Evaluating the impact of eDoS attacks to cloud facilities

Gian-Luca Dei Rossi [1]   Mauro Iacono [2]   Andrea Marin [1]

[1]Università Ca' Foscari Venezia [2]Seconda Università di Napoli

November 18, 2015

The setting

Nowadays the use of cloud computing is widespread

- Infrastructure as a service
- Platform as a service
- Software as a service
- ...

Cloud services providers have to manage capacity within constraints such as

- Performance constraints (SLAs,...)
- Economic constraints (budgets, pricing policies,...)

Economic constraints impose *energy management* policies

- Hardware powered on and off *on demand*
- Policies have to take into account performance constraints
    - Strategies can be complex and at different granularities

eDoS attacks

Cloud facilities may be subject to *Denial of Service* (DoS) attacks

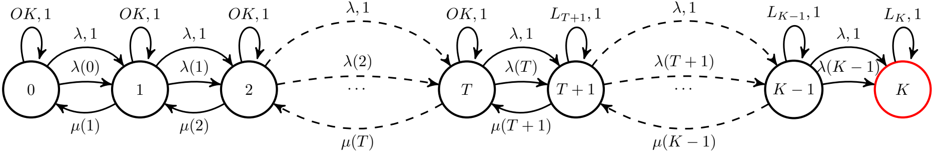
- aiming at degrading performance indices, e.g., average response time, and breaking SLAs
- easy to notice, but not so easy to counteract
- the attacker has a simple and noticeable goal

An *Energy oriented Denial of Service* (eDoS) attack, on the other hand

- aims at the maximisation of energy consumption
- using legitimate workload
- non-disruptive and long-term
  - it should not crash the system
  - it has to be hard to notice
- the attacker has not a feedback on the success of the attack
  - no knowledge about energy management policies of providers
  - lack of a simple correlation between load and energy consumption

We want to model the behaviour of those attacks with respect to different strategies.

A model for cloud infrastructures



Finite set of states $\mathcal{S}_C = \{0, 1, 2, \dots K\}$

- states 0 to T: system dynamically scales its computational power
- states $T+1$ to $K-1$: system cannot scale, performance degradation
- state K: the system has crashed or the attack was discovered

Transitions:

$$\mathbf{C}_0(i,j) = \lambda(i)[j = i+1] + \mu(j)[j = i-1][j \neq K] \quad \text{std. workload and services}$$

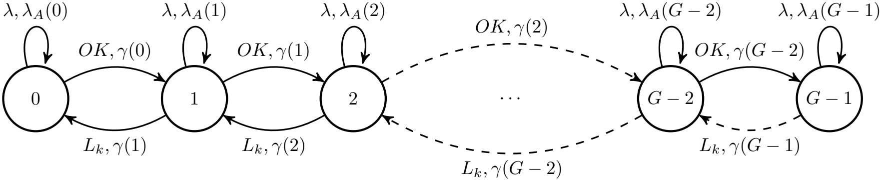
$$\mathbf{C}_{OK}(i,j) = [i = j][i \leq T], \quad 0 \leq i, j \leq K \quad \text{performance are OK}$$

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

$$\mathbf{C}_\lambda(i,j) = [j = i+1] \quad \text{workload from the attacker}$$

Let $p\colon \mathcal{S}_C \to \mathbb{R}^+$, $p(K) = 0$, represent the power spent in each state of the cloud.

A model for e-attackers



Finite set of states $\mathcal{S}_A = \{0, \ldots, G-1\}$
Transitions:

$$\mathbf{A}_\lambda(i,j) = [i = j]\lambda_A(i) \quad \text{attack intensity}$$
$$\mathbf{A}_{OK}(i,j) = \gamma(i)[j = i+1] \quad \text{increase intensity}$$
$$\mathbf{A}_{L_k}(i,j) = \gamma(i)[j = i-1], \quad T+1 \leq k \leq K-1 \quad \text{decrease intensity}$$

Note: A_{OK} and A_{L_k} may vary with respect to the strategy adopted.

Cloud-Attacker interaction

We define the joint model between attacker and cloud using the $G(K+1) \times G(K+1)$ transition matrix

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

The corresponding infinitesimal generator is

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

and the associated Markov chain is $X(t)$

- states of $X(t)$ are pairs (k, g) with $0 \leq k \leq K$ and $0 \leq g \leq G-1$
- we write $|X(t)|_1$ ($|X(t)|_2$) to denote the first (second) component of the pair.

Quantitative Indices

States of $\mathbf{M}$ does not describe an ergodic CTMC

- Once the cloud is in state K (failure or attack detection) it cannot leave
- In the joint model all states (K, g) with $g = 0, \ldots G-1$ form an *absorbing subset* of the states

τ is the r.v. representing the time required by the chain to reach an absorbing state:

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

$\overline{\tau} = E[\tau]$ is the finite *expected time to absorption*.
The energy consumed up to absorption is the r.v. defined as:

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

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.

Exact computation of the indices

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., the DTMC embedded in $X(t)$ reduced to the transient states.
Let $\mathbf{r}$ be the vector s.t. $\mathbf{r}(s) = E[R|X(0) = s]$, computed as

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

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

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

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

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

The computation of $\overline{\tau}$ is analogous, fixing the numerator of $\mathbf{v}$ to 1

Approximate computation

When the attack is very long, $\mathbf{I}-\mathbf{P}$ is almost singular $\Longrightarrow$ **numerical instability**

- We propose an approximation based on *quasi stationarity* theory
- If $\overline{\tau} \gg$ trans. times of $X(t)$, transient part may have a stationary behaviour.

Let $\mathcal{U}$ be the set of the transient states of $X(t)$

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

and $\mathbf{Q}_U = [\mathbf{Q}]_{KG}$ be the infinitesimal generator matrix reduced to the states in $\mathcal{U}$.

Definition

A distribution $\mathbf{u}$ is 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 $\mathbf{q}$.

$\mathbf{Q}_U$ has a unique eigenvalue $-\alpha$ with maximal real part. $\mathbf{q}$ is the unique vector s.t.

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

with $\mathbf{1}^T \mathbf{q} = 1$. $\mathbf{q}$ is the unique distribution that satisfies the Definition above.

Approximate computation: absorption time

Proposition (Time to absorption)

Let $\mathbf{q}$ *be the quasi-stationary distribution of* $X(t)$ *for the subset of states* $\mathcal{U}$*, then:*

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

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

Therefore $\overline{\tau} = \alpha^{-1}$ when the chain at time 0 is q.s. distributed.
In general we cannot make that assumption, however the following results hold

Proposition

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, regardless to the initial distribution of $X(t)$,

$$\overline{\tau} \simeq \alpha^{-1}$$

Approximate computation: energy consumption

The computation of the approximate average energy consumption is given by

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

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 $\eta >> \alpha$.

Since $\mathbf{Q}_U$ is a diagonal dominant M-matrix, the computation of the eigenvalue with the smallest real part can use fast and stable algorithms.

Experimenting with the model

The presented model can be used to

- evaluate the energy consumption of a cloud infrastructure given a (legitimate or not) load
- evaluate the behaviour and the effectiveness of an eDoS attacker using a particular strategy
- evaluate the quality of the quasi-stationarity based approximation

In order to perform those evaluations, we use a MATLAB® custom-made implementation of the described methods.

In the following examples, the initial distribution $\boldsymbol{\pi}(s)$ 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.

Attack strategies

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.
- The attacker moves from state g to state $g-1$ whenever it observes a QoS of type L_k.

Strategy 2
- The attacker moves from state g to state $g+1$ whenever it observes a QoS of type OK.
- The attacker goes back to state 0 whenever it observes a QoS of type L_k.

Strategy 3
- The attacker moves from state g to state $g+1$ whenever it observes a QoS of type OK.
- When a QoS of type L_k is observed, the attacker moves from state g to state $\max(g-k+T,0)$.

Parameters

Parameter	Approx. Validation	Strategies comparison
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: Parameter values for the experiments

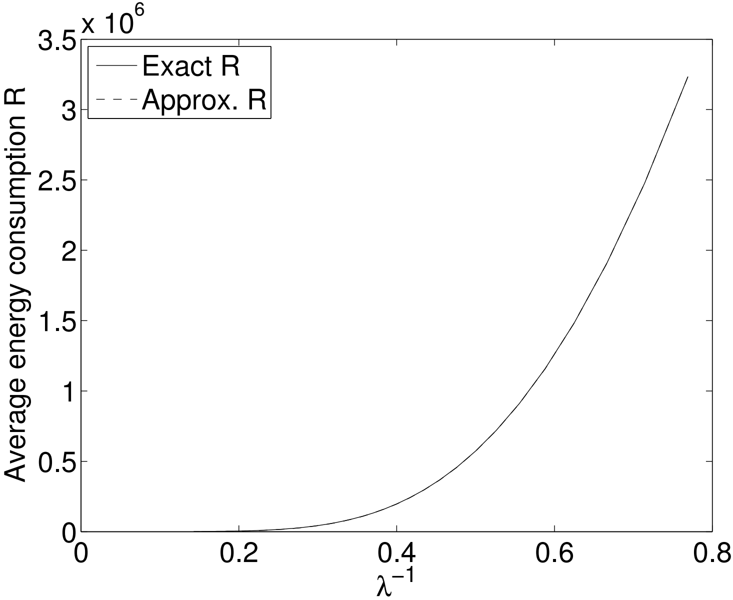


Figure: Exact and approximate computation of $\overline{R}$

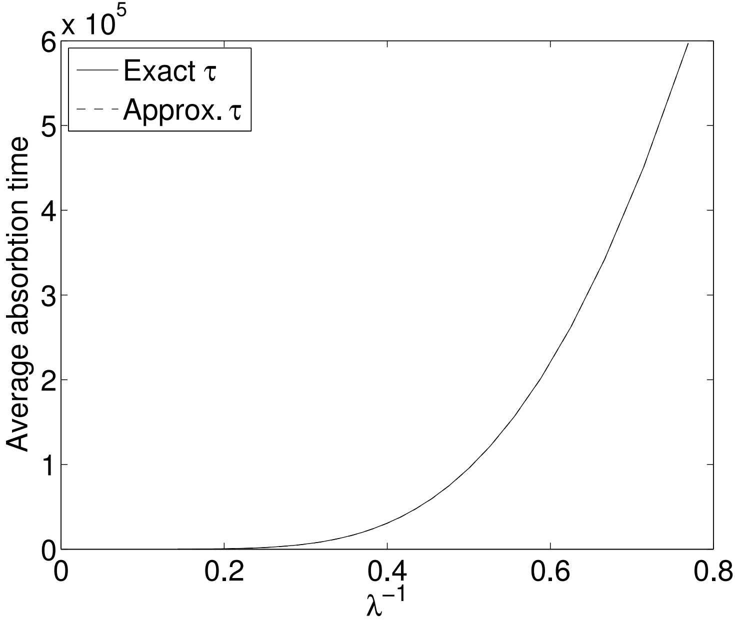


Figure: Exact and approximate computation of $\overline{\tau}$

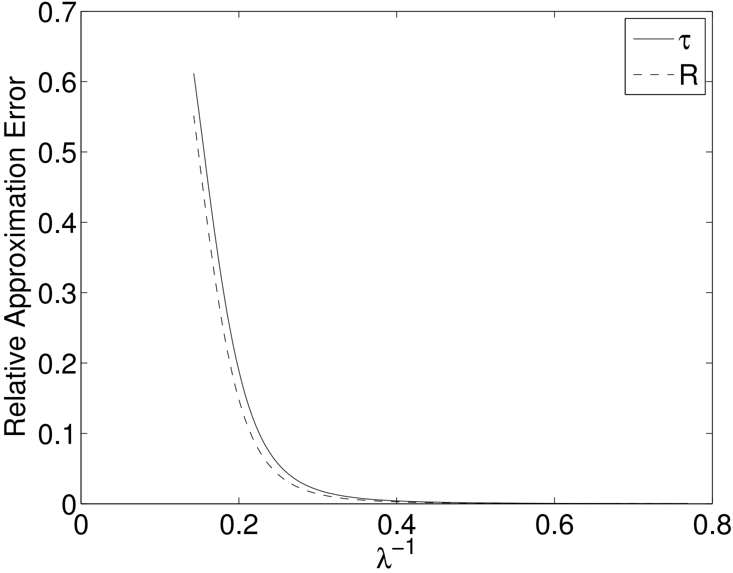


Figure: Relative approximation error for $\overline{R}$ and $\overline{\tau}$, Strategy 1

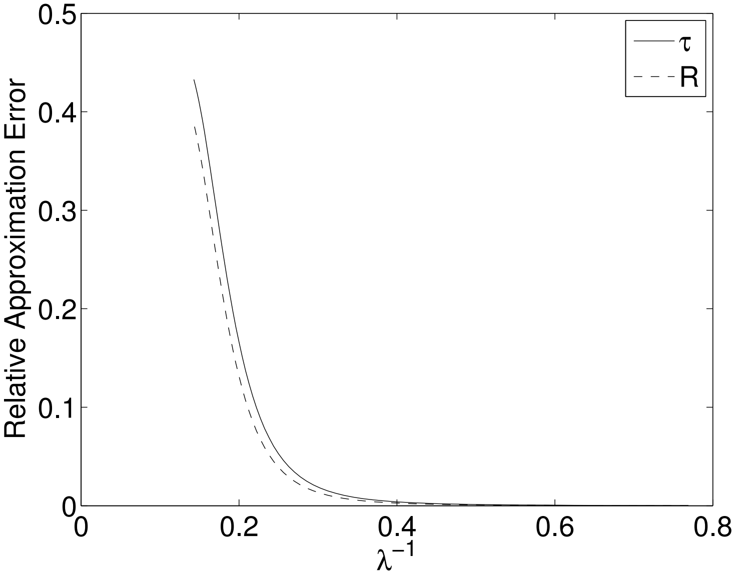


Figure: Relative approximation error for $\overline{R}$ and $\overline{\tau}$, Strategy 2

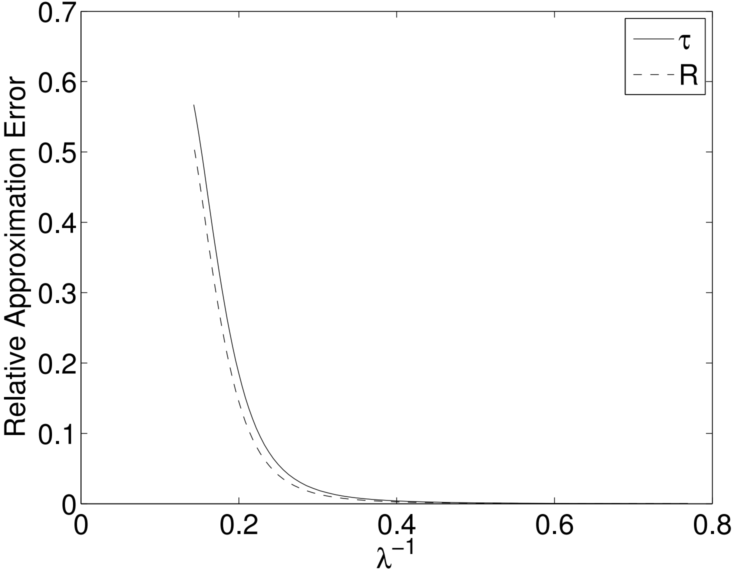


Figure: Relative approximation error for $\overline{R}$ and $\overline{\tau}$, Strategy 3

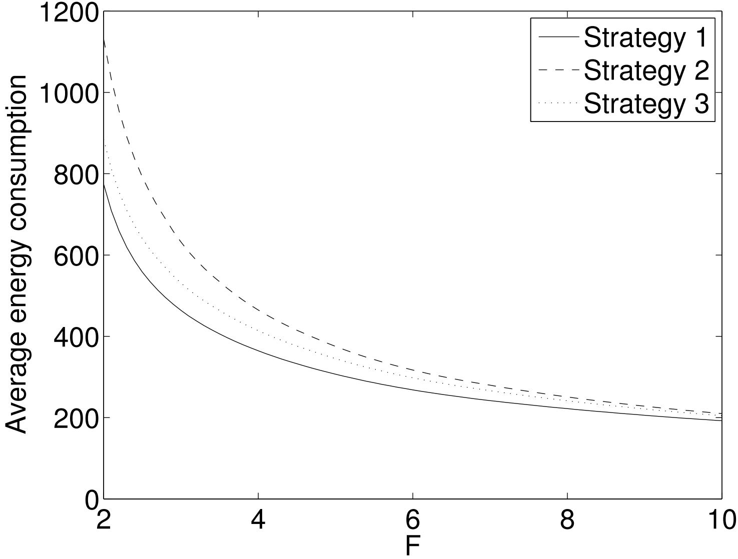


Figure: Computation of $\overline{R}$ for different strategies

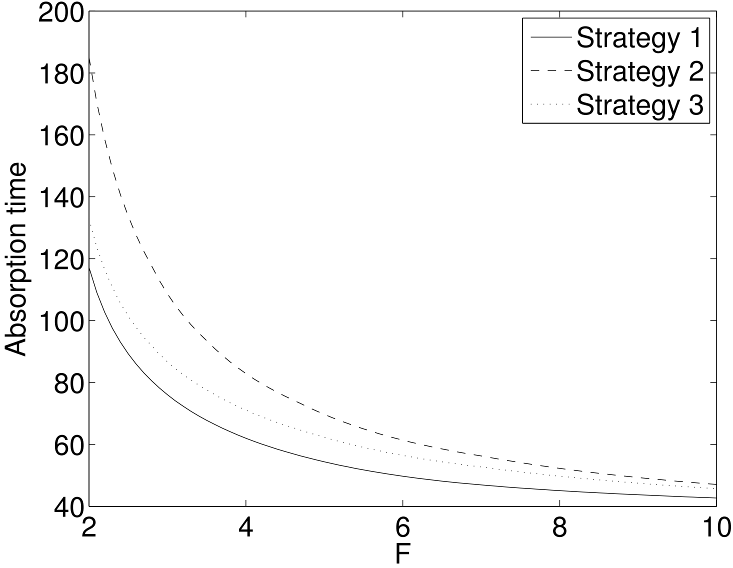


Figure: Computation of $\overline{\tau}$ for different strategies

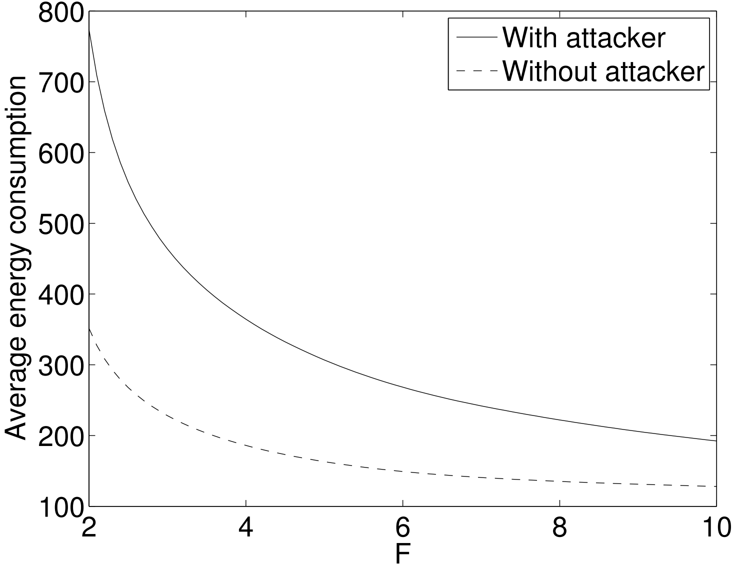


Figure: Comparison of $\overline{R}$ with or without attacker. Strategy 1

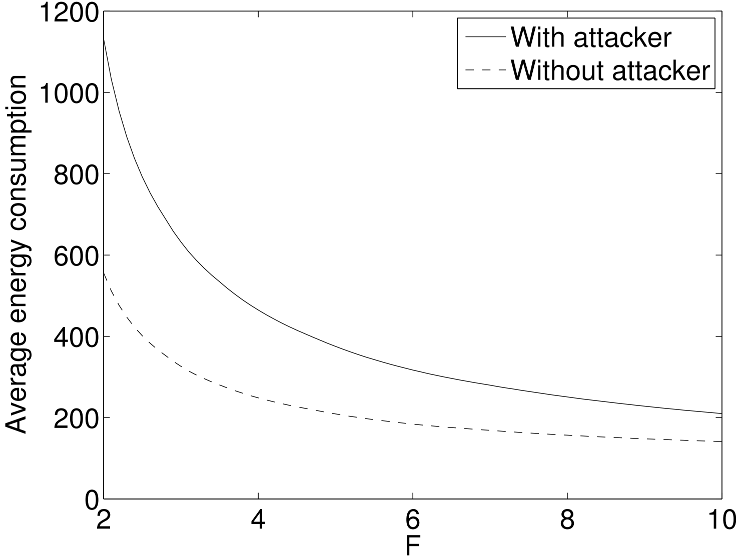


Figure: Comparison of $\overline{R}$ with or without attacker. Strategy 2

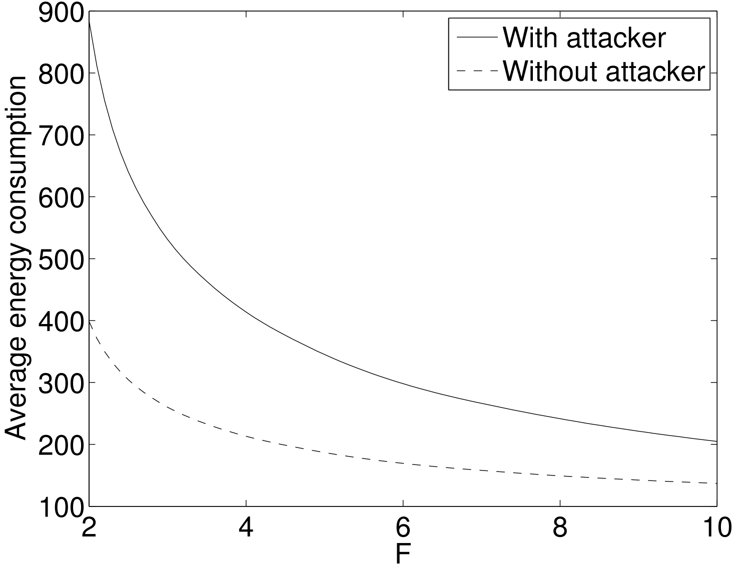


Figure: Comparison of $\overline{R}$ with or without attacker. Strategy 3

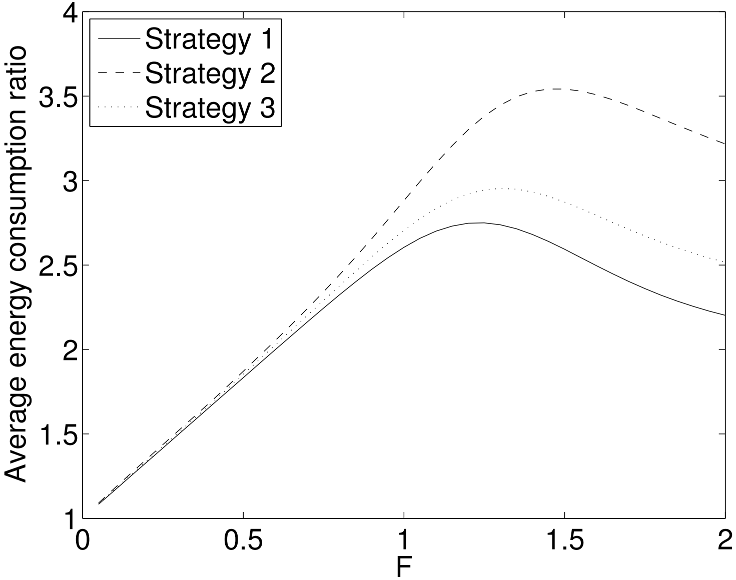


Figure: Ratio between values of $\overline{R}$ with and without attacker, $F \in (0, 2]$

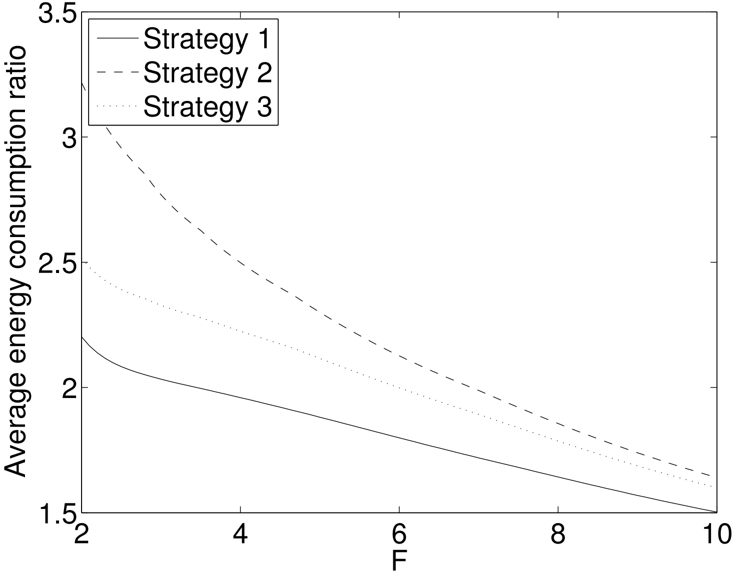


Figure: Ratio between values of $\overline{R}$ with and without attacker, $F \in [2, 10]$

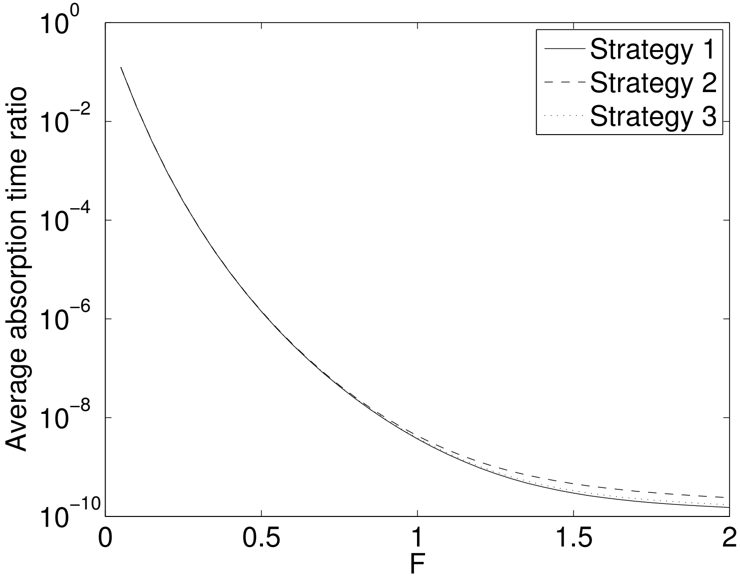


Figure: Ratio between values of $\overline{\tau}$ with and without attacker, $F \in (0, 2]$

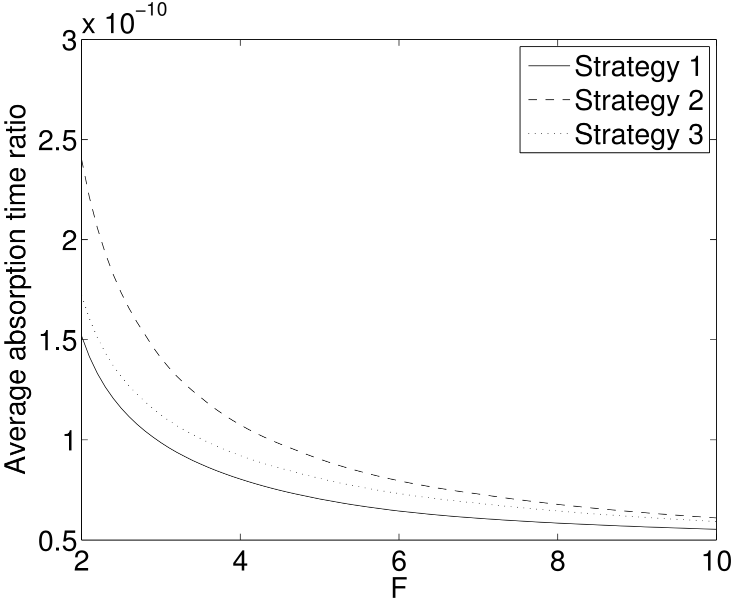


Figure: Ratio between values of $\overline{\tau}$ with and without attacker, $F \in [2, 10]$

Conclusions

- We proposed a Markovian model to study the impact of eDoS attacks to cloud infrastructures.
- We analysed 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.
- We found that low-aggressive strategies of the attackers are more dangerous for the cloud since the do not change significantly the life-time of the systems while they maintain a higher energy consumption.

Future works:

- give a more detailed model of the cloud infrastructure
- give a model for non-coordinated attackers performing a distributed eDoS
- perform a validation of the analysis on real data
- design a statistic approach to estimate the probability of being in presence of an eDoS attack in a cloud infrastructure
- ...

Thanks!

Thanks for your attention

⋮

(even if you slept during the whole presentation)

any question?