On vulnerabilities in EVT-based timing analysis: an experimental investigation on a multi-core architecture

Jamile Vasconcelos, George Lima, Marwan Wehaiba El Khazen, Adriana Gogonel, Liliana Cucu-Grosjean

On vulnerabilities in EVT-based timing analysis: an experimental investigation on a multi-more architecture

Jamile Vasconcelos (jamilevasconcelos@gmail.com)
Federal University of Bahia
George Lima
Federal University of Bahia
Marwan Wehaiba El Khazen
French Institute for Research in Computer Science and Automation
Adriana Gogonel
StatInf
Liliana Cucu-Grosjean
French Institute for Research in Computer Science and Automation

Research Article

Keywords: Measurement-Based Probabilistic Time Analysis (MBPTA), Extremes Value Theory (EVT), Real Time Systems (RTS), Worst Case Execution Time (WCET), Probabilistic WCET (pWCET)

Posted Date: July 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3157836/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Design Automation for Embedded Systems on October 17th, 2023. See the published version at https://doi.org/10.1007/s10617-023-09277-5.

On vulnerabilities in EVT-based timing analysis: an experimental investigation on a multi-more architecture

Jamile de Barros Vasconcelos[1*], George Lima[1*], Marwan Wehaiba El Khazen[2,3], Adriana Gogonel[3], Liliana Cucu-Grosjean[2]

[1]Computer Science Institute, UFBA, Avenida Milton Santos, s/n - Campus de Ondina, PAF 2, Salvador, 40.170-110, Bahia, Brazil.
[2]INRIA, 2 Rue Simone IFF, Paris, 75012, France.
[3]StatInf, 2 Rue Simone IFF, Paris, 75012, France.

*Corresponding author(s). E-mail(s): jamilevasconcelos@gmail.com; gmlima@ufba.br;
Contributing authors: marwan.wehaiba-el-khazen@inria.fr; adriana.gogonel@statinf.fr; liliana.cucu@inria.fr;

Abstract

Hardware architectures based on multiple cores, cache memory and branch prediction usually preclude the application of classical methods for determining execution time bounds for real-time tasks. As such bounds are fundamental in the designing of real-time systems, Measurement-Based Probabilistic Timing Analysis (MBPTA) has been employed. A common choice is the derivation of probabilistic Worst-Case Execution Time (pWCET) via the use of Extreme Value Theory (EVT), a branch of statistics used to estimate the probability of rare events that are more extreme than observations. However, pWCET estimations are usually reported in a controlled or simulated environment. In this paper we apply MBPTA in a real multi-core platform, namely Raspberry Pi 3B, taking into consideration possible interference due to operating system and concurrent activities. The results indicate that although EVT is effective, it does not always produce adequate models and coherent pWCET estimations. As MBPTA is primarily called for when classical methods are not applicable, as it is the case

for the studied platform, the results reported in this paper highlight risks and vulnerabilities when applying MBPTA-EVT for pWCET inference.*

Keywords: Measurement-Based Probabilistic Time Analysis (MBPTA), Extremes Value Theory (EVT), Real Time Systems (RTS), Worst Case Execution Time (WCET), Probabilistic WCET (pWCET)

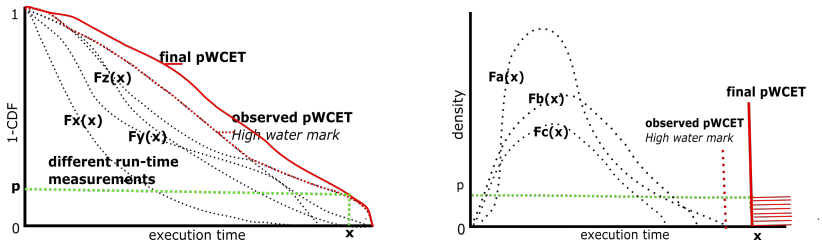
1 Introduction

Real-time systems (RTS) are those whose actions are subject to time constraints, usually defined in terms of deadlines associated with the tasks that perform the system actions. Critical RTS, such as avionics and space systems, must be designed such that all deadlines are met. Information on the worst-case execution time (WCET) for each task is thus needed. In simple architectures, these values of WCET are obtained via classical timing analysis [50], which establishes procedures based on modeling the task code control flow and the hardware characteristics for the worst-case execution scenarios. When complex hardware platforms are in place, however, classical timing analysis becomes more difficult since this type of environment exhibits unpredictable hardware and software behaviors, which are caused by memory cache, pipelines, branch prediction or multi-core processors. This difficulty has motivated the use of Measurement-Based Probabilistic Timing Analysis (MBPTA) [9, 13], supported by the statistical techniques based on Extreme Values Theory (EVT) [11]. The idea behind MBPTA is to collect a sufficiently large execution times sample for the task under analysis running on the target platform and then derive statistical models capable of representing the maximum execution time of a task. Through MBPTA-EVT, instead of estimating a single absolute worst-case value, a probability distribution for the maxima is obtained. This distribution is called probabilistic WCET (pWCET).

Fig 1 illustrates the concept of pWCET through two different graphical representations. Fig 1(a) represents a Complementary Cumulative Distribution Functions (CCDF or $(1 - CDF)$), also known as tail distribution or survival function, and (b) shows the respective Probability Density Functions (PDFs). The black dashed lines represent the distributions obtained from run-time measurements of a program under analysis. The red dashed line shows the observed pWCET distribution and it may not match the distribution for a specific execution scenario. The solid red line highlights the final pWCET distribution, which provides an upper bound for the probabilities of occurrence. The value x in the graph refers to the pWCET estimate for an exceedance probability of p [13].

The pWCET estimate can therefore be considered as a tuple (p, x), where x is the quantile of the distribution of maxima such that the probability of observing a value greater than x is not greater than p. That is, if X is a random variable that models the execution time of the analyzed task, p and x relate to each other as

*This work is an extended version of the conference paper that appeared in DOI: 10.1109/SBESC56799.2022 [48]



(a) 1-CDF (Adapted from [13])

(b) PDF

Fig. 1 Execution times and pWCET distributions of a program.

$p = \Pr\{\max(X) > x\}$, where the probability model for $\max(X)$ is constructed via EVT-based procedures.

Motivation and related work. Due to the relevance of offering execution time guarantees when RTSs make use of complex hardware platforms, there has been immense effort in the field of MBPTA-EVT. Basic concepts and related challenges to use EVT for pWCET inference have been reported [9, 10, 13, 20, 42]. Positive results have been obtained in simulated [32], single-core [4, 33, 41] and other platforms [5, 25]. Some publications have investigated problems and specific characteristics for applying EVT in the domain of timing analysis, such as determining sample sizes [44], issues of representativeness or problems related with dependencies in the collected samples [22, 37, 40]. In some cases, the impact of a specific platform were evaluated [2, 3, 5, 25]. There are results focused on evaluating choices between EVT models or specific distributions [8, 24, 34, 39]. Studies on other specific aspects have been reported, which include modeling dependencies via copulas [6, 36, 37], the use of Goodness-of-Fit (GoF) tests [12, 19, 38] or specific improvements [41].

Despite the large literature in the field, there has not been much attention given to reporting cases where MBPTA-EVT fails in real and complex hardware platforms. Such a kind of studies are reported for simulated environments [32] or based on theoretical grounds [20, 21]. One exception is a recent study [49] that proposes an alternative for EVT applying Markov's inequality using synthetic and real data.

Motivated by the perception that there is a lack of practical experiments that fully apply the MBPTA-EVT in real modern architectures, exposing situations where the technique fails, we propose an experimental study using a Raspberry Pi, version 3B, a multi-core and super-scalar board equipped with branch predictor and multilevel cache. This exhibits several characteristics that preclude the use of classical timing analysis and define scenarios where MBPTA is called for.

Contribution. Our study takes an implementation of 11 algorithms from Malärdalen Benchmark [23] and employs the prescribed EVT-based procedures. Observing the results obtained for these programs when subject to different sets of execution scenarios, we are capable of reporting both success and failure cases for the use of MBPTA-EVT. The main result of this paper is thus sending an alert to the community since we report scenarios where pWCET estimations may not be reliable in a real hardware platform, the kind of platform requiring alternative timing analysis methods.

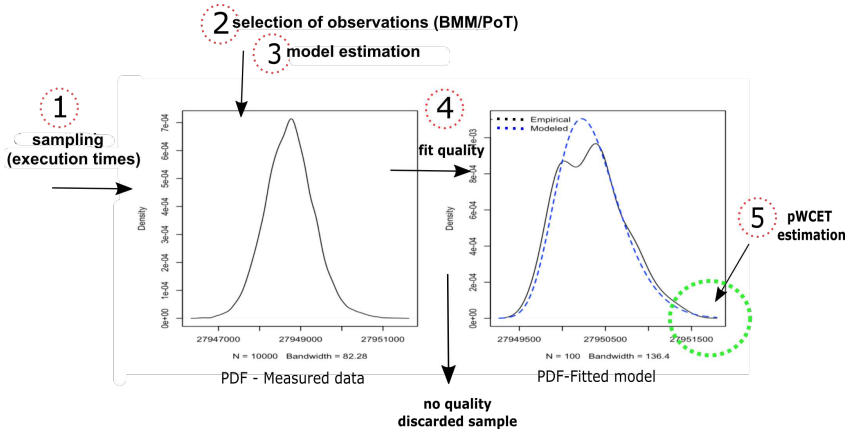


Fig. 2 Steps for applying EVT in MBPTA for pWCET inference.

Paper structure. The remainder of the paper is structured as follows. Section 2 explains basic concepts about EVT and explains how it is applied for the case of MBPTA. Results obtained are reported in Section 3. Section 4 highlights our conclusions.

2 Application of EVT to Timing Analyses

EVT is a branch of statistics widely used in areas such as financial market, meteorology and sports [18, 27, 43, 45]. While classical statistical theory prioritizes the behavior of a distribution around the central moments, the statistical theory of extremes is based on Fisher–Tippett–Gnedenko's Theorem of Extreme Types and models the distribution of the maximum of a sequence of n observations from a random variable X, namely $M_n = \max(X_1 ..., X_n)$. The technique is applied to generate extreme value models for rare events with low probability of occurrence [11].

The EVT pillar model takes $X_1, ..., X_N$, a sequence of N i.i.d. (independent and identically distributed) observations from a continuous random variable and builds a statistical model for its maximum. $(X_i)_{1 \leq i \leq N}$ usually represents the sequence of measurements done for the random variable of interest [11]. In order to model the distribution of the maximum of the observed data, five steps are usually applied, as illustrated in Fig. 2, and explained next for the task execution time of a program taken as the random variable X, such that the distribution of $\max(X)$ can be associated with the pWCET of that program.

2.1 Sampling

Initially the data collection is performed. In our case, the program is run and its execution time is measured in CPU clock cycles. Note that measuring execution time for a piece of code may produce different observations, even if its input is kept the same, due to expected variability from the platform. There are also a series of challenges that can directly affect the measurements.

Sample size. A small sample can cause an incorrect estimate of the parameters of the maximum's distribution. There are some techniques used to define this size [12, 44], but there is no consensus on what this appropriate size should be. Most of the time the sample size is stipulated empirically [32, 40]. In our study, a reasonably large sample size with $N = 10,000$ measurements sufficed for the initial results. Then, larger samples were used in order to evaluate the impact of sample size on the observed results.

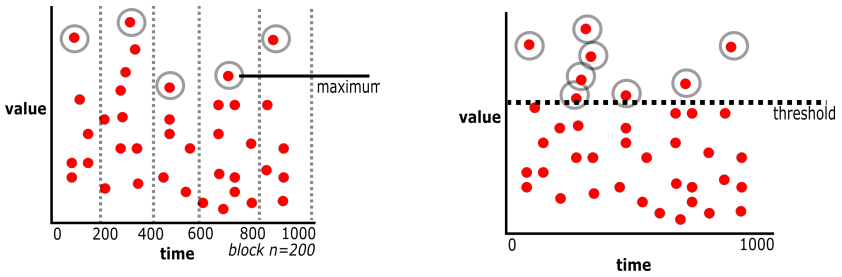
Representativeness. Representative samples are those that reflect the characteristics of the population from which they were extracted. This issue is still an open problem in MBPTA [20] as it is not possible to obtain a single measurement protocol that provides a representative sample of all possible operational scenarios [13]. The elements that can impact measurements for timing analysis are: the program's input data, the varied and possible program execution paths and the possible software and hardware states. Some authors state that a sample is representative if a given number of K collected observations produce a coherent pWCET estimate [35], although it is not clear what coherent estimation might mean. Representativeness issues were not addressed in this work.

Dependencies. Another challenge found during data collection is how to handle dependency relationships that can exist between different components of the system, which may appear as statistical dependency when sampling the data. Such a problem is frequent in RTS and can directly affect measured runtimes. It can happen in different cases, such as, for example, when a program composed of two blocks run in sequence has its execution times impacted by some common and unknown aspect [6], also for situations with execution resulting from operations such as multiplication and division. In this case, depending on the hardware architecture, the instruction may take a fixed time to execute, or it may depend on previous executions produced by sources like *cache* and *pipelines* [8].

For some authors [22, 29, 30, 40, 42] dependency relationships can be classified as: a) stationary, which implies identical distributions between random variables; b) short-range dependence, which focuses on the relationship between measurements that are close in time; and c) long-range, or extreme dependence, which shows a significant correlation in measurements distant in time. In this work, to make sure that an observed measurement does not depend on the previous ones, we start each measurement after resetting the platform.

Adequacy of observations. Some authors apply GoF tests to check if the observations comply with the i.i.d. criteria, such as Kolmogorov-Smirnoff [4, 12, 44], Anderson-Darling [2, 4], Wald-Wolfowitz [4, 12], Ljung-Box [2, 5] among others [40, 41]. However, this action is not mandatory and can be questionable, since there are situations where data meet the i.i.d. criteria but EVT does not properly work, as will be reported in this paper. It is also known that independence conditions can be relaxed, dependencies can be modified during the EVT application process, and it is possible to apply EVT to stationary samples [11].

A less discussed issue is the assumption that the observed distribution in fact belongs to the domain of attraction of an extreme value distribution in the first place,



(a) BMM (b) PoT

Fig. 3 EVT methods for maximums selection.

another requirement for EVT theorems to apply. On the theoretical side of the literature, it was recently shown in [31] that the set of distributions that belong to the domain of attraction of an extreme value distribution, although dense, is meagre in the topological sense. In practice, little is usually done to test the hypothesis on the actual sample. Indeed, two problems may arise: either the underlying distribution of the sample is not in the domain of attraction of any extreme value distribution, or the sample size is not large enough for the convergence to have succeeded.

As will be explained, we have applied the tests by Dietrich and Drees (described and compared in [28]) in order to test this assumption. This procedure were coupled with the standard recommendations based on analyzing GoF visually based on some key graphs like the QQ plots [11].

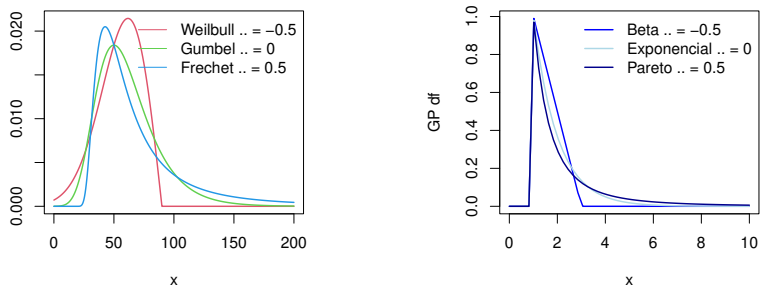
2.2 Selection of observations

A sample of maxima must be obtained from the sample of observed measurements. Two methods can be applied: the Block Maxima Method (BMM) or Peaks over Threshold (PoT) method.

The BMM consists of creating a new sample from the values of the original sample divided into non-overlapping blocks of equal sizes, selecting the maximum value from each block (Fig. 3(a)). Then, EVT procedures estimate the distribution that best explains (or fits) the data given by this sample of maxima. Indeed, the EVT theorems ensure, under certain hypotheses, that the sample's limit distribution belongs to the GEV distribution family, which is determined by three parameters (location - μ, scale - σ, shape - ξ). The shape parameter (also known as the extreme value index) is the one that dictates the behavior of the tail. It can be a Weibull ($\xi < 0$), a Gumbel ($\xi = 0$) or a Fréchet ($\xi > 0$) distribution. The GEV model is represented by the CDF of Eq. (1) and Fig. 4(a) exemplifies the GEV family.

$$F(x) = \begin{cases} exp\left[-(1+\xi\frac{x-\mu}{\sigma})^{\frac{-1}{\xi}}\right] & \text{for } \xi \neq 0 \\ exp\left[-exp(-\frac{x-\mu}{\sigma})\right] & \text{for } \xi = 0 \end{cases} \tag{1}$$

For the PoT method, the sample of maxima is created from observations that exceed a certain predefined threshold (Fig. 3(b)). This sample can be modeled by a member of the GPD family of distributions, which are also defined in terms of three



(a) GEV (b) GPD

Fig. 4 EVT Distribution Families.

parameters, location - μ, scale - σ and shape - ξ, the latter of which represents a Beta distribution ($\xi < 0$), an Exponential distribution ($\xi = 0$) or a Pareto distribution ($\xi > 0$). The GPD model is represented by the CDF of Eq. (2). The location μ can be set to 0 for the distribution of exceedances $Y = X - u > 0$, with u a large enough threshold. Fig. 4(b) exemplifies the GPD family.

$$F(x) = \begin{cases} 1 - [1 + \xi(\frac{x-\mu}{\sigma})]^{-\frac{1}{\xi}} & \text{for } \xi \neq 0 \\ 1 - exp\left(-\frac{x-\mu}{\sigma}\right) & \text{for } \xi = 0 \end{cases} \tag{2}$$

Proper choices for block size, in the case of the BMM, and threshold value, in the PoT, are fundamental for the effective application of the EVT methods [11].

No systematic method exists to set these hyper-parameters, and usually heuristics are used, which is in part what the authors in [49] criticize EVT for. In our work, we used a wide range of block sizes/thresholds and did the GEV/GPD estimations for all of them, before analyzing the quality of the estimations and testing for all the hypotheses needed for the theorems to apply.

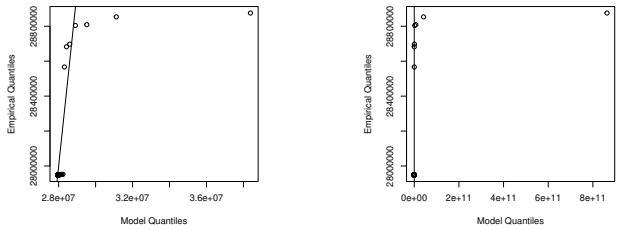
2.3 Model estimation

In this third step, a range of block sizes (resp. thresholds) is chosen, and the parameters of the distribution for GEV (resp. GPD) are estimated for every block size (resp. threshold) before checking the quality of the estimations in the next step. There are several methods for this.

The Maximum Likelihood Estimation (MLE) is the most used method and it estimates the parameters that have the highest probability to produce the observed data. Nonetheless this method may not work when $\xi < -0.5$ [11]. In these cases, another method can be used. In this work we apply the L-Moment (LM) estimator [26].

The LM estimator is based on probability weighted moments and linear combinations of order statistics. It can be used to estimate the parameters of the extreme value distribution. It does not present convergence problems, but it can produce non-symmetrical confidence intervals (CI) and it demands support from computational

[t]



(a) Block $n = 50$ (b) Block $n = 100$

Fig. 5 Failed attempt to fit a GEV model with parameters estimated via MLE for Sample 01.

methods (such as Bootstrap [16]) to estimate CIs. Additionally, the LM estimator is only valid when $\xi < 1$. Therefore, both the MLE and the LM estimator are needed to cover all real numbers for the value of the shape.

2.4 Fit quality checking

In step 4, the model quality, either for GEV or for GPD, is checked. Normally, this process is carried out by visual inspections of some plots, one of which is the Quantile-Quantile (QQ) plot. Through the QQ plot, it is possible to visualize the linear regression of the distribution and observe the values of the empirical quantile against those from the obtained model. If the result follows a straight line, then it has a good fit.

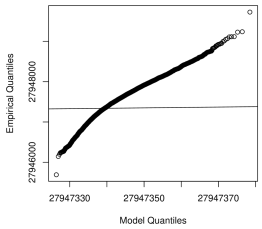
Two samples are taken from our experiments (Section 3) as illustration for our explanation in this and the next sections. They are called Sample 01 and Sample 02, each with $10,000$ observations. The R[1] software and `extRemes`[2] package were the tools used in this work for EVT-based analysis.

Fig. 5 shows two failed GEV fit attempts for Sample 01, for models obtained using block sizes $n = 50$ and $n = 100$. It can be seem that the data completely deviates from the regression line, indicating problems in the quality of the fit for both models.

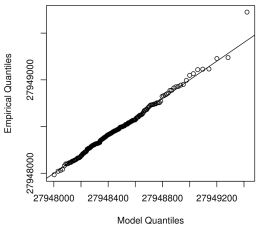
Four Sample 02 GEV fit attempts are presented in Fig. 6. Fig. 6(a) shows the result for a model created without specifying the block size. It can be seen that it is not possible to obtain a good fit. The model in Fig. 6(b) has the block size set to $n = 50$ and a good fit quality can be observed. In Fig. 6(c) and Fig. 6(d), it is possible to perceive the good fit quality for both models (block size set to $n = 100$), but visually a little inferior to the quality obtained in Fig. 6(b). It can also be observed that one model had the parameters estimated via MLE and the another via LM, and that these results show similar behaviors, with very subtle differences, as expected.

GPD models obtained via the PoT method (Fig. 7) produced similar results as those observed for BMM: no fitting for Sample 01 and good fit quality for Sample 02. This suggests a failure of EVT for Sample 01 and a success for Sample 02.

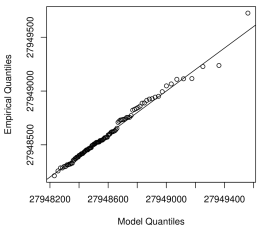
[1] Available at https://cran.r-project.org/.
[2] Further information at https://cran.r-project.org/web/packages/extRemes/extRemes.pdf.



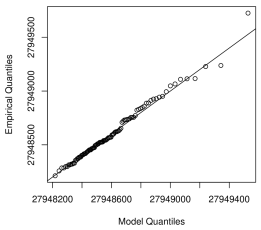
(a) Failed - Parameters via MLE but no block size defined



(b) Fitted - Parameters via MLE and Block $n = 50$

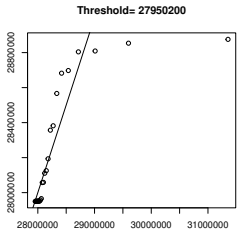


(c) Fitted - Parameters via MLE and Block $n = 100$

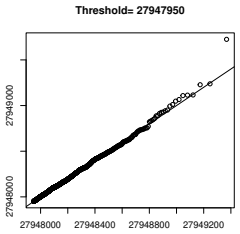


(d) Fitted - Parameters via LM and Block $n = 100$

Fig. 6 GEV fit attempts with parameters estimated via MLE and LM for Sample 02.



(a) Failed - Sample 01



(b) Fitted - Sample 02

Fig. 7 GPD fitting attempts.

Despite the results observed, a single QQ plot does not show a convergence of the distribution of the maximum, it only shows that it happens to be close to an extreme value distribution for a given block size or threshold. Showing convergence requires testing the more constraining hypothesis that the distribution F of the observations is in the *domain of attraction* of some extreme value distribution.

Domain of attraction

Additional tools were used in this work to check the EVT fit quality wrt. the convergence of the maximum distribution: the tests by Dietrich et al. [14] and Drees et al. [15]. The former is available in the `TestEVC1d` R-package and both were implemented by us in Python, following the directions in [28] closely. The purpose of applying the tests was to compare the results between them and with the results of the conventional method of checking QQ plots visually.

The Dietrich and Drees tests check if a distribution F is indeed in the domain of attraction of an extreme value distribution, a hypothesis we may call $\mathcal{H}_0$ required for the EVT theorems to apply. We write $\mathcal{H}_0$: $F \in \mathcal{D}_\xi$ for some $\xi \in \mathbb{R}$. In their conclusion, the authors in [28] specify that the Dietrich test is more suitable for negative shapes ($\xi < 0$) and the Drees test is better for positive shapes ($\xi \geq 0$). This leads us to naturally couple the Dietrich test with the LM estimator (valid for $\xi < 1$), and the Drees test with the MLE ($\xi > -0.5$).

From [28], we give here the two test statistics we implemented in Python, following the instructions given by the authors. First, we extract the $k+1$ largest values of the sample of size N. These are the order statistics, denoted $X_{N-k,N} \leq X_{N-k+1,N} \leq ... \leq X_{N-1,N} \leq X_{N,N}$. Then, the test statistics E_N for the Dietrich test and T_N for the Drees test are given by:

$$E_N = k \int_0^1 \left(\frac{\log\left(X_{N-\lfloor kt \rfloor,N}\right) - \log\left(X_{N-k,N}\right)}{\hat{\xi}_+} - \frac{t^{-\hat{\xi}_-} - 1}{\hat{\xi}_-} \left(1 - \hat{\xi}_-\right) \right)^2 t^\eta dt,$$

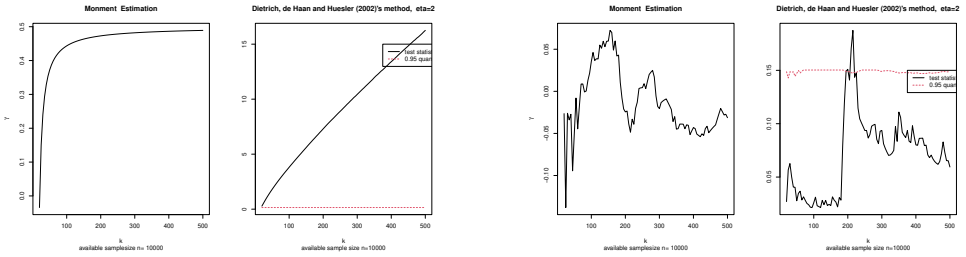
where the authors suggest favoring it for negative shapes ($\xi < 0$), for $\eta = 2$, and where $\hat{\xi}_- = 1 - \frac{1}{2}\left(1 - \frac{\left(M_{k,N}^{(1)}\right)^2}{M_{k,N}^{(2)}}\right)^{-1}$, $\hat{\xi}_+ = M_{k,N}^{(1)}$, for $M_{k,N}^{(j)} = \frac{1}{k}\sum_{i=0}^{k} \left(\log\left(X_{N-i,N}\right) - \log\left(X_{N-k,N}\right)\right)^j$, and

$$T_N = k \int_0^1 \left(\frac{N}{k} \overline{F}_N \left(\hat{a}_{\frac{N}{k}} \frac{x^{-\hat{\xi}} - 1}{\hat{\xi}} + \hat{b}_{\frac{N}{k}} \right) - x \right)^2 x^{\eta-2} dx,$$

where $\eta = 1$ is suggested and positive shapes are favored ($\xi \geq 0$), and where $\hat{\xi}$, $\hat{a}_{\frac{N}{k}}$, and $\hat{b}_{\frac{N}{k}}$ are the MLE estimates for shape, scale and location, respectively.

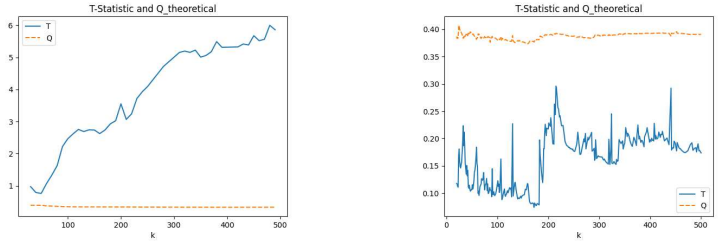
Two last ingredients are required. First, in order to evaluate these integrals, we approximate them with their discrete versions, by dividing the interval $[0, 1]$ uniformly into $M = 100,000$ parts, as recommended in [28]. And second, for the theoretical quantiles, we rely on the tables 1 and 3 (provided in their paper) and use the same linear interpolation method for quantiles that are not given in the tables. Equipped with these statistics and their quantiles under $\mathcal{H}_0$, we apply them to our samples.

In the case of Sample 01 (Fig. 8(a)), it was observed that the results for the Dietrich test are located above the values of the quantile that represents the 95% confidence limit, shown by the red dashed line. For Sample 02 (Fig. 8(b)), the values appear below this limit, confirming the suitability of modeling the sample according to an extreme



(a) Failed - Sample 01 (b) Fitted - Sample 02

Fig. 8 Verification of the fit quality through the Dietrich test.



(a) Failed - Sample 01 (b) Fitted - Sample 02

Fig. 9 Verification of the fit quality through the Drees test.

value distribution. This test reiterates what was observed in the previous subsection with QQ plots.

Similarly, the Drees test that we implemented, the same can be observed for both samples. The T-statistic (blue curve) of the test can be seen to be largely above the 95% theoretical quantile (dashed orange line) for Sample 01 in Fig. 9(a), and well below it in the case of Sample 02 in Fig. 9(b), for values of k ranging from 10 to 500. This means that both tests show that a good fit for the Sample 02 is possible (failure to reject $\mathcal{H}_0$ for Sample 02) and an impossibility for analyzing Sample 01 via EVT (we rejected $\mathcal{H}_0$ for Sample 01).

During our study, samples such as Sample 01 were discarded whereas those similar to Sample 02 were subject to the remaining EVT-based analysis steps.

Procedure for GoF

For every sample, a range of values of k were used, from 10 to 500, in order to apply both tests, and locate a potential region of values of k for which the hypothesis that the sample's distribution belongs to the domain of attraction of an extreme value distribution cannot be rejected.

When no such range of values could be found, the sample was discarded. The QQ plots were still investigated in these cases to compare their results with those of the statistical tests. If a range of valid values for k could be obtained, the next step was to

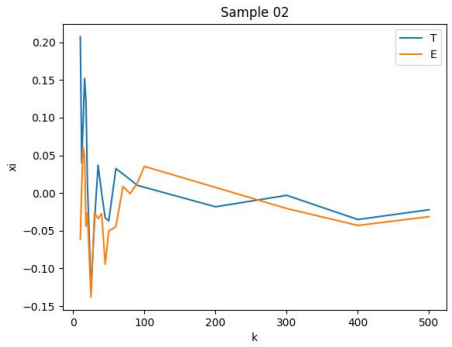


Fig. 10 Estimation of the shape 'xi' in terms of k for Sample 02. The 'T' curve (blue) is the estimation allowed by the Drees test, while the 'E' curve (orange) is the one permitted by the Dietrich test.

Table 1 GEV parameters estimated via MLE and LM for Sample 02 Block $n = 50$ model.

GEV model	**location** μ	**scale** σ	**shape** ξ
MLE Bl. $n = 50$	2.794835e+07	2.242563e+02	-4.109445e-02
LM Bl. $n = 50$	2.794863e+07	2.300160e+02	-5.361614e-02

visually evaluate two types of graphs, one pertaining to the stability of the estimated shape wrt. the choice of k, as in Fig. 10, and the other is the conventional QQ plot. To estimate these, we used the LM estimator for the values of k permitted by the Dietrich test and the MLE for those given by the Drees test. A sample was discarded if no stability in the estimation of the shape was observed, or if the QQ plot showed too much discrepancy. Otherwise, the use of EVT was deemed sound.

In the case of Sample 01 and Sample 02, we have shown the QQ plots and the results of the two statistical tests, and those have already led us to discard Sample 01. In Fig. 10, we show the evolution of the shape estimation as k varies, to drive the point further as to its stability that shows confidence in EVT for this sample.

Finally, with the QQ plots, it was possible to visually compare the fitting quality for the obtained GEV and GPD models for Sample 02 (Figs. 6(b), 6(c), 6(d) and 7(b)), the GEV model seemed to be better fitted. Furthermore, the Block $n = 50$ model seemed to present the best fit among the other tested block size. As a result, for estimating pWCET, the GEV model with Block size of $n = 50$ was favored for this sample, corresponding to 200 extreme values among the $10,000$ observations, which suggests a value of $k = 200$. And indeed, this value is well situated in the graph of Fig 10 and indicates a stable and robust shape close to 0.

Table 1 shows the estimated GEV parameters. As can be seen, the maxima for Sample 02 is distributed according to a Weibull distribution ($\xi = -0.04109445$ via MLE and $\xi = -0.05361614$ via LM). This model will be used in the explanation given in the next section.

Table 2 Return Level (CI 95%) for the GEV model with Block $n = 50$ and parameters via MLE for Sample 02.

Exc. Prob.	95% lower CI	Estimate	95% upper CI
10^{-5}	27948734	27950689	27952644
10^{-6}	27948327	27950996	27953664
10^{-7}	27947835	27951275	27954715
10^{-8}	27947276	27951529	27955783

Table 3 Return Level (CI 95%) for the GEV model with Block $n = 50$ and parameters via LM for Sample 02.

Exc. Prob.	95% lower CI	Estimate	95% upper CI
10^{-5}	27951434	27952277	27953843
10^{-6}	27951464	27952438	27954480
10^{-7}	27951486	27952559	27955101
10^{-8}	27951500	27952649	27955739

2.5 pWCET estimation

After obtaining a model and confirming its goodness of fit, the model can be used for pWCET estimations. Based on the model and on an exceedance probability of interest, the analyst can estimate the corresponding quantile and its associated confidence interval. Tables 2 and 3 give estimated pWCET values and 95%-confidence intervals for different exceedance probabilities via MLE and LM, respectively. Extreme quantiles in the EVT terminology are referred to as return levels.

Taking the exceedance probability of $p = 10^{-5}$ the estimated extreme quantile is $x = 27950689$ clock cycles (from Table 2). That is, it is expected that the observed maximum exceeds x with probability p. Note that this is in line with what has been measured and what predicts an Weibull model (bounded tail), as the maximum measured value in Sample 02 was 27949725, representing a difference of 964 clock cycles.

3 Experiments

In this section we present the results of our experiments obtained by applying the procedures explained in the previous section. The codes implementing 11 benchmarks taken from the Mälardalen WCET Benchmarks [23], are the object of our study. The following benchmarks were evaluated: Insertion Sort, Bubble Sort, Merge Sort, Quick Sort, Cnt, Edn, Fft1, Sqrt, Binary Search, Fibcall and Matmult. This set of programs brings well-structured codes that include operations like nested loops, integer comparisons, bit, float points and arrays manipulation, and array and matrix calculations, which makes it a good case study.

Our experimental environment consisted of a Raspberry Pi board (version 3B) with the Raspbian GNU/Linux 9 OS. This board is a reasonably complex piece of hardware, equipped with several components that bring timing uncertainties such as pipeline, branch prediction and 2-level cache memory. Our focus is not on the timeliness of the system and hence we did not address possible issues related with

real-time capabilities at the OS level. Our main interest is on possible corner scenarios generated by a reasonably complex system that may compromise reliable statistical inference by MBPTA-EVT.

Execution time was measured in clock cycles via a Performance Monitoring Unit (PMU). To do so, we used the Linux performance analysis tool *linux_perf* (v. 4.9.82). For more accurate measurements, the `perf_event_open` system call was used instead of command lines. Despite this, measurement errors are expected to occur, which usually correspond to a few executed instructions excesses.

The collected data were modeled by two random variables: `Cycles`, representing the total number of CPU clock cycles processed during a run, and `Ins`, representing the number of executed retired instructions.

To compose the experiment environment, different scenarios (Tab. 4) were created taking into consideration the variability of three resources: Ethernet, Wifi and Core. The states of network interfaces, Ethernet and WiFi, were set to either On or Off, while the core on which the analyzed program runs could be set to 0 (the same as the OS's core) or 3 (dedicated to the program).

These parameters give rise to 8 possible scenarios, namely F01-F08, and five samples were collected for each scenario, per program. These samples were named **Regular Samples** and were the result of $10,000$ measurements.

In order to observe the experiment in larger samples, five samples with 100k and five with 500k observations were also collected, for two or three scenarios per benchmark. They were named **Extra Samples** and the choice of the extra measurement scenarios was defined based on the results observed in the EVT application in the Regular Samples. In some cases, scenarios where it was observed a high number of good fits were chosen, in others, scenarios with samples with no good fits, and, in others, an arbitrary choice was made.

Hence, a total volume of 700 samples of 10k, 100k or 500k were measured, producing a total amount of $82,400$k observations collected and analyzed.

All programs had fixed input data (single path) to ensure that runtimes were only impacted by the environment. They were grouped into four groups, organized by some of their characteristics [47]:

- Group 01: data memory access for load and write operations (Insert sort, Bubble sort, Merge sort, and Quick sort)
- Group 02: calculations (Fft1, Sqrt, and Fibcall), intensive memory usage
- Group 03: calculations (Compress, Cnt, Edn, and Matmult)
- Group 04: low memory access and computation (Fibcall and Binary Search).

In the first experiment, carried out with Bubble Sort, we first checked whether the i.i.d. requirement holds. To do so, we applied the Kolmogorov Smirnoff (KS) test as follows. From a sample of measurements, other samples were randomly generated with size $m = 150$, for a test power of 99%, and with size $m = 90$, for 95% [7]. This re-sampling process was carried out $10,000$ times for each test power and then KS was applied comparing pairs of generated samples. The test did not indicate that the data collected for all measurements followed different distributions. Further, as each

Table 4 Scenarios created by varying the Ethernet, Wifi and core resources.

	Ethernet		Wifi		Core	
Scenario	On	Off	On	Off	Default (0)	Dedicated (3)
F01		X	X		X	
F02	X		X			X
F03	X		X		X	
F04	X			X	X	
F05		X		X	X	
F06		X	X			X
F07	X			X		X
F08		X		X		X

measurement was carried out independently of the previous run, dependency relations in the observation are not expected to occur.

3.1 Results

The EVT application was carried out according to the steps presented in Section 2. Analyzing data and comparing all the results obtained throughout the experiment, it can be highlighted that:

- The presence of outliers can be noted in some samples. However, the impossibility of employing EVT on this kind of data could not be credited to outliers. Even when outliers were removed from the sample, model fitting was not possible. That is, the underlining distribution seems to be the source of the impossibility of model fitting.
- Increasing the sample size did not increase the amount of good EVT fits. In some cases, larger samples produced more good fits, in others, less, and in others, no change was observed. This indicates that the optimal sample size may be subject to the input data, nature and complexity of the executed code and of the impacts that the environment may cause during the measurement.
- In most cases, when comparing different samples measured for a same scenario, those that presented observations with a lower number of `Cycles` were able to obtain a good EVT fit, while those with the highest numbers could not. There is no relationship between the number of `Cycles` and the possibility or not of obtaining a good fit, however, a lower value of `Cycles` may indicate that the measurement was less exposed to environmental impact, leaving it a more stable and more easily EVT-adjustable distribution.
- The use of performance counters using the `linux_perf` showed to be efficient for the measurements. The low rate of the coefficient of variation (CV) of the variable `Ins` (between 0.00% and 0.19% for all benchmarks) demonstrates that no excessive counts in the number of instructions were taken.
- The stability of `Ins` and the determinism of the executed codes indicate that the oscillation of `Cycles` is due to the impact of the environment.

With respect to the amount of good EVT fits obtained, Tables 5 and 6 present the percentages for the Regular Samples and Tables 7 and 8 for the Extra Samples.

Table 5 Regular Samples - Scenarios that yield samples with good fit, classification by benchmark and group. A sample was considered to have a good fit when at least one of its EVT models (GEV or GPD, via MLE or LM) showed visual suitability in the QQ plot, passed all the statistical tests and produced a coherent pWCET estimation.

Group	Benchmark	Scenario							
		F01	**F02**	**F03**	**F04**	**F05**	**F06**	**F07**	**F08**
01	Insertion Sort					X			X
	Merge Sort				X	X	X	X	X
	Quick Sort	X			X				
	Bubble Sort					X			X
02	Fibcall		X	X	X	X		X	X
	Sqrt								
	Fft1		X				X	X	
03	Edn								
	Cnt	X	X	X	X	X	X	X	X
	Matmult								
04	Binary Search	X	X		X		X		X
	Fibcall		X	X	X	X		X	X

Observing these data considering a single occurrence of *Fibcall*, present in two groups, it can be stated that:

- Despite groupings based on the benchmarks' characteristic similarities, no pattern was observed for EVT good-fit behavior within any group. The only possible exception was for the benchmarks Insertion Sort, Merge Sort and Bubble Sort, from Group 01, which obtained a higher EVT good-fit percentage for scenarios F05 and F08, and no good fit for the others.
- The creation of scenarios varying the Ethernet, Wifi and Core resources was partially effective in impacting measurements. The oscillation of the Core for On and Off did not seem to have caused variations in the `Cycles` of the measurements. However, the measurements set in scenarios F05 and F08 (Ethernet and Wifi Off), produced results with a greater number of good EVT fits. This indicates impacts on measurement produced by Ethernet and Wifi On, even if they are not being requested by the benchmark in execution. This fact, however, was not observed for all tested benchmarks.
- The results show the percentages of good fit observed through the visual analysis of the QQ plot, and confirmed through the application of the Dietrich and Drees tests. They demonstrate very low rates of EVT good fit in general, which indicates unsatisfactory results considering the environment, benchmarks and tested scenarios.

Finally, a low average of good EVT fits was observed in the experiment as a whole. For the Regular Samples, considering all scenarios, the average was 16%. This goes to 40% if we consider only the F05 and F08 scenarios (Ethernet and Wifi Off), however, this represent only 25% of the experiment carried out in the Regular Samples. Tests with the Extra Samples did not show high good-fit percentages either.

The results reinforce that it is not always possible to fit a model using EVT for a given benchmark-platform set, and strengthens the warning about risks and vulnerabilities when applying the technique without performing the proper procedures and fitting quality tests.

4 Conclusion

EVT is an effective statistics branch that has been used to (and indeed can) provide consistent results for estimating probabilistic upper bounds on execution time. However, its theory states that there may be distributions for which EVT is not applicable. As the distributions that come out from the measurements are arbitrary, one must be careful for not validating samples that are not EVT-compliant. In the performed experiment a total of 575 out of 700 samples (with 10k, 100k or 500k observations each) were considered not suitable for EVT. We found that keeping Ethernet or Wifi interfaces on during measurements was a source of problem. This suggests that some background activity related to Ethernet and Wifi could be a cause. Nonetheless, invalid samples were also found in scenarios where Ethernet and Wifi were Off. In fact, in complex platforms, many activities are not controllable. As MBPTA-EVT is mostly called for when dealing with such platforms, extra care when applying EVT is needed.

There are several other aspects not considered in this work. Assessing the meaning of the exceedance probability w.r.t. the analyzed program is an open problem. Data collected from measurements are not necessarily related to data that will be produced when the system is operational and so checking for representativeness is needed. Before getting into these kind of more abstract problems, further investigations into the application of EVT procedures in MBPTA are necessary. The issues reported in this paper, taking into consideration other platform configurations and different coded programs are immediate future research steps. The results we presented can be seen as an alert message, which will certainly motivate deeper investigation in the field.

Acknowledgments. The authors would like to thank Dr. Verônica Lima and MSc. Tadeu Nogueira for their support.

Declarations.

This work have been partially funded by CAPES (Brazil), grant no. 001, by the Inria-UFBA Associated Teams Program under the Kepler project, by the FR ANRT Joint CIFRE Inria and StatInf, grant CIFRE no. 1072_/2020 and by BPI France under the FR PSPC-regions - 2021 STARTREC project. The authors declare they have no financial interests, neither do they have any affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. All data used in the research can be made available under request.

References

[1] The LINUX man-pages. https://man7.org/linux/man-pages/man2/perf_event_open.2.html, 2021 (Accessed in May, 2022).

[2] J. Abella, D. Hardy, I. Puaut, E. Quinones, and F. J. Cazorla. On the comparison of deterministic and probabilistic WCET estimation techniques. Proceedings - Euromicro Conference on Real-Time Systems, pages 266–275, 2014.

[3] J. Abella, E. Quiñones, F. Wartel, T. Vardanega, and F. J Cazorla. Heart of gold: Making the improbable happen to increase confidence in MBPTA. In 2014 26th Euromicro Conference on Real-Time Systems, pages 255–265. IEEE, 2014.

[4] L. F. Arcaro, K. Palma Silva, and R. Silva De Oliveira. On the reliability and tightness of GP and Exponential models for probabilistic WCET estimation. ACM Transactions on Design Automation of Electronic Systems (TODAES), 23(3):1–27, 2018.

[5] K. Berezovskyi, L. Santinelli, K. Bletsas, and E. Tovar. WCET measurement-based and extreme value theory characterisation of CUDA kernels. In Proceedings of the 22nd International Conference on Real-Time Networks and Systems, pages 279–288, 2014.

[6] G. Bernat, A. Burns, and M. Newby. Probabilistic timing analysis: An approach using copulas. Journal of Embedded Computing, 1(2):179–194, 2005.

[7] B. M. Boyerinas. Determining the statistical power of the kolmogorov-smirnov and anderson-darling goodness-of-fit tests via monte carlo simulation. Technical report, Center of Naval Analyses Arlington United States, 2016.

[8] A. Burns and S. Edgar. Predicting computation time for advanced processor architectures. Proceedings - Euromicro Conference on Real-Time Systems, pages 89–96, 2000.

[9] F. J Cazorla, L. Kosmidis, E. Mezzetti, C. Hernandez, J. Abella, and T. Vardanega. Probabilistic Worst-Wase Timing Analysis: Taxonomy and Somprehensive Survey. ACM Computing Surveys (CSUR), 52(1):1–35, 2019.

[10] F. J Cazorla, T. Vardanega, E. Quiñones, and J. Abella. Upper-bounding program execution time with extreme value theory. In 13th International Workshop on Worst-Case Execution Time Analysis. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[11] S. Coles. An Introduction to Statistical Modeling of Extreme Values. Springer, London, 2001.

[12] L. Cucu-Grosjean, L. Santinelli, M. Houston, C. Lo, T. Vardanega, L. Kosmidis, J. Abella, E. Mezzetti, E. Quiñones, and F J. Cazorla. Measurement-Based Probabilistic Timing Analysis for Multi-path Programs. Euromicro Conference on Real-Time Systems, pages 91–101, 2012.

[13] R. Davis and L. Cucu-Grosjean. A Survey of Probabilistic Timing Analysis Techniques for Real-Time Systems. LITES: Leibniz Transactions on Embedded Systems, 6(1):1–53, 2019.

[14] D. Dietrich, L. De Haan, and J. Husler. Testing Extreme Value conditions. Extremes, 5(1):71, 2002.

[15] Drees, H., De Haan, L. & Li, D. Approximations to the tail empirical distribution function with application to testing extreme value conditions. Journal Of Statistical Planning And Inference. **136**, 3498-3538 (2006,10), https://linkinghub.elsevier.com/retrieve/pii/S0378375805000753

[16] B. Efron and R. J Tibshirani. An introduction to the Bootstrap. CRC press, 1994.

[17] P; Francis-Mezger and V. M Weaver. A Raspberry Pi Operating System for Exploring Advanced Memory System Concepts. In Proceedings of the International Symposium on Memory Systems, pages 354–364, 2018.

[18] T. Cristina Soares Garção. Avaliação empírica do risco de mercado: estimação do Value-at-risk pela Teoria dos Valores Extremos. PhD thesis, 2017.

[19] M Garrido and J Diebolt. The ET Test, a Goodness-of-fit Test for the Distribution Tail. Methodology, Practice and Inference, second international conference on mathematical methods in reliability, (September 2007):427–430, 2007.

[20] S. Jiménez Gil, I. Bate, G. Lima, L. Santinelli, A. Gogonel, and L. Cucu-Grosjean. Open Challenges for Probabilistic Measurement-Based Worst-Case Execution Time. IEEE Embedded Systems Letters, 9(3):69–72, 2017.

[21] D. Griffin and A. Burns. Realism in Statistical Analysis of Worst Case Execution Times. OpenAccess Series in Informatics, 15(Wcet):44–53, 2010.

[22] F. Guet, L. Santinelli, and J. Morio. On the Representativity of Execution Time Measurements: Studying Dependence and Multi-Mode Tasks. In 17th International Workshop on Worst-Case Execution Time Analysis (WCET 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[23] J. Gustafsson, A. Betts, A. Ermedahl, and B. Lisper. The Mälardalen WCET Benchmarks: Past, Present and Future. In 10th International Workshop on Worst-Case Execution Time Analysis (WCET 2010). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2010.

[24] J. Hansen, S. Hissam, and G. A. Moreno. Statistical-Based WCET Estimation and Validation. OpenAccess Series in Informatics, 10:1–11, 2009.

[25] R. Heckmann, M. Langenbach, S. Thesing, and R. Wilhelm. The influence of processor architecture on the design and the results of WCET tools. volume 91, pages 1038–1054. IEEE, 2003.

[26] Hosking, J. L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. Journal Of The Royal Statistical Society. Series B (Methodological). **52**, 105-124 (1990), http://www.jstor.org/stable/2345653

[27] W. K Huang, M. L Stein, D. J McInerney, S. Sun, and E. J Moyer. Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions. Advances in Statistical Climatology, Meteorology and Oceanography, 2(1):79–103, 2016.

[28] Hüsler, J. & Li, D. On testing extreme value conditions. Extremes. **9**, 69-86 (2006,11), http://link.springer.com/10.1007/s10687-006-0025-8

[29] M Ross Leadbetter, G. Lindgren, and H. Rootzén. Conditions for the convergence in distribution of maxima of stationary normal processes. Stochastic Processes and their Applications, 8(2):131–139, 1978.

[30] M. R Leadbetter. On a Basis for 'Peaks over Threshold' Modeling. Statistics & Probability Letters, 12(4):357–362, 1991.

[31] Leonetti, P. & Chokami, A. The maximum domain of attraction of multivariate extreme value distributions is small. Electronic Communications In Probability. **27** (2022,1)

[32] G. Lima, D. Dias, and E. Barros. Extreme Value Theory for Estimating Task Execution Time Bounds: A Careful Look. Proceedings - Euromicro Conference on Real-Time Systems, 2016-August:200–211, 2016.

[33] Y. Lu, T. Nolte, I. Bate, and L. Cucu-Grosjean. A New Way about Using Statistical Analysis of Worst-Case Execution Times. ACM SIGBED Review, 8(3):11–14, 2011.

[34] C. Maxim, A. Gogonel, D. Maxim and L. Cucu-Grosjean. HAL Id: hal-00766063 Estimation of Probabilistic Minimum Inter-arrival Times Using Extreme Value Theory. 2013.

[35] D. Maxim, F. Soboczenski, I. Bate, and E. Tovar. Study of the Reliability of Statistical Timing Analysis for Real-time Systems. In Proceedings of the 23rd International Conference on Real Time and Networks Systems, pages 55–64, 2015.

[36] A. Melani, E. Noulard, and L. Santinelli. Learning from Probabilities: Dependences within Real-time Systems. In 2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA), pages 1–8. IEEE, 2013.

[37] V. Nélis, P. Meumeu Yomsi, L. M. Pinho, and G. Bernat. Another look at the pWCET estimation problem. 2014.

[38] Ç. Özari, Ö. Eren, and H. Saygin. A New Methodology for the Block Maxima Approach in Selecting the Optimal Block Size. Technical Gazette, 26(5):1292–1296, 2019.

[39] F. Reghenzani, G. Massari, and W. Fornaciari. The Misconception of Exponential Tail Upper-Bounding in Probabilistic Real Time. IEEE Embedded Systems Letters, 11(3):77–80, 2018.

[40] F. Reghenzani, G. Massari, and W. Fornaciari. Probabilistic-WCET reliability: On the experimental validation of EVT hypotheses. Microprocessors and Microsystems, 77:103135, 2020.

[41] L. Santinelli and Z. Guo. On the Criticality of Probabilistic Worst-Case Execution Time Models. In International Symposium on Dependable Software Engineering: Theories, Tools, and Applications, pages 59–74. Springer, 2017.

[42] 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. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

[43] D. Silva, F. Caeiro, and M. Oliveira. Modelação estatística de valores extremos: probabilidades de excedência, quantis extremais, limite superior do suporte e níveis de retorno no lançamento do disco do atletismo. Revista da Estatística da Universidade Federal de Ouro Preto, 7, 2018.

[44] K. Palma Silva, L. F. Arcaro, and R. Silva De Oliveira. On using GEV or Gumbel models when applying EVT for probabilistic WCET estimation. In 2017 IEEE Real-Time Systems Symposium (RTSS), pages 220–230. IEEE, 2017.

[45] H. Spearing, J. Tawn, D. Irons, T. Paulden, and G. Bennett. Ranking, and other properties, of elite swimmers using extreme value theory. Journal of the Royal Statistical Society: Series A (Statistics in Society), 184(1):368–395, 2021.

[46] V. M Weaver and S. A McKee. Can Hardware Performance Counters be Trusted? In 2008 IEEE International Symposium on Workload Characterization, pages 141–150. IEEE, 2008.

[47] A. Tadeu and G. Lima. On the Selection of Relevant Hardware Events for Explaining Execution Time Behavior. In: 2021 IEEE XI Brazilian Symposium on Computing Systems Engineering (SBESC), pages 1-8. IEEE 2021.

[48] V. Jamile and G. Lima. Possible risks with EVT-based timing analysis: an experimental study on a multi-core platform. In: 2022 IEEE XII Brazilian Symposium on Computing Systems Engineering (SBESC), pages 1-8. IEEE 2022.

[49] Vilardell, S., Serra, I., Mezzetti, E., Abella, J., Cazorla, F. & Castillo, J. Using Markov's Inequality with Power-Of-k Function for Probabilistic WCET Estimation. 34th Euromicro Conference On Real-Time Systems (ECRTS 2022). **231** pp. 20:1-20:24 (2022), https://drops.dagstuhl.de/opus/volltexte/2022/16337

[50] 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., 7(3):53, 2008.

Table 6 Regular Samples - Percentage of samples with good EVT fit by scenario, benchmark and EVT method, expressed in %, where T1 represents good GEV fit and T2 good GPD fit. A sample was considered to have a good fit when at least one of its EVT models (GEV or GPD, via MLE or LM) showed visual suitability in the QQ plot, passed all the statistical tests and produced a coherent pWCET estimation.

		F01		**F02**		**F03**		**F04**		**F05**		**F06**		**F07**		**F08**	
Group	***Benchmark***	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**
01	*Insert S.*	0	0	0	0	0	0	0	0	100	60	0	0	0	0	100	80
	Merge S.	0	0	0	0	0	0	20	20	100	100	20	20	20	20	100	100
	Quick S,	20	20	0	0	0	0	20	20	0	0	0	0	0	0	0	0
	Bubble S.	0	0	0	0	0	0	0	0	80	40	0	0	0	0	40	40
	Good fit	**5**	**5**	**0**	**0**	**0**	**0**	**10**	**10**	**70**	**50**	**5**	**5**	**5**	**5**	**60**	**55**
02	*Fibcall*	0	0	20	20	20	20	40	40	100	100	0	0	20	20	100	80
	Sqrt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Fft1	0	0	20	20	0	0	0	0	0	0	40	40	0	20	0	0
	Good fit	**0**	**0**	**13**	**13**	**7**	**7**	**13**	**13**	**33**	**33**	**13**	**13**	**7**	**13**	**33**	**27**
03	*Edn*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cnt	40	20	20	20	0	20	0	20	100	100	20	20	0	20	100	100
	Matmult	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Good fit	**13**	**7**	**7**	**7**	**0**	**7**	**0**	**7**	**33**	**33**	**7**	**7**	**0**	**7**	**33**	**33**
04	*B. Search*	20	20	40	40	0	0	20	20	0	0	40	40	0	0	20	20
	Fibcall	0	0	20	20	20	20	40	40	100	100	0	0	20	20	100	80
	Good fit	**10**	**10**	**30**	**30**	**10**	**10**	**30**	**30**	**50**	**50**	**20**	**20**	**10**	**10**	**60**	**50**
Total good fit		**7**	**5**	**9**	**9**	**2**	**4**	**9**	**11**	**44**	**36**	**11**	**11**	**4**	**7**	**42**	**38**

Table 7 Extra Samples with 100k observations - Percentage of samples with good EVT fit by scenario, benchmark and EVT method, expressed in %, where T1 represents good GEV fit and T2 good GPD fit. A sample was considered to have a good fit when at least one of its EVT models (GEV or GPD, via MLE or LM) showed visual suitability in the QQ plot, passed all the statistical tests and produced a coherent pWCET estimation.

		F01		**F02**		**F03**		**F04**		**F05**		**F06**		**F07**		**F08**	
Group	***Benchmark***	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
01	*Insertion Sort*	0	0	-	-	-	-	-	-	0	40	-	-	-	-	-	-
	Merge Sort	0	0	0	0	-	-	-	-	80	20	-	-	-	-	-	-
	Quick Sort	0	0	-	-	-	-	-	-	40	80	-	-	-	-	20	40
	Bubble Sort	0	0	-	-	-	-	-	-	0	0	-	-	-	-	-	-
02	*Fibcall*	0	0	-	-	-	-	-	-	60	60	0	0	-	-	-	-
	Sqrt	0	0	-	-	-	-	-	-	-	-	-	-	-	-	60	60
	Fft1	0	0	-	-	-	-	-	-	20	20	0	0	-	-	-	-
03	*Edn*	-	-	0	0	-	-	-	-	-	-	-	-	-	-	20	20
	Cnt	-	-	-	-	0	0	-	-	80	20	-	-	-	-	-	-
	Matmult	-	-	-	-	-	-	-	-	40	40	-	-	-	-	0	0
04	*Binary Search*	40	40	-	-	-	-	-	-	-	-	-	-	-	-	0	20
	Fibcall	0	0	-	-	-	-	-	-	60	60	0	0	-	-	-	-

Table 8 Extra Samples with 500k observations - Percentage of samples with good EVT fit by scenario, benchmark and EVT method, expressed in %, where T1 represents good GEV fit and T2 good GPD fit. A sample was considered to have a good fit when at least one of its EVT models (GEV or GPD, via MLE or LM) showed visual suitability in the QQ plot, passed all the statistical tests and produced a coherent pWCET estimation.

		F01		**F02**		**F03**		**F04**		**F05**		**F06**		**F07**		**F08**	
Group	***Benchmark***	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
01	*Insertion Sort*	0	0	-	-	-	-	-	-	0	0	-	-	-	-	-	-
	Merge Sort	0	0	0	0	-	-	-	-	0	0	-	-	-	-	-	-
	Quick Sort	0	0	-	-	-	-	-	-	60	60	-	-	-	-	0	0
	Bubble Sort	0	0	-	-	-	-	-	-	0	0	-	-	-	-	-	-
02	*Fibcall*	0	0	-	-	-	-	-	-	0	0	0	0	-	-	-	-
	Sqrt	40	40	-	-	-	-	-	-	-	-	-	-	-	-	100	100
	Fft1	0	0	-	-	-	-	-	-	40	40	0	0	-	-	-	-
03	*Edn*	-	-	0	0	-	-	-	-	-	-	-	-	-	-	20	40
	Matmult	-	-	-	-	-	-	-	-	0	0	-	-	-	-	0	0
04	*Binary Search*	20	20	-	-	-	-	-	-	-	-	-	-	-	-	20	20
	Fibcall	0	0	-	-	-	-	-	-	0	0	0	0	-	-	-	-