Valid Application of EVT in Timing Analysis by Randomising Execution Time Measurements

George Lima
Depart. of Computer Science
Federal University of Bahia
Salvador, Bahia, Brazil
gmlima@ufba.br

Iain Bate
Department of Computer Science
The University of York
York, UK
iain.bate@york.ac.uk

***Abstract*—Intrinsic timing uncertainties present in modern hardware platforms have motivated the use of Extreme Value Theory (EVT) to timing analysis, however, the timing behaviour of a task may not entirely fulfil the necessary assumptions. To deal with this difficulty, randomisation at the hardware level has been proposed as a means of facilitating the use of statistical timing analysis. However, it has been shown that hardware randomisation does not solve all the analysis problems and importantly some projects may not wish to change the hardware that is used to support timing analysis. This paper presents an innovative approach, which does not require hardware randomisation or any special system feature, named *Indirect Estimation in Statistical Time Analysis* (IESTA). The main difference is that randomised hardware is performed before software instructions actually executes and is applied to parameters (e.g. cache state) only indirectly linked to timing. In contrast, IESTA adds its randomisation directly to the timing measures without affecting the way the software is executed. The IESTA approach is evaluated by experiments on two real case studies for which execution time measurements are taken from an embedded platform and from a Rolls-Royce Full Authority Digital Engine Controller.**

I. Introduction

Context. In a traditional real-time systems design, the Worst-Case Execution Time (WCET) of each task, namely an upper bound on its execution time, must be known [1]. When modern execution platforms are considered, however, deriving the WCET may not be possible or it may produce values excessively pessimistic. This is the case, for example, when complex memory hierarchies and/or multiple processor architectures are employed, which may introduce too much unpredictability in the system. The challenges in obtaining execution time bounds for such real-time systems have motivated many researchers use statistical techniques so that intrinsic uncertainties associated with complex execution platforms are incorporated into the derived estimates, usually called probabilistic WCET (pWCET): *The probability that the task execution time exceeds its pWCET estimate must not be greater than a threshold established in the system requirements.*

Among the approaches for estimating pWCET, those based on measurements have received special attention: pWCET estimates are derived by statistical analysis applied to data samples collected by measuring tasks' execution time. Many researchers have used Extreme Value Theory (EVT) for this purpose since it offers a set of solid tools capable of modelling the distribution of maxima for random variables (r.v. for short).

Unfortunately, there is no guarantee that EVT can be applicable in general for task execution time data [2]. Consider the example shown in the top-left graph of Figure 1. It shows a scatter plot representing a sample of execution time maxima for one of the case studies considered in this paper, a Binary search algorithm previously studied by Lima *et al.* [3]. The data is clearly discretely distributed and so cannot directly be dealt with by EVT, which assumes that data comes from continuous distributions. Hardware randomisation has been recommended to circumvent this kind of difficulty, *e.g.,* in [4]. However, as shown by Lima *et al.*, even using random replacement policies for caches does not make this example directly EVT-compliant. The case, illustrated in the top-right graph, makes evident that the obtained distribution, although smoother, is still discrete. Another difficulty in applying EVT comes out if observations in the sample are not independent, as it is the case for our second case study, from Rolls-Royce. The approach we describe, named *Indirect Estimation in Statistical Time Analysis* (IESTA), is based on randomising the data and works independently of the hardware being considered. Randomisation by IESTA is able to provide smooth data and to reduce or remove dependency so as to meet the assumptions of EVT-based analysis. An illustration of removing discreteness is given in the bottom graphs of Figure 1, which represent the same data in the graphs directly above following the application of IESTA. An illustration for data dependency will be given later in the paper. In this work we say that data is EVT-compliant if it conforms with EVT assumptions, *i.e.,* comes from a continuous distribution, the sample of maxima is independent and identically distributed (i.i.d.), and the associated distribution of maxima converges to an EV distribution [5], [6]. In short, it means the requirements for a valid application of EVT have been met.

Interestingly, there has been some debate on the use of randomised platforms. Although this recommendation has received some criticism [7], it indeed plays the role of smoothing out possible discreteness of sampled execution time data, making the applicability of EVT more likely [3], [4]. Using randomised hardware devices is not without side-effects,

978-1-5090-5269-1/17/$31.00

though. Studies on cache replacement policies have reported that deterministic approaches usually outperform the random policy [8], [9]. Some specialised deterministic performance-oriented policies have been described, *e.g.,* [10]. Thus, using random caches in a system may negatively impact its performance. Further, requiring devices especially designed for the purpose of time analysability may imply higher manufacture costs; it certainly decreases the number of possible costumers, for instance. Moreover, randomised hardware itself does not suffice to introduce the necessary degree of randomness that allows for EVT-based analysis, as illustrated in Figure 1. Other serious problems associated with hardware randomisation are: it causes greater variability in the actual timing behaviour of the system; it may increase the work needed for functional testing; and, for some applications, the performance of algorithms is affected [11], [12]. Unpredictability also makes debugging challenging [13], [14].

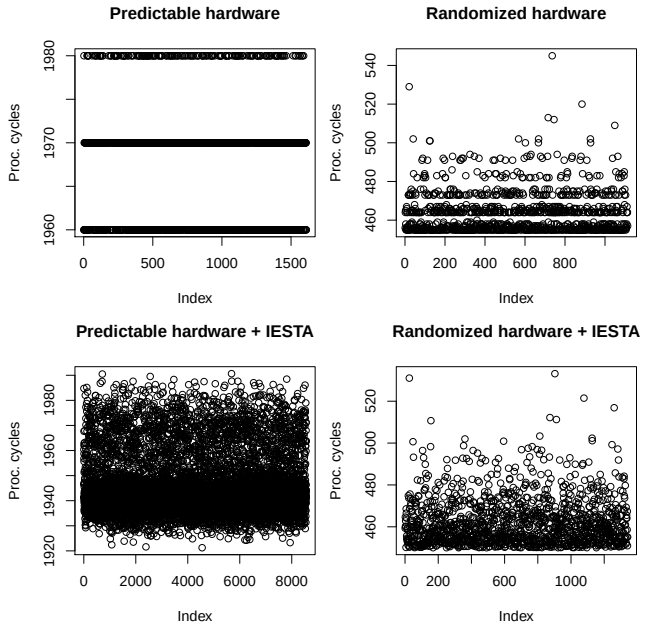

Fig. 1. Effects of randomisation on the sample of maxima w.r.t the execution of the Binary search algorithm on an embedded platform: Highly discrete distribution is observed on predictable architecture; Randomised architectures may not provide the required level of randomisation for ensuring EVT-applicability; IESTA randomises data after measurements potentially making observations in line with EVT assumptions.

Our contribution. Motivated by the aforementioned observations, in this work we answer the following question: is it possible to ensure the applicability of measurement-based statistical timing analysis without artificially introducing extra random effects at the system/hardware level? We answer this question positively and describe IESTA as a procedure capable of producing reliable pWCET estimates via EVT. No special feature at the hardware or system levels is required.

The idea behind IESTA is to accept that measured data from software is unlikely to be continuous or random in nature, and therefore introduce the necessary extra randomisation in the measured data so that it is EVT-compliant independent of the hardware used. We highlight that simply smoothing the observations is overly simplistic as whilst this may reduce the discretization/dependency effects, it would not solve the need of randomness.

The IESTA approach is assessed by experiments presenting two real case studies. We first consider the Binary search algorithm from Malärdalen benchmark [15]. The interest in this simple application is due to the fact that it has recently been shown that EVT cannot be applicable for its analysis even in the presence of randomised hardware [3]. The second case study was taken from a real aircraft engine whose data has been recently used for a path coverage analysis [16] in a highly predictable hardware (*i.e.,* no randomisation is applied). We show the data taken from this previous study is also not EVT-compliant and that it contains dependency relations making statistical analysis harder. IESTA is shown to make EVT applicable in both cases. It should be stressed that IESTA focus is on making execution time data samples EVT-compliant; the problem of obtaining representative data for the studied tasks is not addressed. Clearly, pWCET estimation quality depends on the degree that the measured data represents the task behavior. In this context, the good pWCET estimations we obtain by applying IESTA are also influenced by the fact that both the analysed Malärdalen benchmark code is simple and the analysis carried out by Law and Bate [16] offers good coverage of the code.

Related work. Since the first work on EVT-based timing analysis [17], [18] there has been intensive research in this field; the most relevant results in this area are discussed and cited during the course of the paper. Other measurement-based approaches exist [19], [20], [21]. Those are hybrid in the sense that information on the task code structure is also taken into account. Our focus is on EVT-based timing analysis which takes the analyzed code as a black box. Time measurements are used to estimate the parameters of an Extreme Value (EV) distribution, for which high quantiles are pWCET estimates. In this paper we follow recommendations for using generalized extreme value models, *e.g.,* [3], [5], [22], [23]. When it comes to facilitating the application of EVT for timing analysis, as previously mentioned, hardware randomisation has been the recommended means [4]. To the best of our knowledge, IESTA is the first approach addressing the lack of data randomness from the analysis perspective.

Structure of Paper. Section II gives necessary background on EVT and puts IESTA into its context. Next, the proposed IESTA approach is described in Section III and evaluated in Section IV. Our conclusions are drawn in Section V.

II. Background

Details of the standard EVT procedures we use can be found in specialised textbooks [5]. Here we highlight some key aspects of EVT analysis and put IESTA into its context. In a nutshell, applying EVT-based analysis consists of the following steps: (a) obtaining a sample for the r.v. of interest; (b) obtaining a sample of maxima for the r.v. sampled in (a);

(c) deriving a statistical model that best fits into the sample in (b); (d) model checking; and (e) determining a high quantile (*i.e.,* probabilistic bound) based on the model derived in (c) provided that this model is considered reliable in step (d).

IESTA acts in two stages. First, after step (a) and before step (b), IESTA has the goal of providing a "good" sample for EVT-based analysis. If the sample in (a) is not in line with EVT assumptions, IESTA randomisation scheme is able to generate EVT-compliant data from it. It should be stressed that if the original sample of measurements is good enough, IESTA does not need to be carried out. In the second stage, IESTA acts in step (e) to give a safe upper bound on the actual probabilistic bound on the task execution time. This second stage serves as an indirect estimation.

Step (a) – Execution time sampling. Let X_i, $i = 1, 2, \ldots, n$, be a sequence of r.v. obtained by measuring the execution of a task n times, which is denoted hereafter as $(X_i)_1^n$. The measurements $(X_i)_1^n$ are usually collected with some care since it is required that the sample in step (b) is i.i.d. [5], [3]. In general, sampling is carried out based on the fact that if all machine state variables (*e.g.,* cache, registers *etc.*) are cleared up before each time the analyzed task is set to run and input data for the task is chosen according to some i.i.d. distribution, the i.i.d. assumption for the sample in (b) is highly plausible. For real-time applications, however, this approach might not be possible or easy to implement. Applications that react to global state variables (*e.g.,* feedback control) might not comply with the i.i.d. assumption. Feeding the analyzed task with i.i.d. generated data for the sake of EVT may not be desired either as this does not guarantee the timings of the task conform to i.i.d. and it may reduce the efficiency of achieving sufficient coverage of the task. This is the case of the work by Law and Bate [16], which is taken here as the second case study. They investigated search-based methods to obtain high path coverages during measurements. Even if the i.i.d. assumption holds for the sample of maxima in step (b), there is also the problem of the potential discreteness of its distribution (recall Figure 1), which can prevent EVT from being successfully applied as EV models are continuous [3].

Instead of imposing constraints on the way sampling is carried out, in this paper we assume that both issues, namely data dependency and discreteness, can be present in the sample of execution time measurements. These issues are addressed via randomisation by IESTA. Empirical evidences from our experiments show the effectiveness of doing so.

Randomisation of IESTA (first stage). To address scenarios where $(X_i)_1^n$ is not EVT-compliant, the approach proposed in this paper introduces data randomisation in the form of *data padding*. This is to generate a sequence $(Y_i)_1^n$ from $(X_i)_1^n$ defined as $Y_i = X_i + Z_i$, $i = 1, 2, \ldots, n$, with $(Z_i)_1^n$ being a sequence of r.v. generated by the analyst such that $a \leq Z_i \leq b$. An interval $[a, b]$, defining a dispersion range $b-a$ of $(Z_i)_1^n$ w.r.t. $(X_i)_1^n$, is to be experimentally determined since it depends on the characteristics of $(X_i)_1^n$. Suitable choices for both $[a, b]$ and the distribution that rules $(Z_i)_1^n$ lead to EVT-compliant $(Y_i)_1^n$. In summary, this randomisation step is basically a means of producing a sample in step (a), namely $(Y_i)_1^n$, from the original values in $(X_i)_1^n$ and is only called for when applying EVT taking $(X_i)_1^n$ is not possible.

Step (b) – Sampling the maxima. Selecting the maxima from $(X)_1^n$ (when IESTA is not necessary) or from $(Y)_1^n$ (when IESTA is applied) can be done in different ways [5]. Two classical approaches are Block Maxima (BM) and Peak-over-Threshold (PoT). The former consists of partitioning the sampled data $(X)_1^n$ into equally sized blocks, whose sizes are specified beforehand, and selecting the maximum of each block; whereas the latter selects those values in $(X)_1^n$ above a certain previously defined threshold. Although IESTA is agnostic w.r.t. whether we choose PoT or BM for sampling the maxima, in this paper we focus on the PoT approach. The reason is it is not as robust as BM w.r.t. data dependency [5], which serves as a more difficult test for IESTA considering the dependencies in our case studies. Also, usually PoT is less wasteful when discarding values for obtaining suitable samples of maxima since under BM a single maximum observation is selected per block which reduces the likelihood that difficult to deal with dependencies are not removed. However, it should be stressed that although BM and PoT give rise to different models of extreme events, they are dual under asymptotic conditions and so pWCET estimates based on each of them should be equivalent [5]. Even though PoT is our focus of the evaluation, the evaluation presented does show that IESTA works appropriately for both PoT and BM for our case studies.

Under PoT, the definition of an ideal threshold value (similarly, block size for BM) depends on the sample and must be chosen based on the results of goodness-of-fit tests. Thresholds must be set to high values. Making it too high may cause large confidence intervals, as the number of satisfactory observations in the sample of maxima is reduced, implying models with low significance. In this work, the thresholds were chosen by trial and error with the exact choice of threshold not affecting the validity of IESTA's ability to transform data so that it is EVT-compliant.

Step (c) – Model estimation. An i.i.d. sample of maxima, selected via PoT, should be fitted into the Generalised Pareto (GP) distribution [5] whenever an EV distribution can represent the maximum of a random variable under analysis. For a large enough threshold u, GP asymptotically approximates the distribution function of $(X_i - u)$ conditional on $X_i > u$ with X_i representing an arbitrary value for the r.v. of interest:

$$\Pr\{X_i - u \leq v | X_i > u\} \approx G(v) = 1 - \left[1 + \frac{\xi v}{\sigma}\right]^{-1/\xi} \quad (1)$$

and is defined on $\{v : v > 0 \text{ and } 1 + \xi v/\sigma > 0\}$; with parameters $\sigma > 0$ and $\xi \in \mathbb{R}$ being respectively the scale and shape of G. Based on the sample of maxima, these parameters can be estimated by carrying out numerical procedures, which are found in open source software packages. In this work we use package extRemes [24] available in R [25].

Step (d) – Goodness-of-fitness checking. Once the model parameters are estimated, goodness-of-fitness tests are used for checking the adequacy to the data. Statistical tests to assess

the goodness-of-fit of estimated Generalised Extreme Value (GEV) models are available, *e.g.,* [26]. We have followed recommendations by Coles [5] who establishes the analysis of a set of complementary graphs as a means of model quality checking. The fitting quality is illustrated in this paper via Quantile-Quantile plots (QQ-plots for short) according to which estimated model quantile is plotted against the empirical quantile. A good fitting is shown when the empirical and estimated distributions agree, which is indicated when sampled maximum data appears on or close to the diagonal line. This is illustrated in the graphs on the right of Figure 5. Poor-quality fittings are exemplified by the graphs on the left.

Step (e) – Probabilistic bound derivation. Under the GP framework, if a model can be obtained for a given r.v., it is of interest to determine a value $q(p)$ associated with a probability of exceedance p. That is, $\Pr\{X_i > q(p)\} = p$. From (1) and considering that the extreme value of interest $q(p) > u$,

$$\Pr\{X_i > q(p) | X_i > u\} = \left[1 + \xi \frac{q(p) - u}{\sigma}\right]^{-1/\xi}$$

As $q(p) > u$, $\Pr\{X_i > q(p), X_i > u\} = \Pr\{X_i > q(p)\} = p$, yielding

$$p = p_u \left[1 + \xi \frac{q(p) - u}{\sigma}\right]^{-1/\xi}$$

with $p_u = \Pr\{X_i > u\}$. Solving the above equation leads to

$$q(p) = \begin{cases} u + \frac{\sigma}{\xi}\left[(p_u/p)^{\xi} - 1\right], & \xi \neq 0 \\ u + \sigma \log(p_u/p), & \xi = 0 \end{cases} \quad (2)$$

provided that p is sufficiently small to ensure that $q(p) > u$, which is the case since we are interested in determining an estimation beyond what has been observed. Note that $q(p)$ is the $(1-p)$-quantile of (1) adjusted w.r.t. the chosen threshold u. The value of p_u can be estimated as the ratio between the size of the sample of maxima (values above u) and that of sampled measurements.

Indirect estimation (second stage). Equation (2) gives the probabilistic bound for the sequence $(X_i)_1^n$ or $(Y_i)_1^n$, depending on whether the first stage of IESTA has been applied in step (a). If it has, a probabilistic bound on the maximum of $(X_i)_1^n$ needs to be determined. IESTA does so using the derived quantile for the maximum of $(Y_i)_1^n$. Details on the whole IESTA method is given in the next section.

III. The IESTA method

IESTA is based on adding sufficient padding to the measured execution times so they become EVT-compliant. A more formal justification for the approach is given in Section III-A. Section III-B illustrates its effects. The IESTA procedure is then detailed in Section III-C. Possible limitations on EVT and IESTA are mentioned in Section III-D.

We notice that IESTA padding is different to the padding used in the Enhanced Path Coverage (EPC) approach [27] as it works at the block level whereas IESTA works at the task level. An advantage of IESTA is that the necessary level of instrumentation is reduced. Other randomisation techniques to improve EVT applicability may consist of data shuffling [28], [29], an aspect briefly discussed in Section III-B.

A. Randomisation via data padding

For the sake of notation, we differentiate the true (and unknown) c.d.f. of maxima for a sequence of r.v. $(X_i)_i^n$, denoted $G_X(v) = \Pr\{\max_1^n(X_i) \leq v\}$, from the distribution $G(v)$, expressed as Equation (1). We also denote the quantile functions associated with G and G_X as q and q_X, respectively. If G_X asymptotically converges to some EV distribution, (2) can be used for estimating the pWCET. As this is not always the case (practical examples will be given shortly), we apply an indirect estimation using a different sequence as an estimation means. More specifically, we consider a sequence of r.v. $(Z_i)_1^n$ such that $a \leq Z_i \leq b$, with $a \leq b$ two known constants and define sequence $(Y_i)_1^n$ as $Y_i = X_i + Z_i$, $i = 1, 2, \ldots, n$ such that $(Y_i)_1^n$ is analysable via EVT. By establishing properties relating q_X and q_Y, we then show that we are able to derive bounds on q_X via those of q_Y. This is done in Corollary 1, a consequence of the following property:

Theorem 1: Let $(X_i)_1^n$ be a sequence of r.v. and consider a sequence of r.v. $(Z_i)_1^n$, with $a \leq Z_i \leq b$, $a \leq b$, $i = 1, 2, \ldots, n$. Define the sequence of r.v. $(Y_i)_1^n$ such that $Y_i = X_i + Z_i$. For any value v, the following relation holds:

$$G_X(v - b) \leq G_Y(v) \leq G_X(v - a) \quad (3)$$

Proof: It follows straightforwardly from the fact that $a \leq Z_i \leq b$, implying that $X_i + a \leq Y_i \leq X_i + b$. Thus,

$$\begin{aligned} G_X(v-b) &= \Pr\{\max_1^n(X_i) \leq v - b\} \\ &= \Pr\{\max_1^n(X_i + b) \leq v\} \\ &\leq G_Y(v) = \Pr\{\max_1^n(X_i + Z_i) \leq v\} \\ &\leq \Pr\{\max_1^n(X_i + a) \leq v\} \\ &= \Pr\{\max_1^n(X_i \leq v - a\} = G_X(v - a) \end{aligned}$$

■

Based on Inequality (3), the quantiles of G_Y can be used to bound the quantiles of G_X:

Corollary 1: Let $(X_i)_1^n$ be a sequence of r.v. and consider a sequence of r.v. $(Z_i)_1^n$, with $a \leq Z_i \leq b$, $a \leq b$, $i = 1, 2, \ldots, n$. Define a sequence of r.v. $(Y_i)_1^n$ as $Y_i = X_i + Z_i$. The quantile of G_X is bounded by the quantiles of G_Y as:

$$q_Y(p) - b \leq q_X(p) \leq q_Y(p) - a, \quad 0 < p < 1 \quad (4)$$

Proof: For convenience, let $v = q_Y(p)$ or equivalently $p = G_Y(v)$ for some v. By contradiction, assume that (4) does not hold. This means that

$$q_Y(p) - b > q_X(p) \quad \text{or} \quad q_X(p) > q_Y(p) - a \quad (5)$$

As $0 < p < 1$, $q_X(p)$ exists and so $G_X(q_X(p)) = 1 - p$. Thus, Inequalities in (5) imply that $G_X(v-b) > 1-p$ or that $1 - p > G_X(v - a)$, both of which contradict Theorem 1. ■

Corollary 1 offers a means of bounding $q_X(p)$ by using $q_Y(p)$ instead, introducing a maximum error of $b-a$. In other words, $q_Y(p)$ can be estimated via EVT and $q_Y(p) - a$ is used to bound $q_X(p)$. Thus, estimations on the former are exactly the same as those on the latter minus the shift. That is, if $\max_1^n(X_i)$ converges to an EV distribution, then so will $\max_1^n(Y_i)$. When this is not the case, setting $a < b$ is necessary and distributions of $\max_1^n(Y_i)$ and $\max_1^n(X_i)$ differ. The error estimating $q_X(p)$ increases proportionally to the extent that $\max_1^n(Y_i)$ and $\max_1^n(X_i)$ differs, the difference is bounded by Inequality (4).

To use the results of Corollary 1, we need to construct a suitable sequence $(Z_i)_1^n$. We note the only restriction on Z_i is that its values lie within $[a, b]$. As we also wish to reduce possible dependency relations, we must consider Z_i independent of X_i. Thus, we generate r.v. Z_i according to a well behaved function known to be analysable via EVT.

In the context of all experiments we carried out, it has been observed that generating normally distributed values for Z_i consistently provides better results when compared to using either Uniform or Exponential distributions. By better results we mean obtaining EVT-compliant sequences $(Y_i)_1^n$ with a lower range for interval $[a, b]$, which in turn induces lower pessimism in the pWCET estimates. Other choices for generating Z_i as well as investigation on their properties must be considered in future work. This issue will be further commented in Section III-D. Hereafter, we consider Z_i according to a Normal distribution with mean α and standard deviation β *i.e.,* $Z_i \sim \mathcal{N}(\alpha, \beta^2)$.

In order to keep the values of Z_i between the desired values of a and b with a high probability, the parameters α and β need to be suitably set. We do so in the function of the degree of dispersion $(Z_i)_1^n$ has w.r.t. $(X_i)_1^n$:

Definition 1: The dispersion ratio is the factor between the sizes of intervals within which Z_i and X_i are distributed and is given by

$$\delta = \frac{b-a}{\max_i^n(X_i) - \min_1^n(X_i)} \tag{6}$$

The dispersion ratio δ should be large enough so as to make $(Y_i)_1^n$ analysable via EVT. Too large values, though, may make pWCET estimation too pessimistic. In our experiments, reported later on, we provide empirical evidence indicating that the pessimism introduced is usually small.

From the properties of the Normal distribution, we know that

$$\Pr(\alpha - 5\beta < Z_i < \alpha + 5\beta) = 0.9999994 \tag{7}$$

Hence, if β is set so that $[\alpha - 5\beta, \alpha + 5\beta]$ is close enough to the desired interval $[a, b]$, the values of Z_i are generated within $[a, b]$ with high probability. Using δ defined in (6),

$$\delta = \frac{\alpha + 5\beta - (\alpha - 5\beta)}{\max_1^n(X_i) - \min_1^n(X_i)}$$

which implies that the mean α can be arbitrarily defined; for convenience we set $\alpha = 0$, while β can be

$$\beta = \frac{[\max_1^n(X_i) - \min_1^n(X_i)]\,\delta}{10} \tag{8}$$

B. Effects of randomisation

Statistical analysis of extremes, briefly summarised in Section II, requires an i.i.d. sample of maxima to work with. Data independency trivially holds if the underlying sample, in our case $(X_i)_1^n$ or $(Y_i)_1^n$ after padding, is i.i.d. The independence assumption will also hold if the process of sampling the maxima ensures that the selected values in the sample are sufficiently separated from each other [5] (*e.g.,* via defining a sufficient high threshold u). In general, note that if the sample of maxima converges to an EV distribution, it is guaranteed to be identically distributed; otherwise such a convergence would not be possible [5].

In practical application scenarios the presence of dependent data may need to be addressed. This is the case for financial or environment-related processes, for which data is intrinsically time-dependent. An individual observation can be affected by past events (e.g. due to data or cache state) and this dependency chain can be passed through to the sample of maxima [5]. There are a number of ways of addressing dependency in the analysis of extreme events. For time-dependent chains present in stationary processes, classical means under the PoT approach aim at ensuring independency in the sample of maxima by discarding maximum values close together, which is known as de-clustering [5]. When an EV model is derived for the obtained i.i.d. sample, theoretical results provide means of compensating for the discarded values.

Unlike discarding values of maxima, under IESTA dependency is ensured via padding, which also deals with data discreteness, as illustrated in Figure 1. Figure 2 shows the effects of IESTA for dependencies taking the data set provided by Rolls-Royce identified as ACDT in Table I, one of the analyzed sets in our case study. On the top are scatter plots showing samples of maxima for the measured values $(X_i)_1^n$ and the corresponding padded sequence $(Y)_i^n$, generated for $\delta = 49\%$. The selected values are those above thresholds $u = 900$ and $u = 905.33$, respectively, values chosen to make both samples of maxima be roughly with the same size (around $1\,000$ observations), facilitating a comparative visualisation. As can be seen, padding is capable of removing the discretization observed in the sample $(X)_i^n$.

The other graphs in Figure 2 are Autocorrelation Function (ACF) plots, which indicate the extent to which the ith and jth observations in the same sample of maxima are correlated to each other, with lag representing the distance between them (*i.e.,* $j - i$). Values 1 and -1 mean full positive and negative correlation, respectively. Note the value 1 for lag 0 (jth observation against itself). Horizontal dashed lines represent 95% significance bands. In the second row of Figure 2, ACF plots indicate high autocorrelations indices in the sample of maxima of $(X_i)_1^n$ for lags as high as 40. Also note that the correlation decays slowly until lag 20 from which becomes negative.

After padding autocorrelation is significantly reduced. The last line of the figure shows the same ACF plots for higher threshold values, 927 and 932, respectively. Value 932 was found necessary for reliably estimating the parameters of (1) (as will be seen in Section IV). From both graphs it is clear the role of its threshold in decreasing dependency in the sample of maxima. No patterns or strong autocorrelations are observed.

Another means of reducing dependencies is by generating a random permutation of the data, which has been done in previous work [28], [29]. This data shuffling approach can be justifiable when time-dependent relations, commonly found in other application areas, is not present in the sample of original (or padded) execution time measurements. Nonetheless, note that shuffling does not solve the issue of discreteness. The top-left graph in Figure 1 would still appear with three bands independently of the order data is indexed in the sample and how much data is sampled.

Although IESTA can easily be extended to take shuffled data into consideration, we have not found this necessary. First, during experiments, the padding process was found effective to make data suitable for EVT. Second, when trying to apply padding to shuffled data under IESTA, no benefits in terms of analyzability or pessimism reduction was observed. Third, it might be interesting in future steps of this research to consider models that capture how execution time evolves with time, which might be particularly useful for tasks such those related with feedback-control. Hence, we prefer to take the measurement values as a non-permutable sequence and check the extent IESTA is able to address possible correlation between measured values.

C. Detailed procedure

The following steps specify the proposed procedure for estimating a safe value for $q_X(p)$. It is assumed that the desired probability of exceedance p is given.

(i) Given a sample sequence $(X_i)_1^n$, define a dispersion ratio δ; initially set to a small value, *e.g.,* 1%.

(ii) Generate a sequence of n values Z_i according to some distribution for which the parameters must be computed as a function of the desired value of δ. As previously explained, in this paper we use the Normal distribution. Then define $Y_i = X_i + Z_i$, $i = 1, 2, \ldots, n$, and let $a = \min_1^n(Z_i)$ and $b = \max_1^n(Z_i)$.

(iii) Check whether $(Y_i)_1^n$ can be analysed via EVT. If it cannot, repeat step (ii) after increasing the value of δ;

(iv) Select a sample of maxima for $(Y_i)_1^n$.

(v) Estimate the parameters for the EV model for the selected sample of maxima. If the model is not considered reliable, via goodness-of-fit tests, return to the previous step, searching for better tuning parameters (threshold values or block maxima sizes). Once a reliable model is obtained, compute the desired quantile $q_Y(p)$.

(vi) Estimate pWCET as $q_Y(p) - a$ based on Equation (4).

Steps (iv)-(vi) correspond to steps (b)-(e), as explained in Section II. Steps (i)-(iii) are the the first stage of IESTA within step (a) whereas step (vi) is its second stage.

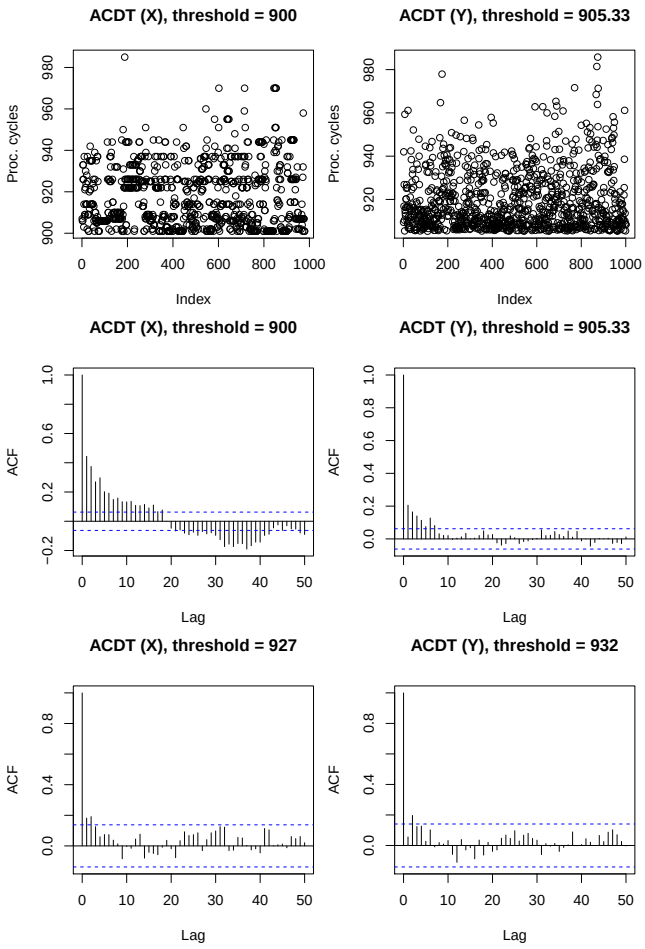

Fig. 2. Graphs showing discreteness and different levels of data dependency present in the ACDT data set. Data padding is observed to reduce dependency relations. High threshold selection make it unlikely dependency appears in the sample of maxima.

As $(X_i)_1^n$ comes from unknown distributions, the resulting sequence $(Y_i)_1^n$ is arbitrary in general. Hence, it is not possible to know beforehand which value of δ makes $(Y_i)_1^n$ EVT-compliant. Step (iii) of the IESTA procedure checks this before going to EV model estimation. We emphasise that when relying only on hardware randomisation, it is not possible to ensure that the measured data is EVT-compliant [3] (recall Figure 1). An advantage of using the IESTA approach is that analysts can control the dispersion ratio employed without imposing restrictions on the underline hardware.

D. Comments on the limitations of IESTA

It is important to emphasise the scope IESTA applies for. An important issue is data representativeness, *i.e.,* if the observed value in $(X_i)_1^n$ does not represent the actual behavior of the task, the estimated pWCET is unlikely to be a safe bound for its WCET. Any timing analysis method based on measurement, including IESTA, has this problem. The focus here is on making a given sample of measurements compliant to EVT.

For the case studies in this paper, we rely on evidence given by previous work [3], [16] that the data is representative. We also note intensive or exhaustive testing combined with appropriate coverage is considered suitable evidence according to most certification standards [30].

Other aspects are related with sequence $(Z)_1^n$. If it is too dispersed w.r.t. $(X)_1^n$, the approximation model (1) is likely to be dominated by the distribution tail of $(Z)_1^n$. This may bring about possible negative effects as for what is predicted by the bounds derived in Section III-A. First, these bounds apply to the actual distributions of maxima for the sequences and Equation (1) is an approximation based on the observed data. As one is usually interested in estimating values far right in the tail, *i.e.,* much further from what has been observed in the sample, estimating high quantiles when $(Z)_1^n$ dominates $(X)_1^n$ may be risky. Second, the bounds in Equations (3) and (4) were derived based on the fact that $a \leq Z_i \leq b$. Using a Normal distribution for generating $(Z)_1^n$ implies that $[a, b]$ is not bounded. Due to these reasons, pWCET estimations should be interpreted with caution when the necessary value of δ to make EVT applicable is too high. For the case studies in this paper such possible negative effects has not being observed. Mechanisms to relate the uncertainties on $[a, b]$, as expressed in (7), the dispersion ratio δ and pWCET estimated via (4) could be necessary. A careful study on applying other distributions for generating $(Z)_1^n$ needs also to be carried out. These aspects should be considered in future research steps.

IV. Evaluation

The overall evaluation of IESTA is based on analysing 9 data sets taken from two case studies with different levels of complexity ranging from a simple Binary search, from Malärdalen Benchmark, to an industrial aircraft control application by Rolls-Royce. These are better characterised in Section IV-A. Using two of these data sets we explain in Section IV-B how IESTA changes the nature of samples and affects model quality. Overall results obtained for the case studies are then given in Section IV-C. In Section IV-D, IESTA is tested against the hardest data sample (in terms of meeting the assumptions of EVT) found in our case studies so as to check analysis robustness. Section IV-E shows the BM approach can also be used with IESTA.

A. Characterisation of the case studies

The data sets we analyse, summarised in Table I, were collected via a series of experiments from previous work [3], [16], which are taken as two case studies. The last column in the table represents the highest execution time value ever observed during extensive experiments, named High Water Mark (HWM).

Case study BS [3]. This consists of a single data set (first line of Table I), which comes from executing the Binary search algorithm from the Malärdalen benchmark [15] in a RISC-V embedded platform, equipped with a single processor with no pipeline. The platform can be set up with or without random cache. For this paper, we considered the no-cache version as there are more challenges for IESTA in achieving EVT-compliant [3]. The measurements were made such that each observation in the sample is independent of the other. As previously illustrated in Figure 1, the distributions for both measured data and the corresponding sample of maxima are discrete. For this case study, the HWM corresponds to the WCET.

TABLE I
Characteristics of the analyzed data. Execution time is given in processor cycles.

Data set	Sample size	Min	Max	Avg	HWM
BS	10 000	740	1 980	1 854.65	1 980
F	291 958	9 651	12 018	9 909.48	12 022
ACDF	268 084	186	308	222.81	314
ACDN	298 773	334	489	379.26	489
ACDP	291 411	451	1 229	656.16	1 230
ACDT	425 052	816	985	829.92	1 011
VCA	804 395	684	2 799	866.71	2 847
VCP	556 548	2 533	5 749	3 534.27	5 802
VCP*	50 050	2 585	5 134	3 280.25	2 847
VCS	560 349	1 712	2 409	2 011.35	2 409

Case study RR [16]. Eight different tasks are considered in this case study. The corresponding data sets are described in Table I. The time measurements were carried out on a cycle accurate simulator for a processor used on Rolls-Royce's aircraft projects. The analyzed tasks present different levels of complexity (as described in [16]) and are from an engine control system from Rolls-Royce. The measurements did not follow the usual approach of simply randomly choosing input data to the tasks. Instead, they were conducted by applying search-based techniques for reliably obtaining high path coverage and high time values during the measurements. In a nutshell, the aim of each of these techniques is to select long execution paths by iteratively maximising a fitness function during the measurements. Simulation annealing was used for optimisation. All but one data set summarised in the table correspond to those applying the BCHLr fitness function[1], which was shown to be more effective by Law and Bate [16]. Data set labeled VCP* is for the same task as VCP but using another fitness function, which will be considered only in Section IV-D. The degree of the dependencies within the sample is based on how the search algorithm works and on the fitness function itself. In general, the resulting samples feature a high-degree of dependencies (recall Figure 2, which illustrates this for ACDT). Data discreteness also shows up due to the nature of the architecture. For the RR data sets, the HWM may not match WCET, which is unknown. HWM is what has been observed during measurements for all fitness functions. Because of this, as can be seen in the table, the maximum observed value in some data sets differ from the HWM [16].

We note that execution time values for one case study cannot be related to the other since they correspond to distinct architectures. The BS case study is being considered here

[1]BCHLr is a fitness function that targets both localised path coverage, i.e. at the function level, and maximises the number of iterations around loops.

because it provides a simple and intuitive example for which EVT cannot be applied even if hardware randomisation is employed [3]. The Rolls-Royce case study provided by Law and Bate serves to validate whether and the extent to which IESTA converts arbitrary measured data for a real industrial system into a form that makes EVT applicable so that sound pWCET estimates can be delivered.

We stress that in most application areas where EVT is applied, the measurement $(X_i)_1^n$ is dependent on how it is taken. To the best of our knowledge, the academic research present to date on timing analysis seems to assume this form of dependency does not exist. In the RR case study the dependencies are significant, as illustrated in Section III-B. This could be exaggerated by the specific way the measurements are gathered. A means of removing those observed autocorrelations is by shuffling $(X_i)_1^n$, as mentioned in Section III-B. This approach has been used by other researchers, e.g. [28], [29]. However, shuffling may not eliminate all possible dependencies as based on discussions with industrialists (including from RR) they will exist anyhow. The key reason is as functions tend to carry state from one cycle to the next, *e.g.,* in the case of feedback-based control systems or more specifically in the case of aircraft engine control as some variables like altitude do not change significantly between successive samples. Data padding and suitably high threshold selections act are complementary mechanisms in this respect.

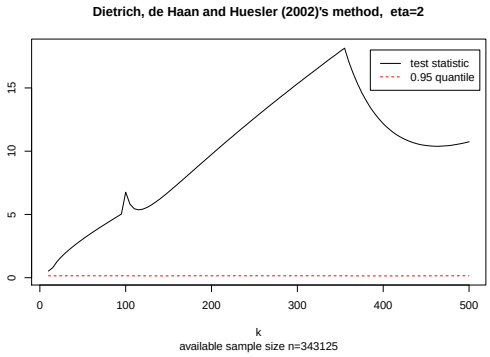

Fig. 3. Test statistic by Dietrich et al. [6] applied to the ACDN data. It indicates that the Hypothesis H_0, that the sequence of r.v. $(X_i)_1^n$ belongs to the domain of attraction of an EV distribution, should be rejected. The test parameter $\eta = 2$ is the best value to be chosen for the test [31].

From the above discussion, one cannot expect to derive reliable models for the data sets under study. Although the graphs shown in Figures 1 and 2 give some intuition about this fact, a more formal treatment can be carried out to check it. Indeed, even if the sample of measurement is continuous and i.i.d., there is no guarantee that it is EVT-compliant. The test by Dietrich et al. [6] can be used for this purpose. This test has been customised and made available as an R package by Hüsler and Li [31], [32]. Informally speaking, this test is based on comparing the values of two quantile functions: one constructed based on the first k largest sampled data at hand; and the other obtained by asymptotic approximation. The hypothesis that the distribution of maxima for the sampled data converges to an EV distribution is rejected if the former function is not consistently smaller than the latter for some range of k. Figure 3 plots the graph for this test. As can be seen, the value of the test statistics for the ACDN data set is clearly above the 95%-quantile function, indicating the non-suitability of the data for EVT-based analysis. None of the data sets described in Table I passed the test.

B. IESTA as a means of achieving EVT-compliance

In this section we use data sets BS and ACDN to illustrate the effects IESTA causes w.r.t. EVT-compliance by empirically checking the corresponding sequences of $(Y_i)_1^n$ for different values of δ. BS and ACDN were observed to be EVT-complaint from values of δ starting from 3% and 6%, respectively, which was checked by running the test by Dietrich *et al.* [6]. Figure 4 illustrates the result for ACDN and can be compared with Figure 3. As the value of the test (solid line) is now consistently below the 0.95-quantile of the reference distribution, the generated sequence $(Y_i)_1^n$ can be considered EVT-compliant.

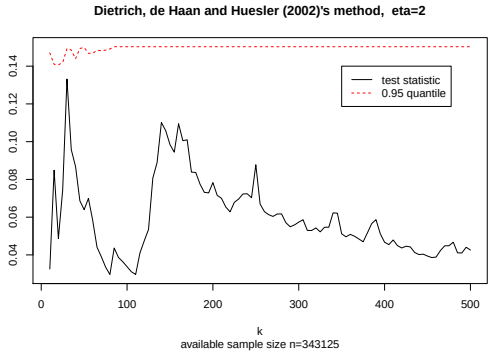

Fig. 4. Test statistic by Dietrich et al. [6] applied to the ACDN data after IESTA randomisation (Normal distribution, $\delta = 6\%$). It shows no argument for rejecting hypothesis H_0, that $(Y_i)_1^n$ belongs to the domain of attraction of an EV distribution.

Figure 5 depicts the goodness-of-fit checking for BS and ACDN based on different values of δ. For $\delta = 1\%$ the derived models are not suitable for explaining the empirical data (distribution of maxima) collected from the measurements. Increasing the dispersion ratio to 3% and 6% makes the derived GP models for BS and ACDN acceptable since what the models predict is followed by what is observed in the samples; and so the respective models can be used for making estimations beyond the range of observations.

C. Estimation results

Table II summarizes the main results from the EVT analysis after applying the IESTA data randomisation method. Each estimated model for which the results are shown in the table was obtained as follows. The procedure described in Section III-C was applied considering $\delta \in [k, 10k]$, with $k = 1\%, 2\%, \ldots$. After checking that IESTA randomisation has provided an EVT-compliant sample, 10 different models

TABLE II

ANALYSIS RESULTS FOR ALL DATA SETS AFTER APPLYING IESTA RANDOMISATION. ESTIMATIONS ARE BASED ON THE GP MODEL FOR EXCEEDANCE PROBABILITY 10^{-4}. TIME VALUES ARE GIVEN IN PROCESSOR CYCLES.

		$(Y_i)_1^n = (X_i)_1^n + (Z_i)_1^n$			pWCET				Estimated EV dist. shape	
Data set	δ	Min	Max	Avg	$q_Y - a$	95%-CI	Pessimism	Threshold	$\hat{\xi}$	95%-CI
BS	3%	723.26	1 990.64	1 854.58	2 011	(2 008, 2 016)	1.57%	1 970	−0.39	(−0.50, −0.29)
F	66%	8 984.83	12 272.95	9 909.73	13 182	(12 884, 13 480)	9.65%	11 650	−0.19	(−0.38, 0.00)
ACDF	7%	181.53	310.73	222.81	316	(314, 318)	0.64%	308	−0.34	(−0.46, −0.22)
ACDN	6%	330.53	491.03	379.23	499	(496, 501)	2.05%	483	−0.07	(−0.15, 0.01)
ACDP	36%	378.71	1310.78	656.15	1 464	(1 442, 1 489)	19.02%	1 240	−0.20	(−0.31, −0.08)
ACDT	49%	779.66	985.79	829.92	1 036	(1 015, 1 057)	2.47%	932	−0.10	(−0.23, 0.04)
VCA	6%	631.58	2 805.42	866.70	2 898	(2 858, 2 936)	1.18%	2 700	−0.15	(−0.30, −0.01)
VCP	32%	2 257.99	5 922.13	3 534.33	6 520	(6 271, 6 768)	12.37%	5 450	−0.07	(−0.29, 0.14)
VCS	31%	1 673.15	2 450.84	2 011.28	2 576	(2 548, 2 664)	6.89%	2 300	−0.11	(−0.17, −0.05)

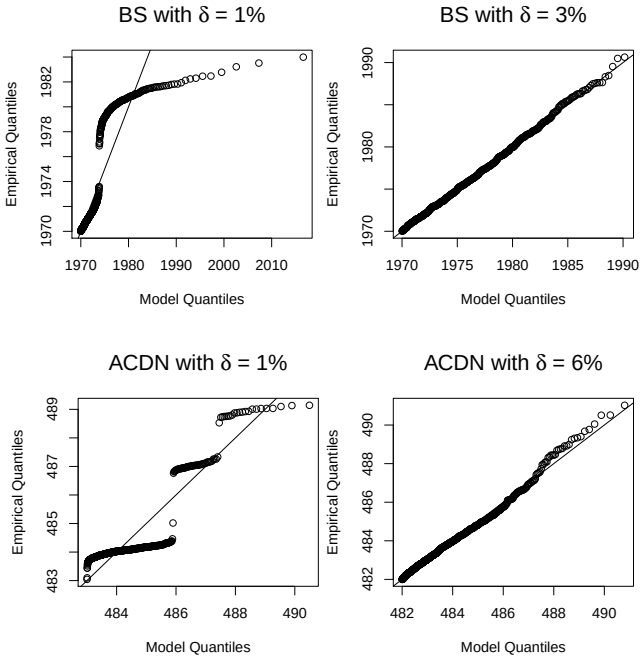

Fig. 5. QQ-plots indicating that model fitting qualities vary with the dispersion ratio δ. Data sets BS and ACDN are used for illustration. The higher the value of δ, the more likely good model fitting is obtained. The level of dispersion depends on data characteristics.

were estimated for the found range of k. Then, the model that gave the tightest confidence interval for its quantile was selected. During this phase, adjustments to the thresholds were made based on the goodness-of-fitness results as described in step (d) in Section II. The pessimism shown in the table is the distance (in percent) of the obtained pWCET estimates from the corresponding HWM, given in Table I. Below we highlight the main observations that can be drawn from Table II.

On the shape of the estimated models. As can be seen, all derived models are estimated with negative shape parameter, confirming previous observations by Lima *et al.* [3], who have found no experimental support to the assumption that maximum task execution times should be modelled by a zero-shape EV distribution. This recommendation was originally suggested by Edgar and Burns [18] and followed by several other researchers. In our case studies, zero-shape EV models are possible for F, ACDN, ACDT, and VCP, for which the confidence intervals for $\hat{\xi}$ admit the null value.

On the value of δ. The effective value of δ was found to vary considerably, from 3% for BS to 66% for F. During the analysis we have observed that the necessary value of δ is mainly linked to the distribution of measurements and could not be predicted beforehand. Interestingly, high values of δ do not necessarily imply the same proportion of induced pessimism, although it makes pessimism more likely. Indeed, pessimism over 2% was obtained for high values of δ. However, comparing pWCET estimated for ACDP and ACDT, the observed pessimism for the former was considerably higher even though δ was set to a much greater value for the latter. Again, as pointed out in Section III-D, results for high values of δ should be taken with care.

On the use of the Normal distribution. Using the Normal distribution for generating the values of $(Z_i)_1^n$ was shown to be effective in providing the necessary randomness without modifying the points in $(X_i)_1^n$ too much. This can be observed by comparing the columns Avg for $(Y_i)_1^n$ and $(X_i)_1^n$ in Tables I and II. Indeed, most values of $(Z_i)_1^n$ are generated around the mean (set to 0 - recall Section III-A), which implies that the mean for both distributions are expected to be the same. Likewise, the effects observed on the minimum and maximum values of $(Y_i)_1^n$ as compared to those of $(X_i)_1^n$ are unlikely to be of the same order as δ. For example, for data set F, which required $\delta = 66\%$, the variations caused by IESTA randomisation on the minimum and maximum values are around 6% and 2%, respectively. In other words, even with high values of δ, the probability that the extremes of the distribution are affected to a great extent is low due to the use of the Normal distribution. This is important since an EV model will capture the distribution of extreme values and so low variations in the maxima imply tightness for the derived values of pWCET. That is, even if δ has a larger value evidence suggests that the pessimism introduced is not significant allowing systems to be resource efficient.

Let us consider the observed differences in pessimism for ACDP and ACDT. For these cases, the maximum values in $(Y_i)_1^n$ are 6.65% and 0.08% higher than $\max_1^n(X_i)$, respectively. This difference is mainly due to the nature of the IESTA randomisation, as explained above. Distinct values can

be observed in different runs of the randomisation process, although $\max_1^n(Y_i) - a > \max_1^n(X_i)$ as stated in Corollary 1.

D. Alternative fitness function for RR case study

The results previously represented for the RR case study have taken data collected via the BCHLr fitness function. BCHLr is a fitness function that tried to maximise the paths covered within a region of code, e.g. a function, as well as maximising the loop count, which has been shown to be more effective in terms of path coverage leading to a more reliable WCET. In this section we are interested in testing IESTA in a more challenging scenario. To do that we chose the hardest to analyze data set provided by Law and Bate, named VCP* in Table I. It was collected by the Ran fitness function, which simply picks $50\,050$ (less than 10% of the sample size for BCHLr) execution time values from all other six fitness functions at random. The maximum value observed in VCP* is $5\,134$, against $5\,749$ for BCHLr. Further, the distribution of measurements was shown to be more difficult to make EVT-compliant. The value of δ had to be set as large as 118% to make EVT-based analysis possible. This caused, as expected, an increase in the estimated pWCET. For exceedance probability $p = 10^{-4}$, pWCET was estimated as $7\,683$ processor cycles with 95%-confidence interval of $(6\,994, 8\,372)$. Despite the large value of δ, this corresponds to a relative small increase in the estimated pWCET: of about 18% in comparison with pWCET of $6\,520$, as reported in Table II; and of 38% with respect to the largest value observed in $(Y_i)_1^n$, which was equal to $5\,682.34$. These observations indicate certain robustness of the IESTA-EVT framework described in this paper. However, as indicated in Section III-D, further research must be carried out so as to detect whether such large values of δ can compromise safety in estimations.

E. Alternative modelling - GEV

Our modelling choice, based on the PoT sampling approach and on the Generalised Pareto distribution, was found to be more effective when compared to the alternative of using the BM approach and the GEV distribution. In this section we give some arguments in favour of our choice based on the observed experimental results. We do not give details as for the GEV model applied to timing analysis, though. Interested readers may refer to Lima *et al.*[3] for more information.

For the sake of argumentation, let us consider VCA as an example since it has the largest sample size. For this case, a reliable GEV model was obtained when the block size was set to $37\,900$. This is explained by the fact that measured data contains dependencies (caused by the fitness function used during sampling – recall Section III-B) and the BM approach is more data wasteful when compared to PoT [5]. Indeed, due to dependencies, the BM approach is bound to discard several high observations (in the same block) in order to produce an i.i.d. sample of maxima, which is necessary for model fitting.

Even though, pWCET estimation by the estimated GEV distribution, equal to $2\,905$ processor cycles, was found equivalent to the one reported in Table II. This is expected since both GP and GEV are considered dual models. However, the estimated confidence interval based on the derived GEV distribution for this case, namely $(2\,861, 3\,041)$, was found to be larger. This was caused by the fact that the sample of maxima by BM contains much fewer observations (21) than that generated by PoT (196). Similar effects were observed for all other RR data sets for which required block sizes were not below $10\,000$. As for BS, GEV estimations was as good as those based on GP since the raw data does not contain dependency.

V. Conclusion

We have described IESTA, an approach to statistical timing analysis capable of providing pWCET estimates even when the underlying execution platform does not deliver the necessary level of randomness and/or measurements are not i.i.d. to allow for reliable EV-model derivation. This capability removes the need of using randomised hardware devices. Experiments have indicated that IESTA is an effective approach independent of the EVT method (i.e. for both GEV and PoT) and that PoT provides better results than GEV. Its only observed side-effect, the extra pessimism incurred, has been experimentally shown to be acceptable. A further benefit is to the analyst researching other issues with applying EVT, *e.g.,* the importance of data representativity, since IESTA isolates such problems from the ones caused by model fitting difficulties related to the absence of necessary data independency and randomness. Another issue to be further investigated is on the extent data can be randomised without compromising the validity of pWCET estimations via EVT. The results presented here offer a good indication that the proposed approach is robust enough. Other case studies and application scenarios should be considered in future research steps.

Acknowledgment

The authors would like to thank Rolls-Royce Control Systems for allowing us to use data originally studied by Law and Bate [16] in our research. Insightful comments by Prof. Alan Burns, Dr. Verônica Lima, Dr. Rob Davis and anonymous reviewers have also contributed to improve the quality of the paper. The first author has been funded by FASPESB (grant number 4932/2015) and CNPq (grant number 456193/2014-6). The second author has received support of the UK EPSRC Project MCCps (EP/P003664/1), EPSRC Research Data Management: No new primary data was created during this study.

References

[1] R. Wilhelm, J. Engblom, A. Ermedahl, N. Holsti, S. Thesing, D. Whalley, G. Bernat, C. Ferdinand, R. Heckmann, T. Mitra, F. Mueller, I. Puaut, P. Puschner, J. Staschulat, and P. Stenström, "The worst-case execution-time problem – overview of methods and survey of tools," *ACM Trans. Embed. Comput. Syst.*, vol. 7, no. 3, pp. 36:1–36:53, May 2008.

[2] D. Griffin and A. Burns, "Realism in statistical analysis of worst case execution times," in *Proc. of 10th Workshop on Worst-Case Execution Time Analysis*, 2010, pp. 44–53.

[3] G. Lima, D. Dias, and E. Barros, "Extreme value theory for estimating task execution time bounds: A careful look," in *Proc. of the 28th Euromicro Conference on Real-Time Systems*, 2016, pp. 200–211.

[4] E. Mezzetti, M. Ziccardi, T. Vardanega, J. Abella, E. Quiñones, and F. Cazorla, "Randomized caches can be pretty useful to hard real-time systems," *Leibniz Transactions on Embedded Systems*, vol. 2, no. 1, pp. 01–1–01:10, 2015.

[5] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, 3rd ed. Springer-Verlag, 2001.

[6] D. Dietrich, L. de Haan, and J. Hüsler, "Testing extreme value conditions," *Extremes*, vol. 9, pp. 75–85, 2002.

[7] J. Reineke, "Randomized caches considered harmful in hard real-time systems," *Leibniz Transactions on Embedded Systems*, vol. 1, no. 1, pp. 03–1–03:13, 2014.

[8] H. Al-Zoubi, A. Milenkovic, and M. Milenkovic, "Performance evaluation of cache replacement policies for the spec cpu2000 benchmark suite," in *Proc. of the 42nd Annual Southeast Regional Conference*, ser. ACM-SE 42. ACM, 2004, pp. 267–272.

[9] S. Altmeyer, L. Cucu-Grosjean, and R. I. Davis, "Static probabilistic timing analysis for real-time systems using random replacement caches," *Real-Time Systems*, vol. 51, no. 1, pp. 77–123, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11241-014-9218-4

[10] A. Jaleel, K. B. Theobald, S. C. Steely, Jr., and J. Emer, "High performance cache replacement using re-reference interval prediction (rrip)," in *Proc. of the 37th Annual International Symposium on Computer Architecture*, 2010, pp. 60–71.

[11] I. Bate and A. Burns, "An integrated approach to scheduling in safety-critical embedded control systems," *Real-Time Systems Journal*, vol. 25, no. 1, pp. 5–37, Jul 2003.

[12] B. Lincoln and A. Cervin, "Jitterbug: A tool for analysis of real-time control performance," in *Proceedings of the 41st IEEE Conference on Decision and Control*, 2002.

[13] R. Williams, *Real-Time-Systems-Development*. Butterworth-Heinemann, 2005.

[14] H. Kopetz, *Real-Time Systems Design for Distributed Embedded Applications*. Kluwer Academic Publishers, 1997.

[15] J. Gustafsson, A. Betts, A. Ermedahl, and B. Lisper, "The Mälardalen WCET benchmarks – past, present and future," in *Proc. of Workshop on Worst-Case Execution Time Analysis (WCET)*, 2010, pp. 137–147.

[16] S. Law and I. Bate, "Achieving appropriate test coverage for reliable measurement-based timing analysis," in *Proc. of 28th Euromicro Conference on Real-Time Systems*, 2016, pp. 189–199.

[17] A. Burns and S. Edgar, "Predicting computation time for advanced processor architectures," in *Proceedings 12th Euromicro Conference on Real-Time Systems. Euromicro RTS 2000*, 2000, pp. 89–96.

[18] S. Edgar and A. Burns, "Statistical analysis of WCET for scheduling," in *Proc. IEEE Real-Time Systems Symposium*, 2001, pp. 215 – 224.

[19] G. Bernat, A. Colin, and S. M. Petters, "WCET analysis of probabilistic hard real-time systems," in *Proc. of the 23rd IEEE Real-Time Systems Symposium*, 2002, pp. 279–288.

[20] G. Bernat, A. Colin, and S. Petters, "pWCET: A tool for probabilistic worst-case execution time analysis of real- time systems," University of York., Tech. Rep. YCS-2003-353, 2003.

[21] "Rapita Systems LTD," http://www.rapitasystems.com.

[22] K. Berezovskyi, F. Guet, L. Santinelli, K. Bletsas, and E. Tovar, "Measurement-based probabilistic timing analysis for graphics processor units," in *Proc. of the 29th International Conference on Architecture of Computing Systems*, F. Hannig, M. J. Cardoso, T. Pionteck, D. Fey, W. Schröder-Preikschat, and J. Teich, Eds. Springer International Publishing, 2016, pp. 223–236.

[23] L. Santinelli, J. Morio, G. Dufour, and D. Jacquemart, "On the sustainability of the extreme value theory for wcet estimation," in *14th International Workshop on Worst-Case Execution Time Analysis*, H. Falk, Ed., vol. 39, 2014, pp. 21–30.

[24] E. Gilleland and R. W. Katz, "New software to analyze how extremes change over time," *Eos*, vol. 92, no. 2, pp. 13–14, 2011.

[25] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org/

[26] A. Zempléni, "Goodness-of-fit test in extreme value applications," Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, München, Discussion paper 383, 2004.

[27] M. Ziccardi, E. Mezzetti, T. Vardanega, J. Abella, and F. J. Cazorla, "EPC: extended path coverage for measurement-based probabilistic timing analysis," in *Proceedings of the IEEE Real-Time Systems Symposium*, 2015, pp. 338–349.

[28] Y. Lu, T. Nolte, I. Bate, and L. Cucu-Grosjean, "A new way about using statistical analysis of worst-case execution times," *SIGBED Rev.*, vol. 8, no. 3, pp. 11–14, 2011.

[29] ——, "A statistical response-time analysis of real-time embedded systems," in *Proceedings of the 33rd Real-Time Systems Symposium*, 2012, pp. 351–362.

[30] P. Graydon and I. Bate, "Realistic safety cases for the timing of systems," *The Computer Journal*, vol. 57, no. 5, pp. 759–774, 2014.

[31] J. Hüsler and D. Li, "On testing extreme value conditions," *Extremes*, vol. 9, pp. 69–86, 2006.

[32] ——, "Testevc1d – r package," 2006. [Online]. Available: http://my.gl.fudan.edu.cn/teacherhome/lideyuan/pdf/TestEVC1d.txt