Designing Cloud Computing Services Resilient to DDoS Attacks

Extended Abstract

Sungjune Park
The Belk College of Business
University of North Carolina at Charlotte
9201 University City Blvd.
Charlotte, NC 28223, USA
supark@uncc.edu

Chandrasekar Subramaniam
The Belk College of Business
University of North Carolina at Charlotte
9201 University City Blvd.
Charlotte, NC 28223, USA
csubrama@uncc.edu

Nam K. Kim
Department of Industrial Engineering
Chonnam National University
77 Yongbong-ro, Buk-gu
500-757 Gwangju, Republic of Korea
freedom@jnu.ac.kr

Won Seok Yang
Department of Business Administration
Hannam University
70 Hannam-ro, Ojeong-dong
306-791 Daejeon, Republic of Korea
wonsyang@hnu.kr

Keywords: distributed denial of service, moving target defense, network security, queueing analysis

Designing Cloud Computing Services Resilient to DDoS Attacks

Introduction

With the availability of various cloud computing platforms and infrastructures, firms are increasingly demanding reliable online services. At the same time, distributed denial-of-service (DDoS) attacks disrupting such services are ever-increasing. A DDoS attack that overwhelms a networked system with a large number of requests in a short period of time can lead to loss of trust and confidence of the customers of the victimized businesses. Among the mechanisms proposed to prevent or mitigate the impacts of DDoS attacks, the moving target defense mechanism is an interesting approach (Jia et al., 2013). The basic principle of the moving target defense is that, on detection of a DDoS attack, the connections to the server subject to DDoS attacks are automatically moved to a different server. This recurring server migration process can be modeled as a queueing system that resets the system after voluntarily flushing out all customers in the system. In this study, we derive the optimal timing of migration for a given attack size distribution and processing capacity and then calculate the system performance using three performance measures in terms of service completion probability, service rate, and service time. We also consider the impact of misclassifying benign client requests into DDoS attacks and vice versa on these performance measures.

A Queueing Model for Moving Target Defense Mechanism

The advantage of the moving target defense approach is that it does not depend on global adoption of the defense mechanisms on the Internet routers or collaboration among the network providers in order to work. In one implementation of the moving target defense described by Wang et al. (2014), the proxy nodes are known only to legitimate clients who have been successfully authenticated and these clients are migrated from the attacked proxy node to an alternative proxy node. Thus, new requests are all filtered out when the proxy nodes are moved, thereby continuing to provide services to authenticated clients. However, the above mechanism still relies on authenticating the clients and keeping track of the authenticated connections. A simple but effective alternative mechanism would be the one in which the server automatically migrates to a new one whenever the volume of client requests reaches a threshold. Such a mechanism only requires simple statistical analysis of traffic and attack data to characterize the network to protect from DDoS attacks. Our approach enables us to defend from any DDoS attacks made of seemingly legitimate service requests, including requests created with spoofed IP addresses and distributed service requests from botnets. We model this mechanism as a new queueing system where DDoS attacks are treated as a batch arrival of service requests.

Our model differs from the previous models in literature and contributes as follows. First, our model can apply to any type of DDoS attack since the trigger for the migration in our model depends on the threshold of service requests (i.e., the network traffic) from clients and not on the explicit detection of DDoS attack signatures. Second, our model uses *voluntary* flushing, a new queuing policy to model online services that migrate from the attacked servers to a new group of servers. Third, our moving target defense approach is simple to configure and implement while achieving a high level of protection from DDoS attacks making it easy to implement in various cloud computing environments.

In our queueing system, client requests are serviced as they enter the service system of an online server. The server is assumed to be able to serve any number of requests entering the system, but the service becomes slower as the requests increase since all the requests in the system share the processing capacity of the online server. As the number of requests increases beyond a certain point, the service becomes too slow, i.e., the wait becomes unacceptable to the clients. Our model considers that a sudden increase in the requests and deterioration of wait time could be caused by a DDoS attack. At this point, we adopt a moving target defense mechanism to abandon all requests currently in the system, regardless of whether they are a legitimate or DDoS attack, and restart the service in a new server. We call this system the *K*-policy system. Compared to the well-known *N*-policy system (Heyman 1968, Yadin and Naor 1963), where the decision is to start the service based on the system size, in the *K*-policy system, the decision is to *stop* the service based

on the system size (K). The system size of a queuing system is typically fixed and represents the maximum number of customers. In our K-policy system, this system size representing the capacity of a queueing system is a decision variable. Although the service needs to be restarted for requests from legitimate customers, in the long run, we can avoid the impact of DDoS attacks and maintain acceptable service time for these requests by setting the optimal system size (K^*) to trigger the flushing.

We derive three performance parameters with the proposed queueing model. The first is *effective service probability* (ESP) denoted by p_e, which is the proportion of customers who successfully complete their services without getting flushed out. The second measure is effective service rate (ESR), denoted by $\bar{\mu}$, which is the average service rate that a legitimate customer receives from the processor-shared server. The third measure is *effective service time* (EST), denoted by $\bar{S}$, which can be interpreted as the average of the total time a legitimate customer may need to successfully complete her service assuming the customer keeps coming back when she gets flushed.

Numerical Analysis

As DDoS attacks usually come in different scales, we examine how the optimal K changes when the attack-size distribution varies. We consider four cases of probability distributions (deterministic, two uniform with high/low variance, and Poisson), for all of which we set the average attack size to 20 (20 times the legitimate client requests). The decision on K is also affected by the frequency of attacks and how quickly the server processes service requests. We create nine attack scenarios with nine different sets of arrival rates for both legitimate and illegitimate (DDoS attacks) customers, service rate, and the expected server utilization. We investigate how the optimal K varies when server utilization designed for legitimate customers is high ($\rho_1 = 0.2$), medium (0.5), or low (0.8) and the frequency of DDoS attacks is low ($\lambda_2 = 0.025$), medium (0.05), or high (0.1).

For all cases and scenarios, p_e is initially low for a very low K due to frequent flushing and increases up to an optimal K (K^*). The value of p_e then decreases beyond this point as the arrival of the DDoS attacks increases the service time unacceptably even for the legitimate customers. We call this *zombie effect* because legitimate customers are affected negatively as though they are part of the DDoS attacks. As the DDoS attacks become more frequent, legitimate customers are likely to suffer more from the zombie effect, due to which K^* decreases. It is interesting to note that for the deterministic case, $K^* = 20$, which is the size of DDoS attack. If $K \leq 20$, the system protects itself from DDoS attacks because their arrivals triggers the system to flush all illegitimate customer requests from DDoS attack. Hence, setting K at the highest possible value ($K = 20$) while maintaining such protection, increases the chance of serving the legitimate customers currently in the system. If $K > 20$, the system may allow a large number of illegitimate customers from a DDoS attack to enter the system. This slows down the services of both legitimate and illegitimate customers significantly. As a result, legitimate customers arriving even after a DDoS attack may get flushed out due to the increased chance of zombie effect. For other cases where attack size is not deterministic, K^* is always less than 20 (the average size of DDoS attacks). The value of K^* tends to decrease as the variance of attack size increases. This could be due to the fact that the zombie effect can be mitigated by flushing the system whenever there is a doubt of a DDoS attack. Higher variance in attack size causes the system to flush earlier to avoid the zombie effect leading to a smaller K^*. The variance of DDoS attack size also affects how the P_e changes. The smaller the variance, the steeper is the declining slope of P_e.

The second performance measure we examined is the effective service rate (ESR). ESR is strictly decreasing as higher K indicates more chances of DDoS attacks sharing the processing capacity. The strictly decreasing ESR is expected since ESR is inversely proportional to the number of customers in the system. The marginal decrease in ESR is greater when the number of customers in the system is smaller. However, if K is large enough that the number of legitimate customers reaching the full capacity is less likely, we see ESR does not change as much. Finally, there is a noticeable drop in ESR, in particular, at $K = 20$ (DDoS attack size) for the deterministic case. This is due to a sudden increase in system size resulting from a DDoS attack. In other cases, similar noticeable decrease in ESR happens in a range around the average DDoS attack size and the range depends on the variance of the attack-size distribution.

The third performance measure, effective service time (EST), is more useful practically for DDoS defense systems than ESP or ESR. Since ESP is concave and ESR is strictly decreasing, EST, which is derived as $\bar{S} = 1/p_e \times 1/\bar{\mu}$ is convex and hence gives a unique K that minimizes EST. When K is too small, the high

number of flushing and restart of service for legitimate customers results in initial high EST. As K increases, EST decreases rapidly due to the decreased probability of flushing. When K is large enough to increase p_e and ESR is insensitive to the difference in K, the optimal K is determined. EST increases slowly after the optimal K value, and increases rapidly where there is a possibility of DDoS attack. For example, for the deterministic case, there is an inflection point at $K = 20$, which is the same as the size of DDoS attack. This implies that it will never be optimal to set K greater than the known size of DDoS attack. The optimal K may seem quite small if we choose EST as the performance measure to optimize. This low K^* is due to no penalty given for not completing services for legitimate customers when they are repeatedly flushed from the system. However, in reality, it is more meaningful to adjust EST by giving appropriate weights to p_e and ESR to account for this penalty. Alternatively, EST may be considered just a constraint, and we may maximize p_e subject to EST being less than a reasonable value.

Our model can be applied to any cloud computing environments where a DDoS attack can be isolated in the old service location while all legitimate requests are redirected to a new server location. In the case of systems where filtering of service requests is necessary for such isolation, our model can be extended to incorporate the filtering-based moving target defense mechanism. While we are redirecting the legitimate requests to the new service location in the cloud, some illegitimate requests from a DDoS attack may also be redirected due to the false positive classification of the filtering system. Such type I error incurs a cost c_1 due to the increased service load to the shared processor in the new location. Every time a legitimate service request is serviced, a revenue of b is generated, but when it is flushed instead due to the K-policy we incur a type II error cost c_2. Also, given a value of K, the probability of a DDoS attack being serviced (p_d) as well as the probability of legitimate request being served (p_e) can be derived. Let q_{ij} be the number of customer type i requests that are classified as type j ($i, j = 1$ for legitimate or 2 for illegitimate) by the filtering system in place. Finally, the expected pay-off from the system can then be calculated as the sum of the entries in table 1.

	Legitimate request	Illegitimate request (DDoS attack)
Serviced	$b(q_{11}p_e + q_{12}p_d)$	$-c_1(q_{22}p_d + q_{21}(1 - p_e))$
Flushed	$-c_2(q_{11}(1 - p_e) + q_{12}(1 - p_d))$	0

Table 1. Expected Payoff Matrix

We are currently in the process of investigating the optimal expected payoff and the impact of filtering capability, including type I and type II errors.

Conclusions

In this paper, we have modeled and investigated a simple, yet effective, mechanism to defend against DDoS attacks. Unlike previous studies requiring explicit detection of DDoS attacks, we use a threshold of service requests (called K-policy) in order to trigger the response. We derive three performance measures, effective service probability, effective service rate, and effective service time, to show that this simple K-policy can still be successful in defending from DDoS attacks and preventing subsequent service deterioration.

References

Heyman, D. P. (1968). Optimal operating policies for M/G/1 queuing systems. *Operations Research*, 16(2), 362-382.

Jia, Q., Wang, H., Fleck, D., Li, F., Stavrou, A., & Powell, W. (2014, June). Catch me if you can: A cloud-enabled DDoS defense. In *Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on* (pp. 264-275). IEEE.

Wang, H., Jia, Q., Fleck, D., Powell, W., Li, F., & Stavrou, A. (2014). A moving target DDoS defense mechanism. *Computer Communications*, *46*, 10-21.

Yadin, M., & Naor, P. (1963). Queueing systems with a removable service station. *Journal of the Operational Research Society*, 14(4), 393-405.