POSSIBLE INTRUSIONS' IP TRACE-BACK IN CLOUD COMPUTING ENVIRONMENT

Swapan Debbarma[1], Anupam Jamatia[2], Nikhil Debbarma[3], Kunal Chakma[4]

Department of Computer Science and Engineering
NIT, Agartala, India
swapanxavier@gmail.com, anupamjamatia@gmail.com,
nikhildb@rediffmail.com, kchax4377@gmail.com

ABSTRACT

Defending against distributed denial-of-service attacks is one of the hardest security problems on the Internet today. One difficulty to-wards these attacks is to trace the source of the attacks as the attackers intentionally use spoofed IP source addresses to disguise from the true origin. The IP Trace-back in cloud environment is like an Advanced Marking Scheme and the Authenticated Marking Scheme that evolved from the probabilistic packet marking scheme (PPM), which allow the victim to trace-back the approximate origin of spoofed IP packets. The techniques feature low network and router overhead, and support incremental deployment. In contrast to previous works, our techniques have significantly higher precision (lower false positive rate) and lower computation overhead for the victim to reconstruct the attack paths under large scale virtual and distributed denial of-service attacks. Furthermore the Authenticated Marking Scheme provides efficient authentication of routers' markings such that even a compromised router cannot forge or tamper markings from other uncompromised router. The aim is to prevent the network from attackers by reconstructing the attacking path.

Keywords : Component; formatting; style; styling; insert.

I. INTRODUCTION

DENIAL-OF-SERVICE (DoS) attacks pose an increasing threat to today's Internet . Even more concerning, automatic attacking tools (such as Tribal Flood Network (TFN),TFN2K, Trinoo, and stacheldraht) allow teenagers to launch widely distributed denial-of-service (DDoS) attacks with just a few keystrokes. Just to name one of the many cases, in February 2000, several high-profile sites including Yahoo, eBay, and Amazon were brought down for hours by DDoS attacks. And real DDoS attacks are often mounted from hundreds or even thousands of hosts. A serious problem to fight these DoS attacks is that attackers use spoofed IP addresses in the attack packets and hence disguise the real origin of the attacks. Due to the stateless and cloud nature of the Internet, it is a difficult problem to determine the source of spoofed IP packets, which is called the *IP traceback* problem in cloud environment. One promising solution, recently proposed by Savage et al., is to let routers probabilistically mark packets with partial path information during packet forwarding. The victim then reconstructs the complete paths after receiving a modest number of packets that contain the marking. We refer to this type of approach as the *IP marking approach by PPM*. From this marking various other more efficient marking schemes were discovered namely distributed packet marking, router based marking, advanced and authenticated marking, adaptive probabilistic marking and many others. However all of the modifications were done on the probabilistic packet marking (PPM). Thus we can say that the evolution started from PPM.

A. *What is Cloud Computing?*

Cloud Computing is a nebulous term covering an array of technologies and services including; Grid Computing, Utility Computing, Software as a Service (SaaS), Storage in the Cloud and Virtualization. There is no shortage of buzzwords and definitions differ depending on who you talk to. The common theme is that computing takes place 'in the cloud' - outside of your organizations network.

B. *Why Cloud Computing?*

Cloud Computing is not mysterious, although it is the highest of high technology. Internet based applications appear daily, such as online slide show creation, online design tools, online mind mapping and file conversion, collaborative software, social media, etc. Some applications are easy to use and others are wanting. Communicators working in high tech are familiar with Cloud Computing, but practitioners in other disciplines might not be. Perhaps the key question for the rest of us is why one would move from PC-based applications to internet-based software? One answer to that is some applications like social media are

only on the internet while others make work easier. A second answer is that it is inevitable communicators will transition to internet-based applications. There are technological, economic and communications reasons for why Cloud Computing is becoming common. Technologically, we use Cloud Computing because we can. Economically, there is less expense, and finally, it makes interactivity easier to achieve with target audiences.

II. MODEL OF DDoS ATTCAK TREE

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math's. etc.

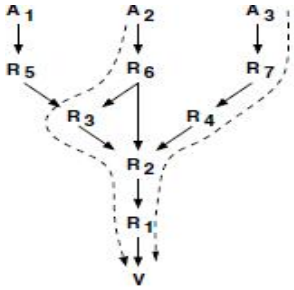


Fig 1 directed acyclic graph showing DDoS attack

The directed acyclic graph (DAG) rooted at V in figure 1 represents the network as seen from a victim V and a distributed DDoS attack from A_2 and A_3. V could be either a single host under attack or a network border device such as a firewall representing many such hosts. Nodes R_i represent the routers, which we refer to as *upstream* routers from V, and we call the graph the *map of upstream routers from* V. For every router R_i, we refer to the set of routers that immediately before Ri in the graph as the *children* of R_i, e.g. R_3;R_6 and R_4 are R_2's children. The leaves $\{A_i\}$ represent the potential *attack origins*, or *attackers*. The *attack path* from A_i is the ordered list of routers between A_i and V that the attack packet has traversed, e.g. the two dotted lines in the graph indicate two attack paths: (R_6;R_3;R_2;R_1) and (R_7;R_4;R_2;R_1). The *distance* of R_i from V on a path is the number of routers between R_i and V on the path, e.g. the distance of R_6 to V in the path (R_6;R_3;R_2;R_1) is 3. The *attack graph* is the graph composed of the attack paths, e.g., the attack graph in the example will be the graph containing the two attack paths (R_6;R_3;R_2;R_1) and (R_7;R_4;R_2;R_1). And we refer to the packets used in DDoS attacks as *attack packets*. We call a router *false positive* if it is in the reconstructed attack graph but not in the real attack graph. Similarly we call a router *false negative* if it is in the true attack graph but not in the reconstructed attack graph. We call a solution to the IP traceback problem robust if it has very low rate.

III. EXISTING PACKET MARKING SCHEMES

The existing packet marking schemes introduced so far are

A. Probabilistic Packet Marking (PPM)
B. Deterministic Packet Marking(DPM)
C. Advanced and Authenticated Packet Marking (APM)
D. TTL based Packet Marking (TPM)
E. Router based Packet Marking(RPM)
F. Nouvel Packet Marking
G. Others

IV. EVOLUTION OF VARIOUS MARKING SCHEME

It has been found that all the packet marking schemes excluding PPM has been formed or formulated by modification of PPM. Thus we can say that probabilistic packet marking is the basic for all other marking scheme. The evolutionary tree can be shown by the diagram below.

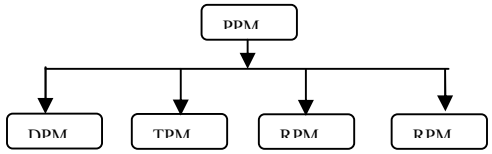


Fig 2 Evolution tree of various marking scheme

These evolution took place to make the marking scheme more efficient and robust. PPM was improved to other marking scheme to overcome the limitations and to trace the attack path faster.

V. PERFORMANCE COMPARISON AMONG VARIOUS MARKING SCHEMES

	PPM	TPM	DPPM	APM
Reaching Probability	Uneq ual	Unequal when deployed in the internet	Equal but is not 1/d, when an attacker craftily spoofs initial TTL values of packets	1/d
Unmarked Probability	>0 (e.g.5 2%)	Greater than 0(e.g.30%) when an attacker craftily spoofs initial TTL values of packets	Greater than 0 (e.g.48%) When an attacker craftily spoofs initial TTL values of packets	0

VI. PROBABILISTIC PACKET MARKING (PPM)

Probabilistic packet marking (PPM) was originally suggested by Burch and Cheswick and was carefully designed and implemented by Savage *et. al.* to solve the IP trace-back problem which can be stated as follows: given a stream of packets arriving at a receiver, identify the source of these packets and the path they took through the network. However, it is apparent that PPM is a general technique (beyond IP trace-back) to communicate internal network information to end-hosts. The basic idea of PPM can be explained using the illustration in Fig.3.Consider traffic flowing on an Internet path from source S_2 to destination D along the path $S_2 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow D$. A subset of the routers in the path has some local information that needs to be communicated to the destination R. (In the figure all

X	V			

routers in the path have some local information that need to be conveyed.) In order to communicate this information, a PPM scheme sets aside a few bits (PPM bits) in the header of IP packets. In the figure we assume that the number of such available bits is 4. Based on its local information, each router transforms the value of these bits as they pass through. The destination infers the local information at intermediate routers using the value of the PPM bits conveyed using a sequence of such IP packet.

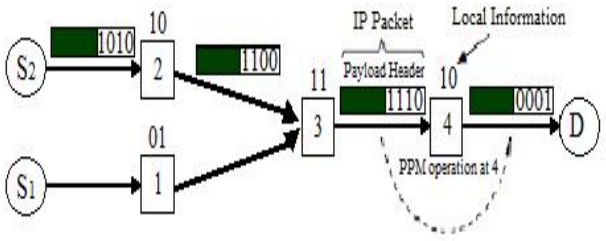


Fig 3:An Example Of Probabilistic Packet Marking

VII. MARKING BY EDGE SAMPLING

Marking in PPM is done by edge sampling algorithm. The basic idea of the IP marking approach is that routers probabilistically write some encoding of partial path information into the packets during forwarding. A basic technique, the *edge sampling algorithm*, is to write *edge* information into the packets This scheme reserves two static fields of the size of IP address, *start* and *end*, and a static *distance* field in each packet. Each router updates these fields as follows. Each router marks the packet with a probability When the router decides to mark the packet, it writes its own IP address into the start field and writes zero into the distance field. Otherwise, if the distance field is already zero which indicates its previous router marked the packet, it writes its own IP address into the end field, thus represents the edge between itself and the previous routers. Finally, if the router doesn't mark the packet, then it always increments the distance field. Thus the distance field in the packet indicates the number of routers the packet has traversed from the router which marked the packet to the victim. The distance field should be updated using a saturating addition, meaning that the distance field is not allowed to wrap. The mandatory increment of the distance field is used to avoid spoofing by an attacker. Using such a scheme, any packet written by the attacker will have distance field greater than or equal to the length of the real attack path. The victim can use the edges marked in the attack packets to reconstruct the attack graph.

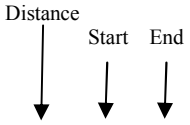


Fig-4:Edge Sampling

limitations of PPM (probabilistic packet marking)

In order to use the 16-bit IP Identification field to store the IP markings, we need an encoding scheme to reduce the storage requirements in each packet. The PPM encoding scheme splits each router's IP address and redundancy information into eight fragments and probabilistically marks the IP packet with one of the eight fragments .This encoding scheme works well with just a single attacker. But in case of a distributed denial-of-service attack, PPM suffers from two main problems: High computation overhead, because it needs to check a large number of combinations of the fragments, Large number of false positives, because the redundancy check is insufficient and the false positives at a closer distance to the victim can cause even more false positives further away from the victim. For example, even in case of a DDoS from 25 distributed attacker sites, PPM takes days to reconstruct the attack graph and results in thousands of false positives. PPM also suffers from the fact that it is not robust against a compromised router. Even worse, a victim cannot even tell that a router has been compromised merely from the information in the packets received. The main challenge is to design an efficient, accurate, and authenticated encoding scheme for IP marking that only uses the 16 bits available from the IP identification field.

VIII. EVOLUTION OF ADVANCED AND AUTHENTICATED MARKING SCHEME FROM PPM

Advanced Marking schemes ,in which new encoding schemes that are efficient and accurate are used even for DDoS attacks originating from over 1000 simultaneous attackers. We observe that if the victim knows the map of its upstream routers, it does not need the full IP address in the packet marking to reconstruct

the attacking graph, and hence the marking scheme can be more communication and computation efficient. we assume the victim has a map of its upstream routers.

Advanced marking scheme

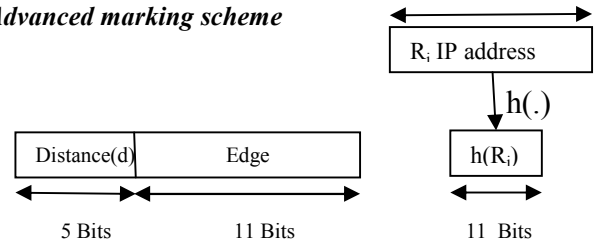
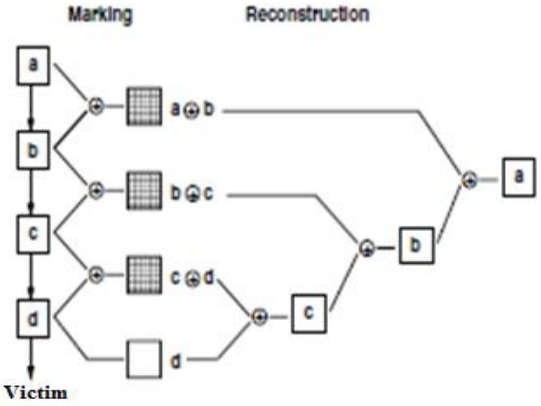


Fig-5: Encoding in advanced marking scheme

In this scheme, we divide the 16-bit IP Identification field into a 5-bit *distance* field and a 11-bit *edge* field. Note that 5 bits can represent 32 hops which is sufficient for almost all Internet paths Marking. Figure 5 describes the marking procedure of Advanced Marking Scheme. Note that we actually use two independent hash functions, h and h0, in the encoding of the routers' IP addresses. h and h0 both have 11-bit outputs. Every router marks a packet with a probability q when forwarding the packet. If a router R_i decides to mark the packet P, it writes $h(R_i)$ into the edge field and 0 into the distance field in packet P. Otherwise, if the distance field is 0 which implies its previous router has marked the packet, it XORs $h0(R_i)$ with the edge field value and overwrites the edge field with the result of the XOR. The router always increments the distance field if it decides not to mark the packet. The XOR of two neighbouring routers encode the edge between the two routers of the upstream router map. The edge field of the marking will contain the XOR result of two neighbouring routers, except for samples from routers one hop away from the victim. Because a _ b _ a = b; we could start from markings from the routers one hop away from the victim, and then hop-by-hop, decode the previous routers, as shown in figure . The reason to use two independent hash functions is to distinguish the order of the two routers in the XOR result

Reconstruction of attacking path

To reconstruct the attack paths, the victim uses the upstream router map Gm as a road-map and performs a breadth-first search from the root. Let's denote the set of edge fields marked with a distance d as d (do not include duplicates). At distance 0, the victim enumerates all the routers one hop away from itself in Gm and checks which routers have the hash value of their IP addresses, h(Ri), matched with the edge fields in 0, and denotes the set of matched IP addresses as S0.Therefore S0 is the set of routers one hop away from the victim in the reconstructed attack graph. Sd denotes the set of routers at distance d to the victim in the reconstructed attack graph. Then for each edge x in d+1, and for each element y in Sd, the victim computes z = x _ h0(y): The victim then checks whether any child Ri of y in Gm has the hash value of its IP address, h(Ri), equal to x. If the victim finds a matched IP address Ru, then it adds Ru to the set Sd+1 (initially Sd+1 is empty). The victim repeats the steps until it reaches the maximal distance marked in the packets, denoted as maxd. Thus, the victim reconstructs the attack graph.



Algorithm for Advanced marking and reconstruction

Marking procedure at router R_i
for each packet P
Let u be a random number from
If u ≤ q then
P.distance ←0
P.edge ← $h(R_i)$
Else
if (P.distance==0) then
P.edge← P.edge XOR $h'(R_i)$
P.distance← P.distance + 1

Reconstruction procedure at victim V
Let S_d be empty for 0≤ d ≤ maxd
For each child R of v in G_m
if h(R) € Ψ_0 then
insert R into S_0
For d:=0 to maxd-1
For each y in S_d
For each x in Ψ_{d+1}
Z=x XOR h'(y)
For each child u of y in G_m
If h(u)=z then
Insert u into S_{d+1}
Output S_d for 0≤ d ≤ maxd

Authenticated marking scheme

A fundamental shortcoming of the advanced marking schemes is that the packet markings are not

authenticated. Consequently, a compromised router on the attack path could forget the markings of upstream routers. Moreover, the compromised router could forge the markings according to the precise probability distribution, preventing the victim from detecting and determining the compromised router by analyzing the marking distribution. To solve this problem, we need a mechanism to authenticate the packet marking. A straightforward way to authenticate the marking of packets is to have the router digitally sign the marking. However, digital signatures have two major disadvantages. First, they are very expensive to compute (a 500 MHz Pentium can only compute on the order of 100 1024-bit RSA signatures per second). Secondly, the space overhead is large (128 bytes for a 1024-bit RSA signature). We propose a much more efficient technique to authenticate the packet marking, the Authenticated Marking Scheme. This technique only uses one cryptographic MAC (Message Authentication Code) computation per marking, which is orders of magnitude more efficient to compute (i.e., HMAC-MD5 is three to four orders of magnitude more efficient than 1024-bit RSA signing) and can be adapted so it only requires the 16-bit overloaded .

Authentication with a MAC

Message Authentication Codes (MAC) such as HMAC are commonly used for two-party message authentication. Two parties can share a secret key K. When party A sends a message M to party B, A appends the message with the MAC of M using key K. When B receives the message, it can check the validity of the MAC. A well-designed MAC guarantees that nobody can forge a MAC of a message without knowing the key. Let f denote a MAC function and fK the MAC function using key K. If we assume that each router R_i shares a unique secret key K_i with the victim, then instead of using hash functions to generate the encoding of a router's IP address, R_i can apply a MAC function to its IP address and some packet-specific information with K_i. Because a compromised router still does not know the secret keys of other uncompromised routers, it cannot forge markings of other uncompromised routers. The packet-specific information is necessary to prevent a replay attack, because otherwise, a compromised router can forge other routers markings simply by copying their marking into other packets. We could use the entire packet content in the MAC computation, i.e. encode R_i as fKi(hP;Rii). But for efficiency, it might also be sufficient to just use the source and destination IP addresses in the packet, i.e. encode R_i as fKi (hsourceIP; destinationIP;Rii). In this case, a compromised router might still be able to forge a marking in a packet by using the same source IP address, but in this case, the vic- MAC computation is very efficient, e.g. a fast workstation can compute around 300; 000 8-byte HMAC-MD5 per second, tim can block traffic coming from this source IP address. (Also the extended scheme in step 2 can reduce the possible number of source IP addresses that the compromised router could use to replay.) Besides the change of using a MAC function with secret keys instead of publicly available hash functions, the marking and reconstruction procedure is similar to the Advanced Mark.

IX. COMPARISON BETWEEN PPM AND ADVANCED AND AUTHENTICATED MARKING SCHEME

- In PPM the network and router overhead is high because of high bit number to represent the IP address of routers where as in advanced and authenticated marking router and network overhead is low because of using hash function and XOR ing the function reducing from 63 bits to 11 bits.
- advanced and authenticated marking supports incremental deployment.
- higher precision and low computation overhead in advanced marking compared to PPM authentication prevents forging or tempering by compromised routers whereas PPM is vulnerable to spoofing.

X. CONCLUSION

With the passage of time and the increase in the number of attack tools and DDoS attacks. resulted in the continuous evolution of packet marking scheme for IP trace-back from basic marking PPM to more efficient and more secure marking schemes like **advanced and authenticated marking** scheme **adaptive packet marking** and others .the evolution process will continue till the most efficient robust and more secure packet marking scheme that has no limitations is developed in future.

REFERENCES

[1] K. Park and H. Lee, "On the effectiveness of probabilistic packet marking for IP traceback under denial of service attack," Proceedings of IEEE INFOCOM (2001).

[2] J. Liu, Z.J. Lee, and Y.C. Chung, "Dynamic probabilistic packet marking for efficient IP traceback," Computer Networks 51(3).

[3] V. Paruchuri, A. Durresi, and S. Chellappan, "TTL based Packet Marking for IP Traceback," Proceedings of IEEE GLOBECOM.

[4] D.X. Song, A. Perrig, Advanced and authenticated Marking Schemes for IP Traceback, Proceedings of IEEE INFOCOMM[C], Anchorage, AK,USA, 2001.pp.878–886 Arnon Boneh, Micha Hofri.The Coupon-Collector Problem Revisited.