MATHEMATICAL MODELS OF IP TRACEBACK METHODS AND THEIR VERIFICATION

Keisuke Ohmori[1], Ayako Suzuki[1], Manabu Ohmuro[1], Toshifumi Kai[2], Mariko Kawabata[1] , Ryu Matushima[1] and Shigeru Nishiyama[1]

[1]NTT Advanced Technology Corp. Systems Development Unit,1-19-3, Nakacho, Musashino-shi, Tokyo, 180-0006, Japan; [2]Matsushita Electric Works, Ltd. Systems Technology Reserch Laboratory, 4-8-2, Shiba, Minato-ku, Tokyo 108-0014, Japan

Abstract: IP traceback is a technology for finding distributed-denial-of-service (DDoS) attackers. Various IP traceback methods have been proposed. When a new method is proposed, a performance comparison with the conventional methods is required. In this paper, mathematical models of ICMP, probabilistic packet marking, hash-based, and Kai's improved ICMP method are proposed. The mathematical models proposed can be applied to arbitrary network topologies, and are applicable for evaluating the performance of a new traceback. The mathematical models are verified by comparing the theoretical values with actual measurements of a network of about 600 nodes.

Key words: ICMP traceback,; Probabilistic packet marking traceback; Hash-based IP traceback; Mathematical model

1. INTRODUCTION

Distributed-denial-of-service attacks (DDoS attack) cause serious damage to the Internet community; programs for implementing such attacks are typically propagated using worms or viruses. Research and development to prevent such attacks is necessary.

IP traceback can look for the attack routes, even if the IP address of the attacker is forged. It is one technology that may be employed to defend a computer system from DDOS attacks. ICMP traceback[1], probabilistic

packet marking traceback[2], and hash-based traceback[3] are typical IP traceback methods; new traceback methods are also being proposed.

When a new traceback method is proposed, we need to compare the performance with the conventional IP traceback methods. Ideally, one would install the conventional IP traceback systems and evaluate the performance; however, the systems are difficult to install. Therefore, performance estimation by mathematical modeling becomes desirable.

The conventional mathematical models[4] apply to simple network topologies such as linear and binary trees. They are not applicable to arbitrary network topologies.

In this paper, we propose mathematical models of typical traceback methods: ICMP traceback (**iTrace**), probabilistic packet marking traceback (**PPM**), and hash-based traceback. We also propose a mathematical model for improved ICMP traceback method, which does not use probabilistic packet sampling. These models can be applied to arbitrary network topologies and we show the validity of the mathematical models by comparing them with actual measurements of a large-scale verification network.

This paper is organized as follows. We propose mathematical models of the IP traceback methods in Section 2. We present the verification method and a verification network in Section 3. We compare the theoretical values with actual measurements in Section 4. Finally we summarize our results and present areas for future research in Section 5.

2. THE MATHEMATICAL MODELS

2.1 Summary of IP traceback methods

First, we present an overview of IP traceback. An IP traceback looks for DDoS attackers by examining the flow of attack packets. An agent of the IP traceback is sent to each router. It generates traceback information, which includes the packets that pass through the router. This traceback information is sent to a collector and is used for the traceback.

An example of a traceback is shown in Fig. 1. V is a victim, and A1, A2, A3, A4 are attackers.The attackers A1, A2, A3, A4 attack the victim V through routers. For example, attack packets from the attacker A1 reach the victim through the edge e6, e3, and e1. Therefore, a trace back to the attacker A1 becomes possible when traceback information about e1, e3, and e6 is generated.

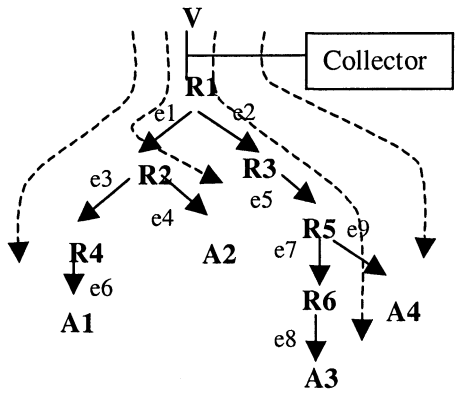

Figure 1. Example of IP traceback

IP traceback methods may be divided into those that use probabilistic packet sampling, such as ICMP and probabilistic packet marking, and those that do not, such as hash-based traceback and improved ICMP traceback.

In ICMP and probabilistic packet marking, traceback information is generated probabilistically for packets, both normal and attack packets. Therefore, the discovery probability of the attackers can be calculated from the generation probability of the traceback information about each edge of the attack routes. For example, the discovery probability Pr (A1 ∩ A2 ∩ A3 ∩ A4) of the four attackers A1, A2, A3, A4 can be calculated with the Eq. (1), using the probability that traceback information is generated for the edge ei.

$$\Pr(A1 \cap A2 \cap A3 \cap A4) = \prod_{i=1}^{9} \Pr(ei) \tag{1}$$

The hash-based traceback and improved ICMP traceback methods are not probabilistic packet sampling methods. They have no collector; a manager controls the agents on the routers.

In hash-based traceback, an agent of each router calculates the hash value of every packet and registers it in a hash table. The manager asks agents whether they routed an attack packet using the hash value of the attack packet. The attackers can be found using the answers from the agents.

In improved ICMP traceback, the manager sends a request packet asking the router to inform it if it routes an attack packet. The agent informs the

manager when the attack packet is routed. The manager decides where to send the next request packets using the link information of the routers.

2.2 ICMP traceback

2.2.1 Outline

An agent which generates traceback information is installed in each router. A traceback information collector is placed just before the victim. The agent generates an iTrace packet with a probability p (usually, one in 20000) for packets destined for the victim. The iTrace packet includes the original packet, and the collector looks for the attackers.

2.2.2 Implementation

Implementation is based on the Internet-Draft[1]. In a normal ICMP traceback, a single iTrace packet including an attack packet means that the attack packet was routed down the edge. To reduce false positives, it was decided that two iTrace packets must include an attack packet before attributing the attack to the edge.

2.2.3 Mathematical model

We find the probability that the agent of a router generates two iTrace packets including the attack packet on an edge. Let p be the probability of generating an iTrace packet and N be the number of attack packets arriving on edge ei; then the probability of generating two or more iTrace packets becomes

$$Pr(ei) = F(N) = 1-(Np(1-p)^{N-1} +(1-p)^{N}) \quad (2)$$

Here $(1-p)^{N}$ is the probability that no iTrace packets are generated and $Np(1-p)^{N-1}$ is the probability that only one iTrace packet is generated. We can calculate the discovery probability $Pr(\prod Aj)$ of all the attackers by using the iTrace packet generation probability on each edge ei of the attack routes.

$$Pr(\prod_{j=1}^{n} Aj) = \prod_{i=1}^{m} Pr(ei) \quad (3)$$

Changing the number of packets N in Eq. (2) and Eq. (3) allows us to calculate the discovery probability for different attack scenarios. Traceback time is calculated from the number of packets found with the formula Eq. (3).

For example, we may apply Eq. (2) and Eq. (3) to the scenario illustrated in Fig. 1. In Fig. 1, the attackers A1, A2, A3, A4 carry out a DDoS attack; each sends *a* attack packets per second. We may calculate the probability that the attackers are discovered after *t* seconds as shown in Table 1. Here the number of packets on edges e1, e2, e5 is twice that of the other edges because two edges join into one. For example, suppose that each attacker sends 1000 attack packets per second. In this case, the traceback takes 96 seconds for the discovery probability to reach 95% for A1, A2, A3, and A4.

Table 1. Example for how to calculate the discovery probability of the attackers in ICMP traceback

Edge	Number of attack packets after t seconds	Probability that two or more iTrace packets are generated at each edge
e1, e2, e5	2at	F(2at)
e3, e4, e6, e7, e8	at	F(at)
the discovery probability of attackers A1, A2, A3, A4		$F(2at)^3*F(at)^5$

2.3 Probabilistic packet marking traceback

2.3.1 Outline

The agent, which marks routed packets, is installed in each router. The collector, which collects marked packets, is arranged just before the victim. A hash value for a packet is stored at each router with probability p (usually, 1/20). The collector can look for the attackers using marked packets sent to the victim.

2.3.2 Implementation

Implementation is based on Song and Perrig's AMS-II(Advanced and Authenticated Marking Scheme-II). The probabilistic packet marking traceback evaluated here divides the 64-bit hash value into 8 individual fragments; one of these is chosen at random and marked. The collector considers an attack packet to have been routed when the 64-bit hash value (16 marked packets) arrives twice.

2.3.3 Mathematical model

We assume that the *d* individual routers are found in a direct route between the attacker and the victim. First, a router Ri marks a packet, and the probability that the other routers do not rewrite the marked packet is calculated.

Figure 2. For mathematical model computation of PPM

For example, in Fig. 2, if router R1 marks one packet, the probability that it is not marked by other routers is $p\,(1-p)^{d-1}$. The hash value is divided into 8 individual fragments and one of those is sent at random. Therefore, the generation probability of a marked packet is p/8. One mark is generated, and the probability F_d, that the mark is not rewritten by other routers, becomes

$$F_d=p(1-p)^{d-1}/8. \tag{4}$$

The probability Pr(ei), that two or more marked packets arrive on edge ei as set of eight individual fragments is

$$Pr(ei)= G(d, N) = (1-(N*F_d(1-F_d)^{N-1}+(1-F_d)^N))^8 \tag{5}$$

Here N is the number of packets, $N*F_d\,(1-F_d)^{N-1}$ is the arrival probability of one marked packet, and $(1-F_d)^N$ is the probability that no marked packet reaches the collector. By deducting the two values from 1, the probability that two or more marked packets arrive at the collector can be calculated. It is raised to the eighth power because eight fragments are necessary for the traceback. The discovery probability of the attackers and traceback time can be found with Eq. (3) in the same way as for ICMP traceback.

We may apply Eq. (3) and Eq. (5) to the scenario illustrated in Fig. 1. In an IP marking system, there are routers which may rewrite marked packets between a router to mark and the victim, unlike the ICMP system. Let *a* be the number of attack packets per second from A1, A2, A3, and A4. The discovery probabilities of the attackers after t seconds are shown in the Table 2.

For example, suppose that each attacker sends 25 packets / sec; then the traceback time takes 65 seconds for the discovery probability to reach 95% for A1, A2, A3, and A4.

Table 2. Example for how to calculate the discovery probability of the attackers in probabilistic packet marking traceback

Edge	d	Number of routed packets after t seconds	Probability of generating two or more marked packets on each edge
e1,e2	1	2at	G(1,2at)
e5	2	2at	G(2,2at)
e3,e4	2	at	G(2,at)
e6,e7,e9	3	at	G(3,at)
e8	4	at	G(4,at)
Discovery probability of A1,A2,A3,A4		$G(1,2at)^2*G(2,2at)*G(2,at)^2*G(3,at)^3*G(4,at)$	

2.4 Hash-based traceback

2.4.1 Outline

The agent of each router registers the hash value of every packet. The manager asks the agents whether attack packets with the same hash value were routed through each router. An example of hash-based traceback is shown in Fig.3. In this example, the manager is tracing A3. The agent of each router returns the answer to manager's inquiry by yes or no.

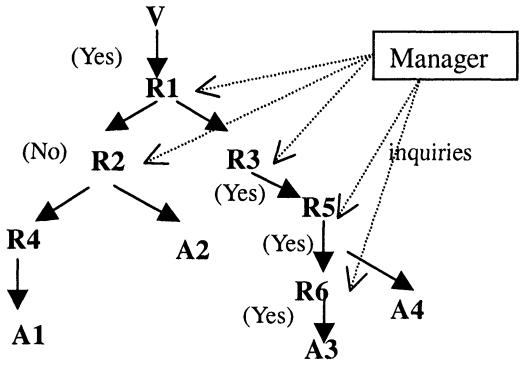

Figure 3. Example of Hash-based traceback

2.4.2 Implementation

We use the system of BBN technologies. Because of hash collisions, we let the hash table size be 14 bits in the evaluated system.

2.4.3 Mathematical model

The traceback time depends on the total number of inquiries from the manager. The number of inquiries increases very much as the attackers increase. It is difficult to derive the mathematical model because the traceback is done by the parallel processing. Therefore, the regression analysis was done based on the measurement data this time.

The measurement data and regression analysis are shown in Fig.4. The measurement data was obtained by the verification item shown in paragraph 4.2 . The number of attackers is changed from 1 to 100.

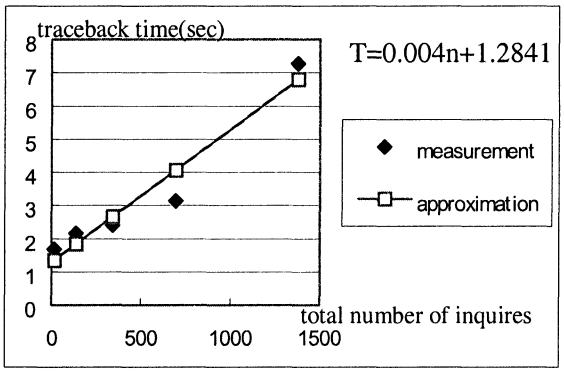

Figure 4. Regression analysis of hash-based traceback time

2.5 Improved ICMP traceback

2.5.1 Outline

Improved ICMP traceback places an agent on each router and has a manager. We show the flow of improved ICMP traceback in Fig. 5. First, the manager sends a traceback request to an agent on the router just before the victim. The traceback request has information about the attack packet. When the agent detects the requested attack packet, it generates a uTrace packet, which includes link information, and sends the uTrace packet to the manager. The manager receives the uTrace packet and sends a traceback request to the routers which link to the previous router. The manager then repeats these transactions.

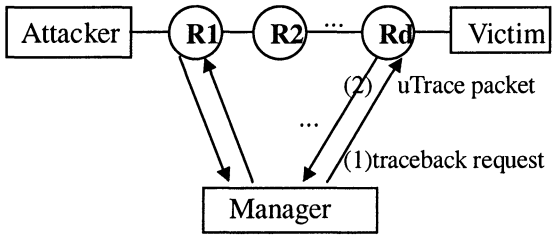

Figure 5. The flow of improved ICMP traceback

2.5.2 Implementation

An agent usually generates an uTrace packet when one attack packet comes on an edge. An exception is the edge just before an attacker. In this case, the agent generates an uTrace packet when two attack packets come on the edge to prevent false positives.

2.5.3 Mathematical model

The traceback time T of an attack route is the following.

$$T = \sum T_{wait(i)} \tag{6}$$

Here $T_{wait(i)}$ is the waiting time that an agent waits for an attack packet on an edge ei. The traceback time is the sum of the waiting times of each router on the attack route. When there are many attack routes, the traceback time is the one route with the longest traceback time.

In Fig.1, when A1, A2, A3, A4 attack with 50 packets per second, the traceback of the attack route from A3 takes the most time. The average waiting time of an attack packet on edges e2, e5, e7 and e8 are 100ms, 100ms, 200ms, 200ms × 2, respectively. The average traceback time is 800ms.

3. VERIFICATION METHOD AND NETWORK

3.1 Verification method

In order to verify our mathematical model against actual measured values, we constructed a verification network. The network had 600 nodes. The network has the following parameters:

- The number of the hops between the victim and the attacker
- The number of attackers, arranged at random.

In the measurement of the verification network, the traceback time was measured when the false negative rate and the false positive rate were within 5% of each other.

3.2 Verification network

The specification of the verification network is shown in the Table 3.

Table 3. The specification of verification network

Item	Specification
Network scale	300 servers, 600 nodes
OS	Linux
Network speed	100Mbps
The number of DoS/DDoS attackers	at most 100
Attack packet amount per a machine	at most 25000 packet/sec
Network topology	a mesh at the core, trees at the edges

When a large-scale network is constructed, we must consider the cost, the setting, securing a power supply, and the problem of heat in the room. In this research, these problems are solved by virtual OS technology. We selected UML (User Mode Linux)[5] because the specified OS is Linux. We assigned 32 MB to each virtual OS. The maximum number of virtual servers is six. We chose not to use the virtual OS for the router because of the limitation of the input and output interface speed of the PC. The structure of the verification network is shown in Fig. 6. We had 380 servers and clients and 110 routers. Each router was a PC router using zebra[6]. The number in parenthesis is the actual number of PCs.

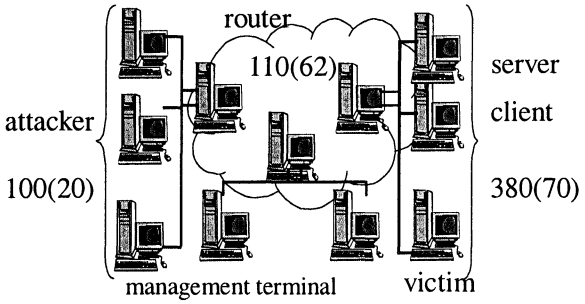

Figure 6. The structure of the verification network

4. EVALUATION

4.1 The number of the hops

We measured the traceback time of a linear topology by changing the number of hops; we used the values 1, 3, 5, 10, 15, and 20. The measurement conditions are shown in Table 4. The number of trials was decided from both the standard deviation in measurement values and the test efficiency.

Table 4. The measurement condition

Traceback method	Number of hops	Number of attack packets per second	Number of trials
ICMP	1,3,5,10,15,20	1250pps	10
PPM	"	50pps	100
Hash	"	50pps	5
Improved ICMP	"	50pps	60

We evaluate the relationship between the number of the hops and traceback time. Actual measurement values and the theoretical values from the mathematical model are shown in the Fig. 7. Character M and T in parentheses mean the measurement values and the theoretical values respectively. The theoretical value is about the same as the actual measurement value. The hash-based and improved ICMP traceback are different from tracebacks where traceback information is generated probabilistically, such as ICMP and PPM traceback. They are very fast, because they can trace back with the arrival of a single attack packet.

Traceback time for these is under 2 seconds and increases with the number of the hops.

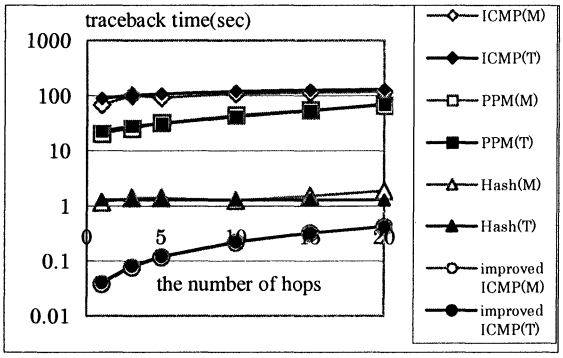

Figure 7. The relation between the number of hops and the traceback time

4.2 The number of attackers

We also measured the change in traceback times in relation to the number of attackers. The numbers of attackers were 1, 10, 20, 50, and 100. We fixed the total number of attack packets sent to the victim. The measurement conditions are shown in Table 5; the topology of the verification network is shown in Fig. 8.

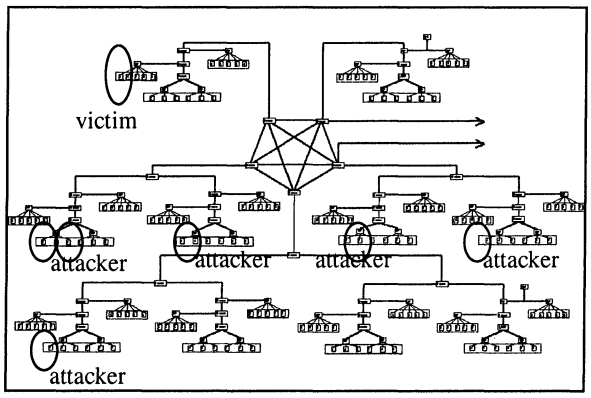

Figure 8. Topology of the verification network

Table 5. The measurement condition

Traceback method	Number off attackers	Total number of attack packets per second	Number of trials
ICMP	1,10,20,50,100	25000pps	10
PPM	"	1000pps	60
Hash	1,10,25,50,100	50pps	5
Improved ICMP	10,20,50,100	100pps	60

The actual measurement values and the theoretical values from the mathematical model are shown in Fig. 9. Character M and T in parentheses mean the measurement values and the theoretical values respectively. The theoretical values are about the same as the actual measurement values. The traceback time for ICMP traceback with 100 attackers is not shown because it took longer than the measurement time limit of 10 minutes.

In the Hash-based and improved ICMP traceback, traceback times were almost the same even when the number of attackers changed, because the maximum number of hops is about the same even if the number of attackers is different.

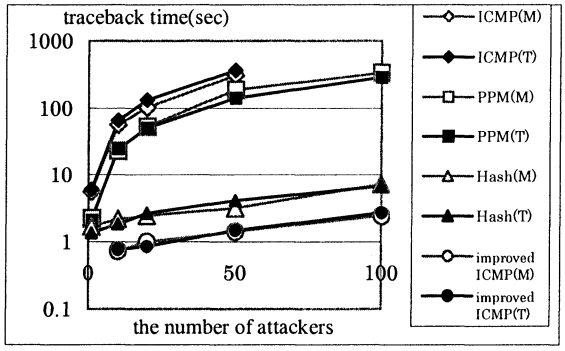

Figure 9. The relation between the number of attackers and the traceback time

5. SUMMARY AND FUTURE RESEARCH

- Mathematical models of IP traceback

We proposed mathematical models of ICMP, probabilistic packet marking, hash-based and improved ICMP traceback methods. In improved ICMP traceback, the manager sends a request packet to inform it if an attack

packet is routed through each router. The agent informs the manager when the attack packet gets routed. We constructed a verification network to verify the mathematical models. We evaluated the relationship between the number of hops and traceback time and the relationship between the number of attackers and traceback time. We confirmed that the theoretical values of the mathematical models are almost same as the values actually measured, and the improved ICMP performance is about the same as hash-based traceback.

The models can be applied to any network topology; therefore, the models can be used for the performance comparison of a new traceback model. The models can also be used to predict the performance of a typical traceback.

- The large-scale verification network

The verification network had 600 nodes made by 152 PCs. We used a virtual OS to increase the number of "machines." In order to prevent performance decline, we chose to use PC routers rather than virtualize.

- Future research

This time, we verified the mathematical models on a verification network with one autonomous system (AS). We plan to evaluate them on a verification network which has multiple ASs.

ACKNOWLEDGEMENTS

National institute of information and Communications Technology (NICT) funded this research (2002 - 2005).

REFERENCES

1. Steven M. Bellovin, "ICMP Traceback Message", Internet Draft, Oct. 2001, http://mark.doll.name/i-d/itrace/obsolete
2. Dawn Xiaodon Song, Adrian Perrig, "Advanced and Authenticated Marking Schemes for IP Traceback", IEEE INFOCOM 2001, http://vip.poly.edu/kulesh/forensics/docs/advancedmarking.pdf
3. Alex C. Snoeren et al., "Hash-Based IP Traceback", Proc. of the ACM SIGCOMM conference 2001, San Diego, CA, Computer Communication Review Vol. 31, No 4, October 2001. http://nms.lcs.mit.edu/~snoeren/papers/spie-sigcomm.pdf

4. Vadim Kuznetsov, Andrei Simkin, Helena Sandstrom, "An evaluation of different IP traceback approaches", ICICS, 2002, 37-48
5. The User-mode Linux Kernel Home Page. http://user-mode-linux.sourceforge.net
6. Zebra Home Page, http://www.zebra.org