, low complexity and efficiently (human-friendly) . The basic idea is to use digital filters. We have three evaluating metrics. The first one is to be detect the Events from the original traffic data as good as or better than the Deviation Score. The second one is to detect the Events with low complexity. The third one is to detect the Events automatically ,which is different from the Deviation Score.

Before showing our idea, the Event definition is " Event is a phenomenon that traffic swinging wildly in some ranges where a network administrator want to know. " Since network systems failures should effect on traffic, we think that the Event should be including Misuse and Anomaly.

To achieve the intended two goals, we propose to detect Events using digital signal processing. The basic idea is to use a digital filter because a digital filter has a characteristic that can isolate some abnormal signal from the original signal. To be able to isolate the Events which network operators want to know, we will show our proposed scheme outline as follows:

Step1. Filtering Raw Traffic data.

In this step, there are two processes: the first one is use high-pass filter to capture the variance of traffic. Because we need the increase traffic information. The second one is use low-pass filter to get rid of the hight frequency part of the output of high-pass filter, so as to capture the normal traffic pattern. Here, because we need the normal traffic information to detect long time Event, we have to add the moving-average of the output of high-pass filter to make a based situation as the normal traffic. The weight of low-pass filter and moving-average of the output of high-pass filter is α. And the result is the output of Decision Function $\mathbf{F(x)}$.

Step2. Event Detection Engine.

In this step, the goal is to design a decision function and a decision rule. Deviation Score also has the same goal. There are two processes: the first one is to design the Decision Function and the second one is to design the Decision Rules. Here, the based idea of filter design[7] is used as reference. There are two tasks in Event Detection Engine:

(1) Design the Decision function

In this process, the designed filters need to have the decision function that can capture the normal traffic.

(2) Design the Decision rules

In this process, one rule has to be established which can isolate the Events information from the original traffic data.

Step3. Auto Threshold Configuration.

In this step, the goal is to design a process which can calculate the parameters automatically. Deviation Score had attempted to do that, but they did not demonstrate clearly how to decide their threshold. We will clearly specify that how to decide parameters automatically. We use a feedback mechanism to set the thresholds automatically. In Deviation Score, the threshold is set by network manager manually.

3.1 Filtering Raw Traffic

In network, the traffic will be effected once Events occurs. So we must design the digital filters which can detect the variation of traffic. Let the time be $0,1,2 \cdots t$, the vectors $x_1, x_2, \cdots x_t$ be traffic data. First, we need to detect the variation of traffic by calculating the input data $H_t = \mid x_t - x_{t-1} \mid$. Among the variation of traffic, we must pass the little sharp variation of traffic so as to make the Detection Engine not go ballistic on detecting Event. We think the low pass filter would be a good choice because the characteristics of low pass filter. In low-pass filter, the parameter is ω_{cutoff} which means the cutting off frequency. The input data is H_t, then the output low-pass filterL_t can be calculated by

$$L_t = H_{t-1} - L_{t-1} * \omega_{cutoff} * \Delta t + L_{t-1}, 0 < \omega < 1 \quad (1)$$

At the same time, we also calculate the average of variation of traffic by using the function

$$A_t = \frac{\sum_{i=0}^{i=N} H_{t-i}}{N} \quad (2)$$

3.2 Event Detection Engine

We design our Decision Function F(x) as

$$\mathbf{F(x)} = \alpha L + (1-\alpha)A, 0 < \alpha < 1 \quad (3)$$

α is a mixing ratio and the Decision rule will be like

$$H_t > \frac{\mathbf{F(x)}}{\lambda}, 0 < \lambda < 1 \quad (4)$$

λ will be a parameter. The decision rule means that when the $H_t > \frac{\mathbf{F(x)}}{\lambda}$, we will interpret the x_i as an Event and record the generating time.

3.3 Auto Threshold Configuration

In the process of Event detection, we can sum up the total of Events. After changing the parameters including α, ω, λ, we can get a deferent of total of Events. In our experiments, we will show that if we change the parameters and print down the total of Events, the remain Events are stronger than those not elected. Based on this idea, the network administrators only need to decide the count of Events, the parameters will be confederated automatically and the Events which have the same S_t or bigger than S_d will be detected. Here the S_t means the Event Strength. We use the $\frac{H_t}{F(x)}$ to express the Event Strength.

$$\frac{H_t}{\mathbf{F(x)}} = \frac{\mid x_t - x_{t-1} \mid}{\alpha L + (1-\alpha)A} \quad (5)$$

Here variance of traffic $\mid x_t - x_{t-1} \mid$ will be changed if the traffic Event occurred. The output of Decision Function $\mathbf{F(x)}$ is the normal traffic and will not be changed Event the Event occurs. When a Event occurs, based on our filters system's frequency response (the designed filters can weaken sudden traffic increase), the output of $\mathbf{F(x)}$ will be not changed so much while the output of $\mid x_t - x_{t-1} \mid$ will be changed to bigger. The stronger a Event occurs, the bigger value of the S_t. So we use $\frac{H_t}{\mathbf{F(x)}}$ to explain the concept of Event Strength.

4 Experiments and Evaluation

Our experimental goal is to evaluate our proposed scheme with Deviation Score. In our experimental environment, we gather the traffic data from three networks. The first one is JGN2 network. In this network, there is not so much traffic data unless someone does a experiment such as video transfer experiment. So our goal is to detect the experiment Event. The second one is experimental network in shiratori lab. The network address is 130.34.38.128/26. Because this network is used to do experiments like mobile ipv6, the traffic data's patterns will be changed frequently. The third one is shiratori laboratory network. The network address is 130.34.209.128/26. Because this network is used to supply the mundane services like web, mail and file sharing service, the traffic data will reflect the mundane network's characteristics. For example, the traffic will have some patterns rhythmically.

4.1 Results of Experiments

We will evaluate three points which are (1) inexact detection result, (2) high computational complexity, (3) requirement of manually set threshold. There are three experiments as follow. We use the JGN2 traffic in the link which is from Tokyo University and Tsuruoka campus of Keio University on 2005/01/18. First we focus on the traffic from 0:00 am to 10:00 am. From 10:00 am, we can see there is a traffic Event until 11:00 am. In this Event, the max of traffic volume is about 3.5 kbps. Here, we could get the detecting results that the Figure 1 shows the result by Deviation Score while the Figure 2 shows the result using the proposed scheme.

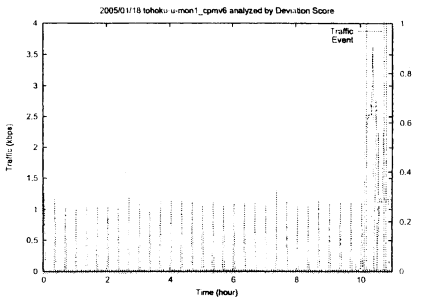

Figure 1: 0:00 - 10:00 analyzed by Devision Score

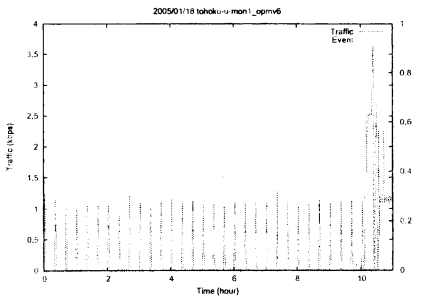

Figure 2: 0:00 - 10:00 analyzed by proposed scheme

But from 11:30 am, there is a traffic Event until 13:30 pm. In this Event, the max of traffic volume is about 250 Mbps. If we just use a simple threshold like "a Event occurs if traffic volume over 3.5 kbps (because we should detect the first Event from 10:00 am)", we will get a lot of Events between 11:30 am to 13:30 pm. Here, we could get the detecting results that the Figure 3 shows the result by Deviation Score while the Figure 4 shows the result using the proposed scheme.

With the detection result, we can see the proposed scheme can detect the Events exactly even the based traffic situation has been changed very high from 3.5 kbps to 250 Mbps while there is a time delay problem shown in the result via Deviation Score.

4.2 Computational Complexity

It is not possible to detect an Event just looking at the traffic volume some time. Since the wavelet's order is order(n), our proposal's order is order(1), the Deviation Score will take more computational power if we want

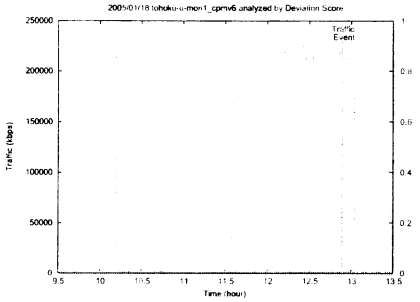

Figure 3: 9:30 - 13:30 analyzed by Devision Score

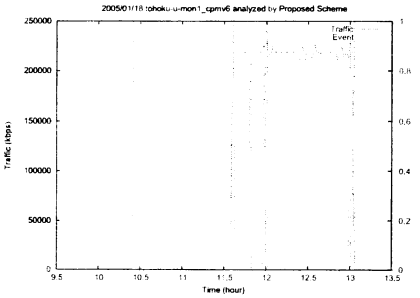

Figure 4: 9:30 - 13:30 analyzed by proposed scheme

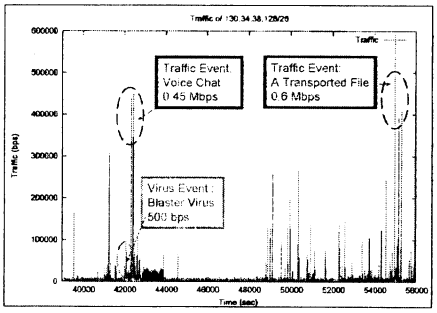

Figure 5: Traffic including virus

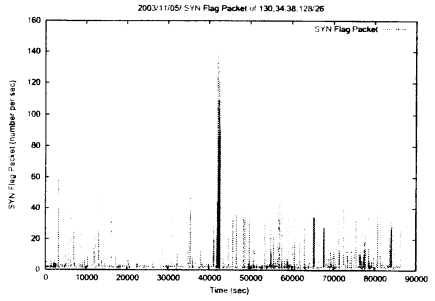

Figure 6: SYN Flag of Traffic including virus

to detect the variation information not only with the traffic volume, but also every protocol such as IP, TCP, UDP ,ICMP etc. and every application such as HTTP, FTP, SMTP, POP3 etc. Let's see an example.

In the Figure 5, we can see the Blaster virus's traffic volume is not bigger than the two traffic Events: Voice Chat and A file transported. In this situation, we think it is difficultly to detect the virus Event just only using the traffic volume. As we all know, the virus blaster a worm that exploits the DCOM RPC vulnerability using TCP port 135. Event the worm targets only Windows 2000 and XP machines, it generates an IP address and attempts to infect the computer that has that address. So we can use the Traffic Volume and SYN Flay Packets to detect this virus. Note: We accentuate the two elements because only use SYN Flay Packets can not determine there is a virus Event or not. Because if a lot of people access the web all together, the SYN Flag Packets will be increased, and the traffic volume too. But if the traffic volume doesn't increase while the SYN Flag Packets increases, the traffic should be suspected. The SYN Flag Packets is shown in the Figure 6. Additionally, we fess up that our scheme also can't detect the virus Event without SYN Flag packet information. Network administrator can find out the virus Event if monitoring the situation of SYN Flag packet all the while. We reaffirm that network administrator can detect the anomalies syn flag quickly if use the proposed scheme. Because our proposed scheme can filtering out the normal syn flag packets.

4.3 Threshold Configuration

Deviation Score has a process to decide the threshold by human so it will take a lot of man-hours. In our proposal, we have a process to decide the threshold by the feedback informations, so our proposal is fuss-free to network administrators. A sample is from shiratori lab 209. It is a normal network for shiratori lab. First, we analysis one day traffic. For example, set the Event count to be 2 and decide the parameters. Next, use the parameters to analysis one week traffic to test the detection result. For actual operating, network administrators don't need to know the parameters' volume.

In the Figure 7, we can see the detected Event is 2. Using the same parameters, we will find 30 Events in the week traffic analysis in the Figure 8. We will merge the shortcoming of Deviation Score and our proposal's characteristics in a chart as follows:

Table 1: Advantage and Drawback of DS and Our Scheme

Metrics	Deviation Score	Our proposal
Exactness	False positive	Reduced
	False negative	Reduced
	Delayed response	Prompt response
Complexity	Order of N	Order of 1
Setting Threshold	Manual	Automatic

In the result, we can see that our proposed scheme

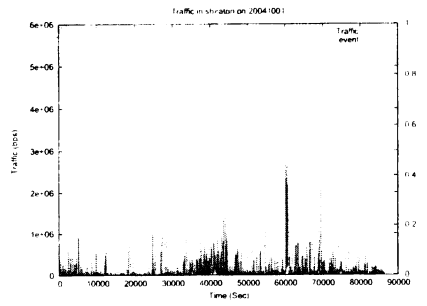

Figure 7: One day traffic analyzed by proposed scheme

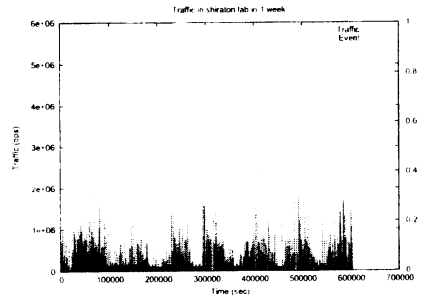

Figure 8: One week traffic analyzed by proposed scheme

can provide higher performance than Deviation Score in higher precision, lower complexity and less task.

5 Conclusion

For efficient network administrations, it is important to know when, where and what happened in their networks. That is why Event Detection algorithm deliberated over time. In this thesis, we introduce some related works especial Deviation Score which use Digital Signal Process. We notice it because it can detect Events from traffic informations with real-time while other method can't.

We also point out the shortcomings of Deviation Score. The most important is that it is difficult to define the threshold in Deviation Score. So network administrators have to decide the threshold by themselves. That means this method will spend network administrators a lot of time to find a fit threshold to detect good results. If they can't find a fit threshold, the detecting result should be false positive or false negative.

To overcome these problems, we proposal an better solution using digital signal processing too. We screw the digital filter to isolate the Event information from the original traffic data. To cut down network administrators' work time, we also use the feedback Event detecting result to implement automatic threshold configuration process. For understanding our proposal well, we clarify the parameters of effect on result of the detection with our proposal.

6 Future works

Our proposal has a simple process which can detect the variation of traffic quickly, so we think it is possible to get more particular Event information if we use the protocol data or application data as the input information. For example, if the total traffic volume is not changed while SYN Flag Packets increasing, we can figure out virus or attack. If the FTP traffic changes, we also can figure out if some one is downloading or uploading some files. Briefly, apply our proposal to every protocol or application is a topic for future works.

References

[1] Hassan Hajji,B. H. Far, and Jingde Cheng," Detection of Network Faults and Performance Problems, " *Internet Conference 2001.*

[2] J. Toelle and O. Niggemann," Supporting intrusion detection by graph clustering and graph drawing, " in Proceedings of Third International Workshop on Recent Advances in Intrusion Detection RAID 2000,Toulouse,France ,October 2000.

[3] K. Fox, R. Henning, J. Reed, and R. Simonian," A neural network approach towards intrusion detection, " Tech. Rep.,Harris Corporation,July 1990.

[4] N. Ye, " A markov chain model of temporal behavior for anomaly detection, "in *Workshop on Information Assurance and Security*,West Point, NY, June 2000.

[5] Paul Barford, Jeffery Kline, David Plonka and Amos Ron, A Signal Analysis of Network Traffic Anomales, *Internet Measurement Workshop 2002.*

[6] J. Jung, B. Krishnamurthy, and M. Rabinovich, " Flash crowds and denial of service attack: Characterization and implications for CDNs and web sites, " in WWW-02, Hawaii, May 2002.

[7] M. Basseville and I. V. Nikiforov," Detection of Abrupt Change: Theory and Application "

[8] JGN2 is an ultra-high-speed network for R&D, supported by TAO (NiCT) http://www.jgn.nict.go.jp