ISSN No. 0976-5697

Volume 5, No. 7, September-October 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Partitioning of Resource Provisions for Cloud Computing Infrastructure against DoS and DDoS Attacks

Emmanuel C. Ogu[1], Alao, O.D.[2], Omotunde, A.A.[3], Ogbonna A.C.[4], Izang A.A.[5]
[1,2,3,4,5]Department of Computer Science and Information Systems,
School of Computing and Engineering Sciences, Babcock University,
Ilishan-Remo, Ogun State. Nigeria.

Abstract: The Internet is growing to become a necessary and indispensable tool for the common man. Security has remained a hindrance to reaping the full, immense benefits of the Internet. Cybercrimes have been on the rise, and are becoming even easier to perpetrate even for the unskilled and naïve users. Denial of Service (DoS) is becoming one of the most common and most devastating of known cybercrimes that leave huge financial losses in its wake, with possibilities ofeven loss of lives in developed countries. This research proposes a model for partitioning of resource provisions in cloud computing infrastructure, as a means of mitigating against DoS attacks, especially at the Application Layer. This model focuses on securing critical computing resources of cloud computing infrastructure by providing different resource pools for servicing requests of applications and processes whose resource consumption levels are within safe limits, and for those that are suspected to be headed for a DoS. A full theoretical framework for this model is established with the use of an algorithm, a flowchart and a functional flow block diagram (FFBD).

Keywords: Virtualization, Cloud Computing, Resource Allocation, Denial of Service, Botnets, Functional Flow Block Diagrams.

I. INTRODUCTION

A **Denial of Service** (DoS) Attack is a class of cybercrime attacks that aims at exhausting critical computing resources (*CPU, Bandwidth and Memory resources*) of servers and computing infrastructure, thus making it impossible for clients / users who have legitimately subscribed to use these resources to have access to them. A Distributed Denial of Service (DDoS) Attack involves the use of compromised machines (known as "zombies") to orchestrate a DoS attack; it is typically achieved by overwhelming target machines with such massive volume of malicious requests or useless traffic harnessed from various compromised sources (acquired over the internet or any other network) that they are unable to respond to legitimate requests from legitimate users in a timely manner [1]. This class of attacks have more recently been targeted towards incapacitating cloud computing infrastructure that deliver various services over the Internet, such that they are unable to service legitimate requests from authorized subscribers. The practice of distribution and distributed coordination greatly amplifies the power and fatality of a DoS attack and greatly complicates the task of detection and defence [2]. Figure 1 illustrates how DDoS attacks are orchestrated.

What distinguishes DoS attacks, especially those that are orchestrated over networks and its distributed form is not so much the content of the packets in the attack traffic, but the sheer, overwhelming nature of the volume of the attack traffic. Because the strength of a DoS attack lies in its volume:

a. Attackers can greatly complicate the task of defence by sending a wide variety of different traffic packets that may be similar to legitimate traffic (in behaviour and certain properties).
b. The traffic volume in a DoS attack must be sufficient enough to successfully overwhelm the resources and provisions of the target. This is why the attacker typically seeks to gain control of zombies in order to channel such massive volume of traffic in the direction of the victim [2].

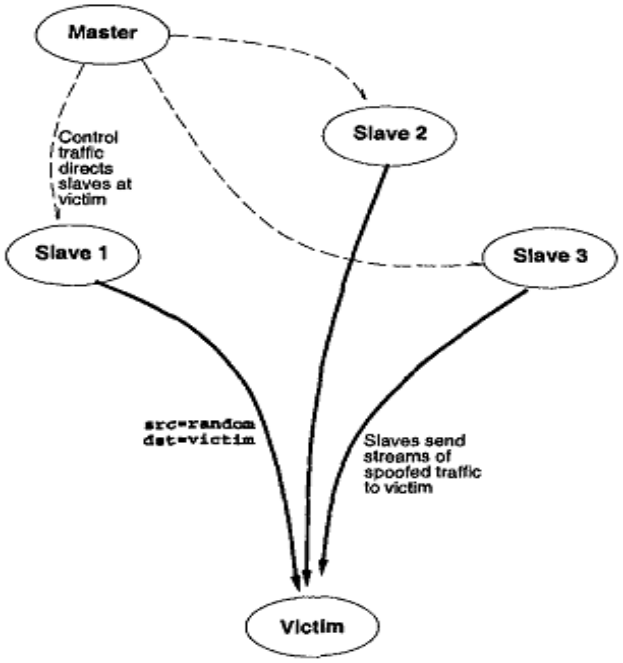

Figure 1: Structure of a typical DDoS Attack [3]

Flash Crowds are a phenomenon that occur when a large crowd of legitimate users try to gain access to a server resource or service at the same time [2]. Flash Crowds can also cause DoS to occur, and in fact go a long way to further complicate the task of detecting and controlling DoS attacks. This is so because flash crowd traffic and DoS attack traffic have certain characteristics in common that have been pointed out by [2] as in table 1 below, and distinguishing them under the rush and load of DoS traffic can be a really difficult task.

DOI: 10.13140/2.1.2259.7129

Table 1: Comparison between Bandwidth Attacks and Flash Crowds [2]

	Bandwidth Attack	Flash Crowd
Network impact	Congested	Congested
Server impact	Overloaded	Overloaded
Traffic	Malicious	Genuine
Response to traffic control	Unresponsive	Responsive
Traffic type	Any	Mostly Web
Number of flows	Any	Large number of flows
Predictability	Unpredictable	Mostly predictable

Cloud Computing (CC) is defined by the United States National Institute of Standards and Technology (NIST) as "a model for enabling *ubiquitous, convenient, on-demand* network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction; having characteristics of on-demand self-service, broad network access, resource pooling, rapid elasticity and payment per usage of various business models." [4]. Cloud computing services are delivered through three standardized service models: the Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and the Software as a Service (SaaS) Models.

The service models specified above directly define the three layers comprised in the core of most modern Cloud Computing Infrastructure. Each of these layers offer the specified types of services to a particular segment of the consumer market while at the same time paying for the services provided by the preceding layer (except the IaaS layer) [5]. This is shown in the figure 2below:

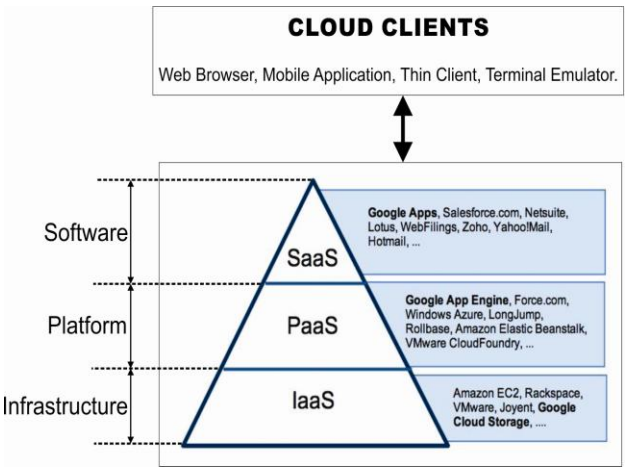

Figure 2: Cloud Computing Layers [6]

Virtualization is the technology that provides the abstraction that Cloud Computing enjoys by taking a physical resource such as a server and dividing it into virtual resources called virtual machines. In case of server consolidation, many small physical servers are replaced by one larger physical server, to increase the utilization of costly hardware resources such as CPU time [7].

II. PROBLEMSTATEMENT

Denial of Service (DoS) attacks have remained a nagging threat to the success of Cloud computing, and has greatly limited its rate of adoption in education, business, economics and commerce; as well as to the rapid development of the Internet.

In recent times, Cybercrimes in the United States alone have been estimated to cost losses of up to $100 billion annually [8].

Generally, aDoS-style Attack is known to be very hard to defend againstbecause most DoS attacks target and rapidly consume network and transport layers resources (and also the resources of the underlying computing infrastructure through the Application-Layer DoS that is currently gaining rapid popularity), where it is difficult to authenticate whether an access, packet, process, application or connection is genuine or illegitimate and malicious [9].

Hence, a more exigent objective would be to direct more focus to the rate at which these accesses, packets, processes, applications and connections consume critical computing resources, the unavailability of which results in a denial of service, as against analysing through and authenticating the various contents of these. This new focus necessitates the need for partitioning provisioned resources for cloud subscribers such that different resource pools are available for servicing requests of applications, processes, packets and connections, whose resource consumption levels are within safe limits, and for servicing those that are noticed to be bothering on critical computing resources, separately.

III. RELATED WORKS AND PAST RESEARCHES

Resource Allocation (RA) for cloud computing is the process of assigning (provisioning) available resources over the internet to various cloud applications that are in need of them. When Resource Allocation is not managed properly, it could result in resource starvation for some critical services that have subscribed to use such resources. Resource Provisioning (RP) seeks to solve this problem by allowing cloud service providers to manage resources for each and every subscribed module. Resource Allocation Strategies (RAS) are concerned with integrating the activities of cloud providers for the utilization and allocation of scarce resources to meet the needs of various cloud applications within the cloud environment. This requires a knowledge of the type and amount of resources needed by each application to deliver on user jobs. Optimal RASs are to consider the order and time of resource allocation, as well as avoiding issues relating to resource contention, scarcity of resources, resource fragmentation, over-provisioning as well as under-provisioning [10].

A **policy** is simply a set of rules for allocation of resources when resource demands exceed resource supplies. Policies could either be made to seek efficient usage through proper allocation of whatever is left, or seek maximal revenue generation (where profit is involved) [11]. Most virtual machine monitor designers and cloud computing service providers fall guilty of policies in the latter jurisdiction (to maximize revenue).

[10], carried out a survey of Resource allocation strategies for cloud computing. These strategies were generally classified based on Execution Time, Policy, Virtual Machine, Gossip, Utility, Hardware Resource Dependency, Auction, Application and Service Level Agreement. Figure 3 below gives a detailed description of this classification:

DOI: 10.13140/2.1.2259.7129

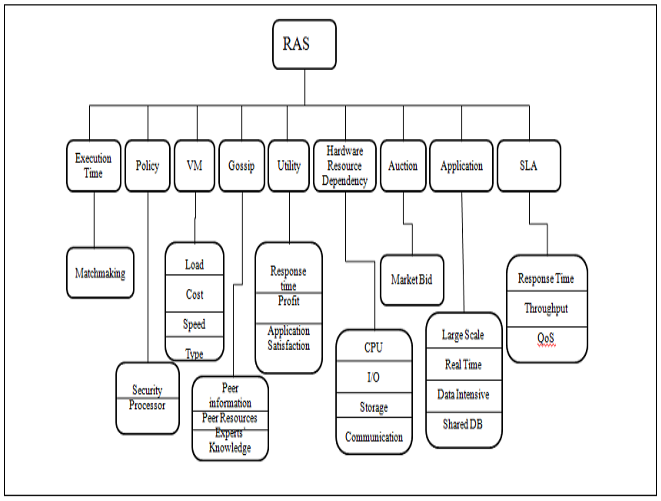

Figure 3: Resource Allocation Strategies in Cloud Computing [10]

[12], carried out a survey on the categories of threats to cloud computing infrastructures and possible solutions to these threats. This research pointed out Denial of Service (DoS) and Distributed Denial of Service (DDoS) attacks as a class of attacks that affect the Availability of Cloud computing services and Infrastructure and proposed a solution to this event in an increase in the number of critical resources that are usually affected by such attacks (CPU time/cycles, Disk Cycles, Memory and Network bandwidth) by what is known as massive overprovisioning. This solution, however, may only be feasible with very few attack zombies in a DDoS attack and very few request packets in a DoS attack. All that is required is simply to sufficiently increase the number of packets that are needed to overwhelm the strength of resources (such information could be gathered at reconnaissance) in a DoS attack or simply to increase the number of attack zombies (in order to generate sufficiently massive packets) in the distributed variant. To what end then, can we keep increasing the amount of critical resources in an attempt to make a host machine resilient enough to denial-of-service attacks and distributed denial-of-service attacks? This literature, however, succeeded in providing a good classification of the threats that can be encountered in Cloud computing.

[13], proposed a solution for combating SlowPOST Denial of Service attacks that target the OSI application layer. Using this method, server resources / threads are partitioned into two different resource pools for handling various requests with differing data rates. A threshold is set, and requests / connections that fell below the threshold are handled by a different resource pool while those that fell above the threshold were handled by another resource pool partition. This solution requires clients to specify additional information – the data rate – which allows the server to estimate the time to receive the full payload carried by the request. Requests or connections were sustained based on their ability to consistently excel in a Periodic Compliance Verification test (in which their entire payload was expected to arrive at the server within a specific time) that is conducted over a pre-configured number of times, else such connections / requests are dropped. An algorithm was provided for implementing this solution, and it shows great promise and ability in being able to counter application layer DoS attacks. However, this technique seemed to focuson analysing packets and their content for data rates for compliance verification rather than focusing on the preservation of critical resources (which has been discovered to be a more exigent concern in the face of a DoS), and it is also possible for both the upper and lower resource pools to be used up at the same time once an attacker is able to gain precise knowledge of the threshold mark through proper reconnaissance. And since the distinguishing strength and devastation of most DoS attacks lie more in its volume than in its content, as has been identified by [2], focusing on the content of the traffic (analysis of data rates / payload delivery timing) rather than on its volume and resource consumption indices in relation to what is left over of the resources, may not be considered best practice in most cases.

IV. MODEL FOR RESOURCE PARTITIONING

This model pre-allocates subscribed resources and allocates a percentage of the subscribed resources (which clients/subscribers agree to, and is documented as part of the Quality of Service [QoS] agreement) as reserves or quarantines that would be used to service attack traffic,applications, processes, packets and connections, that have been observed, by any means of monitoring, to be headed for a DoSwhen the major resource provisions are under attack.

The client or administrator provides as inputs, the total resource allocations provisioned to the subscriber as found in the QoS agreement and inputs the percentage of the total resource provision that is to be left as reserves (also as agreed in the QoS agreement). When these two inputs are provided, the resource partitioning system passes a system call containing these parameters to the software or application that is responsible for resource management (allocating, holding and leasing) in the operating environment. The resource management software (a hypervisor in a virtualized environment or an operating system resource manager in a native environment) would thenavail resources to the major partition, and create another partition out of that as reserves, according to the percentage supplied.

When a process, packet traffic or application has been identified to be straining the provisioned resources, the "culprit" is identified and moved from the main resource provisions into the reserve provisions. In practice, the reserve provisions would serve as a quarantine where the process, packet traffic or application would still be serviced, but at a slower, controlled rate by a much more rationed quota of resources (under the watchful eye of some form of monitor). This "culprit" would remain in this "quarantine" until its resource consumption levels normalize, then it would be transferred back into the quarters of the main resource provisions to continue normal operation.

V. THE ALGORITHM

The algorithm for thismodel system is given below:
Step 1: **Start**resource_partitioning_system
Step 2: **Input**total_client_resource_provision
Step 3: **Input**percentage_reserve_provision
Step 4:**Partition**(reserve_provisions_from_main_resource_provision)
Step 5: **Return**main_resource_provision; reserve_resource_provision
Step 6: **Stop**resource_partitioning_system

DOI: 10.13140/2.1.2259.7129

VI. THE FLOWCHART

The algorithm is illustrated using the flowchart in figure 4 below:

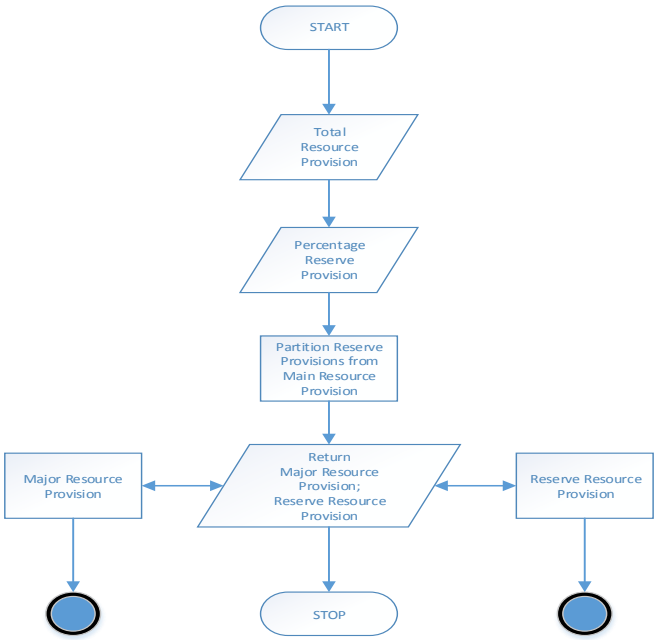

Figure 4: Flowchart for the Resource Partitioning System

VII. THE FUNCTIONAL FLOW BLOCK DIAGRAM

Functional Flow Block Diagrams (FFBDs) are basically a schematic diagram that is used to show / illustrate a process or procedure, broken down into functions and sub-functions [14].

The functional flow block diagram (FFBD) for the resource partitioning system is given in figure 5 below:

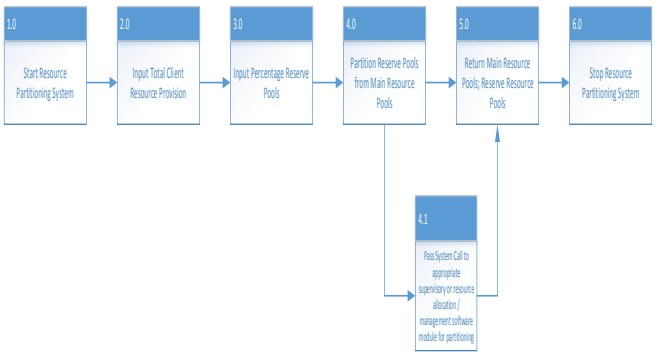

Figure 5: Functional Flow Block Diagram of the Resource Partitioning System

VIII. CONCLUSION

This model for resource partitioning does not seek interest in the content of the traversing packet, process or application; rather it is concerned with the more exigent objective of securing critical (sometimes limited and expensive) resources in cloud computing environments. Hence, it would be able to secure against Application-Layer DoS Attacks that are targeted towards using cloud service applications and web servers to rapidly usurp resources in cloud computing environments.

Flash Crowds would also benefit from this resource allocation model, because flash crowd traffic would not just be discarded summarily, but would still be serviced; only at a slower, better controlled rate.

IX. RECOMMENDATION AND FURTHER STUDY

This resource partitioningsystem is to be implemented at the hypervisor or pre-hypervisor level (essentially a very high privilege level), for virtualized cloud computing infrastructures, or at the operating system level (for native computing infrastructures), for it to possess the adequate permissions necessary for smooth operation. The implementation of this resource partitioning system on some open-source hypervisor such as Xen would also be considered in future researches, as well as an implementation of a resource monitor for the effective enforcement of compliance to threshold limit specifications.

X. REFERENCES

[1]. Sabahi, F. (2011). Virtualization-Level Security in Cloud Computing. Communication Software and Networks (ICCSN), IEEE, 250-254.

[2]. Peng, T., Leckie, C., & Ramamohanarao, K. (2007). Survey of network-based defense mechanisms countering the DoS and DDoS problems. ACM Computing Surveys (CSUR): Article No. 3, 39(1). doi:10.1145/1216370.1216373

[3]. Paxson, V. (2001, July). An analysis of using reflectors for distributed denial-of-service attacks. ACM SIGCOMM Computer Communication Review, 31(3), 38-47. doi:10.1145/505659.505664

[4]. Mell, P., & Grance, T. (September 2011). The NIST Definition of Cloud Computing. Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, United States Department of Commerce. Gaithersburg, MD 20899-8930: National Institute of Standards and Technology. Retrieved January 28, 2014, from http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[5]. Strømmen-Bakhtiar, A., & Razavi, A. R. (2011). Cloud Computing Business Models. Springer Computer Communications and Networks, 43-60.

[6]. Gartner AADI Summit. (2009). Cloud Computing as Gartner Sees it. Gartner's Application Architecture, Development & Integration Summit.

[7]. Gurav, U., & Shaikh, R. (2010). Virtualization – A key feature of cloud computing. Proceedings of the International Conference and Workshop on Emerging Trends in Technology (ICWET 2010) (pp. 227-229). Mumbai, Maharashtra, India.: Association for Computing Machinery.

[8]. Gorman, S. (2013, July 22). Annual U.S. Cybercrime Costs Estimated at $100 Billion; Study Casts Doubt on Previous, Higher Figures. Wall Street Journal Publications.

[9]. Peng, T., Leckie, C., & Ramamohanarao, K. (August 2003b). Protection from Distributed Denial of Service Attack Using History-based IP Filtering. The University of Melbourne, Australia, Department of Electrical and Electronic Engineering. Victoria 3010, Australia: ARC Special

DOI: 10.13140/2.1.2259.7129

Research Center for Ultra-Broadband Information Networks. Retrieved January 31, 2014, from http://ww2.cs.mu.oz.au/~tpeng/mudguard/research/icc2003.pdf

[10]. Vinothina, V., Sridaran, R., & Ganapathi, P. (2012). A Survey on Resource Allocation Strategies in Cloud Computing. International Journal of Advanced Computer Science and Applications (IJACSA), Volume 3(Issue 6), 97-104.

[11]. Shneidman, J., Ng, C., Parkes, D. C., AuYoung, A., Snoeren, A. C., Vahdat, A., & Chun, B. (June, 2005). Why Markets Could (But Don't Currently) Solve Resource Allocation Problems in Systems. Proceedings of the 10th USENIX Workshop on Hot Topics in Operating Systems (HotOS-X), 37-42.

[12]. Nagaraju, K., & Sridaran, R. (2012, September). A Survey on Security Threats for Cloud Computing. International Journal of Engineering Research & Technology (IJERT), Volume 1(Issue 7), 1-10.

[13]. Raghunath, A., Ramachandran, S., Vaidyanathan, S., & Subramania, S. (2013, September). Data Rate Based Adaptive Thread Assignment Solution for Combating the SlowPOST Denial of Service Attack. ACM SIGSOFT Software Engineering Notes, Volume 38(Issue 5), 1-5.

[14]. Blanchard, B., & Fabrycky, W. (2011). Systems Engineering and Analysis (Fifth Edition). New Jersey: Pearson.

DOI: 10.13140/2.1.2259.7129