A Secured Decentralized Cloud Firewall to achieve Resources Provisioning Cost Optimization and QOS

Mallika T M
Dept. Of CSE
GMIT, Bharathinagar

Nandini K
Dept. Of CSE
GMIT, Bharathinagar

Sowndarya M H
Dept. Of CSE
GMIT, Bharathinagar

Likitha R
Dept. Of CSE
GMIT, Bharathinagar

Bhramaramba D S
Dept. Of CSE
GMIT, Bharathinagar

Pradeep B M
Dept. Of CSE
GMIT, Bharathinagar

Abstract*—Cloud computing technique is becoming popular as the next infrastructure of computing platform. Security has become the major concern that people hesitate to transfer their applications to clouds. Concretely, cloud computing platform is under numerous attacks. As a result, it is definitely expected to establish a secured firewall to protect cloud from attacks. By employing embedded Markov chain and Z-transform techniques, we obtain a mean packet response time. However, setting up a secured centralized firewall for a whole cloud data center is infeasible from both performance and financial aspects. In this paper, we propose a secured decentralized cloud firewall framework for individual cloud customers. We investigate how to dynamically allocate resources to feasibility of resources provisioning cost, while satisfying QoS requirement requested by individual customers simultaneously. Our numerical results also indicate that we are able to set up cloud firewall with affordable cost to cloud customers. Through extensive simulations and experiments, we conclude that an M/Geo/1 model results in the cloud firewall *real system much better than a traditional M/M/1 model.

Keywords*— *Cloud computing, firewall, resources allocation, system modeling

1. INTRODUCTION

Cloud computing is becoming popular as the next infra-structure of computing platform in the IT industry [1].

The large volume hardware and software resources pooling and delivered on demand, cloud computing provides rapid elasticity. In this service-oriented architecture, cloud services are broadly offered in three forms: Infrastructure- as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). Cloud computing also brings down both capital and operational expenditure for cloud customers by outsourcing their data and business.

On one hand, traditional attacks such as Distributed Denial of Service (DDoS), viruses and phishing still exist in clouds. The new specific attacks on the computing mechanisms of cloud have also been found, including Economic Denial of Sustainability (EDoS) attack [3], cross Virtual Machine (VM) attack [4], and so on. It is an effective and necessary choice of establishing a cloud firewall to protect cloud data centers from all these attacks. Compared to a cloud platform, traditional firewalls are generally deployed for private networks which host relatively specific services [7], [8].

Only a few work [9], [10] have been done on cloud firewall, and both proposed a centralized cloud firewall. The diversity of heterogeneous services and complex attacks definitely means a large rule set and high packet arrival rate if a centralized firewall is applied for a whole cloud data center. As a result, it is hard to guarantee QoS requirement specified by cloud firewall customers. Moreover, important design factors like packet arrival rateduration and number of rules are customers specific. Therefore, it is more practical to offer a cloud firewall for individual cloud firewall customers.

Generally, a question that arises in setting up a cloud firewall is how to price this service. From cloud firewall customers perspective, they prefer to rent a cloud firewall from firewall providers as cheap as possible. While for cloud firewall providers, their primary goal is financial reward. Therefore, cloud firewall providers need to optimize resources provisioning cost, which offers a chance of lowering the cloud firewall price on behalf of customers without reducing providers' profit. There is an inherent tradeoff between the two goals: to guarantee a pressed response time, large volume of resources should be invested by cloud firewall providers, which in turn increase provisioning cost (and vice-versa). Meanwhile, QoS requirement about the cloud firewall system specified by customers should be satis- fied.

In our paper, we propose a decentralized cloud firewall framework. The cloud firewall is offered by Cloud Service Providers (CSP) and placed at access points between cloud data center and the Internet. Individual cloud customer rents the firewall for protecting his cloud hosted applications. Hosting servers of applications are grouped into several clusters, and resources are then dynamically allocated to set up an individual firewall for each cluster. All these parallel firewalls will work together to monitor incoming packets, and

guarantee QoS requirement specified by cloud customers at the same time. By covering the vast cloud and firewall related parameter space, we formulate the resources provisioning cost.

As aforementioned, the essential issue to achieve a financial balance between firewall providers and customers is to optimize resources cost. In order to conduct the optimization, we need to capture mean packet response time through the firewall system. As widely adopted in cloud performance analysis [5], [11], we employ queuing theory to undertake system model- ing. However, we have to point out that the cloud firewall service times follow a geometric distribution according to rule match discipline.

The contributions of this paper are summarized as follows:

- We propose a decentralized cloud firewall framework for individual cloud firewall customers.
- Resources are dynamically allocated to optimize the provisioning cost, and guarantee QoS requirement specified by customers at the same time.
- We introduce novel queuing theory based model M/ Geo/1 or M/Geo/m for performance analysis of the proposed cloud firewall. By employing Z-transform and embedded Markov chain techniques, a closed- form expression of mean packet response time is derived.
- Extensive simulations and experiments are conducted to verify our analytical model. The simulation results claim that geometric distribution is more suitable for firewall system modeling, and give a deep insight into tradeoff among optimal resources provisioning cost, QoS requirement and packet arrival rate.

2. CLOUD FIREWALL FRAMEWORK

In this section, we first discuss several important characteristics about cloud firewall, and then present our decentralized cloud firewall framework.

2.1 Basic Knowledge About Cloud Firewall

Dynamic packet arrival rate. In general, cloud services are hired by legitimate customers. However, cloud applications are also vulnerable to various attacks, and a long time attack is usually rare as they can easily detected [11]. Therefore, incoming packets to cloud firewalls are composed of long term legitimate packets and bursty attack packets. In addition, packet arrival rate is dynamically changing over the time. Moreover, arrival rate of legitimate packets from benign customers is relatively low, while attack packets for malicious purposes are usually at a high rate. In conclusion, it requires a feasible model to capture the dynamic packet arrival rate in both attack and normal period.

As a main threat to cloud availability [2], here we take DDoS attack for example. Moore et al. [12] indicated that the average DDoS attack duration is around 5 minutes, with the average DDoS attack rate being around 500 requests per second. While Yu et al. [11] presented that the mean arrival rate to an observed e-business site in normal period is lower than 10 requests per second.

On-demand resources provisioning. In order to provide a cloud firewall, firewall service providers should invest various resources to fight against possible attacks. Current CSPs usually pack resources such as CPU, bandwidth and storage into Virtual Machine instances for service. Generally, multiple VM instance types are offered and each type has a limited service capacity for a particular application, which is evident by analysis results in [13].

In our case, VM instances are launched by providers to host the cloud firewall. When packet arrival rate increases, a single VM instance tends to be incapable of handling the massive incoming packets, or the response time will violate QoS requirement specified by customers. According to QoS requirement, packet arrival rate and VM instances service rate, firewall service providers need to invest more resources on-demand by launching additional VM instances. New VM instances can be cloned based on the image file of the original firewall using the existing clone technology [14], [15]. Specifically, firewall providers have to invest different volume of resources in attack and normal period.

Cost and performance tradeoff. There is an inherent tradeoff between the following two goals:

- QoS requirement satisfaction. Mean packet response time requirement specified in QoS should be satisfied.
- Resources provisioning cost optimization. Resources provisioning cost of cloud firewall should be minimized as long as QoS requirement is satisfied.

2.2 A Decentralized Cloud Firewall Framework

As aforementioned, each VM instance has a limited service capacity for a cloud firewall application. Hosting a cloud firewall in a single VM instance (even the most powerful one) tends to be incapable of satisfying customer specific QoS requirement. In other words, it's hard to guarantee response time through a centralized cloud firewall. Therefore, we propose a decentralized framework where several firewall run in parallel. As shown in Fig. 1, hosting servers are grouped into several clusters and a VM instance is launched to host an individual firewall for each cluster. By distributing the packet arrival rate into several parallel firewalls and launching suitable VM instance for each firewall, response time through each firewall can satisfy the QoS requirement.

Suppose that there are M servers in the cloud data center hosting applications of an individual cloud firewall customer. x^n denotes packet arrival rate to these applications in non-attack period

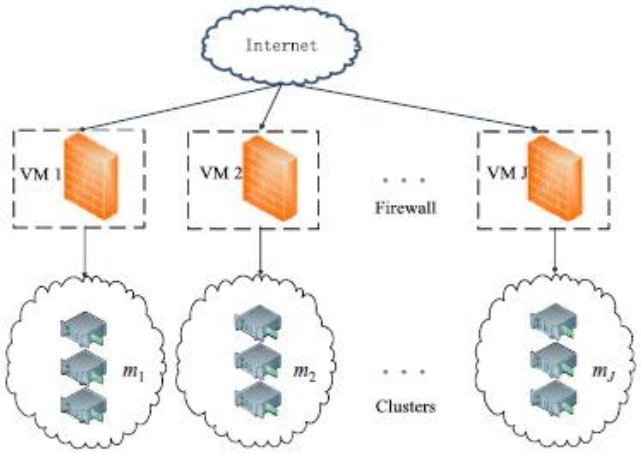

Fig. 1. A decentralized cloud firewall framework.

(superscript n stands for normal or non-attack). The M servers are grouped into J clusters ($m_1^n,\ldots m_j^n,\ldots m_J^n$) for processing legitimate packets. ($V_1^n,\ldots V_j^n,\ldots V_J^n$) denote the set of VM instances which host the parallel cloud firewall for each cluster. ($\lambda_1^n,\ldots \lambda_j^n,\ldots \lambda_J^n$) denote packet arrival rates to each firewall, then

$$\begin{cases} \sum_{j=1}^{J} m_j^n = M \\ \sum_{j=1}^{J} \lambda_j^n = \chi^n. \end{cases} \quad (1)$$

We define the corresponding variables in attack period as follows: x_a denotes packet arrival rate to the hosting servers (superscript a stands for attack), which are grouped into K clusters ($m_1^a,\ldots m_j^a,\ldots m_J^a$) for processing attack packets. ($V_1^a,\ldots V_j^a,\ldots V_J^a$) denote VM instances space, and ($\lambda_1^a,\ldots \lambda_j^a,\ldots \lambda_J^a$) denote packet arrival rates to each firewall. Similarly we have,

$$\begin{cases} \sum_{k=1}^{K} m_k^a = M \\ \sum_{k=1}^{K} \lambda_k^a = \chi^a. \end{cases} \quad (2)$$

3. RESOURCES PROVISIONING COST OPTIMIZATION

We first formulate resources provisioning cost. As firewall service rate modeling is critical to resources provisioning cost optimization, we establish a mathematical model according to cloud firewall rule matching discipline and derive that system service times follow geometric distribution.

3.1 Resources Provisioning Cost

Let T^n denote the unit time interval that CSPs charge VM instances. T^a denotes average attack duration in T^n. For simplicity, the scenario that various types of attacks occur with unequal attack rate and attack duration is not covered in this paper. In fact, our model can be easily extended to this general case.

Our primary goal is to optimize resources provisioning cost, while satisfying QoS requirement at the same time. It is intuitive that resources provisioning cost for our proposed cloud firewall depends on packet arrival rate. Given x^a and x^b, it further relies on how many clusters (J and K) are formed. Moreover, it is determined by VM instance configuration for the parallel firewalls. In order to cover the vast cloud firewall related parameter space, the resources provisioning cost is formulated as follows:

Minimize

$$T^n \sum_{j=1}^{J} p_j^n + T^a \sum_{k=1}^{K} p_k^a. \quad (3)$$

Subject to

$$\forall j \in [1, J], \begin{cases} \lambda_j^n \le \mu_j^n \\ \overline{r_j^n} \le \Delta T. \end{cases} \quad (4)$$

$$\forall k \in [1, K], \begin{cases} \lambda_k^a \le \mu_k^a \\ \overline{r_k^a} \le \Delta T. \end{cases} \quad (5)$$

Here p^n_j and p^a_k denote unit price of VM instance V^n_j in non-attack period and V^a_k in attack period, respectively (If the two VM instances are of the same type, then $p^n j = p^a_k$). μ^n_j and μ^a_k denote service rate of the two VM instances when running the cloud firewall, which are in terms of packets per second (pps) and will be given later. r^n_j and r^a_k are response time through firewall for cluster m^n_j and m^a_k in non-attack and attack period respectively, and they also will be given later. ΔT is an acceptable response time threshold specified in firewall customers QoS requirement.

The objective function (3) is to minimize resources provisioning cost for our proposed cloud firewall. Equations (4) and (5) are the conditions that have to be met when configuring VM instances for each firewall in non-attack and attack period, respectively. Concretely, QoS requirement constraint has to be met, and arrival rate to each firewall should be less than its service rate to keep the system in a stable state.

In general, CSPs specify a limitation of concurrent VM instances that are available to an account [4]. For example, this threshold is 20 in Amazon EC2. As a result, we can simply iterate J and K to find the optimal solution to equation (3). In each iteration, a greedy algorithm is applied to get an optimal cost for each J and K. Concretely, we rank the VM instances in ascending order according to service rate μ then choose VM instance which satisfies QoS requirement ΔT with least μ.

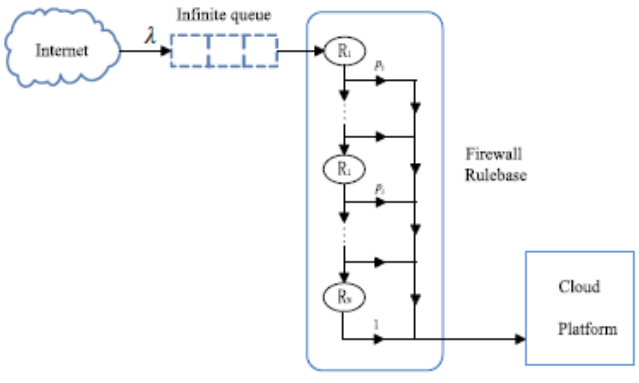

Fig. 2. Flow chart for firewall rule matching.

These obtained VM instances are just the VM configuration that leads to the optimal cost for each given J and K. Finally, by minimizing these optimal costs over J and K, we are able to find an optimal solution to Equation (3).

In order to simplify the calculation, we assume that packet arrival rate to each firewall is proportional to number of servers included in the cluster. Then the mean packet arrival rate to firewall for cluster mnj and mak are given by,

$$\lambda_j^n = \frac{\chi^n m_j^n}{M}, \quad (6)$$

$$\lambda_k^a = \frac{\chi^a m_k^a}{M}. \quad (7)$$

4. PERFORMANCE EVALUATION

In this section, we first validate our analytical model and investigate basic parameter settings of the proposed cloud firewall. Then the tradeoff among resources provisioning cost, QoS requirement and packet arrival rate is thoroughly studied.

4.1 Analytical Model Validation

In the following experiments, we take VM instances offered by Amazon EC2 for calculation. Two pricing options for VM instances are offered by Amazon EC2: on-demand and reservation. To capture the dynamic resources provisioning and allow for request of VM instances at any time, we employ the on demand pricing option.

We first have to give a sensible estimation of service rate of each VM instance, which is determined by N, T, m and p according to Equation (13). Here N is set 1,000 and p = 1/N. As cloud firewall should be transparent to users, we assume response time through each cloud firewall is in granularity of millisecond (which is reasonable according to analysis results in [5]). As a result, rule matching time T should be in granularity of microsecond. In this paper, T is set as 27 μs. Service rate of VM instances are listed in Table 1.

As discussed previously, we use average packet response time through the cloud firewall as a key metric for our performance evaluation. First, we are interested in the comparison between our M/Geo/1 model and the general M/M/1 model. The experimental results are shown in Fig. 4. It is easy to find that our M/Geo/1 model outperforms the M/M/1 model as it matches the simulation results much better. In other words, it's more reasonable to assume the firewall service rate follows a discrete geometric distribution than a continuous exponential distribution. Here λ is set at most 50 packets per second as service rate of the small VM instance is 58.

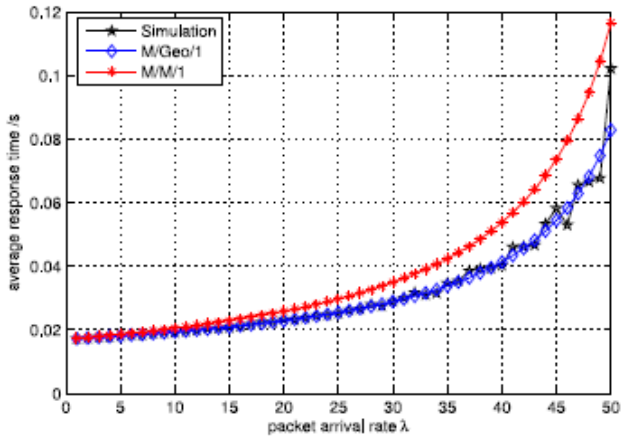

Fig. 4. Comparison of M/Geo/1 and M/M/1.

For an M/Geo/m queue, its closed-form response time is approximately given, which is a decision variable to resources provisioning cost optimization. Therefore, we have to check whether this approximation is reasonable. We simulate the relationship between average request response time and attack rate, and compare simulation results with analytical results derived for M/Geo/m. The simulation is conducted for an extra large VM instance, i.e., m = 8.

Both analytical and experimental results are illustrated in Fig. 5. Our M/Geo/m model is confirmed by the simulation results that the mean experimental response time fluctuates around the expectation obtained from Equation. As can be seen from Fig. 5, the average response time smoothly increase when attack rate grows. However, as offered load ρ→1, the response time increase sharply. The reason for this sharp increase is that the arrival packets to the extra large VM instance reaches its maximum processing rate of 1/t, which is approximately 468 packet per second (pps).

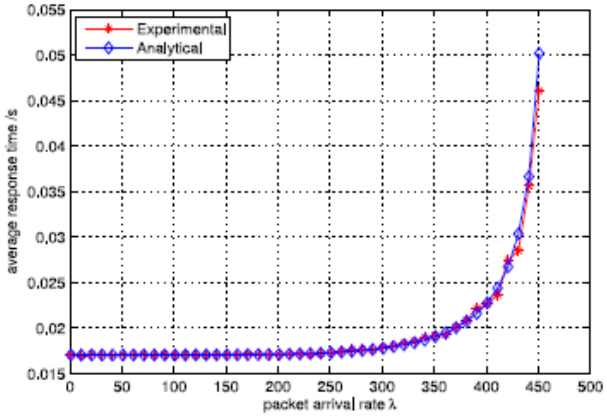

Fig. 5. Validation of approximate closed-from mean packet response time for M/Geo/m.

The results also confirm our earlier claim that a centralized firewall for a whole cloud platform is impractical. It is rather easy for packet arrival rate in attack period to exceed service rate of VM instances. A much larger N (e.g., 10,000 rules) makes it even worse. Based on these results, we claim that the proposed decentralized cloud firewall is necessary and feasible.

4.2 Firewall Parameter Settings

Cloud customers usually have personalized requirements for firewall, which is mainly due to that different applications are of varying degrees vulnerable to various attacks. For example, an e-business website is highly likely more vulnerable to phishing attack compared to a news site due to that cloud customers earn much more money from the former. Therefore, rule set in cloud firewalls differs for cloud customers. In this section, we aim to find the relationship between N, p and mean packet response time. In Fig. 6, we show the relationship between number of rules N and average response time of a cloud firewall. Here p=1/5, 000 and λ is set at most 90 pps due to that service capacity of the extra large VM instance is now approximately 93 pps according to Equation. From our analytical model, it's expected that more rules decrease service capacity and result in more time to process arrival packets of a given attack rate, which is confirmed by the simulation results.

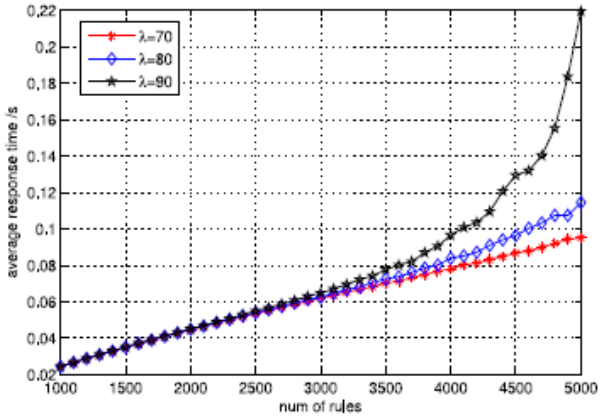

Fig. 6. The relationship between number of rules and mean packet response time.

Fig. 7 exhibits the impact of rule matching probability p against average response time. λ is set as 90. It's easy to find that a larger matching probability p leads to less response time, which means firewall service providers are encouraged to put rules easier to match on top of the rule list in cloud firewall to satisfy firewall service customers QoS requirement.

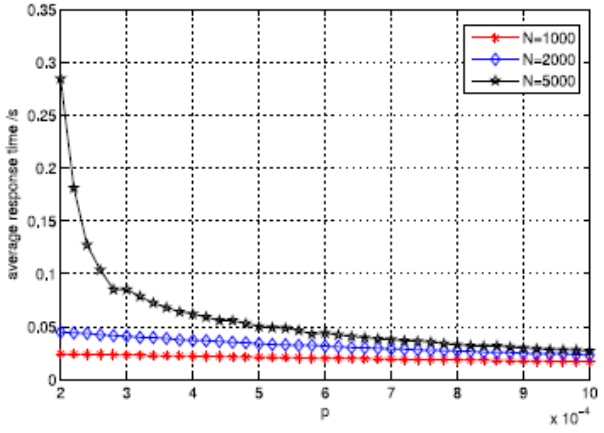

Fig. 7. The relationship between rule matching probability and mean packet response time.

5 FURTHER DISCUSSION

To the best of our knowledge, this paper is an early work to discuss resource provisioning cost optimization in the context of cloud firewall. As a new research field, there are many other mathematical tools available to address this optimization problem, such as game theory [25], integer linear programming [24] and stochastic programming [26]. Due to the limitations of knowledge, time and space, we have only employed queuing theory in this paper.

Our analytical model assumes that packet arrivals to the cloud server follow Poisson distribution, and the service times follow Geometric distribution. For certain types of network traffic, assuming Poisson arrivals is feasible [27]. However, for general traffic like Ethernet, their arrivals do not always follow a Poisson distribution but are rather bursty or heavy-tailed [28], [29]. Also, the assumption that all rules share the same matching probability is hard to meet in reality. Considering different matching probability and non-Poisson distribution will make a closed-form analytical solution intractable. To address these limitations, Discrete Event Simulation (DES) can be employed [30].

6 CONCLUSION AND FUTURE WORK

In this paper, we point out that it's impractical to establish a firewall for a whole cloud data center. However, cloud service providers possess a potential to provide cloud firewalls for individual cloud customers. In view of this challenge, we propose a decentralized cloud firewall framework, where several firewall running in parallel to guarantee QoS requirement. As resources are dynamically allocated in cloud firewall, we investigate how to optimize the resources provisioning cost. We establish novel queuing theory based model for performance analysis of the proposed cloud firewall, where firewall service times are modeled to follow geometric distribution. Extensive simulations confirm that M/Geo/1 reflects the cloud firewall real system better than traditional M/M/1. Besides, it is feasible to set up firewall for individual cloud hosted services with an affordable cost to cloud customers.

As future work, we first plan to improve the decentralized framework to capture more personalized details in application level. Second, we would like to propose a pricing model for the cloud firewall, which helps to achieve a financial balance between provider and customer. Real cloud environment experiments for the proposed cloud firewall are also expected in the near future.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," Commun. ACM, vol. 53, no. 4, pp. 50–58, 2010.

[2] Z. Xiao and Y. Xiao, "Security and privacy in cloud computing," IEEE Commun. Surveys Tuts., vol. 15, no. 2, pp. 843–859, May 2013.

[3] C. Hoff. (2008). Cloud computing security: From ddos attack (distributed denial of servie) to edos (economic denial of

sustainability) [Online]. Available: http://www.rationalsurvivability.com/ blog/?p=66.

[4] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: Exploring information leakage in third-party compute clouds," in Proc. 16th ACM Conf. Comput. Commun. Security, 2009, pp. 199–212.

[5] K. Salah, K. Elbadawi, and R. Boutaba, "Performance modeling and analysis of network firewalls," IEEE Trans. Netw. Serv. Manage., vol. 9, no. 1, pp. 12–21, Mar. 2012. Fig. 9. Tradeoff among resource provisioning cost, QoS requirement and attack rate. TABLE 2 Optimal Resources Provisioning Cost and Corresponding VM Instance Configuration LIU ET AL.: A DECENTRALIZED CLOUD FIREWALL FRAMEWORK WITH RESOURCES PROVISIONING COST OPTIMIZATION 629

[6] D. Rovniagin and A. Wool, "The geometric efficient matching algorithm for firewalls," IEEE Trans. Dependable Secure Comput., vol. 8, no. 1, pp. 147–159, Jan./Feb. 2011.

[7] A. X. Liu and F. Chen, "Privacy preserving collaborative enforcement of firewall policies in virtual private networks," IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 5, pp. 887–895, May 2011.

[8] A. X. Liu, "Firewall policy change-impact analysis," ACM Trans. Internet Technol., vol. 11, no. 4, p. 15, 2012.

[9] A. R. Khakpour and A. X. Liu, "First step toward cloud-based firewalling," in Proc. IEEE 31st Symp. Reliable Distrib. Syst., 2012, pp. 41–50.

[10] S. Yu, R. Doss, W. Zhou, and S. Guo, "A general cloud firewall framework with dynamic resource allocation," in Proc. IEEE Int.Conf. Commun., 2013, pp. 1941–1945.

[11] S. Yu, Y. Tian, S. Guo, and D. Wu, "Can we beat ddos attacks inclouds?" IEEE Trans. Parallel Distrib. Syst., in press, 2014.

[12] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage,"Inferring internet denial-of-service activity," ACM Trans. Comput. Syst., vol. 24, no. 2, pp. 115–139, 2006.

[13] U. Sharma, P. Shenoy, S. Sahu, and A. Shaikh, "A cost-aware elasticity provisioning system for the cloud," in Proc. IEEE 31st Int. Conf. Distrib. Comput. Syst, 2011, pp. 559–570.

[14] R. Wartel, T. Cass, B. Moreira, E. Roche, M. Guijarro, S. Goasguen, and U. Schwickerath, "Image distribution mechanisms in large scale cloud providers," in Proc. IEEE 2nd Int. Conf. Cloud Comput. Technol. Sci., 2010, pp. 112–117.

[15] J. Zhu, Z. Jiang, and Z. Xiao, "Twinkle: A fast resource provisioning mechanism for internet services," in Proc. IEEE INFOCOM, 2011, pp. 802–810.

[16] J. Cao, K. Hwang, K. Li, and A. Zomaya, "Optimal multiserver configuration for profit maximization in cloud computing," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1087–1096, Jun. 2012.

[17] H. Khazaei, J. Misic, and V. B. Misic, "Performance analysis of cloud computing centers using m/g/m/m+ r queuing systems," IEEE Trans. Parallel Distrib. Syst, vol. 23, no. 5, pp. 936–943, Apr. 2012.

[18] H. Khazaei, J. Misic, V. Misic, and S. Rashwand, "Analysis of a pool management scheme for cloud computing centers," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 5, pp. 849–861, May 2012.

[19] H. Khazaei et al., "Performance of cloud centers with high degree of virtualization under batch task arrivals," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 12, pp. 2429–2438, Dec. 2013.