

A Model-Based Approach to Anomaly Detection Trading Detection Time and False Alarm Rate

Charles F. Gonçalves^{*†}, Daniel S. Menasché[§], Alberto Avritzer[‡], Nuno Antunes^{*}, Marco Vieira^{*}

^{*}University of Coimbra, CISUC, DEI, Portugal - {charles,nmsa,mviera}@dei.uc.pt

[†]Information Governance Secretariat, CEFET-MG, Brazil - charles@cefetmg.br

[‡]eSulab Solutions, Princeton, New Jersey - beto@esulabsolutions.com

[§]Federal University of Rio de Janeiro, Brazil - sadoc@dcc.ufrj.br

Abstract—The complexity and ubiquity of modern computing systems is a fertile ground for anomalies, including security and privacy breaches. In this paper, we propose a new methodology that addresses the practical challenges to implement anomaly detection approaches. Specifically, it is challenging to define normal behavior comprehensively and to acquire data on anomalies in diverse cloud environments. To tackle those challenges, we focus on anomaly detection approaches based on system performance signatures. In particular, performance signatures have the potential of detecting zero-day attacks, as those approaches are based on detecting performance deviations and do not require detailed knowledge of attack history. The proposed methodology leverages an analytical performance model and experimentation, and allows to control the rate of false positives in a principled manner. The methodology is evaluated using the TPCx-V workload, which was profiled during a set of executions using resource exhaustion anomalies that emulate the effects of anomalies affecting system performance. The proposed approach was able to successfully detect the anomalies, with a low number of false positives (precision 90%–98%).

Index Terms—anomaly detection, security, modeling, virtualization

I. INTRODUCTION

Complex computing systems, such as cloud solutions [1], are ubiquitous. Such ubiquity, in turn, is a potentially fertile ground for security and privacy breaches [2]–[5]. Efficiently detecting and mitigating such attacks is an important step to counter the threat that they pose to the existing IaaS systems and, more broadly, to the virtualization culture that supports a significant fraction of today’s systems [6].

The design of intrusion detection systems (IDSs) for detecting anomalies, such as zero-day attacks and advanced persistent threats (APTs) [7] in virtualized environments, poses several domain-specific challenges [8], [9]. In particular, (i) it is challenging to comprehensively define normal behavior in a diverse cloud environment, (ii) malicious attackers may adapt their behavior to fit the domain definition of “normal behavior”, and (iii) data availability on anomalies at cloud environments, which would be key for training, is hard to obtain [10], [11]. To tackle those challenges, we focus on anomaly detection approaches based on system performance signatures. In particular, **performance signatures have the potential of detecting zero-day attacks** [8], [9], as those approaches are based on detecting performance deviations and do not require detailed knowledge of attack history [12].

In this paper, we propose a methodology for anomaly detection based on performance deviations caused by anomalies in complex virtualized systems. **The proposed tuning of the**

anomaly detection mechanism leverages an analytical performance model and experimentation, and allows to control the rate of false positives in a principled manner [8]. After a careful analysis of every kind of transaction in the target system, the methodology profiles the system operation under normal conditions for its key transactions. Then, during system operations performance monitoring, performance deviations from the baseline are reported as anomalies.

To validate the proposed methodology, we ran an extensive experimental campaign using the TPCx-V workload [13], which is representative of a large virtualized infrastructure that supports a business that relies on transactional systems. Fault injection was used to emulate the effects of anomalies, e.g., due to attacks, impacting system performance. Experience and practice show that injecting the effects of faults and attacks is an effective way to check systems dependability [14], [15].

The experiments showed the applicability and effectiveness of the proposed anomaly detection methodology. In our experiments, it was possible to detect most of the performance deviations, with a low number of false positives (precision of 90% and 98% for the worst and best configurations). In addition, given the model-based nature of the solution, it is amenable to what-if analysis so as to trade between the rate of false positives and detection time.

In summary, the paper’s contributions are the following:

(i) An **analytical model to support anomaly detection designs**, which allows conducting principled parameterization. The model can be used to cope with the tradeoff between time to detect an anomaly and the rate of false alarms (Section III).

(ii) An **experimental assessment of the methodology in practice** using a representative system. We established the feasibility of detecting anomalies based on non-intrusive user-level performance metrics that are available in production environments (Sections IV and VI).

(iii) A **model-driven principled mechanism design** that allows revisiting the experimental results and conduct what-if analysis to assess different performance metrics of the considered anomaly detection algorithms as a function of the parameterization (Section V).

The remainder of the paper is organized as follows. Section II covers related work, followed by our contributions in Sections III–VI as indicated in the summary of contributions above. Finally, Section VII concludes the paper.

II. RELATED WORK

In this section, we revise related literature indicating how the current work relates to prior art.

A. Anomaly detection for cybersecurity

An approach for anomaly detection consists in running sequential hypothesis tests [16], [17]. In [17], sequential hypothesis tests are used for the detection of malicious port scanners. The authors have developed a link between the detection of malicious port scans and the theory of sequential hypothesis testing. They have also shown that port scanning can be modeled as a random walk. The detection algorithm matches the random walk to one of two stochastic processes, which correspond to malicious remote hosts or to authorized remote hosts. The approach considered in this paper is similar in spirit to that considered in [16], [17], as *our analytical results are derived from a birth-death Markov chain*. Such Markov chain can be interpreted as a random walk through buckets which fill as the system degrades, and empty as the system recovers (see Section III-C).

A number of previous works have considered anomaly detection approaches using performance signatures [18]–[21]. In [19] an approach for the mitigation of worm epidemics in tactical Mobile Ad-Hoc Networks (MANETs) using performance signatures (response time) and software rejuvenation was introduced. The work in [20] introduced a framework that detects anomalous application behavior using regression-based models and application performance signatures. Then, [21] builds on top of previous work on performance signatures [18]–[20] and proposes an anomaly detection approach based on performance signatures based on CPU, I/O, memory and network usage for the detection of security intrusions.

B. Bucket algorithm and sequential decision making

The performance of signature-based intrusion detection systems relies on intrusion detection algorithms that account for workload variability to avoid a high rate of false positive alerts. An example of such workload-sensitive algorithms is the Bucket Algorithm (BA) that was introduced in [18] and is presented in detail in Section III.

In Section III we revisit the BA mechanism, and present an analytical model that is instrumental to parameterize the BA from experimental data. A statistical analysis of the behavior of a family of BAs has been described on [22], without accounting for the tradeoff between detection time and false alarm rate. *One of the goals of this paper is to fill that gap*. In the previous cited research [18], [19], simulations were used to support the analysis of the BAs algorithms. In contrast, in this work we introduce an analytical model of the BAs algorithms that can be used to support anomaly detection designs, and an experimental assessment of the methodology in practice using a representative system.

III. ANOMALY DETECTION MECHANISM AND MODEL

In this section we describe the anomaly detection mechanism considered in this paper followed by the proposed analytical model.

A. Anomaly Detection Mechanism

The bucket algorithm for anomaly detection based on performance degradation works by continuously measuring the throughput, \bar{x} , and maintaining B buckets of depth D each. Samples are added to and removed from buckets as a function of the history of most recent throughput measurements, as shown in Fig. 1, wherein each ball corresponds to a throughput sample. The scalar value b is a pointer to the current bucket, $b = 1, \dots, B$ and d is the number of recent throughput samples stored in the current bucket, $d = 0, 1, \dots, D$.

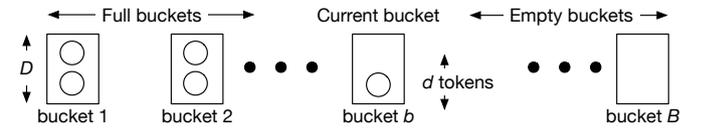


Fig. 1. System of buckets representing the dynamics of the anomaly detection algorithm. D and B must be properly parameterized for adequate operation.

Let μ be the baseline average throughput, and σ be the baseline standard deviation. Both μ and σ can be derived from the execution of controlled experiments without anomalies (i.e. **golden runs**). The pointer b to the current bucket is used to determine the current target throughput, which is given by $\bar{x} = \mu - (b - 1)\sigma$. Once the current bucket overflows (resp., underflows), the target throughput is shifted upward (resp., downward) by one standard deviation. When all buckets overflow the algorithm detects a performance degradation and triggers an anomaly alarm. The performance degradation detection algorithm, that we will refer hereinafter as Bucket Algorithm (BA), is given as follows:

Initialization: $\{b \leftarrow 1; d \leftarrow 0\}$, with all buckets empty.

Main loop: for each sample \hat{x} of throughput, execute the steps below.

- 1) if $\hat{x} < (\mu - (b - 1)\sigma)$ then $\{d \leftarrow d + 1\}$, throughput smaller than reference value, add token to current bucket;
- 2) else do $\{d \leftarrow d - 1\}$, throughput larger or equal than reference value, remove token from the current bucket;
- 3) if $(d > D)$ then do $\{d \leftarrow 0; b \leftarrow b + 1\}$, current bucket overflow, go to next bucket;
- 4) if $((d < 0)$ and $(b > 1))$ then do $\{d \leftarrow D; b \leftarrow b - 1\}$, current bucket underflow, go to previous bucket;
- 5) if $((d < 0)$ and $(b == 1))$ then do $\{d \leftarrow 0\}$ all buckets empty, system recovered from transient performance degradation;
- 6) if $b > B$, all buckets overflow, trigger performance degradation alarm.

The performance degradation detection algorithm can be tuned by varying the bucket depth, D , and the number of buckets, B . The larger the product $D \times B$ the smaller the rate of false alarms but the longer it takes for the algorithm to detect the performance degradation.

B. Hypothesis Testing

The system administrator continuously considers two alternative hypothesis: (i) null hypothesis H_0 corresponding to a situation where there is no anomaly taking place and (ii) alternative hypothesis H_1 meaning that there is an anomaly,

e.g., the system is under attack. Then, the key quantities of interest can be defined as a function of H_0 and H_1 . To simplify presentation, in what follows time is measured in number of collected samples.

Definition 1: The mean time until a false alarm under H_0 is denoted by $A_B(D)$.

As discussed in the following section, $A_B(D)$ is given by the mean time to reach the absorbing state of a Markov chain characterizing the bucket algorithm. When $B = 2$, we provide closed-form expressions for $A_B(D)$.

Definition 2: A lower bound on the number of samples until a true positive under H_1 is denoted by L . Assuming all buckets are initially empty, we let $L = BD$.

Definition 3: The probability of false alarm under H_0 is the probability that an alarm is triggered outside an anomaly, $f_B(D) = \mathbb{P}(R < T)$, where R is a random variable with mean $A_B(D)$ characterizing the time until an alarm is triggered, and T is a random variable with mean $1/\alpha$ characterizing the time until an anomaly occurs.

In this paper, except otherwise noted we assume that $f_B(D)$ depends on R and T only through their means.

Definition 4: The expected cost of a given system parameterization is a weighted sum of the probability of false alarms, computed under H_0 , and a lower bound on the number of samples to detect an anomaly, computed under H_1 ,

$$C(\mathbf{p}, w, D, B, \alpha) = BD + wf_B(D). \quad (1)$$

Table I summarizes the notation introduced in this section. Additional details about how to estimate $A_B(D)$ and $f_B(D)$ are provided in Sections III-C and III-D, respectively. Then, the cost function (1) (Definition 4) will be instrumental to parameterize the bucket algorithm in Section III-E.

TABLE I
TABLE OF NOTATION

variable	description
B	number of buckets
b	current bucket, $b = 1, \dots, B$
D	maximum bucket depth
d	current depth of bucket b , $d = 0, \dots, D$
$A_B(D)$	mean time to false alarm, under H_0 (no anomaly), i.e.: mean number of collected samples to reach absorbing state
$f_B(D)$	probability of false alarm
F	target probability of false alarm
α	anomaly rate
p_i	probability that sample adds ball to bucket, when $b = i$

C. Analytical Model

Simple algorithms to detect anomalies, such as the bucket algorithm, can be tuned using first principles. The larger the depth of the bucket, the lower the false alarm probability, but the longer it takes for a true positive to be identified. To simplify the analysis, we **work under the assumption** that anomalies will change the throughput distribution, and will always be detected. However, the number of samples to detect the anomaly may vary depending on the depth of the bucket. Our **second key simplifying assumption** is that the number of samples to detect the anomaly is much smaller than the number of samples collected before getting a false alarm. *The two assumptions above are mild, and should typically hold in real settings* as the time until a false alarm in practical systems

should be much longer on average than the time until a true positive [7], [19]. Then, we aim at answering the following question: what is the smallest bucket depth to produce a false alarm probability upper bounded by a given threshold?

Next, we introduce a discrete time birth-death Markov chain (DTMC) to characterize the behavior of the BA. State (b, d) of the Markov chain corresponds to the setup wherein there are d balls in bucket b , and D balls in buckets $b - 1, \dots, 1$.

Each transition of the DTMC corresponds to the collection of a new sample. Such sample causes the system to transition from state (b, d) to one of its two neighboring states. Let p_i be the probability that the number of balls at bucket i increases after a new sample is collected. Then, $p_i = \mathbb{P}(\hat{x} > \bar{x} + (i - 1)\sigma)$, for $1 \leq i \leq B$. The entries of the transition probability matrix are readily obtained from Fig. 2.

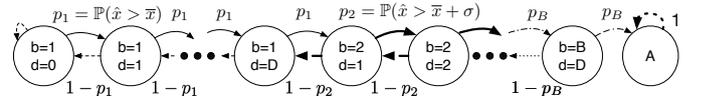


Fig. 2. Discrete time Markov chain characterizing the behavior of the BA. Each transition corresponds to the collection of a new sample.

Once the terminal absorbing state is reached an alarm is triggered (state A in Figure 2). The number of samples collected until absorption accounts for a tradeoff between the mean time until (a) a false alarm, in the absence of anomalies, and (b) detection, in the presence of an anomaly. Larger values of bucket depth D favor the reduction of the former but increase the latter.

Let $\tilde{A}_B(D; p_1, p_2)$ be the time until absorption, measured in number of collected samples, accounting for B buckets of depth D each. We denote its mean by A_B , $\mathbb{E}(\tilde{A}_B) = A_B$. Under the hypothesis of no anomaly, \tilde{A}_B is the time to a false alarm. We derived in [23] a closed-form expression for A_B , which is instrumental to handle tradeoffs in the choice of the bucket depth D as illustrated in the upcoming sections. In particular, for $B = 2$, the resulting expression is given by

$$A_2(D; p_1, p_2) = A_2^{(1)}(D; p_1, p_2) + A_2^{(2)}(D; p_1, p_2) \quad (2)$$

where $A_B^{(i)}$ is the mean time to fill the i -th out of B buckets,

$$A_2^{(1)} = \Delta_1 (\delta_1 - D) \quad (3)$$

$$A_2^{(2)} = \Delta_2 (\delta_2 - D) + \Delta_1 \left(\frac{1 - \rho_1^{D+1}}{\rho_1^{D+1}} \right) \delta_2 \quad (4)$$

and

$$\rho_i = \frac{1}{p_i} - 1, \quad \Delta_i = \frac{1 + \rho_i}{1 - \rho_i}, \quad \delta_i = \frac{1 - \rho_i^{-D}}{\rho_i - 1}. \quad (5)$$

Our experimental results indicate that $B = 2$ suffices in the considered scenarios (see Section IV). For this reason, in the remainder of this paper all numerical results derived from the proposed analytical model are reported letting $B = 2$, making use of equations (2)-(5). In what follows, we illustrate how to leverage the proposed model to estimate the probability of false alarms.

D. Modeling the Probability of False Alarms

Next, we leverage the proposed model to estimate the probability of false alarms. To that aim, we assume that anomalies, e.g., due to attacks, arrive according to a Poisson process with rate α . Recall that $f_B(D)$ denotes the probability of a false alarm (Definition 3). In what follows, we derive expressions for $f_B(D)$ under different assumptions on the distribution of $\tilde{A}_B(D)$.

Assuming that $\tilde{A}_B(D)$ can be roughly approximated by a constant, and that the time between anomalies is exponentially distributed with mean $1/\alpha$,

$$f_B(D) = e^{-A_B(D)\alpha}. \quad (6)$$

Alternatively, if we approximate $\tilde{A}_B(D)$ by an exponential distribution,

$$f_B(D) = \frac{1/A_B(D)}{1/A_B(D) + \alpha} = \frac{1}{1 + A_B(D)\alpha}. \quad (7)$$

In the expressions above, we made the dependence of f_B and A_B on the bucket depth D explicit as one of our goals is to study the relationship between D , f_B and A_B . The closed-form equations (6) and (7) are instrumental to get insights about the interplay between the different model parameters. In particular, as D increases A_B increases and f_B decreases (Definition 1), but the time to detect an anomaly increases (Definition 2). As indicated in the sequel, the equations above allow us to find the minimum D such that $f_B(D)$ is below a given threshold. Then, in Section IV we experimentally validate that the values of D obtained through the proposed model produce the desired probability of false alarms in realistic settings.

E. Parameterization of the Anomaly Detection Mechanism: a Model-Driven Optimization Approach

Next, we show how to use the proposed model and the obtained expressions of probability of false alarm for the purposes of running statistical hypothesis tests to determine whether there is an ongoing anomaly in the system.

Given a target false alarm probability, denoted by F , the system administrator goal is to determine the optimal number of buckets and bucket depth so as to minimize the lower bound on number of samples to detect an anomaly, L , while still meeting the target false alarm probability.

PROBLEM WITH HARD CONSTRAINTS:

$$\min L = BD \quad (8)$$

$$\text{subject to } f_B(D) \leq F \quad (9)$$

In what follows, we assume that B is fixed and given. Then, as $f_B(D)$ is strictly decreasing with respect to D , the constraint above will be always active and the problem translates into finding the minimum value of D satisfying the constraint. The problem above is similar in spirit to a Neyman-Pearson hypothesis test, for which similar considerations apply, i.e., the optimal parameterization of the test is the one that satisfies a constraint on the false alarm probability.

Alternatively, the problem above can be formulated through the corresponding Lagrangian,

PROBLEM WITH SOFT CONSTRAINTS:

$$\min \mathcal{L}(D) = BD + w(f_B(D) - F) \quad (10)$$

where w is the Lagrange multiplier. The Lagrangian naturally leads to an alternative formulation of the problem, wherein the hard constraint in (9) is replaced by a soft constraint corresponding to the penalty term $f_B(D) - F$ present in the cost Lagrangian. The Lagrangian is a cost function, motivating Definition 4. Note that as wF is a constant, minimizing (10) is equivalent to minimizing (1).

IV. EXPERIMENTAL SETUP AND FAULT MODEL

To illustrate and validate the methodology presented in Section III, we ran an experimental campaign using the TPC Express Benchmark V [13] (TPC_x-V), as described in Sections IV-A and IV-B. A fault injection approach was used to emulate the effects of performance affecting security intrusions, as described in Section IV-C. Then, the model-based calibration of the anomaly detector is reported in Section V.

A. System Under Test

The TPC_x-V is a publicly available, end-to-end benchmark for data-centric workload on virtual servers. The benchmark kit provides the specification, implementation, and tools to audit and run the benchmark. Details can be found in [24], [25]. TPC_x-V models many features commonly present in cloud computing environments such as multiple Virtual Machines (VMs) running at different load demand levels, and significant fluctuations in the load level of each VM [24]. We use the workload and software provided by the TPC_x-V [25] to emulate a context closely related to a real-world scenario of brokerage firms that must manage customer accounts, execute customer trade orders, and be responsible for the interactions of customers with financial markets.

The goal of TPC_x-V is to measure how a virtualized server runs database workloads, using them to measure the performance of virtualized platforms, specifically the hypervisor, the server hardware, storage, and networking. The minimal deployment of the TPC_x-V comprises four groups of three VMs, representing four different subsystems. Table II summarizes the considered experimental setup. In Table II, tpc- gn refers to a VM of group n . Each group was defined according to the benchmark recommendations [13].

The TPC_x-V workload is made up of **12 types of transactions** that are submitted for processing at multiple databases (market, customer, and broker) following a specified mix of transactions per load phase. A typical run consists of **10 distinct load phases of 12 minutes each**. Transactions simulate the stock trade process. When a trade finishes, a transaction named *Trade-Result* is issued. The primary performance metric for the benchmark is the business throughput (tps_v). It represents the number of completed *Trade-Result* transactions per second.

TABLE II
EXPERIMENTAL SETUP

VM	MB	vCPU	VM	MB	vCPU	VM	MB	vCPU
tpc-g1a	1024	1	tpc-g1b2	4096	4	tpc-g1b1	8192	2
tpc-g2a	1024	1	tpc-g2b1	12288	2	tpc-g2b2	6144	6
tpc-g3a	1024	1	tpc-g3b1	16384	2	tpc-g3b2	8192	8
tpc-g4a	1024	2	tpc-g4b1	20400	2	tpc-g4b2	10240	8
tenant A	1532	2	tenant B	1532	2	tenant C	1532	2
dom0	1962	4	tpc-driver	1882	2			

B. Experimental Setup

Our *experimental setup* is a deployment of the TPC_X-V over two physical servers. The first server is a Dell PowerEdge R710 with 24 Cores, 96 GB RAM, and 12 TB disk, and is managed by a Xen hypervisor (4.4.1). It has a privileged domain (dom0) with a dedicated VM, and 15 additional VMs. One set of 12 VMs is dedicated to TPC_X-V , with four groups of three VMs each (*tpc-xxx*). Another set of 3 VMs corresponds to an additional group running the compromised system. The second server (2 Cores, 8GB RAM and 1 TB disk) runs the same software and a single VM used for the driver component as prescribed on the TPC_X-V specification. The details of each VM are described in Table II.

C. Tools Set and Fault Model

Each **single experiment lasts roughly 4 hours** comprising a minimum of 2h as demanded by the benchmark specification and additional 2h to track the restoration of the full environment to its initial state after performance degradation. Initial state restoration is achieved by rebooting the servers and recovering the system and databases, including the restoration of all virtual disks.

Next, we describe the considered fault model. Our goal is to account for resource exhaustion faults [26]. To that aim, we make use of the Stress-NG module [27], which was designed to “exercise various physical subsystems of a computer as well as the various operating system kernel interfaces”. In particular, we have defined three reference configurations to emulate resource exhaustion anomalies, e.g., due to attacks.

(H): A high intensity workload consists of I/O intensive tasks executed by eight parallel processes for 300 seconds.

(L): A low intensity workload comprises the execution of ten intervals, where each interval consists of 15 seconds of I/O intensive tasks executed by two parallel processes followed by 15 seconds in idle mode.

(Ls): A shorter low intensity workload is similar to a low intensity workload configuration except that it comprises the execution of three intervals.

Recall from Section IV-A that TPC_X-V comprises 10 load phases with diverse load demands. We emulated resource exhaustion anomalies on phases 4 and 6, which correspond to phases wherein the usage of physical resources and the reference tps_V reach their maximum values, respectively. Combining the observations above, we have a total of six fault models, which we will refer to using the corresponding phase number followed by the reference configuration: 4H, 4L, 4Ls, 6H, 6L, and 6Ls. An execution of the system without fault emulation is referred to as a *golden run*. The model

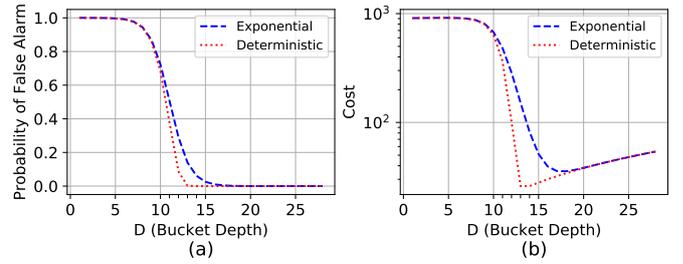


Fig. 3. Model based calibration of anomaly detector: (a) probability of false alarm and (b) cost given by Equation (1) as a function of bucket depth.

based calibration of the anomaly detector introduced in the upcoming section is solely based on *golden runs*, whereas the results presented in Section VI leverage the calibrated anomaly detector together with the fault model introduced above.

V. MODEL BASED CALIBRATION OF ANOMALY DETECTOR AND COUNTERFACTUAL ANALYSIS

This section provides insights on the experimental results using the proposed model. Our goals are to (a) illustrate the applicability of the model in the real experimental setting described in the previous section; and (b) indicate how the model can be used to trade between contending aspects such as false positive rate and time to detect anomalies.

To exemplify the general process, we focus on one of the twelve transactions referred to in Section IV-A, namely the TRADE_LOOKUP transaction, for which we identify that $p_1 = 0.46$ and $p_2 = 0.71$ considering the emulation of the system without anomalies (i.e., considering golden runs). We assess the expected number of samples until a false alarm, obtained from (2), letting $B = 2$ and D vary between 1 and 30. For $D = 15$, for instance, the number of samples until a false alarm as estimated by the model already surpasses 10^7 .

Fig. 3(a) shows the probability of false alarm as a function of the bucket depth. Fig. 3(a) accounts for a fault model wherein the mean time between anomalies is $1/\alpha = 5 \times 10^5$ samples. The dashed (resp., dotted) line corresponds to the exponential (resp., deterministic) approximation for the time between anomalies, corresponding to (7) (resp., (6)). As the bucket depth increases, the probability of false alarm decreases. For $D \geq 12$, the probability of false alarm is close to zero.

As discussed in Section III, there is a tradeoff between the probability of false alarm and the time to detect anomalies once they occur. To cope with such a tradeoff, we consider both approaches introduced in Section III-E, namely the hard and soft constraint problems. Under the hard constraint problem, a target probability of false alarm is determined, and the minimum value of D that satisfies such target is sought. For instance, if we set $F = 0.03$ in (9) then the minimum values of D satisfying the constraint are $D = 15$ and $D = 13$ under the exponential and deterministic fault models, respectively.

Next, we assess the cost $C(p, w, D, B, \alpha)$ introduced in Definition 4. Figure 3(b) shows how the cost varies as a function of D , letting $B = 2$, $p_1 = 0.46$, $p_2 = 0.71$ and $\alpha = 2 \times 10^{-6}$. To generate the plots, we let $w = 909$, which corresponds to the Lagrange multiplier of the constrained

problem under the deterministic model (see also (10)). In that setting, the optimal bucket depth equals $D = 13$ (see dotted line in Figure 3(b)), which is in agreement with the result presented in the previous paragraph. Note that for the exponential model the minimum cost is attained at $D = 18$ (dashed line in Figure 3(b)), which is slightly larger than $D = 15$ obtained in the previous paragraph. This is because the Lagrange multiplier corresponding to the exponential model is $w = 57$. Using such a smaller weight favors a reduction in the optimal bucket depth to $D = 15$, again in agreement with the results discussed in the previous paragraph.

Take away message and engineering implications: the analysis presented in this section is instrumental to perform what-if counterfactual analysis and execute utility-driven model parameterization. If the system administrator implements global countermeasures against attacks, for instance, it is expected that the rate of anomalies will decrease. In that case, the bucket depth can be adjusted accordingly using the approach introduced above.

The results presented in this section are also instrumental to reverse engineer the utility function subsumed by existing systems. As indicated in the following section, letting D vary between 12 and 15 performed well in the considered real scenarios. The analysis presented above shows that a system operating with $B = 2$ and $D = 15$ is optimal, for instance, in case 5×10^5 samples are collected inbetween anomalies and $w = 57$, leading to a false positive probability of roughly 0.03. Knowing that this is the case, one can tune the utility function, e.g., to account for anomalies that occur at different rates, and verify when/if the parameters of the bucket algorithm should be adjusted.

VI. EXPERIMENTAL ASSESSMENT OF THE CALIBRATED ANOMALY DETECTOR IN FACE OF FAULTS

In what follows, we provide additional experimental evidence of the effects of the bucket depth D on different system metrics. Motivated by the model-based analysis presented in Section V, we focus most of our attention on values of maximum bucket depth D varying between 12 and 15. Our goals are to (a) analyze the residual effects that can arise after the anomalies; and (b) assess the effectiveness of the proposed anomaly detection approach through two case studies and accounting for standard performance metrics as detailed next.

Our results are discussed using three metrics widely adopted in classification assessment [28], namely precision, recall and F-measure, which are defined as a function of true positives (TP), false positives (FP) and false negatives (FN) as follows,

$$\text{Pr} = \frac{TP}{TP + FP}, \quad \text{Re} = \frac{TP}{TP + FN}, \quad \text{F1} = \frac{2 \times \text{Pr} \times \text{Re}}{\text{Pr} + \text{Re}}.$$

Precision (Pr) measures the impact of FP in the method's positive prediction. Recall (Re) reflects the sensitiveness of the algorithm, capturing the fraction of correct predictions. F-measure (F1) is the harmonic mean of precision and recall.

A. Residual Effects and Survivability Analysis

In our experiments we observed that the number of bucket overflows after an attack ends was significantly greater than in other non-attack periods. This phenomenon is typically studied

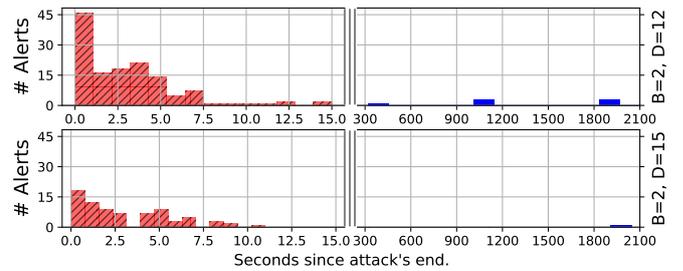


Fig. 4. Post-attack alerts distribution for bucket configuration with $B=2$ and $D=[12,15]$. Note that x scale is cropped to improve readability.

in the realm of survivability analysis, which focuses on system behavior from failure up to full recovery, including transient performance degradation [29], [30]. In this section, we refer to the event of an overflow of the B -th bucket as an *alert*, and explicitly distinguish alerts from alarms. Intuitively, after an alarm is triggered during an attack, the set of alerts caused by transient effects should not trigger additional alarms.

Fig. 4 shows the number of alerts as a function of time. Following the survivability perspective, time is measured in seconds after an attack ends. Note that a significant fraction of alerts occurs a few seconds after the attack, which suggests that those alerts are due to residual effects of attacks.

Next, we propose a simple heuristic to determine when a set of alerts should be aggregated into a single alarm. To that aim, let δ be the meantime until the first alarm is triggered during an attack phase. Table III reports how δ varies as a function of D for two of the twelve transactions considered in our workload. The values of δ are relatively stable across transactions and bucket depths. Note that δ increases as D grows, up to $D = 12$. Correspondingly, as D grows the overall number of alerts decreases (see Definition 2). Together, Figure 4 and Table III suggest the following heuristic: after an alert at time t_0 , any additional alerts during the interval $[t_0, t_0 + c\delta]$ are due to residual effects (e.g., emptying of queues and recovery of error states) and are discarded. In our experiments we let $c = 3$. It is also worth noting the slight decrease in δ when D varies from 12 to 15. We are currently investigating such a decrease, noting that it may not be statistically significant as the number of alerts decreases from 78 and 88 to 65 and 71, respectively, when D varies from 12 to 15 for the two transactions in Table III. Together, Figure 4 and Table III suggest the following heuristic: any alert that occurs at time $t < \delta$ is due to residual effects (e.g., emptying of queues and recovery of error states).

Fig. 5 shows the number of residual alerts as a function of D , for the six fault models introduced in Section IV-C. The proposed heuristic yields residual alerts under high intensity faults (6H and 4H, corresponding to blue dashed lines and red dotted lines). In addition, the number of residual alerts decreases as D grows, as the the larger the value of D the

TABLE III
MEAN TIME TO FIRST ALARM (δ) DURING THE ANOMALY INJECTION (IN SECONDS)

Transaction	D=6	D=9	D=12	D=15
TRADE_LOOKUP	31.03	50.06	69.18	61.63
MARKET_WATCH	36.08	54.11	60.38	59.51

higher is the tolerance for transient faults.

Under the scenarios considered in this paper, the proposed heuristic *accurately classified all alerts due to residual effects as spurious, and did not misclassify any non-residual alert and the heuristic is subsumed under all the reported results in the sequel.*

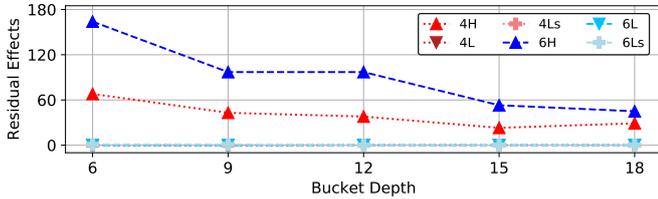


Fig. 5. Distribution of the residual effects by failure mode and bucket depth.

B. Experimental Assessment of Parametrization Impact on Performance

Next, our goals are to (a) assess the performance of the *BA* over the six considered fault models and (b) indicate how its parametrization affects performance.

We consider the same bucket depth D across all operations. Anomalies are detected per transaction and per VM group. Noting that 9 out of the 12 transactions referred to in Section IV-A turned out to be representative, and that we have 4 groups of VMs (4 first rows of Table II), the bucket algorithm counts with 9×4 sets of buckets, one set for each transaction-group pair. Each set of buckets comprises $B = 2$ buckets.

According to the cost function and the fault model considered in Section V, the optimal bucket depth D resides between 12 and 15. Table IV reports the performance metrics obtained from our experimental campaign, divided into two groups corresponding to $D = 12$ and $D = 15$. The first line of each group accounts for *all* fault models, and the subsequent six lines correspond to the six fault models described in Section IV-C.

TABLE IV
PERFORMANCE METRICS, INCLUDING RESIDUAL EFFECTS COUNTS (RE),
PRECISION, RECALL AND F-MEASURE (F1)

		<i>TP</i>	<i>FN</i>	<i>FP</i>	<i>RE</i>	<i>Pr</i>	<i>Re</i>	<i>F1</i>
$B = 2$ & $D = 12$	All	108	18	12	135	0.90	0.86	0.88
	4H	20	1	0	38	1.00	0.95	0.98
	4L	21	0	3	0	0.88	1.00	0.93
	4Ls	9	12	1	0	0.90	0.43	0.58
	6H	21	0	2	97	0.91	1.00	0.95
	6L	21	0	4	0	0.84	1.00	0.91
	6Ls	16	5	2	0	0.89	0.76	0.82
$B = 2$ & $D = 15$	All	82	44	2	76	0.98	0.65	0.78
	4H	20	1	0	23	1.00	0.95	0.98
	4L	16	5	0	0	1.00	0.76	0.86
	4Ls	1	20	1	0	0.50	0.05	0.09
	6H	21	0	0	53	1.00	1.00	1.00
	6L	17	4	1	0	0.94	0.81	0.87
	6Ls	7	14	0	0	1.00	0.33	0.50

Table IV indicates that the proposed parametrization indeed yields a small number of false alarms (column *FP*), as suggested by the analytical model, and that the F-measure is typically greater than 0.78 (two notable exceptions being

under fault models 4Ls and 6Ls). In addition, Table IV also shows that *BA* performance varies as a function of the anomaly intensity and algorithm configuration. In particular, letting $D = 12$ or $D = 15$ the algorithm is effective to detect high intensity anomalies (4H and 6H) and low intensity anomalies of long duration (4L and 6L). In those scenarios, the performance of the anomaly detector under $D = 12$ and $D = 15$ is similar, suggesting robustness of the solution with respect to its parametrization.

To deal with short anomalies of low-intensity (4Ls and 6Ls), we found that D must be fine tuned. In the 6Ls scenario, setting $D = 12$ is key to control the number of false negatives. Indeed, $D = 12$ produces an F-measure of 0.82 which significantly outperforms an F-measure of 0.5 under $D = 15$. In the most challenging setup 4Ls, we must vary D in a broader range beyond 12 and 15 to detect short bursts of faults. Such observation, in turn, motivates a transaction-based parametrization of the anomaly detector in those settings.

In this work we consider the same value of D for all transactions in order to show the effectiveness of the proposed approach in its simplest configuration. Our preliminary results (not shown in this paper) indicate that allowing a distinct parametrization per-transaction suffices to deal with scenarios such as 4Ls. The decision between tuning the parameters in a system wide manner or in a per-transaction basis must balance between simplicity and effectiveness, and we leave its detailed experimental analysis as subject for future work.

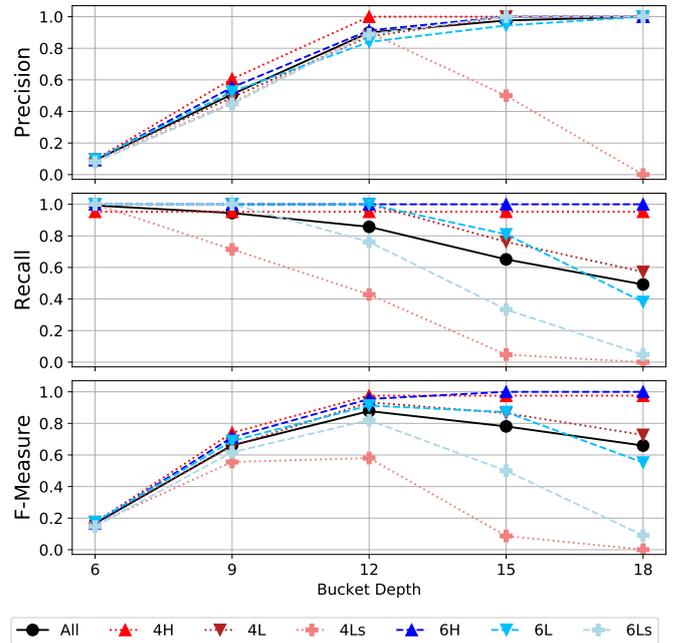


Fig. 6. Performance metrics under the six fault models.

Fig. 6 shows how precision, recall and F-measure vary as a function of D , for D varying between 6 and 18. Larger values of D favor higher tolerance to performance variability under normal conditions. Therefore, the number of *false positives* (FPs) decreases and precision increases as D grows. Recall (*Re*), in contrast, decreases and the number of *false negatives* (FNs) increases as D grows, as such growth produces longer detection times. The F-measure balances precision and recall,

typically reaching its maximum value between $D = 12$ and $D = 15$ under the considered setups, which is in agreement with the results obtained through the analytical model parametrization in Section V.

C. Take away message and engineering implications

Whereas the model-based parametrization from Section V is based on the cost function (1) to be minimized (Figure 3(b)), the F-measure yields an utility function to be maximized (Figure 6). In essence, *both the cost and utility functions capture the fundamental tradeoff between detection time and false alarm rates* and are complementary to each other. While the experimental approach serves for validation purposes and to explain system behavior in retrospect, after experiments are executed, the model-based approach is key to perform what-if counterfactual analysis and for predictive purposes.

VII. CONCLUSION

In this work, we presented a methodology for anomaly detection based on performance degradation, e.g., caused by security attacks at complex virtualized systems. The approach leverages an analytical model to find the optimal parametrization of an anomaly detector in a principled way.

Our experimental assessment indicates the method's effectiveness by injecting resource exhaustion attacks in a virtualized system. Results show that it is possible to detect anomalous behavior using the throughput of the business transactions with an average precision of 90% and recall of 86%. Our experimental results also bring awareness about residual effects of high-intensity fault loads, which may persist after active attacks have been interrupted.

We believe that the analytical and experimental contributions presented in this work advance the state of the art providing novel perspectives towards the classical and fundamental tradeoff between detection time and false alarm rates faced in the optimal design of anomaly detection mechanisms.

For future research, we intend to extend our experiments with fault models representing other types of attacks, and to cope with a transaction-oriented system parametrization.

ACKNOWLEDGMENT

This work was funded by CEFET-MG/Brazil, eSulab Solutions, CAPES, CNPq and FAPERJ, and by the Portuguese Foundation for Science and Technology (FCT) through the Ph.D. grant SFRH/BD/144839/2019, and the project METRICS (agreement no POCI-01-0145-FEDER-032504), within the scope of the project CISUC - UID/CEC/00326/2020 and by European Social Fund, through the Regional Operational Program Centro 2020. It was also supported SPEC RG Security Benchmarking (Standard Performance Evaluation Corporation; <http://www.spec.org>, <http://research.spec.org>).

REFERENCES

- [1] "83% of enterprise workloads will be in the cloud by 2020." [Online]. Available: <https://tinyurl.com/forbescloud2018>
- [2] Intel, "Unexpected page fault in virtualized environment advisory," 2019, <https://tinyurl.com/intelfault>.
- [3] DigitalOcean, "DigitalOcean reply to Intel security advisory," 2019, <https://hup.hu/index.php/node/166970>.
- [4] M. Wallschläger, A. Gulenko, F. Schmidt, O. Kao, and F. Liu, "Automated anomaly detection in virtualized services using deep packet inspection," *Procedia Computer Science*, vol. 110, pp. 510–515, 2017.
- [5] A. Gulenko, M. Wallschläger, F. Schmidt, O. Kao, and F. Liu, "Evaluating machine learning algorithms for anomaly detection in clouds," in *IEEE Conference on Big Data (Big Data)*, 2016, pp. 2716–2721.
- [6] B. Hayes, "Cloud computing," *Communications of the ACM*, vol. 51, no. 7, pp. 9–11, 2008.
- [7] M. Grottko, A. Avritzer, D. S. Menasché, L. P. de Aguiar, and E. Altman, "On the efficiency of sampling and countermeasures to critical-infrastructure-targeted malware campaigns," *SIGMETRICS Performance Evaluation Review*, vol. 43, no. 4, pp. 33–42, 2016.
- [8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.
- [9] A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer, and B. D. Payne, "Evaluating computer intrusion detection systems: A survey of common practices," *ACM Computing Surveys (CSUR)*, vol. 48, no. 1, p. 12, 2015.
- [10] F. Erlacher and F. Dressler, "Fixids: A high-speed signature-based flow intrusion detection system," in *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2018, pp. 1–8.
- [11] —, "Testing IDS using GENESIDS: Realistic mixed traffic generation for IDS evaluation," in *Proceedings of the ACM SIGCOMM 2018 Conference on Posters and Demos*, 2018, pp. 153–155.
- [12] G. Fernandes Jr, J. J. P. C. Rodrigues, L. Carvalho, J. F. Al-Muhtadi, and M. Proença Jr, "A comprehensive survey on network anomaly detection," *Telecommunication Systems*, vol. 70, pp. 447–489, 2019.
- [13] *TPC Express Benchmark™ V (TPCx-V) Specification*, Transaction Performance Council (TPC), 04 2019.
- [14] J. Arlat, M. Aguera, L. Amat, Y. Crouzet, J.-C. Fabre, J.-C. Laprie, E. Martins, and D. Powell, "Fault injection for dependability validation: A methodology and some applications," *IEEE Transactions on software engineering*, vol. 16, no. 2, pp. 166–182, 1990.
- [15] J. Arlat, Y. Crouzet, J. Karlsson, P. Folkesson, E. Fuchs, and G. H. Leber, "Comparison of physical and software-implemented fault injection techniques," *IEEE Transactions on Computers*, vol. 52, no. 9, pp. 1115–1133, 2003.
- [16] J.-W. Ho, M. Wright, and S. Das, "Fast detection of mobile replica node attacks in wireless sensor networks using sequential hypothesis testing," *IEEE Trans. Mobile Computing*, vol. 10, no. 6, pp. 767–782, 2011.
- [17] J. Jung, V. Paxson, A. W. Berger, and H. Balakrishnan, "Fast portscan detection using sequential hypothesis testing," in *IEEE Symposium on Security and Privacy, 2004*. IEEE, 2004, pp. 211–225.
- [18] A. Avritzer, A. Bondi, and E. J. Weyuker, "Ensuring stable performance for systems that degrade," in *International Workshop on Software and Performance (WOSP)*. ACM, 2005, pp. 43–51.
- [19] A. Avritzer, R. G. Cole, and E. J. Weyuker, "Using performance signatures and software rejuvenation for worm mitigation in tactical manets," in *6th International Workshop on Software and Performance*. New York, NY, USA: ACM, 2007, p. 172–180.
- [20] L. Cherkasova, K. Ozonat, N. Mi, J. Symons, and E. Smirni, "Automated anomaly detection and performance modeling of enterprise applications," *ACM Trans. Computer Systems*, vol. 27, no. 3, nov 2009.
- [21] A. Avritzer, R. Tanikella, K. James, R. G. Cole, and E. Weyuker, "Monitoring for security intrusion using performance signatures," in *first joint WOSP/SIPEW International Conference on Performance Engineering*. ACM, 2010, pp. 93–104.
- [22] A. Avritzer, A. B. Bondi, M. Grottko, K. S. Trivedi, and E. J. Weyuker, "Performance assurance via software rejuvenation: Monitoring, statistics and algorithms," in *DSN*, 2006, pp. 435–444.
- [23] C. Gonçalves, A. Avritzer, D. Menasché, M. Vieira, and N. Antunes, "Tuning the bucket algorithm parameters through its birth-death analysis," Technical report, 2020. [Online]. Available: <https://eden.dei.uc.pt/~charles/medcommnet/>
- [24] A. Bond, D. Johnson, G. Kopczyński, and H. R. Taheri, "Architecture and performance characteristics of a postgresql implementation of the TPC-E and TPC-V workloads," in *5th TPC Technology Conference*, 2013, pp. 77–92.
- [25] —, "Profiling the performance of virtualized databases with the tpcx-v benchmark," in *7th TPC Technology Conference*, 2015, pp. 156–172.
- [26] N. Gruschka and M. Jensen, "Attack surfaces: A taxonomy for attacks on cloud services," *IEEE Conf. on Cloud Computing*, pp. 276–279, 2010.
- [27] Ubuntu, "Stress NG," 2019, <https://kernel.ubuntu.com/~cking/stress-ng/>.
- [28] M. J. Zaki and J. Wagner Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.
- [29] P. E. Heegaard and K. S. Trivedi, "Network survivability modeling," *Computer Networks*, vol. 53, no. 8, pp. 1215–1234, 2009.
- [30] Y. Liu and K. Trivedi, "Survivability quantification: The analytical modeling approach," *International Journal of Performability Engineering*, vol. 2, no. 1, pp. 29–44, 2006.