Evaluating the impact of eDoS attacks to cloud facilities

Gian-Luca Dei Rossi
Università ca' Foscari Venezia
DAIS
Via Torino 155, Venezia, Italy
deirossi@dais.unive.it

Mauro Iacono
Seconda Università di Napoli
DSP
Viale Ellittico 31, Caserta, Italy
mauro.iacono@unina2.it

Andrea Marin
Università ca' Foscari Venezia
DAIS
Via Torino 155, Venezia, Italy
marin@dais.unive.it

ABSTRACT

The complexity of modern cloud facilities requires attentive management policies that should encompass all aspects of the system. Security is a critical issue, as intrusions, misuse or denial of service attacks may damage both the users and the cloud provider including its reputation on the market. Disruptive attacks happen fast, cause evident and short term damages and are usually the result of operations that are hard to disguise. On the other hand, Energy oriented Denial of Service (eDoS) attacks aim at producing continuous minor damages, eventually with long term consequences. These long lasting attacks are difficult to detect. In this paper we model and analyse the behaviour of a system under eDoS attack. We study the impact in terms of cloud energy consumption of an attack strategy previously proposed in the literature and compare it with other strategies that we propose. Our findings show that the strategy previously proposed in the literature, based on keeping the cloud close to saturation, is not optimal (from the point of view of the attacker) in presence of non-constant workload and that there is a trade-off between the aggressiveness of the attacker and the duration of the attack in order to maximise the damage.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Performance,Security

Keywords

Security, Energy Denial Of Service, Markovian models

1. INTRODUCTION

Data centres are complex and dynamic environments which can adapt their behaviours according to some external conditions such as the current workload in order to offer services which respect the Service Level Agreement (SLA) signed with the customers. Although most of the complexity of a cloud system arises in order to offer computing power or other services to the subscribers, the management and support infrastructures are equally critical for the sustainability and the efficiency of cloud infrastructures. Clearly, from the end-user point of view, a cloud offers high quality services whenever the functional and non-functional aspects of the desired computation satisfy the agreements and the SLA. Nevertheless, from the point of view of the administration of the cloud, other aspects play a vital role. Among these we mention the energy consumption, which is one of the important costs that a cloud service supplier has to support. This is witnessed by the many efforts that the research community has devoted to the definition of administration policies which tend to reduce the energy consumption while maintaining reasonable levels for the Quality of Service (QoS), see, e.g., [8, 13, 9] just to mention a non exhaustive list of recent works. Informally, the idea behind these works is that the computational power of the cloud infrastructure is reduced when the workload is light so that the overall power consumption is reduced. When the workload intensity increases the "sleeping" components are waken up so that the QoS perceived by the users is still acceptable, at the cost of an increased power consumption. A similar duality between users' and cloud administration's need does generally characterise the common perception of security problems. The end-users' concerns are devoted to preventing attacks to exposed system services which are generally disruptive and evident, but the cloud management has to consider also other equally dangerous and important menaces that target the resource management architecture. Clearly, the scale of cloud systems is an additional complexity factor which makes the design of suitable countermeasures a challenging task. Indeed the vulnerability of cloud systems depends on two fundamental characteristics: first, they apply, to some extent, autonomic and user driven resource allocation policies and a number of self-managed mechanisms providing elasticity, based on reconfiguration; secondly, they are open to Internet access, serving a large number of users that demand services which can be designed, offered and managed by the cloud users who signed an agreement about the QoS requirements.

The importance and exposure on the Internet of cloud computing facilities make them a very attractive target for malicious users. In this paper we present a stochastic model to study the impact on a cloud computing facility of an En-

VALUETOOLS 2015, December 14-16, Berlin, Germany

DOI 10.4108/eai.14-12-2015.2262650

ergy oriented Denial of Service (eDoS) attack. This is a specially subtle, non disruptive attack that aims at increasing the energy consumption of a system by injecting legitimate workload with malicious purposes. The success of an eDoS attack depends on two factors: the gap between the ideal power consumption (i.e., without malicious users) and the effective power consumption, and the length of the attack. An aggressive attacker may inject a fictitious workload with high intensity with the aim of maximizing the power dissipation but as a consequence the risk of passing from a eDoS attack to a DoS attack is high. In these cases the cloud administration may take countermeasures to defend the system from DoS attacks, or the system may be overloaded. On the other hand, a less aggressive attacker may decide to reduce the power dissipation with the objective of extending the attack duration, since the eDoS will less likely become a DoS attack. In this paper we investigate this trade-off and compare different strategies that can be employed by the attackers. Our findings show that some of the strategies proposed in the literature [6] produce on average more damages than others, and show how the balance between the attacker aggressiveness and the attack duration could be handled. We believe that these results can be useful for the cloud infrastructures to define statistical methods for recognizing eDoS attacks.

The paper is organised as follows: the next section presents eDoS attacks and related works, Section 3 describes the model and Section 4 gives the algorithms for the computation of the performance indices. Section 5 presents the different attacker strategies that we consider and compare their effects on the energy consumption of the system. Finally, Section 6 gives some final remarks.

2. EDOS ATTACKS TO CLOUD SYSTEMS

The most common security attacks to networked systems aim at taking over or taking down the target. In the first case, the goal is to gain administrative privileges in the system, to take its control and use all or part of its resources for own purposes. This kind of attack requires a very good knowledge of the target and a strong expertise, and is generally detected by the administrator after a while because of irregular behaviours of the system or different usage patterns. In the second case, it is possible to flood the system with connection requests by means of a high number of irregular handshaking messages from a (group of) computer(s), and saturate the target possibility of processing regular requests obtaining a (Distributed) Denial of Services attack - (D)DoS. This kind of attack simply requires what is needed to saturate the incoming request bandwidth or computational power of the target, and it is immediately detected by the administrator because of its strong impact on the operations and the network connection. However, more subtle attacking techniques may be used such that those aiming at damaging the system in a hardly detectable way for instance by targeting the operating cost. This can be achieved by forcing the system to consume more energy and hence to need a higher cooling power, to shorten the lifetime of its hardware components because of the overload [11]: it is the case of eDoS attacks that allow malicious users to abuse the system and to force it to spend more energy than needed for the normal workload by injecting fictitious jobs with minimal impact on normal usage patterns. Although the instantaneous overload may be minimal, the effects on the long term cause additional costs, that may significantly affect the ability of a provider to stay on the market or to pay back the investments. As first, adding a low but continuous overhead implies additional costs for the energy needed for providing computing power, but also raises the energy expenditure for the cooling subsystem. This is obtained in small scale by preventing single components from entering low consumption states, and in big scale by forcing the infrastructure to use more computing nodes than actually needed [4]; moreover, the storage subsystem could also be involved, with more additional energy requirements, and potentially additional network related costs, as in many architectures computing and storage nodes are separated and more network traffic implies more routing activities. As second, all components are overused: consequently, maintenance operations will be more frequent and parts will need to be replaced earlier than expected [12].

During a eDoS, the attacker bases its strategy on the request of services, that cannot be distinguished from other requests, with the double aim of monitoring the response of the system and slightly increasing the resources that are needed to serve them. According to the results of the monitoring, the attacker decides the request rate so to raise the workload without making it evident, and continues raising at given time intervals until the malicious workload reaches a top level, beyond which the risk of being detected becomes too high. Such an attack spans over a significantly long time, so that the additional workload produces significant effects on the long distance and becomes less and less distinguishable from normal workload patterns. Eventually, a well designed eDoS attack can exploit the elasticity mechanisms of the cloud infrastructure, causing additional damage. The first examples of similar attacks are presented in [7] and [18], while a good survey on the topic is [12] and [6] presents a proof of concept and an experimental performance evaluation.

Detecting an eDoS is generally difficult, as there is need for knowledge about temporal behaviour of energy usage and for good hypotheses about the regularity of the workload: while for batch systems or scientific oriented computation facilities a fairly good knowledge of the workload is conceptually achievable, the characterization of the workload of a typical cloud infrastructure is less easy to be available. The low rate nature of eDoS contributes to lowering the effectiveness of usual traffic monitoring based detection techniques [5].

In order to model the behaviour of a cloud system during a eDoS attack, we chose a Markov chains based approach.

3. A MODEL FOR THE EVALUATION OF EDOS ATTACKS

In this section we introduce a Markovian model for the analysis of eDOS attacks. It consists of two cooperating components: one modelling the behaviour of the cloud system and the other the attacker policy. Both the models are abstractions of much more complicated systems but they catch the salient aspects which are important to allow us to compare the impact of different policies adopted by the attackers. As for the notation, we use the Iverson's brackets, i.e., $[C]$ is 1 if the Boolean proposition C is true, 0 otherwise and the standard Kronecker's algebra symbols to specify the cooperation

between the models.

3.1 A model for the cloud infrastructure

We model the cloud infrastructure with a finite set of states

$$\mathcal{S}_C = \{0, 1, 2, \dots K\}$$

with $K > 1$. Each state corresponds to a different power intensity used by the cloud infrastructure to serve the requests. Let $p(k) \in \mathbb{R}^+$ be the power spent in state $k \in \mathcal{S}_C$. We assume $p(k)$ to be non-decreasing bounded function of k for $k \in [0, K-1]$ and $p(K) = 0$. We partition the states into three disjoint classes:

- States from 0 to $T > 0$ denote the situation in which the cloud infrastructure works properly, i.e., by scaling its computational power (and hence reducing or augmenting its power consumption) the system can fulfill the desired service level agreement (SLA).
- States from $T + 1$ to $K - 1$ denote the situation in which the quality of service deteriorates since the cloud system is at its full computational power and is not able to improve the quality of service.
- State K denotes the situation in which the system crashes due to an excess of workload (e.g., the eDOS attack has degenerated into a DOS attack). Alternatively, we can see state K as the situation in which the countermeasures to defend the system from DOS attacks are taken and hence the effectiveness of the eDOS attack is over.

We abstract out the details of a real-world cloud behaviour and model its behaviour as a continuous-time random walk on the line with an absorbing barrier (i.e., state K). This implies that only transitions between adjacent states are allowed and that the residence time in a state is exponentially distributed. These assumptions can be relaxed at the cost of a higher computational complexity for the model analysis. Using the Iverson's bracket notation, we can describe the transition rate matrix of the cloud as follows:

$$\mathbf{C}_0(i,j) = \lambda(i)[j = i+1] + \mu(j)[j = i-1][j \neq K]\,, \quad (1)$$

where $i, j \in [0, K]$. Notice that both $\lambda(i)$ and $\mu(j)$ are state dependent. The former value represents the standard workload of the cloud while the latter the service rate. In our scenario we have $\mu(j) = \mu(T)$ for $j \in [T+1, K]$, i.e., the service rate does not increase with the system load when we are in the phase in which we observe the QoS degradation. For each state of the cloud model, the attacker can observe a certain QoS. However, all the states in $[0, T]$ satisfy the SLA and hence are indistinguishable from the point of view of the attacker. We represent this fact with a $K+1 \times K+1$ matrix where the diagonal elements have value 1 if the SLA are satisfied:

$$\mathbf{C}_{OK}(i,j) = [i = j][i \leq T]\,, \quad 0 \leq i, j \leq K\,. \quad (2)$$

Let L_k be the QoS observed by the attacker when the cloud model is in state k with $k \in [T+1,\ K-1]$ and let

$$\mathbf{C}_{L_k}(i,j) = [i = j][i = k]\,, \quad T+1 \leq k \leq K\,, \quad (3)$$

and $i, j \in [0, K]$. Finally, we introduce a matrix that describes the reaction of the cloud to the workload generated by the attacker. Essentially this is indistinguishable, from the point of view of the cloud, from the standard requests. The matrix consists of 0s and 1s, where 1 denotes the transition corresponding to an injection of work from the attacker. The rate of this transition is specified by the attacker.

$$\mathbf{C}_\lambda(i,j) = [j = i+1]\,. \quad (4)$$

3.2 A model for the attacker

In this section we model the behaviour of the attacker described in Section 2. We recall that the attacker can only observe the QoS offered by cloud in order to decide the intensity of the workload that will be required in order to achieve a eDOS attack. The attacker consists of G states

$$\mathcal{S}_A = \{0, \dots, G-1\}\,,$$

each of which generates a workload for the cloud $\lambda_A(g)$, $g \in [0, G-1]$. Without loss of generality we assume $\lambda_A(g_1) \geq \lambda_A(g_2)$ if $g_1 \geq g_2$. Let $\mathbf{A}_\lambda$ be a $G \times G$ matrix defined as:

$$\mathbf{A}_\lambda(i,j) = [i = j]\lambda_A(i)\,. \quad (5)$$

The attacker observes the QoS offered by the cloud which can be $OK, L_{T+1}, \dots, L_{K-1}$ and decides the transition to a new state. Intuitively, given that the attacker is in state g_1, if a QoS denoted by OK is observed then a transition to a state $g_2 \geq g_1$ is performed, whereas if a QoS L_k is observed then a transition to state $g_2 \leq g_1$ is performed. Let $\gamma(g)$ be the rate at which the attacker decides the intensity of the workload given that it is in state g, and let $\mathbf{A}_{OK}$, $\mathbf{A}_{L_k}$, $k \in [T+1, K-1]$ be the matrices that describe the reaction of the attacker to an observed QoS. For instance reasonable settings for these matrices could be:

$$\begin{aligned} &\mathbf{A}_{OK}(i,j) = \gamma(i)[j = i+1]\,, \quad (6)\\ &\mathbf{A}_{L_k}(i,j) = \gamma(i)[j = i-1]\,, \quad T+1 \leq k \leq K-1\,, \end{aligned}$$

and $0 \leq i, j \leq G-1$. In this case the attacker moves only between adjacent states trying to find an equilibrium that maximises the cloud power consumption while maintaining the QoS acceptable. A different strategy consists in drastically reduce the workload as soon as the attacker perceives a deterioration of the QoS. This model aims at reducing the probability of failing the attack (i.e., the cloud model goes to state K). This behaviour can be modelled as follows:

$$\begin{aligned} &\mathbf{A}_{OK}(i,j) = \gamma(i)[j = i+1]\,, \quad (7)\\ &\mathbf{A}_{L_k}(i,j) = \gamma(i)[j = 0]\,, \quad T+1 \leq k \leq K-1\,, \end{aligned}$$

3.3 The cooperation between attacker and cloud models

Now we define the joint model between attacker and cloud by means of the Kronecker's algebra. Its transition rate matrix $\mathbf{M}$ is:

$$\mathbf{M} = \mathbf{C}_0 \otimes \mathbf{I}_G + \mathbf{C}_{OK} \otimes \mathbf{A}_{OK} + \sum_{k=T+1}^{K-1} \mathbf{C}_{L_k} \otimes \mathbf{A}_{L_k} + \mathbf{C}_\lambda \otimes \mathbf{A}_\lambda\,, \quad (8)$$

where $\mathbf{I}_G$ is the identity matrix with size G. The corresponding infinitesimal generated is defined as:

$$\mathbf{Q} = \mathbf{M} - \mathrm{diag}(\mathbf{M}\mathbf{1})\,, \quad (9)$$

where $\mathbf{1}$ is a column vector of all 1 and diag transforms a column vector $\mathbf{v}$ into a diagonal matrix whose elements in position (i,i) are the elements $\mathbf{v}(i)$.

3.4 Quantitative indices

We observe that the states of $\mathbf{M}$ does not describe an ergodic CTMC. Indeed, once the cloud model is in state K it cannot leave. Therefore, in the joint model described by transition matrix $\mathbf{M}$ all the sates (K,g) with $g = 0,\dots G-1$ are associated with a failure of the attack and represent an absorbing subset of the states. Let $X(t)$ be the CTMC whose infinitesimal generator is $\mathbf{Q}$ as defined by Equation (9). A state of $X(t)$ is a pair (k,g) with $0 \le k \le K$ and $0 \le g \le G-1$. We write $|X(t)|_1$ ($|X(t)|_2$) to denote the first (second) component of the pair. Let τ be the r.v. representing the time required by the chain to reach a state of the class (K,g) where $g \in [0, G-1]$:

$$\tau = \inf\{t \ge 0 | X(t) = (K,g)\,, g \in [0,G-1]\}\,.$$

By definition of $\mathbf{M}$ (and hence of $\mathbf{Q}$) when the transition rates are strictly positive τ is finite with probability 1. Hence $\overline{\tau} = E[\tau]$ is the finite expected time to absorption. The energy consumed up to absorption is the r.v. defined as:

$$R = \int_0^\infty p(|X(t)|_1)dt\,, \tag{10}$$

since $p(K) = 0$ by definition. Since $p(k)$ is bounded then $P\{R < \infty\} = 1$ and we define $\overline{R} = E[R]$ as the expected energy consumed by the cloud before the absorption.

4. PERFORMANCE INDICES

In this section we give effective methods for the computation of the performance indices. We observe that the standard approach based on the computation of the expected time and reward to absorption is not feasible for studying attacks with long duration since the matrix inversion which is required becomes a numerically unstable operation. For this reason we resort to an approximation technique which is based on the notion of quasi-stationary distribution. In Section 5 we show the quality of this approximation by comparing it with the exact results in regions in which the numerical stability problem of the exact problem does not arise yet.

4.1 Exact numerical computation

The computation of $\overline{R}$ and $\bar{t}$ can be performed in a standard way (see, e.g., [14, Ch. 10]). Let $\mathbf{M}' = [\mathbf{M}]_{KG}$ be the transition rate matrix formed with the first $K \cdot G$ rows and columns of $\mathbf{M}$, and let $\mathbf{P}$ be defined as:

$$\mathbf{P} = \left([\mathrm{diag}(\mathbf{M}\mathbf{1})]_{KG}\right)^{-1}\mathbf{M}'\,,$$

i.e., $\mathbf{P}$ is the transition matrix of the discrete time Markov chain (DTMC) embedded in $X(t)$ reduced to the transient subset of states and hence $\mathbf{P}$ is sub-stochastic and can be inverted [14]. Let $R_s = E[R|X(0) = s]$ with $s \in [0,K-1] \times [0,G-1]$, and let $\mathbf{r}$ be the column vector whose component s is R_s. We have:

$$\mathbf{r} = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{v}\,, \tag{11}$$

where $\mathbf{v}$ is a column vector whose s-th component is

$$\mathbf{v}(s) = \frac{p(|s|_1)}{\sum\limits_{\substack{j \in [0,K]\times[0,G-1] \\ j \ne s}} q_{sj}}\,.$$

Let $\boldsymbol{\pi}(s)$ be the column vector with the initial distribution, then $\overline{R}$ is:

$$\overline{R} = \boldsymbol{\pi}^T\mathbf{r}\,. \tag{12}$$

The numerical computation of $\overline{\tau}$ is analogous, the difference relies on the definition of vector $\mathbf{v}$ which has to be replaced by $\mathbf{w}$ in Equation (11) where:

$$\mathbf{w}(s) = \frac{1}{\sum\limits_{\substack{j \in [0,K]\times[0,G-1] \\ j \ne s}} q_{sj}}\,,$$

and hence $\mathbf{r}' = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{w}$ and

$$\overline{\tau} = \boldsymbol{\pi}^T\mathbf{r}'\,. \tag{13}$$

4.2 Approximation of the performance indices for long absorption times

In practice, when the duration of an attack is very long the evaluation of the performance indices by means of Equation (11) is unfeasible because matrix $\mathbf{I} - \mathbf{P}$ is almost singular. Therefore, we propose an approximation method that relies on the theory of *quasi stationarity* [3]. Intuitively, when the expected time to absorption is much greater than the CTMC's transition times and the finite transient subset of states is irreducible, it may be the case that the transient part, conditioned to the fact that the absorbing states are not visited, reaches a stationary behaviour. Based on this intuition, the theory of quasi-stationary CTMCs has been developed to study the extinction times in population models (see, e.g., [3, 2, 1]).

Let us consider the CTMC $X(t)$ above defined and let

$$\mathcal{U} = \{(k,g) : k \in [0,K-1] \wedge g \in [0,G-1]\}\,,$$

be the set of transient states and $\mathbf{Q}_U = [\mathbf{Q}]_{KG}$ be the infinitesimal generator matrix reduced to the states in $\mathcal{U}$. Vector $\mathbf{a}$ defined as

$$\mathbf{a} = -\mathbf{Q}_U\mathbf{1}\,,$$

has non-negative entries that represent the sum of the transition rates from any transient state in $\mathcal{U}$ to one of the absorbing states. Note that since we are not interested in the behaviour of the chain once it reaches a state $s = (K,g)$ for arbitrary g we can simplify our exposition by assuming a unique absorbing state $s = (K,\cdot)$. With this simplification the generator of $X(t)$ can be written as:

$$\begin{bmatrix} \mathbf{Q}_U & \mathbf{a} \\ \mathbf{0} & 0 \end{bmatrix}.$$

Definition 1. *Let τ be the time to absorption of $X(t)$, i.e., then a distribution $\mathbf{u}$ is said to be quasi-stationary for $X(t)$ if*

$$Pr_{\mathbf{q}}\{X(t) = s | \tau > t\} = \mathbf{q}(s)\,,$$

where $Pr_{\mathbf{q}}$ denotes that the distribution of $X(0)$ is $\mathbf{q}$.

Henceforth, we assume that the states in $\mathcal{U}$ form a single communicating class. Matrix $\mathbf{Q}_U$ has a unique eigenvalue $-\alpha$ with maximal real part and $\alpha \in \mathbb{R}^+$ with geometric multiplicity 1 [15, 10, 17]. Therefore, there exists a unique pointwise positive vector $\mathbf{q}$ such that

$$\mathbf{q}^T \mathbf{Q}_U = -\alpha \mathbf{q}^T ,$$

with $\mathbf{1}^T \mathbf{q} = 1$ and $\mathbf{q}$ is the unique distribution that satisfies Definition 1. The following proposition plays a pivotal role in our approximation because it gives an exact numerical method to derive the distribution of the time to absorption from a quasi-stationary distribution.

PROPOSITION 1 (TIME TO ABSORPTION [10, 17]). *Let $\mathbf{q}$ be the quasi-stationary distribution of $X(t)$ for the subset of states $\mathcal{U}$, then we have:*

$$Pr_{\mathbf{q}}\{\tau > t + \Delta_t | \tau > t\} = e^{-\alpha \Delta_t} \quad t, \Delta_t \geq 0 . \tag{14}$$

i.e., the absorption time from a quasi-stationary distribution is exponentially distributed with parameter given by the highest (negative) real (left) eigenvalue of $\mathbf{Q}_U$.

As a consequence we simply have $\overline{\tau} = \alpha^{-1}$ when the chain at time 0 is distributed according to a quasi-stationary distribution.

In our case, the initial state of $X(t)$ is not necessary a quasi-stationary distribution since this would require the attacker to choose its initial behaviour from a random state chosen according to a distribution whose computation requires a deep knowledge of the cloud architecture that is not available. However, for large values of τ we assume that by the time of absorption the chain has spent enough time in its quasi-stationary distribution in such a way that the warm up period duration is negligible. Indeed, the following result holds [17].

PROPOSITION 2. *Let $\mathbf{w}$ be any probability distribution over $\mathcal{U}$, then*

- $\lim_{t\to\infty} Pr_{\mathbf{w}}\{\tau > t + \Delta_t | \tau > t\} = e^{-\alpha \Delta_t}$;
- $\lim_{t\to\infty} Pr_{\mathbf{w}}\{X(t) = s | \tau > t\} = \mathbf{q}(s)$.

Therefore, for large absorption times we approximate $\overline{\tau}$ as:

$$\overline{\tau} \simeq \alpha^{-1} , \tag{15}$$

regardless to the initial distribution of $X(t)$. Finally, we have:

$$\overline{R} \simeq \alpha^{-1} \sum_{s \in \mathcal{U}} p(|s|_1) \mathbf{q}(s) . \tag{16}$$

In practice the precision of the approximation depends on the spectral gap η between α and α_2, where α_2 is the eigenvalue with the next largest real part after α:

$$\eta = Re(\alpha_2) - \alpha .$$

The convergence of the initial distribution of $X(t)$ to the quasi-stationary distribution is fast if $\eta >> \alpha$.

It remains to address the problem of determining the highest left eigenvalue of $\mathbf{Q}_U$. We note that since $\mathbf{Q}_U$ is a square matrix with components in $\mathbb{R}$ the set of right and left eigenvalues are the same. Moreover, finding the eigenvalue with the highest real part of $\mathbf{Q}_U$ corresponds to finding the eigenvalue with the smallest real part of $-\mathbf{Q}_U$. Indeed, $-\mathbf{Q}_U$ is a M-matrix [17]. Moreover $\mathbf{Q}_U$ is also diagonal dominant, therefore the computation of the eigenvalue with the smallest real part can be performed with one the algorithms described in [16] whose numerical accuracy is higher than the standard QR decomposition for tiny values of α (values of 10^{-12} are handled with precision of 10^{-16}) even for ill-conditioned cases.

5. EXPERIMENTS

The goal of this section is to use the model defined in Section 3 and the analysis techniques introduced in Section 4 to evaluate the effects of the eDoS attacks and compare different strategies for the attacker. We propose three strategies following the lines of [6]:

Strategy 1 The attacker moves from state g to state $g+1$, i.e., it increases the arrival intensity at the cloud system whenever it observes a QoS of type OK and it has not already reached the maximum intensity it could generate. Conversely, the attacker moves from state g to state $g-1$ whenever it observes a QoS of type L_k and it is not already at state 0. This is the same behaviour that is described by Equation (6).

Strategy 2 As in the previous strategy, the attacker moves from state g to state $g+1$ whenever it observes a QoS of type OK and it has not already reached the maximum intensity it could generate. However, when it observes a QoS of type L_k, the attacker goes back to the state 0. This is the behaviour described by Equation (7).

Strategy 3 The attacker, like in the previous strategy, moves from state g to state $g+1$ whenever it observes a QoS of type OK and it has not already reached the maximum intensity. When a QoS of type L_k is observed, the attacker moves from state g to state $\max(g-k+T, 0)$. This approach is less aggressive than the one of Strategy 1, but more than the one of Strategy 2.

In Section 5.1 we validate the accuracy of the approximation method based on quasi-stationary distributions and in Section 5.2 we study the performance indices (expected energy consumption and expected time to absorption) for the three strategies. Specifically, we evaluate the *damage* of an attack as the ratio between the expected energy consumption by considering the effects of the attacker and the expected energy consumption (in the same time interval) of the cloud without the attacker in the same time interval. The latter index is computed by computing the quasi stationary distribution of the CTMC underlying the cloud model by conditioning on the fact that the absorbing states are not visited. Finally we consider the *impact* of an attack as the ratio between the expected time to absorption in presence of the attacker and the expected time to absorption without the attacker.

The parameters of the experiments shown in Sections 5.1 and 5.2 are given by the respective columns of Table 1. In-

terval values, such as those of λ for Section 5.1 and of F for Section 5.2, indicate that the experiment is carried on over sampled points in that range. All the computations were carried out using MATLAB.

5.1 Validation of the approximation

In Section 3 we observed that, for long absorption times $\overline{\tau}$, $\overline{R}$ and $\overline{\tau}$ can be approximated using Equations (16) and (15), respectively. Now we evaluate the accuracy of those approximations for several sets of parameters. We show some significant plots that summarise our validation procedure. We chose an arbitrary set of parameters, given by Table 1, whose names are coherent with those given in Section 3. The denominator of parameter $\gamma(g)$, i.e., the frequency at which the attacker decides to change its state according to the observed behaviour of the cloud, is chosen in order to allow the attacker to gain a statistically significant number of observations. In this section the parameters of the attacker are fixed, while the independent arrival rate to the cloud λ, thus its initial load, varies within the given interval. The power consumption of the cloud is proportional to the number of active servers, thus $p(k) = \min(k, T)$.

In Figures 1 and 2 we plot the exact and approximate values of $\overline{R}$ and $\overline{\tau}$, respectively, using Strategy 1. In Figures 3, 4 and 5 we show the relative error between the exact and the approximate value of $\overline{R}$ and $\overline{\tau}$, for the set of parameters given by Table 1 and the strategies 1,2 and 3 described below, respectively. In those and all the following examples, the initial distribution $\boldsymbol{\pi}(s)$ in Equations (12) and (13) is assumed to be

$$\boldsymbol{\pi}(s) = \begin{cases} \boldsymbol{\pi}(s)_{[\mathbf{C}]_K}\left(\left\lfloor \frac{s}{G} \right\rfloor\right) & \text{if } s \mod G = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\boldsymbol{\pi}(s)_{[\mathbf{C}]_K}$ is the stationary distribution of the cloud $\mathbf{C}$, conditioned on the fact that the absorbing states have not been visited, considered in isolation.

We have studied our model in a region in which the expected time to absorption can still be studied in the exact way but is long enough to make the approximating approach accurate. For longer expected absorption times the exact method based on the matrix inversion would not be feasible any more, due to the numerical instability of the matrix inversion required by Equation (12) and (13). The figures clearly show that, as the inter-arrival time increases, making the mean absorption time longer, the relative error of the approximation rapidly decreases. This is coherent with the claims of Section 3.

5.2 Comparison of different attack strategies

In this section we compare the aforementioned three strategies of the attacker with respect to the effects on the average absorption time $\overline{\tau}$ and the average energy consumption $\overline{R}$ for the cloud system model. As shown in Table 1, the parameters for those experiments are partially different from those for the experiments of Section 5.1. In particular, we chose a reference λ for the independent arrivals to the cloud, and we varied the scale of the attack intensity, here represented by a factor F. We parametrise the attack intensity $\lambda_A(g)$ for a state g of the attacker such that, after an initial phase in which the attacker does not perform requests, the attacker

Parameter	Sec. 5.1	Sec. 5.2
K	20	20
T	14	14
G	6	6
λ	$[1.3, 7.0]$	1
μ	1.2	0.5
$\gamma(g)$	$\mu/30$	$\min\left(\max\left(\lambda_A(g), \lambda\right), T\mu\right)/30$
$\lambda_A(g)$	Fg	$Fg\mu$
F	0.8	$[2.0, 8.0]$
$p(k)$	$\min(k, T)$	$\min(k, T)$

Table 1: Parameter values for the experiments of Section 5

at first chooses a request rate slightly lower (here by a factor of 0.8) than the service rate of a single Cloud server, which is μ. The choice for the transition rate $\gamma(g)$ between a state of the attacker g and its adjacent states $g-1$ and $g+1$ (if they exists) is determined by $\min\left(\max\left(\lambda_A(g), \lambda\right), T\mu\right)/30$. Here, as stated before, the denominator represents the fact that, in a real system, we want to collect enough samples (here 30) on the QoS data to have statistical significance. The numerator, here, represents the rate at which we could expect to collect those data, i.e., how many service completions the attacker could observe for its own requests. This quantity is capped by the maximum throughput of the cloud $T\mu$, and is given either by the throughput λ of the cloud in isolation or by the rate at which the attacker performs requests $\lambda_A(g)$.

In Figure 7 we show how the mean absorption time for the same set of parameters varies according to the chosen strategy, while in Figure 6 the same comparison is made for the expected energy consumption. In Figure 8, Figure 9 and Figure 10, the expected energy consumption of the cloud system without an attacker is compared to the one having an attacker using strategies 1, 2 and 3 respectively. In those last examples, the energy consumption of the cloud in isolation is computed over the mean absorption time $\overline{\tau}$ for the model with the attackers (which explains why apparently the model in isolation shows a different behaviour for different attacker strategies). Figures 11 and 12 show the ratio, for the previously described strategies, between the mean energy consumption of the model including an attacker and the energy consumption, over the same time frame, of the cloud in isolation. For the first of these Figures the range chosen for F is $(0, 2]$, in order to show that the ratios are not monotonic and that they may have a different maximum for different strategies. Finally, Figures 13 and 14 show, for each of the aforementioned strategies, the ratios between the average absorption time of the model with and without the attacker, in the range $F \in (0, 2)$ and $F \in [2, 10]$, respectively. The figures show that, for this set of parameters, Strategy 2, i.e., the one with the least aggressive behaviour, is the most effective in inducing the cloud system to consume more energy, followed by Strategy 3. Strategy 1, i.e., the approach suggested in the literature, is the most aggressive but the least effective of these strategies, especially for a low load factor. Although the comparison between Figures 6 and 7 seems to suggest that those performance differences exist due to the differences in absorption time, Figures 11 and 12 show that, even for ratios between consumptions normalised over the same time, Strategy 2 is indeed better.

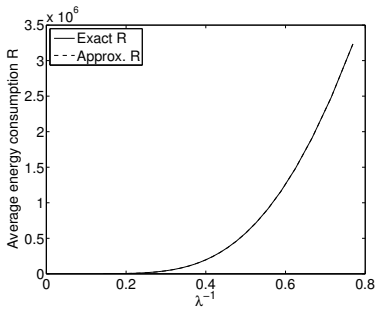

Figure 1: Exact and approximate computation of $\overline{R}$

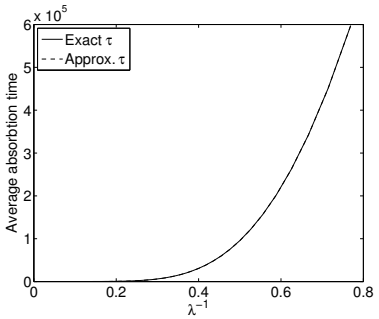

Figure 2: Exact and approximate computation of $\overline{\tau}$

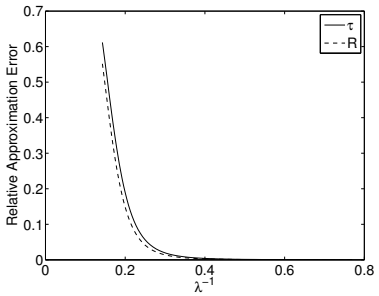

Figure 3: Relative approximation error for $\overline{R}$ and $\overline{\tau}$, Strategy 1

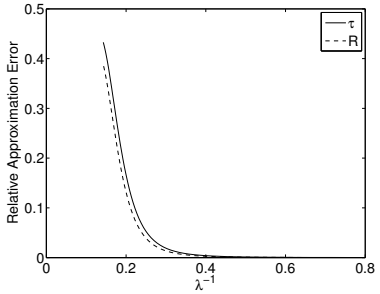

Figure 4: Relative approximation error for $\overline{R}$ and $\overline{\tau}$, Strategy 2

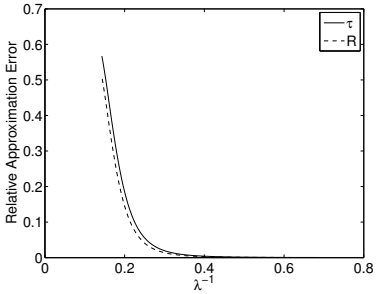

Figure 5: Relative approximation error for $\overline{R}$ and $\overline{\tau}$, Strategy 3

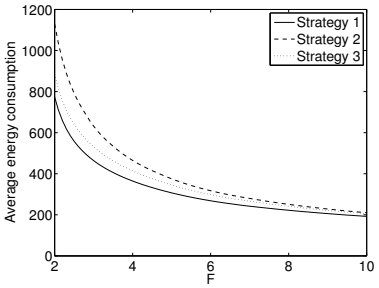

Figure 6: Computation of $\overline{R}$ for different strategies

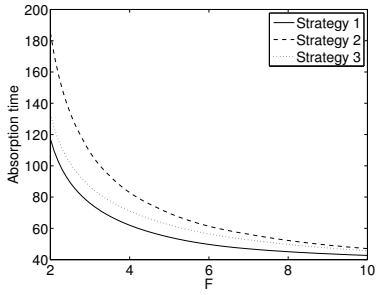

Figure 7: Computation of $\overline{\tau}$ for different strategies

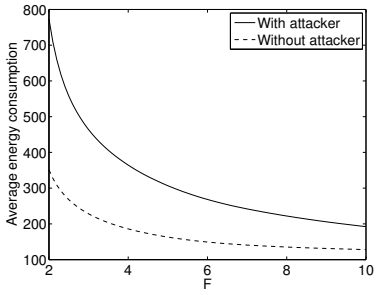

Figure 8: Comparison of $\overline{R}$ with or without attacker. Strategy 1

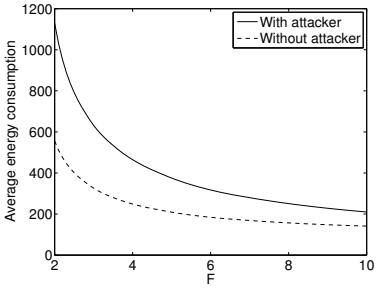

Figure 9: Comparison of $\overline{R}$ with or without attacker. Strategy 2

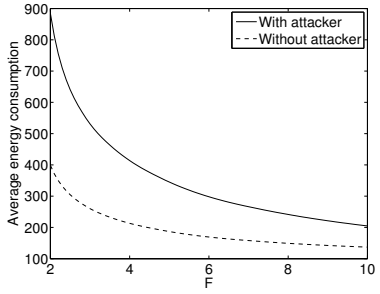

Figure 10: Comparison of $\overline{R}$ with or without attacker. Strategy 3

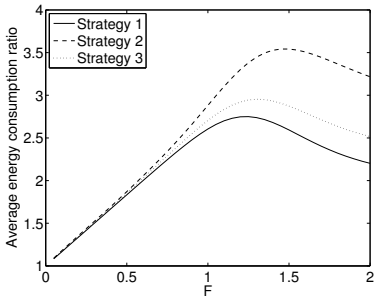

Figure 11: Ratio between values of $\overline{R}$ with and without attacker, $F \in (0, 2]$

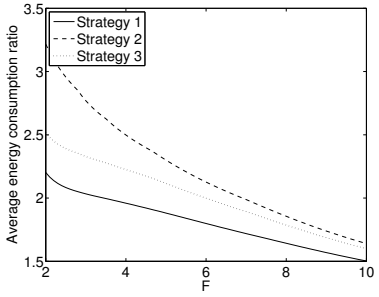

Figure 12: Ratio between values of $\overline{R}$ with and without attacker, $F \in [2, 10]$

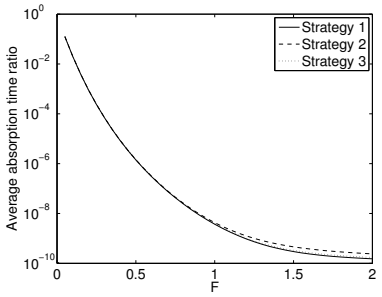

Figure 13: Ratio between values of $\overline{\tau}$ with and without attacker, $F \in (0, 2]$

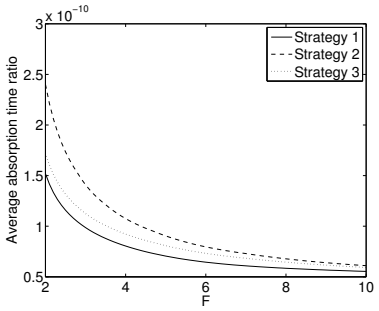

Figure 14: Ratio between values of $\overline{\tau}$ with and without attacker, $F \in [2, 10]$

6. CONCLUSION

In this paper we have proposed a Markovian model to study the impact of eDoS attacks to cloud infrastructures. These kind of security attacks target the management infrastructure of the cloud by injecting fictitious workload that increases the system energy consumption. The proposed analysis is based on the evaluation of the mean time to absorption and on the expected cumulated rewards in a CTMC describing the attacker strategy and the cloud state. We gave numerically stable methods to compute (or approximate for long-lasting attacks) the performance indices that allow us to evaluate the impact of an attack. Our findings show that low-aggressive strategies of the attackers are more dangerous for the cloud since they do not change significantly the life-time of the systems while they maintain a higher energy consumption. Future works include a formulation of a more detailed model of the cloud infrastructure and a validation of the analysis. Based on these works we plan do design a statistic approach to estimate the probability of being in presence of an eDoS attack in a cloud infrastructure.

7. REFERENCES

[1] J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero. The maximum number of infected individuals in SIS epidemic models: computational techniques and quasi-stationary distributions. *J. of Comput. Appl. Math.*, 233:2563–2574, 2010.

[2] A. D. Barbour. Quasi-stationary distributions in markov population processes. *Adv. Appl. Prob.*, 8:296–314, 1976.

[3] M. S. Barlett. *Stochastic Population Models in Ecology and Epidemiology*. London: Methuen, 1960.

[4] G. Ciardo, M. Gribaudo, M. Iacono, A. Miner, and P. Piazzolla. Power consumption analysis of replicated virtual applications in heterogeneous architectures. In *ITAIS 2015*, pages –, to appear 2015.

[5] V. Durcekova, L. Schwartz, and N. Shahmehri. Sophisticated denial of service attacks aimed at application layer. In *ELEKTRO, 2012*, pages 55–60, May 2012.

[6] M. Ficco and F. Palmieri. Introducing fraudulent energy consumption in cloud infrastructures: A new generation of denial-of-service attacks. *Systems Journal, IEEE*, PP(99):1–11, 2015.

[7] J.-C. Lin, F.-Y. Leu, and Y.-P. Chen. Analyzing job completion reliability and job energy consumption for a heterogeneous mapreduce cluster under different intermediate-data replication policies. *The Journal of Supercomputing*, 71(5):1657–1677, 2015.

[8] V. J. Maccio and D. G. Down. On optimal policies for energy-aware servers. In *Proc. of MASCOTS*, pages 31–39, 2013.

[9] I. Mitrani. Managing performance and power consumption in a server farm. *Annals OR*, 202(1):121–134, 2013.

[10] I. Nasell. Extinction and quasi-stationarity in the verhulst logistic model. *J. of Theoretical Biology*, 211:11–27, 2001.

[11] F. Palmieri, S. Ricciardi, and U. Fiore. Evaluating network-based dos attacks under the energy consumption perspective: New security issues in the coming green ict area. In *Broadband and Wireless Computing, Communication and Applications (BWCCA), 2011 International Conference on*, pages 374–379, Oct 2011.

[12] F. Palmieri, S. Ricciardi, U. Fiore, M. Ficco, and A. Castiglione. Energy-oriented denial of service attacks: an emerging menace for large cloud infrastructures. *The Journal of Supercomputing*, 71(5):1620–1641, 2015.

[13] C. Ren, D. Wang, B. Urgaonkar, and A. Sivasubramaniam. Carbon-aware energy capacity planning for datacenters. In *Proc. of MASCOTS*, pages 391–400, 2012.

[14] G. Rubino and B. Sericola. *Markov chains and dependability theory*. Cambridge Press, 2014.

[15] E. Seneta. *Non-negative matrices and Markov chains*. Springer, 1981.

[16] A. Sule Alfa, J. Xue, and Q. Ye. Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix. *Mathematics of computation*, 71(237):217–236, 2001.

[17] E. A. van Doorn and P. K. Pollett. Survival in quasi-death process. *Linear algebra and its applications*, 429:776–791, 2008.

[18] Z. Wu, M. Xie, and H. Wang. On energy security of server systems. *Dependable and Secure Computing, IEEE Trans. on*, 9(6):865–876, Nov 2012.