Combining Speak-up with DefCOM for Improved DDoS Defense

Mohit Mehta, Kanika Thapar, George Oikonomou
Computer and Information Sciences
University of Delaware
Newark, DE 19716, USA

Jelena Mirkovic
Information Science Institute,
University of Southern California
Los Angeles, CA 90292, USA

***Abstract*—This work combines two existing defenses against distributed denial-of-service (DDoS) attacks DefCOM and Speak-up resulting in a synergistic improvement. DefCOM defense organizes existing source-end, victim-end and core defenses into a collaborative overlay to filter DDoS floods. Source networks that do not participate in DefCOM, often receive poor service and their traffic is severely rate-limited. This is because core nodes in DefCOM that perform filtering lack cheap algorithms to differentiate legitimate from attack traffic at line speed they must conservatively assume all high-rate traffic from legacy to be attack. Thus, in its attempt to mitigate DDoS, DefCOM ends up denying service during attacks to any legitimate hosts that reside in legacy networks. Speak-up is a recently proposed defense, which invites all clients of the DDoS victim to send additional payment traffic, with the assumption that attack machines are already sending close to their full capacity. Clients that send a lot of payment traffic are considered legitimate and whitelisted. Speak-up is relatively cheap to deploy at the clients and the DDoS victim, but since payment traffic needs to be sent continuously, this creates additional congestion to the victim, which is undesirable.**

We combine Speak-up and DefCOM into a synergistic defense that addresses the shortcomings of the individual defenses and confirms the success of collaborative protection against DDoS attacks. Speak-up is integrated with core defenses in DefCOM and whitelists clients based on their payment traffic. Legitimate clients in legacy networks can thus be detected and served. Further, since Speak-up is implemented in the core payment and attack traffic do not reach the victim and any undesirable congestion effects are localized to the vicinity of legacy networks.

Keywords: DDoS, DefCOM, defense by offense, Speak-up I. INTRODUCTION The number of Internet security incidents has grown exponentially in the last two decades [3], which has drawn research communitys attention towards developing novel security systems. Distributed denial-of-service (DDoS) attacks have been identified as the most expensive computer crime for victim organizations [4]. DDoS defense goals involve (1) accurate detection of the attack, (2) rate-limiting of traffic and (3) differentiating attack traffic from legitimate traffic so that while attack packets are blocked, legitimate hosts obtain the service of the victim even during the period of attack. DefCOM [1] is a distributed collaborative defense that attempts to harvest the strengths of existing defense mechanisms in order to achieve these defense goals. It makes use of victim-end, source-end and core defense mechanisms to perform attack detection, traffic differentiation and rate-limiting respectively. Nodes communicate using an automatically-built overlay. They collaborate by exchanging messages, marking packets as high or low priority and prioritizing traffic during attack. Victim-end defenses play the role of alert generators they detect the attack and disseminate the alerts to all DefCOM participants. DefCOM makes an assumption that differentiating legitimate from attack traffic is CPU and memory-intensive and should be done by source-end defenses, which play the role of classifiers. These defenses can usually ascertain that some traffic outgoing from their network is legitimate. They mark all verified-legitimate traffic with a high priority mark, and other traffic with a low priority mark. Core nodes act as rate-limiters; they limit the traffic to the victim but implement weighted fair sharing of the limited resource according to traffics priority marks. Legitimate traffic from source networks that participate in DefCOM by deploying a classifier reap benefits, since their traffic reaches the victim during an attack. Victim is also well protected since rate limiters drop excess traffic preventing congestion. Legitimate clients in legacy networks, however, receive poor service. Rate-limiters conservatively and severely rate-limit all traffic coming from these networks because they lack means to verify if this traffic is legitimate or attack. This results in a denial of service to legacy source networks during attacks. Another DDoS mitigation technique, Defense by Offense [2] is a currency-based defense against application-level DDoS attacks. It uses a mechanism called Speak-up in which the server encourages all clients to send more traffic to it, in the form of payment traffic, so that they get a better representation at the server. Traffic from end-hosts is prioritized depending on the amount of payment traffic received from them during the attack. The idea behind this is that since the attacking hosts are already using most of their upload bandwidth, they will be able to send only a small amount of payment traffic and thus, traffic from legitimate hosts gets a higher priority as these hosts are able to send a considerably higher volume of payment traffic. We make use of the idea of Speak-up to enhance handling of traffic from legacy networks by rate-limiters in DefCOM. Our assumption is that legacy networks that do not deploy a source-end defense system coupled with a DefCOM classifier, may have legitimate clients that are willing to deploy Speak-up individually. This assumption makes sense as Speak-up can be implemented on an end machine under a users control while classifiers need to be deployed inline at the network

level, which requires administrative privileges. DefCOMs rate-limiters enhanced by Speak-up ask all clients that send traffic without DefCOM marks (which indicates that they reside in legacy networks) to send payment traffic to them. Depending on the amount of payment traffic received from an end host, its traffic is classified as high-priority, low-priority or is dropped. Hosts from legacy networks that send high payment traffic are thus classified as legitimate and DefCOMs service is improved by the addition of Speak-up. All payment traffic is dropped at the rate-limiters. Thus, combination of DefCOM and Speak-up overcomes the limitations of using Speak-up alone, since it reduces congestion introduced by payment traffic. The use of Speak-up in DefCOM also exemplifies the ease of integrating existing defenses into its collaborative architecture. The rest of this paper is organized as follows. Section II describes in brief the working of DefCOM and Speak-up followed by a description of our combined system. Section III presents the experimental results that illustrate the improved operation of the combined defense over its parts. Section IV discusses some attacks that may present a challenge to Speak-up, and thus to our combined defense, and proposes ways to mitigate them. Section V discusses related work and we conclude in Section VI. II. DEFCOM AND SPEAK-UP

We describe the working of DefCOM and Speak-up before going on to explain their combination. Figure 1 illustrates the topology that we will use throughout the paper. Nodes 1 through 5 are the end-hosts that use the service provided by node 0 which will be the victim node in this topology during periods of attack. We will be using the end-hosts as either legitimate or attack clients depending on the type and rate of traffic they send to node 0. Nodes 6 through 9 are legacy routers in the Internet and will take on more functionalities as DefCOM is deployed in the topology. Bandwidths for each link are specified in the figure.

A. DefCOM DefCOM [1] makes use of a distributed mechanism to defend against DDoS attacks. Motivation for DefCOM came from the observation that three necessary DDoS defense functionalities attack detection, rate limiting and differentiation of legitimate from attack traffic are most successful when performed at victim end, core and source end respectively. Existing defenses are mostly localized and those that are distributed require participants to deploy new hardware, ignoring any existing defenses. DefCOM attempts to organize existing defenses into a collaborative overlay, harvesting functionalities already offered by them and adding a thin communication layer. There are three types of DefCOM functionalities that can be added to existing routers or defense nodes: alert generator, classifier and rate limiter. A single physical node may deploy multiple DefCOM functionalities. An alert generator relies on any existing defense to detect an attack and then propagates the attack alert to other DefCOM nodes using the overlay. The alert contains the victims IP address and a desired rate limit to be enforced by any inline DefCOM node. After this message, each DefCOM node observes the flow of traffic to the victims IP and sends invitations to form a peering relationship to the victim. These invitations are intercepted by any downstream DefCOM node and a traffic tree is formed dynamically, with the victim as root, and rate limiters or classifiers as leaves. A classifier assumes that any existing defense coupled with it has the ability to distinguish between legitimate and attack traffic, frequently via excessive statistics gathering and using legitimate clients behavior models, and to communicate this to the classifier. The classifier marks verified-legitimate packets with a high-priority mark and all other traffic with a low-priority mark. Classifiers, however, have a somewhat weak economic model since they are deployed to protect communication with a remote server during infrequent DDoS attacks. They require an inline deployment and may interfere with network operation if they malfunction. This may deter a wide deployment of classifier functionality and we assume that a significant portion of networks on the Internet will be legacy networks. Rate limiters help in reducing incoming traffic to the victim during the attack by running the weighted fair sharing algorithm (WSFA) to prioritize traffic based on the marks attached to it, while obeying the rate limit from the attack alert. Priority marks are based on an exchange of a secret key between pairs of DefCOM peers, and cannot be faked by an attacker. The rate-limiter functionality is deployed in all inline DefCOM nodes. Figure 2 shows the operation of DefCOM under attack. Nodes 2, 4, 5 are attackers attempting to bring down node 0 by flooding it with traffic. Legitimate clients of node 0 are nodes 1 and 3. Rate limiter functionality is deployed on all routers 6, 7, 8, 9. Alert generator deployed at node 9 detects that node 0 is under attack and floods alert messages, with rate limit equal to 2Mb, to all DefCOM nodes. Node 6 has a classifier deployed at it and is thus able to separate node 1s legitimate traffic from node 2s attack traffic. Node 1s traffic is marked as high-priority node 2s traffic is marked as low-priority. Rate-limiting is then applied with weighted fair sharing of the limit according to traffic priorities. Since the high-priority weight is much larger than the low-priority one all legitimate traffic is usually allowed through. Node 7 is a leaf rate limiter. Attack nodes 4 and 5 transmit at a considerably high rate, which makes the aggregate traffic from nodes 3, 4, 5 higher than the limit specified by the alert generator. According to DefCOMs algorithm (see [1] for more details), such aggressive traffic is considered attack and dropped. As a result, in its attempt to mitigate DDoS on victim node 0, node 7 ends up denying service to the legitimate client 3 that is in a legacy network and shares the path with attack traffic before reaching any DefCOM node. Our work will help node 3 take individual action to ameliorate its situation, by adding a relatively cheap speak-up functionality to nodes 3 and 7. B. Speak-up Defense by Offense [2] makes use of Speak-up, a currency-based approach with bandwidth as the currency, against application-level DDoS. A victimized server encourages all clients, resources permitting, to automatically send higher volumes of traffic in the form of payment traffic in a congestion-controlled manner. This defense deploys a thinner before the server, which prevents the server overload and implements a virtual auction. It opens separate payment

channels with the clients to receive payment traffic and, when the server is ready, it forwards a service request of the client for which the highest amount of payment traffic has been received to the server. All payment traffic is dropped at the thinner. Speak-up is based on the assumption that the attackers are already using most of their upload bandwidth. According to the authors [2], recently conducted research [15] states that an average bot has 100Kbps bandwidth and that only 5Figure 3 illustrates the working of Speak-up. As in our previous example nodes 2, 4 and 5 are attackers flooding node 0, and nodes 1 and 3 are its legitimate clients. Nodes 6, 7, 8, 9 are legacy routers with no extra functionality. Defense by Offense assumes that the server is overprovisioned with regard to bandwidth, so payment traffic should not overload the servers network before it reaches the thinner. Under attack, thinner at node 0 asks all clients to send payment traffic to it. Since nodes 2, 4 and 5 are attackers, they are already using a major portion of their upload bandwidth for attack and can only send a small amount of payment. Nodes 1 and 3, on the other hand, use only a small portion of their upload bandwidth and are able to send higher volumes of payment traffic getting their requests serviced. Speak-up manages to protect legitimate clients and provide them with the victims service during attacks. However, as payments are calculated by the thinner in front of the server, payments from all clients travel all the way up to the victim through the core and are dropped only then consuming resources and potentially causing congestion on the entire path. It would be highly desirable if payment traffic could be processed closer to the sources and our combined system facilitates this. By processing the payment traffic close to the source of attack, and differentiating legitimate from attack clients, we are also able to drop attack traffic at the edge networks and prevent flooding of the core by attack traffic too. C. Combined system We propose to overcome the limitations of DefCOM and Speak-up by integrating Speak-up with DefCOMs rate limiter functionality. This helps rate limiters detect legitimate traffic from legacy clients. At the same time, payment traffic is processed at the rate limiters, far from the victim, and is dropped along with the attack traffic thus saving downstream bandwidth. Payment is collected for each payment period of 5 seconds, it is processed at the end of the payment period and the decision whether the traffic is legitimate or attack is used for the next period. The length of the payment period is decided by balancing the tradeoff between overreaction to traffic fluctuations and a speedy response during attacks. Keeping the payment period too short could affect legitimate clients that occasionally send traffic in an aggressive manner as they would not be able to send enough payment for that period. Such clients would be misclassified as attackers and their traffic would be dropped in the next period. Keeping the payment period long could be misused by the attackers to their advantage. They can send high payment traffic during this period and attack at full strength in the next period, when their traffic would be marked for high-priority handling. We address this attack in Section IV. Based on the amount of payment received during the current period; regular (non-payment) packets of that client, are marked in the next payment period according to the Table 1. Thus, an attacker using at least half of its upload bandwidth to attack would be able to send payments reflecting only half of its upload bandwidth i.e. 0-50

Payment reflecting available bandwidth for this client The clients traffic mark for next payment period 75-10050-750-50Table 1: Client differentiation based on payment

Figure 4 explains the working of DefCOM combined with Speak-up. All the nodes have the same functionalities as described earlier in DefCOM section (II-A) except for node 7 that now has the ability to distinguish between legitimate and attack clients by using Speak-up. Node 9 detects the attack on node 0, an alert is propagated and the traffic-tree is formed. On receiving the alert, classifier at node 6 starts to mark traffic from node 1 with a high-priority mark while traffic from node 2 is marked with a low-priority mark. Node 7, on receiving the alert, asks all clients that send traffic without marks to the victim to send payment traffic to node 7. Legitimate client 3 is using only a small portion of its upload bandwidth and is able to send payments reflecting an available upload bandwidth of more than 75III. COMPARISON AND EVALUATION

In this section, we illustrate the improved functioning of DefCOM and Speak-up in the combined system over the functioning of individual systems. All experiments were implemented using the same experiment setup used in [1]. DefCOM, Speak-up and their combination were implemented on Linux routers as loadable kernel modules. They were run on real machines in the Emulab testbed [14], and we generated live legitimate and attack traffic to be handled by these defenses. Legitimate and payment traffic were created by establishing multiple telnet-like sessions between legitimate clients and the victim. This traffic is sent over TCP and is sensitive to drops due to congestion or due to a defenses actions. Attack traffic was created by sending a high-volume of TCP data packets to the victim, using the raw socket functionality. We chose to send TCP packets to make the attack traffic resemble the legitimate traffic as much as possible. Our results would be similar if we used any other type of flooding attack that generates high-volume traffic from individual attack machines. The alert generator is coupled with a simple mechanism that detects an attack if one of the following rules become true: (Rule 1) the ratio of incoming to outgoing TCP packets is higher than 3, (Rule 2) total incoming traffic rate is larger than the bottleneck link bandwidth during the last 3 seconds. This attack detection is simple but sufficient to detect attacks in our experiments. We couple D-WARD [12] with a classifier as a source-end defense, like it was done in [1]. D-WARD prevents outgoing DoS attacks by keeping statistics on incoming and outgoing packet counts for each TCP connection established with the victim. It classifies TCP connections with low sent-to-received packet ratio as legitimate. The weighted fair share algorithm (WFSA) in rate limiters has two traffic classes: high-priority and low-priority, and uses ideas from core-stateless fair queuing [13] to divide bandwidth between them. Weights assigned to classes are 0.9 for high-priority and 0.1 for low-

priority. A. DefCOM with and without Speak-up

Experiments with DefCOM use the topology and setup from the Figure 2. Experiments with combined DefCOM and Speak-up use the topology and setup from the Figure 4. Figure 5 shows the traffic reaching the victim server from each client for DefCOM-only tests. Classifier at node 6 marks packets from node 1 with high-priority marks while traffic from attack node 2 is rate limited and marked with low-priority marks. Traffic from nodes 3, 4 and 5 is completely cut off by the rate limiter 7. This results in an obvious denial of service to the legitimate node 3. As the rate limit specified by the alert generator is 2Mb (250,000 Bytes), and node 1s traffic does not completely exhaust it, some low-priority traffic from node 2 reaches the server. Figure 6 shows the traffic reaching the victim server from each client, for the combined DefCOM/Speak-up system. The enhanced rate limiter at node 7 protects legitimate traffic from node 3 during the attack, while traffic from nodes 4 and 5 is blocked. High-priority-marked traffic from node 1 (marked by the classifier) and node 3 (marked by the enhanced rate-limiter) dominates the 2Mb link between node 9 and node 0 during the attack. To further show the benefit of using Speak-up in DefCOM we remove the classifier at node 6 and deploy a rate limiter enhanced by Speak-up at this node. Traffic reaching the victim server from each client is shown in Figure 7. The enhanced rate limiter at node 6 drops traffic from attack node 2 as it is unable to send enough payment traffic (¿50B. The combined system versus Speak-up-only We used the topology and setup from Figure 3 to evaluate Speak-up-only defense and compare it to our combined system from Figure 4. For this comparison, in tests involving our combined system we deployed Speak-up at node 6 in addition to node 7, making them enhanced rate limiters. Figure 8 shows the total traffic (legitimate, attack and payment) reaching the victim node 0. For Speak-up-only, because payments from all clients and the attack traffic, travel all the way up to the victim node 0, to be dropped by the thinner placed in front of the server, the traffic during the attack is considerably high. In the case of our combined system, payment and attack traffic are dropped at enhanced rate limiters and only legitimate traffic reaches the server. Thus, DefCOM with Speak-up prevents the unnecessary consumption of network resources by payment and attack traffic. IV. SECURITY OF THE COMBINED SYSTEM Enhanced rate limiters mark packets for a client based on the payment received from it during the current payment period. One way for attackers to misuse the combined system is to send attack traffic at a moderate rate, and thus able to send payment traffic reflecting available bandwidth of 50-75Another way an attacker could attempt to bypass defensive action is distribute his attack so that attack machines send at an extremely low rate and thus can afford high levels of payment traffic. Such machines would be classified as legitimate and their traffic would compete with truly legitimate traffic because both streams would carry high-priority marks. According to the Defense by Offenses authors [2] such highly-distributed attack is beyond the means of the majority of attackers because it requires a larger botnet than observed in todays DDoS incidents. Finally, intelligent attackers could try to misuse the fact that sending sufficient payment traffic during a payment period buys a client preferential treatment in the next period. The attackers can send enough payment traffic by aborting attacks during a payment period and resuming them in the subsequent period. This either results in periodic pulses at the victim, if all attack machines are synchronized or in a continuous attack when groups of attack machines interleave their activity periods. We modified our rate limiter implementation to detect and handle this type of attacks. We start from an observation that payments from a given attack machine will fall alternately into the low range (see Table 1) and into the medium or high range. The enhanced rate limiter handles this behavior by keeping a history for all clients it sees during the attack. A client whose payments switch from either range (¿75We use the topology and setup from Figure 4 to illustrate handling of pulsing attacks. Node 4 is a smart attacker that performs a pulsing attack on victim node 0. Without the history of clients for the attack duration, the rate limiter with Speak-up at node 7 would let traffic from node 4 to go through with a high-priority mark every second payment period. This is shown in Figure 10, where node 4 is able to have its traffic delivered to the victim each 10 seconds. The magnitude of this attack would be much greater if more attack machines participated. Figure 11 shows how keeping a history for all clients during the attack helps the rate limiter to detect and stop a pulsing attack. Once node 7 notices that node 4 has switched its payments twice, it cuts-off all the traffic from node 4 for the rest of the attack. We keep a threshold of 2 for such behavior to give the benefit of the doubt to legitimate clients that might switch their payments due to bursty traffic once during an attack. This is again a tradeoff between responding too soon and potentially hurting legitimate clients (if threshold were 1) and responding too late (if threshold were ¿ 2).

IV. RELATED WORK

DDoS defenses, similar to DefCOM, that follow a distributed approach include SOS [5], Pushback [6], TVA [7] and COSSACK [8]. SOS uses access points close to source networks to verify legitimate users and send their traffic on the overlay to secret servlets that tunnel it to a distributed firewall protecting the victim. SOS provides good protection to the server but the traffic experiences a delay because it is routed on the overlay. Pushback enables routers to identify high-bandwidth aggregates that contribute to congestion and rate limit them. Pushback inflicts collateral damage when attackers are collocated on the same path as legitimate traffic. Further, it does not work in non-contiguous deployment and cannot detect attacks that do not congest core routers. TVA uses server-issued capabilities to differentiate between legitimate and suspicious traffic. Routers help create capabilities, rate limit new capability requests and give highest priority to capability-carrying traffic, then to request traffic and finally to legacy traffic. However, TVA is always active and its processing and memory cost are high. COSSACK forms a multicast group of defense nodes that are deployed at source and victim networks and cooperate in filtering the attack. It is

,however, unable to handle attacks from legacy networks that do not deploy COSACK defense mechanisms.

Defense by Offense is the only currency-based DDoS defense that uses bandwidth as the currency. Similar currency-based approaches such as [9, 10] describe the solution to DoS attacks on servers computational resources where the clients send fixed number of copies of their messages and the server only processes a fixed fraction of the messages received, thus, reducing the impact of the attackers. V. CONCLUSION In this paper, we combine two existing DDoS defenses DefCOM and Speak-up to achieve a synergistic effect. The flexible architecture of DefCOM enables the integration of existing defenses into it. We show that by using Speak-up at DefCOMs rate limiter nodes we can protect legitimate clients that reside in legacy networks, and ensure that their traffic reaches the victim of the attack, while we drop attack traffic. By dropping all payments along with attack traffic at the leaf rate limiters, Speak-up in DefCOM works better than in isolated deployment where payment and attack traffic consume resources on the entire path to the victim.