A general cloud firewall framework with dynamic resource allocation

Conference Paper · June 2013

DOI: 10.1109/ICC.2013.6654807

4 authors, including:

Shui Yu
Deakin University
246 PUBLICATIONS **5,573** CITATIONS
SEE PROFILE

Wanlei Zhou
Deakin University
499 PUBLICATIONS **9,651** CITATIONS
SEE PROFILE

Song Guo
The Hong Kong Polytechnic University
737 PUBLICATIONS **17,206** CITATIONS
SEE PROFILE

A General Cloud Firewall Framework with Dynamic Resource Allocation

Shui Yu, Robin Doss, Wanlei Zhou
SIT, Deakin University
Victoria, Australia
{syu, rchell, wanlei}@deakin.edu.au

Song Guo
SCSE, The University of Aizu
Aizuwakamatsu, Japan
sguo@u-aizu.ac.jp

***Abstract*—Cloud is becoming a dominant computing platform. However, we see few work on how to protect cloud data centers. As a cloud usually hosts many different type of applications, the traditional packet level firewall mechanism is not suitable for cloud platforms in case of complex attacks. It is necessary to perform anomaly detection at the event level. Moreover, protection objects are more diverse than the traditional firewall. Motivated by this, we propose a general framework of cloud firewall, which features event level detection chain with dynamical resource allocation. We establish a mathematical model for the proposed framework. Moreover, a linear resource investment function is proposed for economical dynamical resource allocation for cloud firewalls. A few conclusions have been extracted for the reference of cloud providers and designers.**

***Index Terms*—Anomaly detection, cloud firewall, resource investment.**

I. Introduction

In this paper, we present a general firewall framework for cloud computing platforms. Different from its ancestors, cloud computing platforms, e.g. cloud data centers, usually host a large number of heterogeneous services, such as web services, e-business services, social network services, and video services. As it always is, these services are under numerous types of malicious attacks, such as distributed denial of service (DDoS) attack, virus, information phishing, and privacy breach. As a result, it is definitely expected to have a cloud firewall to protect the hosted services of cloud data centers. At the same time, the tasks and workload of a cloud firewall is more complex and heavier that its counterpart in a traditional specific domain or a local network network. Motivated by this, we propose a general framework of cloud firewall to meet the needs of the new computing platform in this paper.

Today, cloud computing has become one of the fastest growing sectors in the IT industry all over the world [1]. Despite the promising business model and hype surrounding cloud computing, security is the major concern for businesses shifting their applications to clouds [2], [3].

Before the emergence of cloud computing, a firewall is usually installed for a local area network, which offers relative simple and limited services. The task for a traditional firewall is to pass or drop network packets based on its policies or rules. Research has been done on traditional firewall from different aspects, such as efficient algorithms for rule matching [4], firewall policy anomaly detection [5], and firewall performance modeling and analysis [6]. A traditional firewall usually based on information of a TCP or IP packet, such as, protocol number, source and destination IP addresses, and ports [4]. Therefore, the tasks for a traditional firewall is relatively narrow and simple, and it is vulnerable to complex attacks. For example, a traditional firewall can be easily cheated by DDoS mimicking attacks [7]. However, a cloud data center usually hosts various services of different cloud customers. It is highly possible that one IP is in one hosted service's black list, but in another hosted service's white list. As a result, the traditional packet level based filtering mechanism is not suitable for clouds, and needs to be significantly extended in order to meet the requirements from cloud environment.

This new computing model also attract many existing and new attacks. The traditional attacks still exist for cloud hosted services, such as viruses, distributed denial of service, and phishing. On the other hand, we have witnessed new specific attacks on the business model and computing mechanism of clouds. For example, the resource provision model, "pay-as-you-go" triggers a transformed form of DDoS attack, Economic Denial of Sustainability (EDoS) attack [8], [9], where a large number of bots act as benign users to "enjoy" the service of the victim to financially bankrupt a cloud customer. The diversity of hosted serviced dramatically increases the difficulty and complexity of security defence in cloud.

As a new business model and computing platform, cloud related research has attracted a lot of attention from network security community. Lua and Yow [10] proposed the use of an intelligent fast-flux swarm network on top of the current Internet to mitigate the attack volumes of a DDoS attack. Du and Nakao [11] presented the concept of DDoS defense as a network service, offered by cloud providers using their abundant resources. Chen et al. [12] proposed an on-demand security architecture for cloud computing against possible DDoS attacks. Sqalli et al. [9] designed a two step mitigation scheme against economic denial of sustainability attacks.

To the best of our knowledge, we have not seen any discussions on cloud firewall. We note that it is emergent to extend the traditional concept of firewall for cloud data centers to promote this booming business model.

At the same time, cloud platform offers defenders profound resource to fight against possible attacks, which is not possible

for other computing platforms. Moreover, defenders enjoy a luxury feature of dynamical resource allocation in cloud environment. A a result, we can dynamically and economically invest on cloud firewall according to actual needs.

In this paper, we present a general framework for cloud firewall based on the unique needs of cloud. Our proposal is also based on the unique dynamic resource allocation feature that cloud platform offers us. In our proposal, we establish a detection chain, and each node of the chain performs one specific detection (e.g. a type of DDoS attack or a virus based on signature). Hence a success of detection follows a geometric distribution in probability. Each detection node of the chain has a service rate. We establish a mathematical model for the framework, and further extract a number of performance metric of the system for readers' reference.

This paper makes the following contributions.

- We propose a general framework for cloud firewall. We point out that the current firewall mechanism is not suitable for complex cloud platforms, and needs to be significantly extended to meet the requirements of the new computing platform. An event based detection chain mechanism is designed for cloud firewall with dynamic resource allocation.
- We establish a mathematical model for performance evaluation of the proposed framework with a emphasis on economical resource investment. The model offers cloud providers a tool of resource investment in order to meet the needs of their businesses. Furthermore, our model paves a foundation for further research in this topic.

The rest of this paper is organized as follows. Related work is discussed in Section II. In Section III, we present the framework of cloud firewall. Dynamic resource allocation for cloud firewall is discussed in Section IV. We conducted a simple evaluation on performance of the proposed framework in Section V. Further discussions are presented in Section VI. Finally, we summarize the paper and present possible future work in Section VII.

II. Related Work

There are plenty of work has been done in terms of traditional firewall. One topic is the efficiency of firewall with the aim to ensure that the firewall does not become a bottleneck for a given system. Rovniagin and Wool [4] modeled the firewall packet matching problem as a mathematical point location problem, and proposed a Geometric Efficient Matching (GEM) algorithm. Their experiments indicated that the proposed algorithm is more space efficient for rule based firewalls. Hu, Ahn and Kulkarni [5] considered the quality of policy of firewall configuration, and presented a firewall policy anomaly management framework, which employed a rule based segmentation technique to identify policy anomalies and derive effective anomaly resolutions. Salah, Elbadawi, and Boutaba [6] proposed a embedded Markov chain based mathematical model for rule-based firewall performance analysis, especially against DDoS attacks. They presented close form expressions of a number of important performance metrics, such as mean throughput, service time, CPU utilization. As majority of the similar work, they assumed the arrival rate following the Poisson distribution and the service times are independent and exponentially distributed.

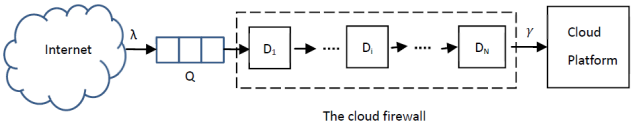

Fig. 1. The framework of the dynamical firewall for cloud platforms

We have seen new type of attacks targeting on cloud platforms. A variation of DDoS attack in cloud computing is the Economic Denial of Sustainability (EDoS) attack [9] or Fraudulent Resource Consumption (FRC) attack [8]. If the billing mechanism for cloud customers is "pay-as-you-use", botnet owners can create a large number of fake users to intensively consume the service of the targeted cloud customer. For example, the existing flash crowd mimicking attacks [13], [14] on a popular news web site is an excellent example. As a result, the bill for the targeted cloud customer will increase dramatically until the victim suspends her service or is bankrupted. On the other hand, if a cloud customer fixed her cost of renting for the resource of her hosted services, then a DDoS attack is effective to disturb it, or even shuts down her services.

There are some security related work on cloud security. Chen et al. [12] proposed on-demand security architecture to offer different services for different needs in cloud computing environments. This includes three inputs: risk of network access, service type and security level. Based on the mechanism of cloud computing, it is a good idea as it meets different requirements of users. In order to deal with EDoS attacks, Sqalli et al. [9] proposed a white and black list based filtering scheme to block malicious service requests. Amazon developed cloudWatch [15], a tool to monitor her cloud resources and mitigate EDoS attacks to her cloud customers. In addition, there are also a number of surveys on general security issues in cloud computing, such as [2] and [3].

III. The Framework of Cloud Firewall

In this section, we present the proposed framework of cloud firewall, and establish a mathematical model for the detection chain mechanism, and further extract a number of performance metrics.

As shown in Figure 1, the proposed cloud firewall is between the Internet and a cloud platform. All incoming requests will be examined by detectors in a sequence until a detector reports positive. Further actions will be taken, e.g. drop or block related requests.

In general, assume we have N different detectors $D_i (i = 1, 2, \ldots, N)$, each of them aims at one specific anomaly, such as viruses, different DDoS attacks, and information phishing. The packet arrival rate is λ, and the throughput rate is represented as γ.

It is generally accepted that request arrival follows the Poisson distribution, which is defined as follows.

$$Pr[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}, k=0,1,.... \tag{1}$$

Equation (1) describes the probability of k arrivals for a given time interval, and the mean of arrivals is λ.

For each detector $D_i(i = 1, 2, \ldots, N)$, we denote its service rate as μ_i^0. We can easily obtain the average service time of the N detectors as

$$E[\mu^0] = \frac{1}{N}\sum_{j=1}^{N}\mu_j^0. \tag{2}$$

In the case that we have no knowledge of the anomaly distribution, then we suppose the N detectors share the same probability of positive detection, p $(p > 0)$, then we know the positive detection follows the geometric distribution. Let random variable X be the number of trials, then the probability that we obtain the first positive detection at $k(k \in \mathbb{N})$ is expressed as follows.

$$Pr[X = k] = p_k = (1-p)^{k-1}p. \tag{3}$$

The probability cumulative distribution function is

$$F[X \leq k] = 1 - (1-p)^k. \tag{4}$$

The mean of the geometric distribution is

$$E[X] = \frac{1}{p}. \tag{5}$$

From a system point of view, we care about the average of system service time, which is

$$\mu = \sum_{k=1}^{N}\left(p_k \cdot \sum_{j=1}^{k}\mu_j^0\right). \tag{6}$$

Based on the Wild theorem, we can rewrite equation (6) based on equation (2) and (5).

$$\mu = E[X]E[\mu^0]. \tag{7}$$

In order to find close form solutions for our studied objects, we further suppose the service time of N detectors is i.i.d, and follows exponential distribution. This approximation is practical and commonly used in performance evaluation. Let μ_e be mean of the service time of the detectors, then equation (7) can be further expressed as

$$\mu = \frac{1}{p}\mu_e. \tag{8}$$

As a result, we can model the cloud firewall as a M/M/1 queue with arrival rate λ and service rate μ. Taking advantage of the conclusions of queueing theory [16], we can extract the key performance metrics that we are interested.

First of all, the the probability of the system stays state q_k (namely, there are k requests in the system) is

$$\begin{cases} p_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{p\lambda}{\mu_e}, \\ p_k = \left(\frac{\lambda}{\mu}\right)^k p_0 = \left(\frac{p\lambda}{\mu_e}\right)^k 1 - \frac{p\lambda}{\mu_e}. \end{cases} \tag{9}$$

The probability density of the time in system is

$$Pr(T = t) = (\mu - \lambda)e^{-(\mu-\lambda)t}, \tag{10}$$

for $t > 0$.

The average time spent in the firewall system is

$$T = \frac{p}{\mu_e - p\lambda}. \tag{11}$$

Based on equation (11), we obtain the average system throughput

$$\gamma = \frac{1}{T} = \frac{1}{p}\mu_e - \lambda. \tag{12}$$

The average number of requests in the system, $\overline{K}$,

$$\overline{K} = \sum_{j=1}^{K} k p_k. \tag{13}$$

In terms of firewall, we are interested in CPU utilization, which is also referred as to *carried load*. We denote this metric as U_{cpu}, and it can be calculated as

$$U_{cpu} = \gamma\mu = \frac{1}{p}\gamma\mu_e. \tag{14}$$

Let ρ be the busy rate of the studied system. From a system viewpoint, in order to make the system stable, the following has to be met.

$$\rho = \frac{p\lambda}{\mu_e} < 1. \tag{15}$$

IV. Dynamic resource allocation for cloud firewall

In this section, we focus on how to economically and dynamically allocate resource to meet the requirement of a cloud firewall.

A. Resource investment function

In order to measure the investment for a expected quality of service, we define a resource investment function $R(x)$ with respect to a system requirement x. For simplicity, we suppose $R(x)$ is a linear and non-decrease function. Then we have the following properties of this investment function.

$$\begin{cases} R(x) = 0, & x = 0 \quad (a) \\ R(x) \leq R(y), & 0 \leq x \leq y \quad (b) \\ R(ax + by) = aR(x) + bR(y), & a, b \in \mathbb{R}. \quad (c) \end{cases} \tag{16}$$

In this cloud firewall case, our system requirement x is the average time in system of requests, T, which is defined in equation (11). In order to avoid our cloud firewall becoming a bottleneck, the time in system for a request has to be limited.

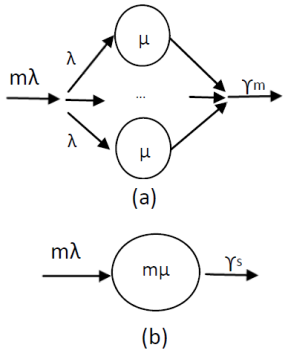

Fig. 2. The two options for resource investment for cloud firewall

Let ΔT be an acceptable threshold for T, then our resource investment problem becomes an optimization issue as follows.

$$\begin{aligned} &minimize R(T) \\ &s.t. \\ &T \leq \Delta T. \end{aligned} \tag{17}$$

B. single detection chain vs multiple parallel detection chains

Suppose a single detection chain with a given resource is work well for an arrival rate λ. As the request for a cloud platform is dynamic. Let $m \in \mathbb{R}$ and $m > 1$ be the *traffic strength*. Intuitively, when arrival rate increase to $m\lambda$, we should invest more resource to handle the traffic. One problem rises naturally, when the request arrivals increases, how should we invest our resource?

In general, there are two options to deal with this case: 1) clone $\lceil m-1 \rceil$ parallel detection chains based on the original detection chain; 2) keep the original detection chain, but increase the service capability of each detector to $m\mu$. We show these two options in Figure 2. We are interested about which one is a better investment.

In order to conduct this comparison, we use service rate as a study object in $R()$. When the arrival rate increase to $m\lambda$. The resource that we need for the multiple detection chain strategy is $R(m\mu)$. Let x be the expected service rate of the the single detection chain strategy, then,

$$\frac{1}{x - m\lambda} = \Delta T = \frac{1}{\mu - \lambda}. \tag{18}$$

It is easy for find that

$$x = \mu + (m-1)\lambda. \tag{19}$$

As $\lambda < \mu$, combining this with equation (19), we have

$$x < m\mu. \tag{20}$$

Based on the property (b) of investment function (16), we have

$$R(x) < R(m\mu). \tag{21}$$

Namely, for a given performance metric,the resource needed by the single detection chain is less than that by the multiple detection chain strategy.

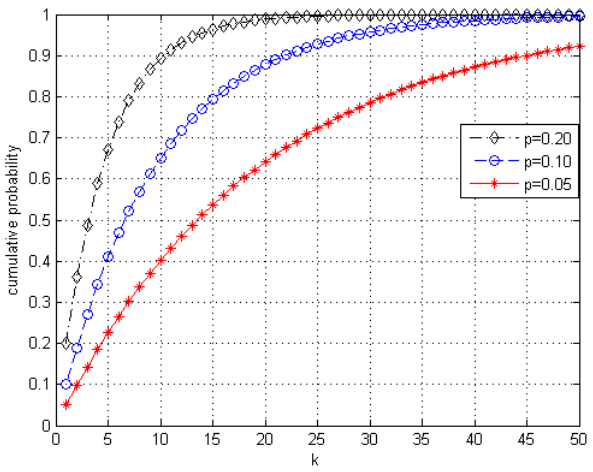

Fig. 3. The cumulative probability of a success detection using the single detection chain cloud firewall under different success probability p.

Furthermore, we expect to know for the same resource investment, how much gain we can obtain from a single detection chain strategy compared with the multiple detection chain strategy. We denote the system throughput for the two strategies as γ_m (multiple detection chains) and γ_s (single detection chain), respectively, and their average time in system for a request as T_m and T_s, respectively.

As shown in Figure 2 (a), the average time in system for the multiple detection chain is equivalent to one of its parallel detection chain. Based on equation (11), we have

$$T_m = \frac{1}{\mu - \lambda}. \tag{22}$$

At the same time, based on Figure 2 (b), we obtain the average time in system for a single detection chain as

$$T_s = \frac{1}{m\mu - m\lambda} = \frac{1}{m} T_m. \tag{23}$$

Therefore, we obtain

$$\gamma_s = m\gamma_m. \tag{24}$$

As we knew $m > 1$, therefore, for the same resource investment, the single detection chain strategy outperforms the multiple detection chain strategy m times.

V. PERFORMANCE EVALUATION

In this section, we conduct performance evaluation in a couple of aspects for the proposed single detection chain cloud firewall.

First, we are interested to see how many detection node should we go through to achieve a positive result, which is denoted by the cumulative probability function with respect to the success probability p. We show the result in Figure 3.

From Figure 3, we obtain a relationship between the positive detection and the number of detection nodes. For example, in order to achieve a 90% probability of a positive detection, we

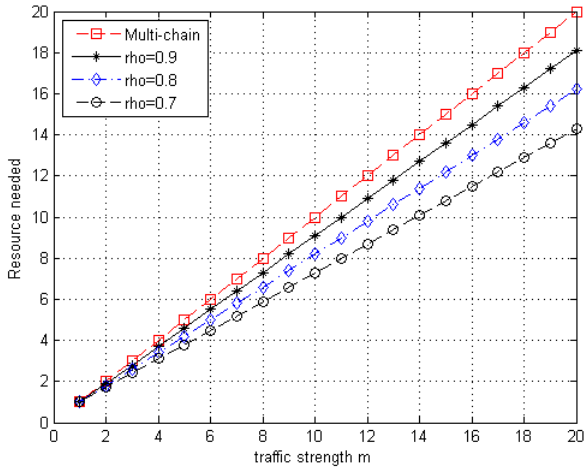

Fig. 4. The comparison on resource investment against different busy rate ρ (rho).

need go through around 45, 22, and 10 detection nodes for $p = 0.05$, 0.10, and 0.20, respectively.

We are also interested in how much should we invest to meet the delay constrain. As we know $R(T)$ is a function depending on m, p, λ and μ_e. Based on equation (15), we know ρ represents the impact of p, λ and μ_e. As a result, we can obtain the resource investment against traffic strength m for a given busy rate ρ. The result is shown in Figure 4.

From Figure 4, we see the linear relationship of investment function, and more request traffic, more resource investment. Moreover, for a give traffic strength m, a smaller ρ system requires less resource.

VI. Further Discussion

To the best of our knowledge, this is one of the early work in the topic. We took convenient assumptions to formulate the problem and obtain solutions. In order to get closer to real applications, there are plenty improvement to be explored further. We list a couple of them as follows.

- Improvement in system modeling. In this paper, similar to many of the previous work, we model our system as a M/M/1 queue. However, things are more complex in practice, and a G/G/1 model is general for all possible scenarios.
- Further investigation on the resource investment function. We suppose the investment function is a linear function, however, this need to be further observed and explored.
- Cost for dynamic resource allocation. In this paper, we do not count the cost of dynamical resource allocation. If the cost is sufficiently big, we can not ignored it in our resource investment function, which will make it more complex to process.

VII. Summary and Future Work

In this paper, we point out the limitation of the traditional firewall for cloud platforms, and propose a general framework for cloud firewall. The framework is an event level detection chain and resources can be dynamically allocated according to the needs on the fly. A mathematical model is established for the framework, and a linear investment function is also presented for dynamical resource allocation purpose. Based on the model, we found that a single detection chain with stronger service capability is better than a multiple detection chain with weaker service rate. In addition, a low busy rate system request less resource to achieve the same filtering performance than a system with a high busy rate. These findings are of values for cloud firewall designers and users.

As future work, we will further explore this topic in various aspects, such as generalizing the model, closely observing real cloud firewall systems to obtain real data. Moreover, the probability distribution of anomaly is not equivalent with a high probability in practice, therefore, we should arrange the order of detectors in the detection chain according to their probability in order to improve the efficiency. How to guarantee the quality of service of cloud firewall is an important topic to explore, the recently invented network calculus is probably a good tool to serve the topic.

References

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A berkeley view of cloud computing," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28, Feb 2009.

[2] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," *J. Network and Computer Applications*, vol. 34, no. 1, pp. 1–11, 2011.

[3] R. Bhadauria, R. Chaki, N. Chaki, and S. Sanyal, "A survey on security issues in cloud computing," *CoRR*, vol. abs/1109.5388, 2011.

[4] D. Rovniagin and A. Wool, "The geometric efficient matching algorithm for firewalls," *IEEE Trans. Dependable Sec. Comput.*, vol. 8, no. 1, pp. 147–159, 2011.

[5] H. Hu, G.-J. Ahn, and K. Kulkarni, "Detecting and resolving firewall policy anomalies," *IEEE Trans. Dependable Sec. Comput.*, vol. 9, no. 3, pp. 318–331, 2012.

[6] K. Salah, K. Elbadawi, and R. Boutaba, "Performance modeling and analysis of network firewalls," *IEEE Transactions on Network and Service Management*, vol. 9, no. 1, pp. 12–21, 2012.

[7] S. Yu, S. Guo, and I. Stojmenovic, "Can we beat legitimate cyber behavior mimicking attacks from botnets?" in *Proceedings of the INFOCOM*, 2012.

[8] J. Idziorek, M. Tannian, and D. Jacobson, "Insecurity of cloud utility models," *IT Professional*, no. PrePrints, 2012.

[9] M. H. Sqalli, F. Al-Haidari, and K. Salah, "Edos-shield - a two-steps mitigation technique against edos attacks in cloud computing," in *UCC*, 2011, pp. 49–56.

[10] R. Lua and K. C. Yow, "Mitigating ddos attacks with transparent and intelligent fast-flux swarm network," *IEEE Network*, no. July/August, pp. 28–33, 2011.

[11] P. Du and A. Nakao, "Ddos defense as a network service," in *NOMS*, 2010, pp. 894–897.

[12] J. Chen, Y. Wang, and X. Wang, "On-demand security architecture for cloud computing," *Computer*, vol. 99, no. PrePrints, 2012.

[13] S. Yu, W. Zhou, W. Jia, S. Guo, Y. Xiang, and F. Tang, "Discriminating ddos attacks from flash crowds using flow correlation coefficient," *IEEE Transactions on Parallel Distributed Systems*, vol. 23, no. 6, pp. 794–805, 2012.

[14] A. El-Atawy, E. Al-Shaer, T. Tran, and R. Boutaba, "Adaptive early packet filtering for protecting firewalls against dos attacks," in *Proceedings of the INFOCOM*, 2009.

[15] CloudWatch, http://aws.amazon.com/cloudwatch/.

[16] L. Kleinrock, *Queueing Systems*. Wiley Interscience, 1975, vol. I: Theory.