Profiling-as-a-Service in Multi-Tenant Cloud Computing Environments

Kuai Xu, Feng Wang
Arizona State University
Email: {kuai.xu, fwang25}@asu.edu

Lin Gu
Hong Kong University of Science and Technology
Email: lingu@cse.ust.hk

***Abstract*—Cloud computing integrates data, applications, users and servers on a vast scale and enables a global optimization of computing resources. However, due to security threats from both outside and inside the cloud, security remains as a significant challenge and obstacle in the wide adoptions of cloud computing paradigms. To enhance the security of networks, applications and data in the cloud, this position paper proposes to develop a profiling-as-a-service architecture to characterize, understand and profile network traffic at multiple layers in the multi-tenant cloud computing environment: network routers, hypervisors, virtual instances and applications. The proposed architecture will not only provide an in-depth understanding on traffic patterns of cloud tenants, but also enhance the security of cloud computing by collaboratively detecting and filtering unwanted traffic towards cloud instances.**

I. Introduction

Cloud computing integrates data, applications, users and servers on a vast scale and enables a global optimization of computing resources. However, due to security threats from both outside and inside the cloud, security remains as a significant challenge and obstacle in the wide adoptions of cloud computing paradigms. For example, the in-cloud experiments in [3] demonstrate the vulnerabilities associated with shared virtual machines (VM) on the same physical host and the possibility of mounting cross-VM side-channel attacks to collect information from the target VMs. Given the magnitude and diversity of security threats towards cloud computing, it is crucial to develop effective solutions to ensure the security and high-availability of data, applications, and networks for cloud tenants.

To enhance the security of cloud computing, the central challenges are i) the vast amount of network traffic in the cloud and the diversity of cloud tenants, ii) the variety of security threats that include traditional threats towards cloud tenants and emerging threats brought by the cloud computing paradigm, iii) the launching points of the attacks from inside and outside the cloud; and iv) the "untrusted" nature of the multi-tenant cloud computing environment [4]. Many recent research have been conducted on new data center architecture [5], [6], [7] and network traffic measurement in cloud computing [8]. However, there has been little attempt to profile network traffic of cloud instances. Existing techniques for cloud computing security such as access control lists or firewalls are widely deployed on data center routers and virtualization servers, however they are insufficient for securing cloud instances as cloud computing tenants face a variety of security challenges such as intrusion attempts, port scanning, and denial of service attacks from outside the cloud as well as from inside the cloud, e.g., cloud providers or other cloud tenants.

In this paper we propose a *profiling-as-a-service* architecture to analyze and characterize network traffic of cloud instances at multiple layers in the multi-tenant cloud computing environment: 1) border routers of cloud networks, 2) hypervisors of virtualization servers, 3) virtual instances (VMs), and 4) applications. Traffic profiling has recently become an essential technique for securing and managing backbone and edge networks, e.g., building normal and anomalous network behavior profiles [9], [10], detecting traffic obfuscation and encryption [11], and accurate identification of network applications [12], [13]. Our proposed layered approach builds hierarchical traffic profiles for cloud instances, and provides an in-depth understanding of network traffic towards cloud instances. The proposed architecture consists of four system components that build upon each other to establish *profiling-as-a-service* in the cloud: i) a layered approach of profiling network traffic of cloud instances, ii) behavior models and structural models based on communication patterns of cloud instances, iii) a collaborative solution for detecting unwanted traffic in the cloud, and iv) a profiling-aware sampling algorithm for improving the robustness of the proposed architecture during sudden traffic surges caused by anomalous events.

The goal of the profiling service is to provide an in-depth understanding of traffic patterns for cloud tenants and to enhance the security of cloud computing by collaboratively detecting and filtering unwanted traffic towards cloud instances. To demonstrate the feasibility of the proposed architecture in the real cloud computing environment, we will design, implement and evaluate a prototype system.

The remainder of this position paper is organized as follows. Section II briefly discusses related work. Section III presents the proposed architecture. Section IV concludes this paper and discusses our ongoing work towards developing the profiling architecture in cloud computing.

II. Related Work

The success and challenges of cloud computing have recently draw broad attentions from the networking and system research community. In a view of cloud computing [14],

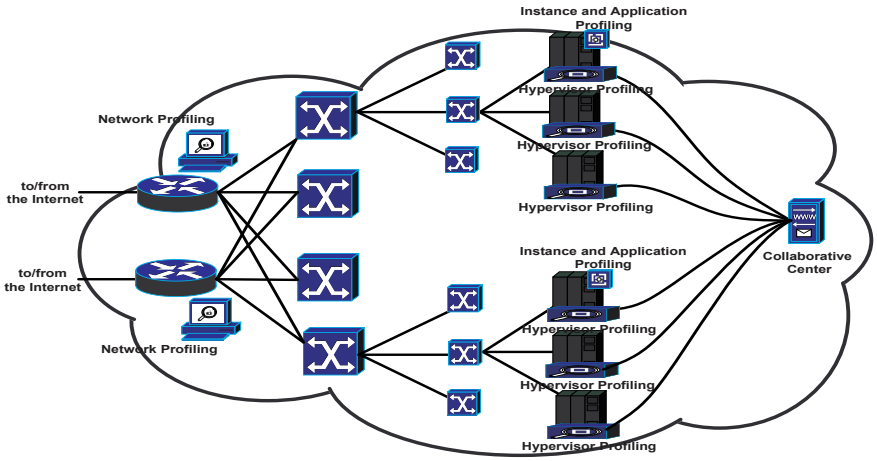

Fig. 1. Architecture of layered profiling in cloud computing.

Armbrust et al. summarize top 10 obstacles and research opportunities for cloud computing. As the infrastructure of cloud computing, the data centers play an increasingly important role in the cloud computing environment. Commercial cloud providers such as Amazon Web Services [15] also provide overviews of the physical and operational security processes for ensuring the security of networks, data and applications for cloud customers.

In light of the potential attacks and threats towards cloud computing, security has become one of the major concerns for the adoptions of cloud computing [16]. The recent work [3] introduces the vulnerabilities with shared virtual machines (VM) from cloud computing providers and demonstrates the feasibility of mounting cross-VM side-channel attacks to gain information from the target VMs. In [17] Ertaul et al. survey security challenges in cloud computing environment, while Chen et al. identified two new facets to cloud computing security [18], namely "the complexities of multi-party trust considerations" and "the ensuring need for mutual auditability". [19] first identifies security concerns on multiple layers arising in cloud computing, and subsequently outlines a policy-based security approach for cloud computing through defining security polices at various layers including networking, storage, systems management and applications.

Network traffic profiling has been extensively studied in the recent years for understanding network traffic in Internet backbone networks and edge networks [9], [12], [10]. For example, [9] builds behavior profiles of end hosts and network applications using traffic communication patterns without any presumption on what is normal or anomalous, while in [12] the authors study the host behaviors at three levels with the objective to classify traffic flows using packet header information only. [10] creates a traffic profile for each network prefix through behavior analysis of aggregated traffic. Unlike these work, this paper attempts to build traffic profiles across all layers in the multi-tenant cloud computing environments for improved security and management in the cloud.

III. Architecture of Profiling-as-a-Service in the Cloud

A. Architecture of Layered Profiling in Cloud Computing

In light of the security and privacy challenges of cloud computing, it becomes increasingly important for cloud customers to know *what happens to their cloud instances* in the multi-tenant cloud computing environment managed by a third-party cloud provider. Towards this end, we propose to build the *profiling-as-a-service* architecture for establishing hierarchical traffic behavior profiles of cloud instances at multiple layers — border routers of cloud networks (network profiling), hypervisors of virtualization servers (hypervisor profiling), virtual instances (instance profiling), and applications (application profiling). Figure 1 illustrates a schematic diagram of the layered *profiling-as-a-service* architecture, in which each layer analyzes cloud traffic from a different perspective and thus provides a unique insight on traffic patterns of cloud instances.

Figure 2 shows the hierarchical relationships of these four levels in the cloud hierarchy — network-level, hypervisor-level, instance-level and application-level. The lower two layers, network and hypervisor profiling, analyze network flows and unwanted traffic of all cloud instances at the routers and hypervisors with coarse granularity, while the higher two layers focus on the fine-grained traffic analysis for individual instances or their applications. Thus, in the proposed *profiling-as-a-service* architecture, each layer provides a unique view to traffic profiles and behavior patterns of cloud instances.

The major intuition of the layered profiling approach lies in that each layer of the cloud computing environment provides a unique perspective on network traffic of cloud instances. For example, in the experiments of mounting cross-VM side-channel attacks [3] the step of network probing for building cloud cartography utilizes wget scan to determine the liveness of EC2 instances, and such activities would leave distinct traffic footprints at each level of the cloud hierarchy and lead to different traffic profiles at four levels. Hence, the combined insights from each layer lead to an in-depth understanding

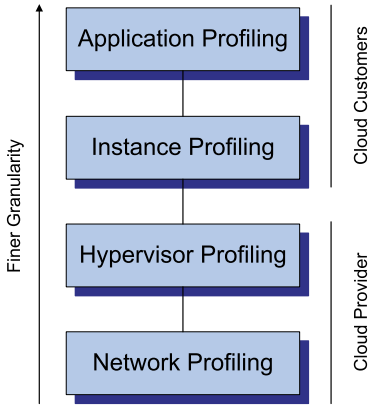

Fig. 2. Hierarchical relationships of the profiling layers.

on the traffic patterns of cloud instances. We could in turn use the complementary profiles in these layers to build a comprehensive and correlated traffic profile of cloud instances. The practical applications of the first two layers include i) understanding network-level traffic patterns of cloud instances and ii) correlating cloud-wide unwanted traffic towards cloud instances, while the applications of the latter two layers include i) revealing the overall traffic patterns and end-user access behaviors of cloud instances and ii) generating application-specific traffic patterns or detecting malicious packet payload to cloud applications.

B. Designing the Profiling-as-a-Service Infrastructure

In the rest of this section we will describe how to profile network traffic at each layer and address the inherent challenges of the *profiling-as-a-service* architecture, such as large volume of network traffic during denial of service attacks. Specifically, we will address these following problems: i) how to build the *profiling-as-a-service* architecture in the cloud networks, ii) what traffic features should be included in traffic profiles, iii) how to profile network traffic and behavior patterns in normal conditions and anomalous events, and iv) how to correlate profiles from distributed hypervisors and instances in the cloud for collaborative security monitoring.

1) Profiling Traffic at Network-Level: The first step of the *profiling-as-a-service* infrastructure focuses on the incoming or outgoing traffic observed at the border routers that connect cloud networks to the Internet. Profiling traffic at the network-level provides a broad view of traffic patterns of cloud instances. However, due to the sheer volume of network traffic to/from thousands of cloud instances, it remains a daunting task to analyze vast network traffic in the cloud networks. Therefore, the size of cloud traffic data calls for lightweight and efficient algorithms to make sense of these traffic and to generate meaningful traffic summaries of cloud instances at the cloud network level. Towards this end, we plan to explore entropy concepts from information theory and histogram analysis to analyze the distribution of traffic features for cloud instances at the network level.

Network profiling analyzes the incoming and outgoing traffic of border routers at the cloud networks, and builds high-level traffic summaries of of all instances through network flows. These profiles can be used to correlate common threats, such as worms and network-wide scanning, and to detect and mitigate volume-based traffic anomalies, such as denial of service attacks.

2) Profiling Traffic at Hypervisor Level: Hypervisor profiling collects, analyzes and correlates "unwanted traffic" captured by firewalls deployed on the hypervisor layer of virtualization servers that support multiple instances. In addition, we establish a central collaborative center that communicates with distributed hypervisors in the cloud, correlates "unwanted traffic" filtered by distributed hypervisor firewalls, and reports the aggregated trends of security threats to all cloud instances. Hence, hypervisor profiling could become a powerful technique for detecting low-volume attacks such as scanning activities and penetration attempts towards cloud instances.

One of the key technologies that drive the cloud computing paradigm is the use of virtualization, which allows multiple instances (also called virtual machines) to run on the same physical machine. These multiple instances are isolated from each other through the hypervisor layer (also called virtualization layer). The hypervisor layer arbitrates and manages CPU, physical memory and I/O devices among multiple instances running on the same machine. All data packets from or to cloud instance pass through the hypervisor layer, as this layer resides between the physical network interface and the virtual network interface of cloud instances. Therefore, the host-based firewall is often deployed at this layer to filter "unwanted traffic" towards the cloud instances using pre-defined firewall policy rules configured by cloud customers. For example, Amazon EC2 allows cloud customers to configure security polices to define certain firewall rules at the hypervisor layer for accurately identifying and filtering the inbound "unwanted traffic" to the cloud instances [20].

In this study, we propose to harvest unwanted traffic from distributed hypervisors in the cloud and to establish a central collaborative center that collects and analyzes unwanted traffic from distributed hypervisors in the cloud computing environment. The idea of this collaborative center is inspired by DShield, a cooperative network security community portal site that collects firewall logs for analyzing the trends and emerging threats of the exploit behaviors [21]. Specifically, we will develop a distributed measurement framework, where each hypervisor has a client program that communicates with a server running on the collaborative center. Once a hypervisor detects unwanted traffic towards one or more instances, the hypervisor will summarize and generate the traffic signature, and then sent to the central collaborative center through the reporting client. Prior studies have shown that the collaborative principles have a wide range of applications in network measurement [22] and security monitoring [23].

Given the volume of unwanted traffic in the Internet background radiations due to vulnerability scanning, worm prop-

agations, system penetration attempts, DoS attacks, and other exploit activities [24], it is not surprising that cloud instances receive a large amount of unwanted traffic. Thus a major challenge of designing and implementing such a collaborative solution for distributed hypervisors in the cloud computing environment lies in the size and diversity of the "unwanted traffic" collected and processed by distributed hypervisors and the central collaborative center. Thus an important research problem in hypervisor profiling is how to develop efficient collaborative algorithms to share and leverage unwanted traffic collected on distributed hypervisors for reducing unwanted traffic towards cloud instances. To prevent distributed hypervisors and the collaborative center from being overwhelmed by a large amount of unwanted traffic, we propose to use two-layer counting bloom filter technique [25] at hypervisors and the collaborative center to reduce the size of unwanted traffic reports by identifying the most aggressive attackers from all source IP addresses of unwanted traffic.

3) Profiling Traffic at Instance Level and Application Level: Network and hypervisor profiling in the proposed *profiling-as-a-service* architecture discover behaviors patterns of cloud instances and unwanted traffic towards the cloud network, respectively. However, both steps lack the visibility of all network traffic towards cloud instances and the applications running on them. To gain a complete picture of network traffic towards cloud instances and their applications, it becomes very necessary to profiling traffic at instance and application levels.

Instance profiling is interested in three important aspects of cloud instances: user, traffic, and performance. The user profiling is focused on the access patterns of end systems on the Internet that communicate with cloud instances, while traffic profiling at the instance level studies traffic characteristics of the cloud instance, such as traffic distributions across ports or applications, and temporal traffic patterns. Similar to the continuous profiling infrastructure deployed at Google data centers [26], the performance profiling aims in quantifying system performance such as CPU and memory utilization, and input/ouput throughput, and end-to-end performance such as network latency and packet losses. As illustrated in Figure 3, instance profiling runs independently on multiple virtual instances that are hosted on the same physical machine.

In addition to profiling traffic at instance level that studies the aggregated network traffic of cloud instances, we will also perform profiling traffic at application level and investigate the application-specific semantics and contents of network traffic. We plan to use the graphic models to represent traffic activity of cloud instances and their applications. The fine-grained traffic analysis on application level provides valuable insights for application diagnosis and troubleshooting, network management and capacity planning.

The challenges of instance profiling and application profiling are i) feature selection for instance or application profiling; 2) payload and content analysis without baseline signatures or prior knowledge of normal or abnormal traffic patterns. To address these challenges, we will employ temporal analysis and feature selection algorithms for instance and application profiling, and explore algorithms in data mining and machine learning for detecting unknown exploit traffic, e.g., emerging worms or virus. The ultimate goal of profiling traffic at instance level and application level is to complement traffic profiles of network profiling and hypervisor profiling and to build a comprehensive traffic behavior profile for each cloud instance by summarizing its traffic behavior and application activities with the full packet traces.

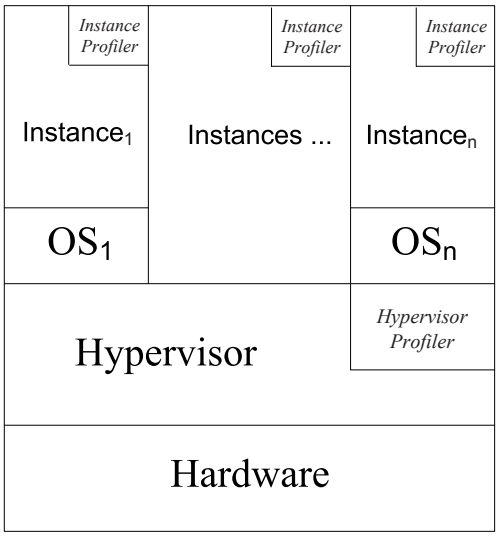

Fig. 3. Hypervisor profiling and instance profiling on the physical machine

4) Profiling High-Volume Network Traffic: An operational challenge of the *profiling-as-a-service* architecture across all levels is the sudden traffic surges during unusual events such as denial of service attacks [27], flash crowds [28], or worm outbreaks [29]. The sheer volume of network traffic during these events introduces significant system pressure for the profiling architecture that runs on commodity PCs with limited CPU and memory capacity. At the same time, it is vital for the profiling architecture to function during these events, since traffic profiles generated during these periods will provide key insights and valuable information for effective response and forensic analysis.

Sampling is a widely deployed technique to reduce system resource consumptions in network traffic monitoring. Traditional sampling approaches include random packet sampling, random flow sampling, smart sampling and sample-and-hold algorithms. However, previous studies [30] have shown that these existing sampling algorithms, albeit significantly reducing resource usage, bring non-trivial accuracy losses on the traffic feature distributions as well as on other traffic statistics. Our preliminary analysis also finds that these sampling algorithms could lead to inaccurate traffic profiles of cloud instances at all levels during sudden traffic surges, although they substantially bring down the CPU and memory usage. To enhance the robustness and accuracy of the profiling-as-a-service architecture under these stress conditions, it becomes very necessary to develop novel sampling algorithms that not only reduce the system resource usage, but also retain the

accuracy of traffic profiles across all levels.

A key lesson from studying traditional sampling approaches in our preliminary analysis is that the cloud instances involved with anomalous events such as denial of service attacks usually receive vast amount of network traffic. Profiling traffic behavior of these instances during these events does not require a very large number of traffic flows, since their feature distributions likely remain the same even with a small percentage of sampled traffic flows. On the other hand, the profiles of other hosts with much less traffic are very sensitive to the number of sampled traffic. Based on this insight, we plan to develop new profiling-aware sampling solutions that limit the number of sampled traffic flows for instances or applications with a large amount of traffic, but adaptively samples on the rest of instances or applications when the profiling system is faced with sudden explosive growth in the number of traffic flows or packets caused by anomalous events such as denial of service attacks or worm outbreaks.

IV. Conclusions and Ongoing Work

In this position paper, we proposed to develop a *profiling-as-a-service* infrastructure in the multi-tenant cloud computing environment, and build traffic profiles of cloud instances at multiple layers for gaining an in-depth understanding of network traffic in the cloud. To demonstrate the operational feasibility and the practical applications of the proposed *profiling-as-a-service* architecture, we are currently in the process of designing and implementing a prototype profiling system. We will evaluate the prototype with real data center traffic collected from a large Internet service provider and existing packet traces of denial of service attacks and worm propagations. Through a variety of experiments, we will show how the traffic profiles generated at different layers could aid in security monitoring, customer profiling, and traffic engineering for cloud instances. In addition, we plan to test the instance profiling and application profiling on existing commercial cloud computing platforms such as Amazon EC2, Microsoft Azune, or Google AppEngine. We will study the system utilizations, such as CPU and memory usage of the profiler at each layer, and examine how fast traffic profiles could be generated at each layer and be used by network operators to real-time traffic analysis or attack mitigation.

References

[1] C. News, "DDoS attack hobbles major sites, including Amazon," http://news.cnet.com/8301-30684_3-10421577-265.html.

[2] W. S. Journal, "Google Introduces New Security Measures After Cyber-Attack," http://online.wsj.com/article/BT-CO-20100112-716798.html.

[3] 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. of ACM Conference on Computer and Communication Security (CCS)*, November 2009.

[4] N. Santos, K. Gummadi, and R. Rodrigues, "Towards Trusted Cloud Computing," in *Proceedings of USENIX Workshop On Hot Topics in Cloud Computing (HotCloud)*, June 2009.

[5] M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," in *Proceedings of ACM SIGCOMM*, August 2008.

[6] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers," in *Proceedings of ACM SIGCOMM*, August 2009.

[7] A. Greenberg, J. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, and P. Pat, "VL2: A Scalable and Flexible Data Center Network," in *Proceedings of ACM SIGCOMM*, August 2009.

[8] T. Benson, A. Anand, A, Akella, and Ming Zhang, "Understanding Data Center Traffic Characteristics," in *Proceedings of SIGCOMM WREN Workshop*, August 2009.

[9] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Internet traffic behavior profiling for network security monitoring," *IEEE/ACM Transactions on Networking*, vol. 16, pp. 1241–1252, December 2008.

[10] H. Jiang, Z. Ge, S. Jin, and J. Wang, "Network Prefix-level Traffic Profiling: Characterizing, Modeling, and Evaluation," *Computer Networks*, 2010.

[11] M. Iliofotou, B. Gallagher, T. Eliassi-Rad, G. Xie, and M. Faloutsos, "Profiling-by-Association: A Resilient Traffic Profiling Solution for the Internet Backbone," in *Proceedings of ACM CoNEXT*, December 2010.

[12] T. Karagiannis, K. Papagiannaki and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," in *Proc. of ACM SIGCOMM*, 2005.

[13] Y. Hu, D.-M. Chiu, and J. Lui, "Profiling and identification of P2P traffic," *Computer Networks*, vol. 53, no. 6, April 2009.

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

[15] Amazon, "Amazon Web Services," aws.amazon.com.

[16] D. Owens, "Securing Elasticity in the Cloud," *Communications of the ACM*, vol. 53, no. 6, pp. 46–51, June 2010.

[17] S. S. L. Ertaul and G. Saldamli, "Security Challenges in Cloud Computing," in *Proceedings of International Conference on Security and Management*, July 2010.

[18] Y. Chen, V. Paxson, and R. Katz, "Whats New About Cloud Computing Security?" *Technical Report No. UCB/EECS-2010-5, University of California at Berkely*, January 2010.

[19] M. Yildiz, J. Abawajy, T. Ercan, and A. Bernoth, "A Layered Security Approach for Cloud Computing Infrastructure," in *Proceedings of International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN)*, December 2009.

[20] Amazon, "Amazon Web Services: Overview of Security Processes," http://aws.amazon.com/security.

[21] DShield.org, "Cooperative Network Security Community - Internet Security," http://www.dshield.org/.

[22] W. Liu and R. Boutaba, "pMeasure: A peer-to-peer measurement infrastructure for the Internet," *Computer Communications*, vol. 29, no. 10, pp. 1665–1674, June 2006.

[23] S. Katti, B. Krishnamurthy, and D. Katabi, "Collaborating Against Common Enemies," in *Proceedings of ACM SIGCOMM Internet Measurement Conference*, October 2005.

[24] E. Wustrow, M. Karir, M. Bailey, F. Jahanian, and G. Houston, "Internet Background Radiation Revisited," in *Proceedings of ACM SIGCOMM Conference on Internet Measurement*, November 2010.

[25] B. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422 – 426, July 1970.

[26] G. Ren, E. Tune, T. Moseley, Y. Shi, S. Rus, and R. Hundt, "Google-Wide Profiling: A Continuous Profiling Infrastructure for Data Centers," *IEEE Micro*, vol. 30, no. 4, pp. 65–79, 2010.

[27] N. World, "DDoS attack against Bitbucket darkens Amazon cloud," http://www.networkworld.com/community/node/45891.

[28] J. Jung, B. Krishnamurthy, and M. Rabinovich, "Flash Crowds and Denial of Service Attacks: Characterization and Implications for CDNs and Web Sites," in *Proceedings of International World Wide Web Conference*, 2002.

[29] C. Zou, L. Gao, W. Gong, and D. Towsley, "Monitoring and Early Warning for Internet Worms," in *Proceedings of ACM CCS*, 2003.

[30] J. Mai, A. Sridharan, C.-N. Chuah, H. Zang, and T. Ye, "Impact of Packet Sampling on Portscan Detection," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 12, pp. 2285 – 2298, December 2006.