Limited Use Cryptographic Tokens in Securing Ephemeral Cloud Servers

Gautam Kumar, Brent Lagesse
University of Washington Bothell,
Computing and Software Systems,
Bothell WA,
USA
{*gautamk, lagesse*}*@uw.edu*

Keywords: Security, Moving Target Defence, Cryptography, Cloud Architecture

Abstract: Many enterprises and consumers today are dependent on services deployed on Infrastructure as a Service (IaaS) cloud providers. Such cloud deployments can have hundreds of virtual servers running. Each virtual server needs to have access to sensitive information such as database passwords and API keys. In such as scenario, verifying that a large number of servers have not been compromised is an arduous task. In this paper we propose an architecture which limits the extent to which an attacker can exploit a compromised server in a large scale cloud deployment. To achieve such a limitation we propose the use of hash chains as an authentication mechanism for virtual server with a Central Trusted Authority (CTA) acting as a proxy to sensitive resources. This architecture shifts the requirement of security validation from hundreds of public facing servers to a few servers without public interfaces which comprise the CTA. Since hash chains offer an inherent limitation in their use, our architecture leans towards using ephemeral virtual servers, thus also providing a moving target defence.

1 Introduction

According to Microsoft (Harms and Yamartino, 2010), more than 85% of work done by enterprise IT departments is towards infrastructure maintenance. Cloud providers offer a solution to this by minimizing the effort needed to purchase and maintain physical hardware which is a unique advantage to enterprises. But cloud infrastructure comes with its own set of unique challenges and one of the primary concerns is security. Many cloud providers offer computational and storage resources on virtualized shared hardware to maximize utilization. Thus the essence of securing cloud systems is using multiple layers (Panwar et al., 2011) of security to increase an attacker's cost for taking over the system. One of the emerging layers of security is moving target defence (Evans et al., 2011). So in this paper on of our one of our focuses is to increase the cost to an attacker for launching a zero-day attack using moving target defence. According to Bilge & Dumitras (Bilge and Dumitras, 2012), on average, zero day attacks remain undetected for 10 months. Once zero-day attacks are disclosed, malware authors begin to utilize them multiple orders of magnitude more frequently; however, there is often a significant delay between notification and the deployment of patches fixing the zero-day exploit. One of the goals of this project is to reduce the risk of systems when we are not aware of the existence of the vulnerability.

Consider the situation where a large service provider has a number of public facing servers with access to a private, remote database. In the case that one of those public facing servers is compromised by an adversary, the data in the private database will be vulnerable for the entire duration that the attacker has access to the server. The goal of our work is to drastically reduce the length of time that the attacker has access. to sensitive information *even when we do not know that the attack has occurred*.

In this paper we propose an implementation of moving target defence using ephemeral servers and a central trusted authority which acts on behalf of an ephemeral server and proxies requests to sensitive resources such as database servers, caching servers and REST end points. Hash chains are used as an authentication mechanism by the central trusted authority. We take advantage of the limited use property of hash

chains to authenticate ephemeral servers for a limited period of time. Hash chains are used as opposed to a timer as it reduces link-ability and provides cryptographic guarantees for its limited use.

2 Background

Cloud infrastructure today is characterised by three primary classes of products. Infrastructure as a Service or IaaS, Platform as a Service or PaaS and Software as a Service or SaaS. IaaS providers take on the responsibility of maintaining and securing physical hardware while organisations using IaaS are expected to maintain and secure their virtual hardware and also their application code. Our research in this paper is primarily focused on providing a mechanism to improve the security for organisations who use IaaS providers.

Hash Chains: Leslie Lamport (Lamport, 1981) first proposed the use of hash chains in his paper on a method for secure password authentication over an insecure medium. One of the properties of hash chain-based authentication is that it has a limited use lifespan. We leverage this property in our system to limit the damage that any single compromised machine can cause.

Threats: According to OWASP, "Sensitive data exposure" is the one of the most critical types of security threat in web applications as of 2013 (Wichers, 2014). Sensitive data exposure refers to the unintended exposure of sensitive information such as passwords, social security numbers and date of birth. In the context of a cloud system, sensitive information may also include database credentials, email server and REST API keys. These credentials are used to authenticate cloud servers with third party services within or outside the private cloud network. Credentials are usually stored as part of configuration files or in a server's memory. Such credentials have a very long life time and are sometimes valid of years. This increases the risk of data exposure when an attacker has a lengthy period of time within which they can discover and exploit such credentials.

A report by Risk Based Security (Risk Based and Security, 2014) notes that the number of data leaks has dramatically increased from 2012 to 2013, to the tune of $812 million. Leaking of credentials as mentioned earlier, could potentially grant access to malicious entities within protected environments such as a corporate network. Sensitive data exposure can also be a consequence of other threats such as cross site scripting (XSS) and Code Injection both of which are considered highly critical threats based on OWASP's classification (Wichers, 2014).

3 Proposed architecture

Our proposed architecture consists of three primary components. The Client Facing Server (CFS), the Central Trusted Authority (CTA) and Sensitive Resources that we're trying protect. This is described in figure 1.

Client Facing Servers are any servers which interact with entities on a public network such as Load balancers, reverse proxies and application servers. CFSs are assumed to be the most vulnerable to exploits by malicious entities. To offer moving target defence and increase the cost to an attacker, CFSs have a limited lifetime based on the length of the hash chain. When a hash chain expires the CFS is expected to shutdown because it can no longer authenticate itself with the CTA.

The *Central Trusted Authority* is an entity within the corporate or private network which acts on behalf of a CFS when a sensitive resource is requested. To act on behalf of a CFS the CTA proxies requests to resources such databases and external APIs and includes authenticating information required by the resources. The authenticating information such as DB Passwords and API Keys are not available to CFS. The CTA is not publicly available and the number of CTAs in operation is relatively small compared to the number of CFS, it is assumed that the CTA is trustworthy as it will be easier to harden and monitor the limited types of activity on CTAs. CTAs authenticate CFS using hash chains as explained in section 3.2, and figure 6 elaborates on the architecture of the CTA.

3.1 Assumptions

Client facing servers are the servers which are exposed outside the private cloud network environment. These client facing servers could potentially be load balancing servers, compute servers. The client facing servers are assumed to be ephemeral. This is common in many cloud deployments (Vaquero et al., 2011) and contributes to the moving target nature of the security architecture. Companies such as Netflix expect this behaviour with their chaos engineering architecture (Basiri et al., 2016). This allows for higher reliability of their server infrastructure. Client facing servers are expected to shut down after their hash chain expires. This contributes to the ephemeral nature and also to moving target defence of the overall system.

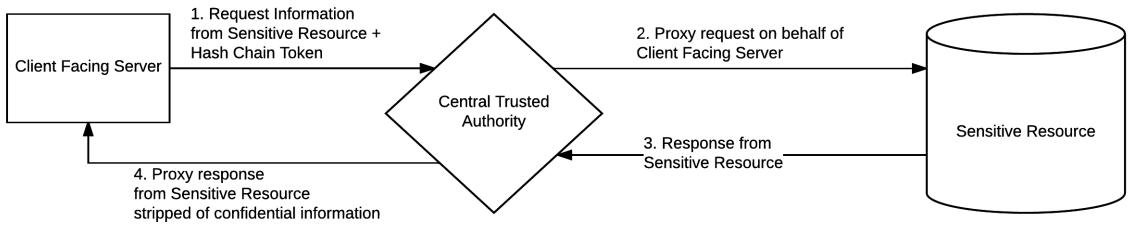

Figure 1: Architectural overview of the system. This figure describes the three primary modules involved. The client facing server, The central trusted authority and a sensitive resource.

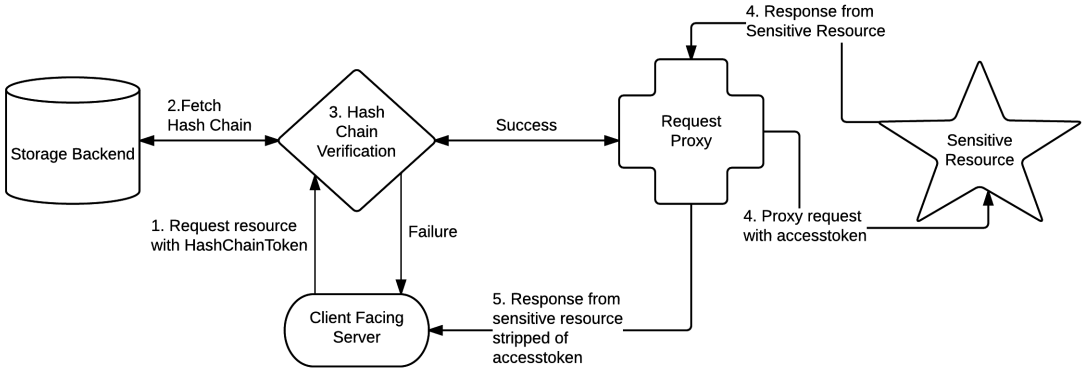

Figure 2: Architecture of the Central Trusted Authority. The CTA consists of a storage backend, a hash chain verifier and a request proxy.

3.2 CTA Architecture and Implementation

The Central Trusted Authority consists of three primary components, A hash chain verifier, A storage back-end and a request proxy. The CTA performs three roles within the system, which are

- Create new hash chains
- Verify hash chains
- Proxy requests to resources

Creating a new hash chain: Hash chains are created by iteratively hashing a secret key K, n number of times. After the hash chain is created the CTA stores $H^n(K)$ in the storage backend and returns K and n to the client facing server. The secret key K is not stored by the CTA. The client facing server can now use the the secret key n number of times. This process is detailed in Algorithm 1.

Hash chain verification: Hash chains are used to authenticate client facing server requests which require access to a sensitive resource. The hash chain verification is detailed in algorithm 2.

Data: Hash Chain secret K and Hash chain length N
Result: Hash Chain $H^N(K)$

$i \leftarrow 1$;
$H^1(K) \leftarrow H(K)$;
while $i <= N$ **do**
  $i \leftarrow i+1$;
  $H^i(K) \leftarrow H(H^{i-1}(K))$;
end
return $H^i(K)$ *where i equals N* ;

Algorithm 1: Generating a Hash Chain

4 Performance

Hash chain creation: In our testing hash chain initialization showed a linear increase, $O(n)$, in time complexity based with hash chain length (n) . This is in line with our expectations for hash chain implementations.

In our proposed architecture hash chain length affects CFS uptime based on the equation,

Data: Authentication key $H^{i-1}(K)$ from the client
Result: Response from sensitive resource
Let $H^i_{client} = H(H^{i-1}(K))$;
Let $AuthenticationData = fetch(H^i_{client})$;
if *AuthenticationData exists in storage backend* **then**
    replace H^i_{client} with $H^{i-1}(K)$ in storage backend ;
    fetch and return response from sensitive resource ;
else
    return authentication failure ;
end

Algorithm 2: Verification of Hash Chain authentication

$$Uptime = \frac{Hash\ chain\ length}{Requests\ per\ second}$$

where $Uptime$ is the maximum time in seconds that a CFS can effectively operate, $Hash\ chain\ length$ is the size of the hash chain provided to the CFS at startup while $Requests\ per\ second$ is the average number of requests that a CFS needs to make to a CTA.

Hash chain verification and request proxy: Hash chain verification is a significantly faster process, $O(1)$, when compared to hash chain creation. Thus performance of the CFS is impacted merely by the delay introduced due to the request proxy.

Our testing reveals that an additional delay in the range of 0.05 to 0.15 seconds is introduced by using Hash chain verification and a simple implementation of request proxy as illustrated by figures 3 and 4.

Testing environment: Our testing environment was based in AWS. The CTA was a pool of EC2 machines behind a elastic load balancer. Load tests were conducted using multiple CFSs located in the same AWS region as the CTA. The payload requested by the CFS was a file located on S3.

4.1 Server Pool

Creating new virtual servers to act as CFSs can be a slow and expensive process; however, by creating a pool of unused CFSs to act as quick replacements, we can minimize delay for customers. The size of the CFS pool needs to be balanced based on the time to initialize a new virtual server, the size of the hash chain, and the rate at which requests are made.

The relationship between pool size, hash chain length and requests per second can be illustrated by

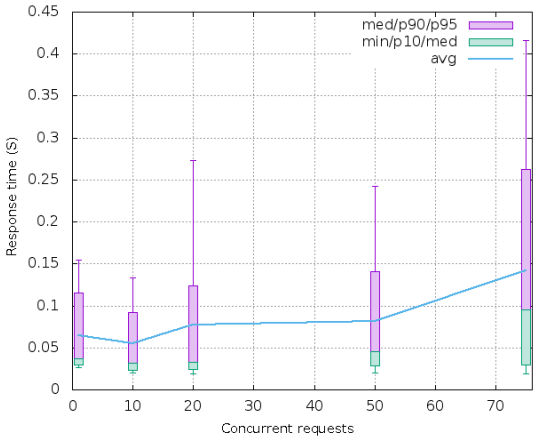

Figure 3: Response time for requests made to a resource through the CTA with a SHA-256 hash chain. The X-axis is the number of concurrent requests made and the Y-axis describes the response time in seconds

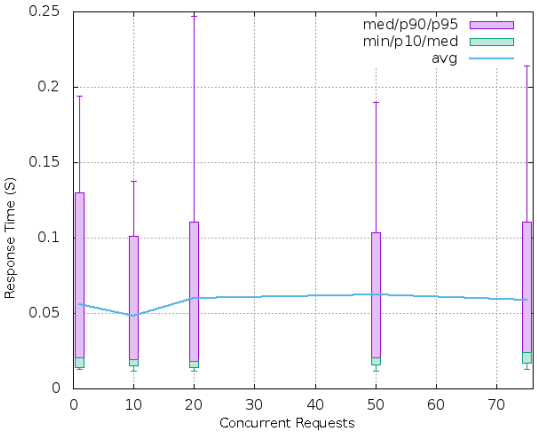

Figure 4: Response time for requests made to a resource directly without authenticating with the CTA. The X-axis is the number of concurrent requests made and the Y-axis describes the response time in seconds

equation 1.

$$r_{pool} = r_{boot} - \frac{r_{req}}{n} \tag{1}$$

Equation 1 describes how the change in pool size relates to the length of the hash, the boot rate, and the

n	Length of hash chain
r_{req}	Request rate
r_{boot}	Boot rate
r_{pool}	Pool size change rate

Table 1: Table of terms for equation 1

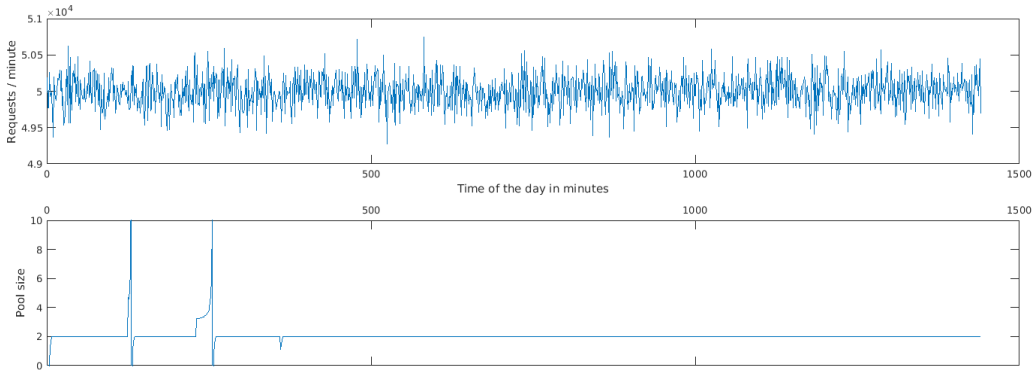

Figure 5: The upper plot shows a random request rate generated by a poisson distribution is with a lambda of 50,000 and the lower plot describes the pool size based on equation 1

request rate. When the number of requests that have been authenticated using a hash chain reaches the length of the hash chain, the VM using it is no longer functional and must be shut down and replaced. Our system keeps a pool of VMs in reserve so that this process is transparent to the user and no delay is noticed. System designers can use this equation to minimize the number of VMs that have to be held in reserve to sufficiently handle a variable load over the course of time. Boot rate is defined as the average number of VMs that can be instantiated in a particular period of time. Request rate is the average number of requests that come in during a time period. This value can change overtime as services will have different peak usage times during a day.

As an example consider that a company receives 50,000 requests per minute on an average day. They want to replace their VMs every 10 minutes, so each VM should have a hash chain with a length of 5000. They would have to boot 1 VM every 5 minutes. Figure 5 shows how our system would keep a small pool of VMs that would not consume significant resources, but would also never run out when the request rates are sampled from a poisson distribution with a lambda of 50,000.

Part of our future work is to design an adaptive algorithm that solves the multi-criteria optimization problem that manages the creating of new VMs based on learned traffic patterns.

5 Security Threats

The proposed architecture of using a Central Trusted Authority (CTA) to proxy requests on behalf of all the clients places the CTA as a single point of failure. We deem this an acceptable trade-off as the CTA is not a public facing system and only serves the purpose of verifying hash chains and proxying requests. As a result, hardening its defences against possible threats becomes an easier problem. In our current implementation, the CTA is a single hardened server; however, we are currently working on a distributed implementation.

Our architecture proposal does not prescribe any particular hardware / VM co-location for any components in the CTA. Thus each component can be implemented on separate machines or the CTA can be instantiated as a whole within a single VM. In this section we detail the potential security threats against each component and possible mitigation strategies.

5.1 Malicious Request Proxy

The request proxy component acts on behalf of the client facing server to make a call to a sensitive resource such as an API endpoint. The request proxy is also responsible for attaching authentication information such as an API Key when contacting the API endpoint. Upon receiving a response from the sensitive resource the request proxy strips out sensitive information such as API Keys and Refresh Tokens before forwarding the response to the client facing server. Under this threat scenario the request proxy is assumed to be untrustworthy and potentially malicious despite the our earlier argument that this case is unlikely.

Chen et al (Chen et al., 2012) propose a solution to this problem of a Malicious proxy using trusted hardware such as Trusted Platform Module (TPM) or the IBM 4758 cryptographic coprocessor (Parno et al., 2010). The CTA executable would also verified by a trusted third party to operate correctly as a proxy as described in (Parno et al., 2010).

There are three primary attacks that a malicious

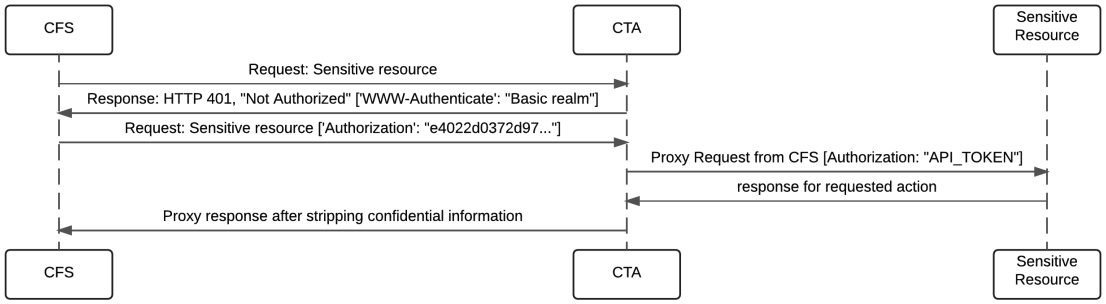

Figure 6: Sequence diagram describing the authentication and proxying capabilities of the CTA. CFS refers to Client Facing Server while CTA refers to Central Trusted Authority

actor may perform on the CTA. First, the attacker may try to expose the sensitive information stored, such as API keys and DB passwords, from the CTA using vulnerabilities in the CTA executable. Second, the attacker could potentially modify the requests / responses which are being proxied by the CTA. Third the attacker could potentially launch a reboot attack to inject a malicious executable after attestation to carry out one of the above attacks.

Prior work (Libert and Vergnaud, 2008) (McCune et al., 2008) on preventing reboot-attacks can be leveraged to impede the attacker's ability to inject a malicious executable. The proposed architecture also allows for the CTA to be placed behind secure corporate firewalls to further limit the risk of a malicious take over by an attacker. Further work is needed to fully secure the CTA against a malicious proxy.

6 Discussions

6.1 Providing moving target defence

Green et al (Green et al., 2015) identify Unpredictability, Vastness and Revocability as some of the criteria for evaluating moving target defences. In this section we try define how our implementation conforms to these criteria.

Unpredictability: Cryptographic hash functions (Rogaway and Shrimpton, 2004) are designed to be collision resistant and comply with the avalanche effect and thus the generated sequence is highly unpredictable in nature. Hash chains can compound this effect by sequential application of the cryptographic hash function.

Vastness: A secure cryptographic hash function such as SHA-256 has a vast state space as the output could be any of 2^{256} different values. As a result, the likelihood of an attacker guessing the correct value is negligible.

Revocability: The proposed architecture allows individual chains or a group of chains to be revoked merely by purging them from the database. Further group / hierarchical revocability can be achieved using Merkle hash tree implementations as a means of Hash chain generation.

6.2 CTA Alternative

In this paper we have presented our solution with a centralized CTA and argued that in some cases this approach is sufficient. In the case that this solution does not work for an application, we are also exploring a DHT-based proxy though we have not yet implemented it. In the DHT-based solution, there would be a large number of proxies in a DHT. Each proxy in the DHT would be responsible for servicing a subset of requests from CFS. In this solution, the DHT would be keyed on the hash chain hash value, so the requests would be routed to a different proxy every time. By using a DHT-based solution, an adversary would only be able to compromise a subset of the requests, further limiting its ability to intercept targeted information or read widespread sensitive requests.

6.3 Future Work

Formulating learning algorithms to create an adaptive balance between server lifespan and hash chain length to optimize computational resources in a cloud system, this can be derived from the equations we have derived in this paper and current work into load balancing in cloud systems (Randles et al., 2010). Such an adaptive system could be used to choose an ideal hash chain size and CFS pool in real time based on the performance requirements of a cloud system.

Timed Release Cryptography (Chalkias and Stephanides, 2006) can be integrated as a hash chain renewal mechanism to potentially increase the lifespan of a server after a certain cool off period. Implementing Time Release Cryptography would enable our architecture to accommodate systems where a constant pool of ephemeral servers are not available and existing servers can be populated with a batch of hash chains which can only be accessed after a certain period of time.

A Merkle hash tree implementation middle-ware can potentially provide hierarchical authentication authorization capabilities to the CTA as illustrated by Yi & Wang (Yi and Wang, 2012). In the scenario of an information leak, Merkle trees would allow administrators to black list either a single or a hierarchy of hash chains to safeguard the system.

7 Related work

There have been other systems proposed which offer moving target defence using a temporary address or a temporary authentication mechanism. Many technology companies have real world implementations of the Central Trusted Authority architecture similar to our proposal. In this section we present the most relevant proposals and discuss how our proposed architecture differs.

Dunlop et al (Dunlop et al., 2011) leverage the vast address space (2^{128}) of IPv6 to move the source and destination IP addresses mid-session based on a pre-agreed pattern to limit an attacker's ability to intercept or interfere with a TCP session. The technique proposed by Kampanakis (Kampanakis et al., 2014), however, operates at the network level by using an SDN's ability to vary the address space and the route taken by packets to increase the cost for an attacker and thus providing moving target defence.

Active authentication proposed by (Yiu et al., 2011; Aksari and Artuner, 2009; Li et al., 2014) offer a way to verify a user's identity based on their behaviour thus eliminating the need for hard to remember passwords. Active authentication is a form of moving target defence where the authenticating factor is constantly changing in a hard to predict manner, thus significantly increasing the cost for an attacker to reproduce the authenticating factor.

Confidant (lyft, 2015) is a library maintained by Lyft, a transportation network company based out of San Francisco. Confidant provides an implementation of the Central trusted authority server with encryption at rest, authentication and authorization handled by AWS's Key Management Service, KMS and a storage backend of DynamoDB. This system lacks the hash chain based approach that we use and as a result does not provide the additional moving target defence that raises the adversary's cost of launching an attack. Another limitation of this implementation is the inability to deploy a Confidant instance on a different cloud IaaS provider besides Amazon's AWS.

8 Conclusion

In this paper we propose an architecture which increases the cost of exploiting vulnerabilities for an attacker with minimal impact on performance by providing moving target defence using a pool of Client Facing Servers (CFS) which only operate for a relatively short period of time. Our architecture also significantly hampers an attacker's ability to steal sensitive configuration information from by centralizing configuration within a Central Trusted Authority.

We believe this approach can be used in cloud systems to limit the risk of a compromised client facing server. Security-related software bugs are constantly being discovered and exploited, so while we may not be able to deploy a system that will never be compromised, we can deploy defences that limit the effectiveness of an attacker, even when they are utilizing a zero-day attack.

Acknowledgements

The work presented in this paper was partially supported under US National Science Foundation Grant DGE-1419313. The author would like to thank Victoria Wettmarshausen from UW Bothell's Writing & Communication Center for her feedback on the structure and presentation of this paper.

REFERENCES

Aksari, Y. and Artuner, H. (2009). Active authentication by mouse movements. In *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, pages 571–574. IEEE.

Basiri, A., Behnam, N., Rooij, R. d., Hochstein, L., Kosewski, L., Reynolds, J., and Rosenthal, C. (2016). Chaos Engineering. *IEEE Software*, 33(3):35–41.

Bilge, L. and Dumitras, T. (2012). Before we knew it: an empirical study of zero-day attacks in the real

world. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 833–844. ACM.

Chalkias, K. and Stephanides, G. (2006). Timed release cryptography from bilinear pairings using hash chains. In *Communications and Multimedia Security*, pages 130–140. Springer.

Chen, R., Reznichenko, A., Francis, P., and Gehrke, J. (2012). Towards statistical queries over distributed private user data. In *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 169–182.

Dunlop, M., Groat, S., Urbanski, W., Marchany, R., and Tront, J. (2011). MT6d: A Moving Target IPv6 Defense. In *2011 - MILCOM 2011 Military Communications Conference*, pages 1321–1326.

Evans, D., Nguyen-Tuong, A., and Knight, J. (2011). Effectiveness of Moving Target Defenses. In Jajodia, S., Ghosh, A. K., Swarup, V., Wang, C., and Wang, X. S., editors, *Moving Target Defense*, number 54 in Advances in Information Security, pages 29–48. Springer New York. DOI: 10.1007/978-1-4614-0977-9_2.

Green, M., MacFarland, D. C., Smestad, D. R., and Shue, C. A. (2015). Characterizing Network-Based Moving Target Defenses. In *Proceedings of the Second ACM Workshop on Moving Target Defense*, MTD '15, pages 31–35, New York, NY, USA. ACM.

Harms, R. and Yamartino, M. (2010). The economics of the cloud. *Microsoft whitepaper, Microsoft Corporation*.

Kampanakis, P., Perros, H., and Beyene, T. (2014). SDN-based solutions for Moving Target Defense network protection. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2014 IEEE 15th International Symposium on a*, pages 1–6.

Lamport, L. (1981). Password authentication with insecure communication. *Communications of the ACM*, 24(11):770–772.

Li, F., Clarke, N., Papadaki, M., and Dowland, P. (2014). Active authentication for mobile devices utilising behaviour profiling. *International journal of information security*, 13(3):229–244.

Libert, B. and Vergnaud, D. (2008). Tracing Malicious Proxies in Proxy Re-encryption. In Galbraith, S. D. and Paterson, K. G., editors, *Pairing-Based Cryptography Pairing 2008*, number 5209 in Lecture Notes in Computer Science, pages 332–353. Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-85538-5_22.

lyft (2015). Confidant: Your secret keeper. A library to store and retrive senstive configuration within a central trusted authority encrypted at rest using Amazon KMS. https://github.com/lyft/confidant.

McCune, J. M., Parno, B. J., Perrig, A., Reiter, M. K., and Isozaki, H. (2008). Flicker: An execution infrastructure for TCB minimization. In *ACM SIGOPS Operating Systems Review*, volume 42, pages 315–328. ACM.

Panwar, A., Patidar, R., and Koshta, V. (2011). Layered security approach in cloud. In *3rd International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2011)*, pages 214–218.

Parno, B., McCune, J. M., and Perrig, A. (2010). Bootstrapping trust in commodity computers. In *2010 IEEE Symposium on Security and Privacy*, pages 414–429. IEEE.

Randles, M., Lamb, D., and Taleb-Bendiab, A. (2010). A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing. In *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 551–556.

Risk Based and Security (2014). An Executives Guide to 2013 Data Breach Trends. *Presentation, Risk Based Security*.

Rogaway, P. and Shrimpton, T. (2004). Cryptographic Hash-Function Basics: Definitions, Implications, and Separations for Preimage Resistance, Second-Preimage Resistance, and Collision Resistance. In Roy, B. and Meier, W., editors, *Fast Software Encryption*, number 3017 in Lecture Notes in Computer Science, pages 371–388. Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-25937-4_24.

Vaquero, L. M., Rodero-Merino, L., and Buyya, R. (2011). Dynamically Scaling Applications in the Cloud. *SIGCOMM Comput. Commun. Rev.*, 41(1):45–52.

Wichers, D. (2014). OWASP Top-10 2013. *OWASP Foundation, February*.

Yi, X. and Wang, W. (2012). The Cloud Access Control Based on Dynamic Feedback and Merkle Hash Tree. In *2012 Fifth International Symposium on Computational Intelligence and Design (ISCID)*, volume 1, pages 217–221.

Yiu, M. L., Lo, E., and Yung, D. (2011). Authentication of moving kNN queries. In *2011 IEEE 27th International Conference on Data Engineering*, pages 565–576. IEEE.