Dealing with uncertainty in pWCET estimations

FEDERICO REGHENZANI, Politecnico di Milano
LUCA SANTINELLI, ONERA and Airbus Defence and Space
WILLIAM FORNACIARI, Politecnico di Milano

The problem of estimating a tight and safe Worst-Case Execution Time (WCET), needed for certification in safety-critical environment, is a challenging problem for modern embedded systems. A possible solution proposed in last years is to exploit statistical tools to obtain a probability distribution of the WCET. These probabilistic real-time analyses for WCET are however subject to errors, even when all the applicability hypotheses are satisfied and verified. This is caused by the uncertainties of the probabilistic-WCET distribution estimator. This article aims at improving the measurement-based probabilistic timing analysis approach providing some techniques to analyze and deal with such uncertainties. The so-called region of acceptance model based on state-of-the-art statistical test procedures is defined over the distribution space parameters. From this model, a set of strategies is derived and discussed, to provide the methodology to deal with the trade-off safety/tightness of the WCET estimation. These techniques are then tested over real datasets, including industrial safety-critical applications, to show the increased value of using the proposed approach in probabilistic WCET analyses.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; **Real-time systems**; • **Computing methodologies** → *Uncertainty quantification*; • **Hardware** → *Statistical timing analysis*;

Additional Key Words and Phrases: pWCET, probabilistic real-time, embedded systems

ACM Reference format:
Federico Reghenzani, Luca Santinelli, and William Fornaciari. 2020. Dealing with uncertainty in pWCET estimations. *ACM Trans. Embedd. Comput. Syst.* 0, 0, Article 0 (2020), 24 pages.
https://doi.org/10.1145/3396234

1 INTRODUCTION

In recent years the increasing computational power demand of applications leads to the evolution of computing platforms towards complex processor architectures and sophisticated system components. To overcome the single-core performance barrier, today's processor manufactures have introduced several advanced features like multi-/many-core, complex pipelines, multi-level caches, memory prefetcher, and many others. Unfortunately, this makes the problem of computing the Worst-Case Execution Time (WCET) with traditional static timing analyses extremely difficult [27], thus limiting the use of modern architectures in some classes of embedded systems, e.g. in safety-critical systems. To reduce the design cost of embedded systems, industry has recently looked at Commercial-Off-The-Shelf (COTS) platforms. The use of COTS components in real-time applications is challenging and adds another layer of complexity in WCET estimation, due to the numerous sources of unpredictability of these platforms [12]. In fact, COTS platforms are built

1539-9087/2020/0-ART0 $15.00
https://doi.org/10.1145/3396234

with average performance in mind and not intended to provide an execution time upper-bound, needed by safety-critical embedded systems.

To obtain a valid and tight WCET estimation, traditional static analyses rely on the detailed knowledge of the hardware and on the task control flow graph. The WCET estimation process requires the exploration of all the input and processor state spaces to identify the worst-case scenario. When the analysis is performed in modern architectures, this process usually demands an infeasible quantity of computational power to carry out the result in a reasonable time, due to the aforementioned complexity issues. Alternatively, considerable approximations are introduced to reduce the exploration space and to make the estimation process feasible, however this leads to an extremely pessimistic, and therefore unusable, WCET result.

To cope with the traditional WCET static analysis problems, measurement-based approaches have been proposed [35]. The main advantage of such techniques is that they do not require an accurate model of the hardware and of the workload, thus overcoming the issues of static timing analysis. The task execution time is measured several times across different inputs and processor states. Then a deterministic or probabilistic analysis is performed on the sampled time trace. In particular, the probabilistic approaches have been recently developed with the help of the Extreme Value Theory (EVT), that is the cornerstone of probabilistic real-time computing [4] [18]. This statistical theory is a well-assessed mathematical proved method used to model the *extreme behaviour* of a statistical distribution, i.e. the values in the tails of that distribution. Common applications of this theory include financial risk assessment and natural disaster prediction. EVT is briefly described later in Section 2.1. The overall process that exploits EVT to obtain a probabilistic distribution of the WCET (probabilistic Worst-Case Execution TIme - pWCET) is called Measurement-Based Probabilistic Timing Analyses (MBPTA). However, measurement-based methods are affected by several issues intrinsic of the estimation phase. Among them, the *input representativity* is still a substantial open problem: how can we be sure that the inputs we provide to our system are able to gather representative time measurements to correctly describe the WCET behaviour? In this work, we voluntarily omit this aspect which is approached in other lines of work [1, 2, 22]. We assume to explore a representative set of inputs and, consequently, that the time measurements are representative of the real execution time. Instead, we focus on the uncertainties that affect the parameters of the estimated EVT distribution and on how they impact the quality of the WCET estimation. The omit made is to be able to focus on other aspects of MBPTA, and push forward its maturity in different needed directions.

Contributions. Even assuming that all the hypotheses of MBPTA are true, the estimation of the probability distribution of WCET is naturally subject to uncertainty. This is due to the fact that the number of samples used to estimate the distribution of execution times is necessarily finite. This paper wants to deal with this uncertainty, proposing a methodology to study the errors affecting the probabilistic WCET estimation and to accordingly deal with the trade-off safety/tightness. To the best of our knowledge, this problem has not been adequately tackled by the probabilistic real-time community yet and it lacks of a precise mathematical formulation. Please note that this paper is not a "statistical work" in the sense that we do not advance any statistical theory. Instead, we focus on how already available and well-assessed statistical concepts can be applied to the pWCET problem, and how they can affect the pWCET reliability from a real-time point of view.

Organization of the paper. After presenting the necessary background (Section 2), we systematically define the space of uncertainties with its relative properties (Section 3). Then, methods to upper-bound and lower-bound these uncertainties are proposed together with the analysis on their effects on WCET (Section 4). By exploiting the defined mathematical tools, we propose some strategies to deal with the tradeoff safety/tightness of the WCET (Section 5). Finally, the proposed tools are tested on industrial datasets to show their benefits when used with MBPTA (Section 6).

1.1 State of the art

In the last decade, several works on both the theoretical and the practical aspects of probabilistic real-time have been published. A couple of comprehensive surveys has been recently published [13] [5]. The probabilistic approach is conceived in 2001-2002 by Edgar et al. [18], and Bernat et al. [4]. From that moment onward, although the methodology has improved substantially, several challenging problems still remain open [22]. Some of these challenges are in common between deterministic and probabilistic measurement-based methods and others are shared by static analyses. The categorization among static and measurement-based, deterministic and probabilistic analyses has been provided by Abella et al. [1], focusing on certifiability aspects.

Concerning the MBPTA approach, that is the subject of this work, a general overview of the methods has been presented in [10] and [44]. Initially, the trend has been to propose randomized architectures to fulfill the EVT requirements [6, 29]; MBPTA has also applied to some industrial case studies [19, 48] and probabilistic-energy estimation [39]. Lately, a generalization of the EVT approach has been proposed [30] with the relaxing of some of the EVT requirements; the focus here is on the EVT applicability, i.e. the satisfaction of EVT hypotheses, to improve the pWCET reliability in realistic real-time systems[1]. A selection of statistical tests to verify these hypotheses has been proposed in [40]. A recent work [14] described how MBPTA and epistemic variability are related and how they impact on pWCET. The work most similar to ours is [46], in which the authors build a confidence region on the EVT distribution. As subsequently discussed in Section 3.4, the authors tried to select the best distribution model through an empirical evaluation. Despite having some practical applications, its empirical nature makes necessary more rigorous mathematical formulations and discussions. The paper of Civit et al. [7] is another example contribution seeking for the best distribution to represent pWCETs. In opposition to that, our approach is not limited to the exponential version of the pWCET and exploits statistical testing to increase the final pWCET reliability.

2 BACKGROUND

This section aims at providing the reader the necessary background on extreme statistics, how it could be applied to derive a probabilistic-WCET (pWCET), and which are the reliability implications of using such statistical techniques for pWCET estimations.

2.1 Extreme value theory

The EVT has been proposed at the beginning of the 20th century, to overcome the limits of the well-known Central Limit Theorem. This theorem provides indeed information on the mean value of the distribution and no information can be inferred on the tail values, i.e. the extreme values. Given a sequence of independent and identically distributed (i.i.d.) random variables $\mathscr{X}_1, \mathscr{X}_2, ..., \mathscr{X}_n$, the EVT process looks for the distribution of the following cumulative distribution function (cdf): $F(x) = 1 - P(x > \max(\mathscr{X}_1, \mathscr{X}_2, ..., \mathscr{X}_n))$. This distribution converges to well-known forms, independently on the original distribution of random variables $\mathscr{X}_i$. The fundamental cornerstone of this statistical theory is the following theorem:

Theorem 2.1 (Fisher-Tippett-Gnedenko theorem [20] [23]). *There exist two constants a_n and b_n such that:* $\lim_{n\to\infty} F(a_n x + b_n) = G(x)$, *where $G(x)$ is the cdf of the extreme value distribution $\mathcal{G}$ that can assume only three forms: the Gumbel, the Weibull, or the Fréchet distribution.*

[1]With realistic real-time systems it is intended real-time systems that are "real": implemented and used in industrial application. This in opposition to unrealistic real-time systems which most of the time are assumed for theoretical analyses/results.

In the '70, it has been proved [32] that these three distributions can be generalized in one single form called Generalized Extreme Value (GEV) distribution:

$$G(x) = \begin{cases} e^{-e^{\frac{x-\mu}{\sigma}}} & \xi = 0 \\ e^{-[1+\xi(\frac{x-\mu}{\sigma})]^{-1/\xi}} & \xi \neq 0 \end{cases}$$

The GEV distribution has three parameters: the *location* μ, the *scale* σ and the *shape* ξ. The sign of ξ determines the distribution class: if $\xi > 0$, the GEV converges to the Fréchet distribution; if $\xi < 0$, it converges to the Weibull distribution; and if $\xi \to 0$, it converges to the Gumbel distribution.

In order to estimate the GEV distribution parameters, it is possible to use the Block-Maxima (BM) approach: the input data are filtered in order to obtain the significant measures for the distribution tail only. In particular, selecting a block size B we define the following sequence of m random variables $\mathscr{X} = \left\{ \mathscr{X}_1^{BM}, \mathscr{X}_2^{BM}, ..., \mathscr{X}_{n/B}^{BM} \right\}$ where $\mathscr{X}_k^{BM} = \max\left(\mathscr{X}_{B(k-1)+1}, ..., \mathscr{X}_{B(k-1)+B} \right)$. Using the sequence of maxima $\mathscr{X}$ it is possible to run any well-known estimator, e.g. the Maximum Likelihood Estimator, to obtain the estimation of the GEV parameters (μ, σ, ξ).

For completeness, it is necessary to cite the alternative method to Block-Maxima to estimate the extreme distribution: Peak-over-Threshold (PoT). This approach applies a filter on the input measurements based on a predefined threshold u: $\mathscr{X} = \{\mathscr{X}_i \text{ s.t. } \mathscr{X}_i > u, \forall i\}$. This *set of exceedances* can be used to estimate a different distribution called *Generalized Pareto Distribution (GPD)*. GEV and GPD are asymptotically equivalent. In this work we consider only the GEV distribution class, however the proposed analyses and methods are still valid and general enough to be applicable to GPD distributions as well.

2.2 Probabilistic-WCET

In probabilistic WCET estimation with measurement-based approaches, the EVT is exploited to estimate the WCET by using a sequence of measurements of the task execution time. The input of the EVT estimation process, i.e. the realization of previously defined random variables $\mathscr{X}_1, \mathscr{X}_2, ..., \mathscr{X}_n$, is the sequence of the measured execution times of a real-time task.

The output of probabilistic approaches is not a single WCET value, rather it is a statistical distribution and, in particular, the distribution which results from the EVT process. This distribution is the so-called probabilistic-WCET (pWCET) and it relates the multiple worst-case values with their associated probability of being exceeded. The pWCET is a generalization of the classical deterministic WCET as a distribution able to upper bound task execution timing behaviors [14] [15].

The pWCET is usually expressed with its complementary cumulative distribution function (ccdf):

$$p = 1 - F(\overline{\text{WCET}}) = 1 - P(\mathscr{X} \leq \overline{\text{WCET}}) = P(\mathscr{X} > \overline{\text{WCET}})$$

where $F(x)$ is the cumulative distribution function (cdf). It is possible to use this notation, chosen a $\overline{\text{WCET}}$ value, and to obtain the probability of violation p, i.e. the probability of observing execution times larger than $\overline{\text{WCET}}$. Alternatively, it is possible to compute the WCET at a given probability of violation $\overline{p}$ by using the notation $\text{WCET} = F'(\overline{p})$, where $F'(\cdot)$ is the inverse complementary cumulative distribution function (iccdf).

2.3 EVT hypotheses and pWCET reliability

The EVT requires to fulfill some requirements in order to produce a valid estimation of the distribution tail. In particular, two main theoretical hypotheses must be satisfied: the fact that the

random variables are independent and identically distributed (i.i.d.)[2] and the Maximum Domain of Attraction (MDA). In probabilistic real-time computing, the i.i.d. hypothesis on execution times is mainly influenced by the processor state space and by the presence of a multi-path control flow graph in the task under analysis. Regarding the MDA, this hypothesis requires that the distribution of the input measurements has to be in the domain of attraction of one of the three extreme value distributions of a GEV or GPD. It is harder to find a direct relationship of MDA with the computing system. The MDA hypothesis is satisfied for the large majority of continuous distributions while this is not true for the discrete distribution [43], e.g. the Poisson distribution is not in the MDA of any extreme value distribution form.

In addition to these two hypotheses, the input representativity problem still represents the major barrier to the use of probabilistic real-time in safety-critical systems. This article assumes i.i.d. and input representativity to be valid and, consequently, the EVT method as applicable. We focus on the MDA hypothesis and on how the Goodness-of-Fit (GoF) tests can help in improving the estimated pWCET distribution. The goal of GoF tests is to check the validity of the MDA hypothesis, i.e. they verify if the estimated distribution $G(x)$ fits the $\mathscr{X}$ set or not. A GoF test allows us to detect ill-formed distributions, e.g. due to the violation of MDA hypothesis, an error during the estimation routine, a wrong selection of the block size B, or a too small sample size. The most commonly used GoF tests for EVT distributions are the Chi-Squared (CS), the Kolmogorov-Smirnov (KS), the Cramer-von Mises (CvM) and the Anderson-Darling (AD) ones [26]. To obtain the *reject*/*non-reject* result, a statistical test usually computes a value from the data, called *statistic*, and compares it against a tabular data called *critical value*. When the chosen test rejects the null hypothesis, the estimated distribution $\mathcal{G}$ is not valid and the analysis must stop. In this case, in fact, the obtained pWCET is not representative of the real WCET distribution.

3 REGION OF ACCEPTANCE

The estimation of the extreme value distribution $\mathcal{G}$, which results from the EVT process, is naturally subject to errors: the necessary condition to obtain the exact distribution for any estimator algorithm is to have infinite measurements, that is clearly not realistic. To this extent, the estimator routine provides us the GEV parameters tuple $(\overline{\mu}, \overline{\sigma}, \overline{\xi})$ that can be rewritten as $(\mu^{\circledast} + \epsilon_\mu, \sigma^{\circledast} + \epsilon_\sigma, \xi^{\circledast} + \epsilon_\xi)$, where $(\mu^{\circledast}, \sigma^{\circledast}, \xi^{\circledast})$ is the exact unknown distribution and the symbols $(\epsilon_\mu, \epsilon_\sigma, \epsilon_\xi)$ represent the unknown errors in our estimation. The goal of the GoF tests previously described is to detect these uncertainties and to reject the estimated distribution when the errors are excessively high. However, because of the finite number of measurements, the GoF test is also imperfect, i.e. it is not able to reject the distributions when $(\epsilon_\mu, \epsilon_\sigma, \epsilon_\xi)$ are low enough to be undetected. For this reason, it exists a multi-dimensional cloud of points in the GEV (or GPD) parameters space that represents the distributions which are not rejected by the GoF test. In this section we explore this region and how its statistical properties affect the reliability of our pWCET estimation.

3.1 Definitions and basic concepts

From now on, we assume that we have already performed the EVT estimation of the pWCET distribution: the input set of the execution time measurements has been filtered by the BM approach to obtain a set $\mathscr{X}$, from which we have estimated the GEV distribution $\mathcal{G}$. We often refer to this set as the *time trace* of the execution time measurements.

Before formally defining the uncertainties of the GEV parameters space, we specify the following helper function:

[2]The i.i.d. hypothesis can be relaxed less strict hypotheses [44], however this discussion is out of scope of this paper.

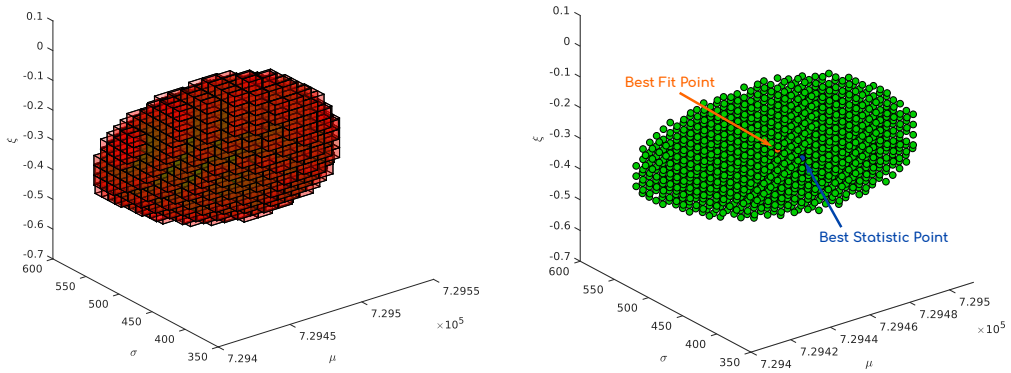


Fig. 1. The Region of Acceptance $R(\mathscr{X})$ plotted for a real set of time measurements $\mathscr{X}$: (a) by color representing the value of statistics $D(\cdot)$ (color scale red-green: the red color identifies values near the critical value, the green color identifies – in the middle of the region – values far form it); (b) by green dots where the $T(\cdot) = 0$ and the highlighted BFP and BSP points.

Definition 3.1 (Test result function). Given a certain statistical test identified by its statistic function $D(\cdot)$, a time trace $\mathscr{X}$, an estimated distribution $\mathcal{G}$, and the critical value[3] CV, its *test result function* is defined as:

$$T(\mathscr{X}, \mathcal{G}) := \begin{cases} 1 & \text{if } D(\mathscr{X}, \mathcal{G}) \geq CV(\mathcal{G}) \\ 0 & \text{if } D(\mathscr{X}, \mathcal{G}) < CV(\mathcal{G}). \end{cases} \quad \triangleleft$$

This definition is a formal notation to state the result of a statistical test: the rejection of the null hypothesis ($T(\mathscr{X}, \mathcal{G}) = 1$) – i.e. the values $\mathscr{X}$ show a strong evidence that have not been drawn from the distribution $\mathcal{G}$ – or the non-rejection of the null hypothesis ($T(\mathscr{X}, \mathcal{G}) = 0$) – i.e. the values $\mathscr{X}$ have been *probably* drawn from the distribution $\mathcal{G}$. This definition can be used to identify the region of points in the parameter space of the GEV distributions for which the test accepts the null hypothesis. Note that this is an abuse of the common notation of hypothesis testing. In statistics we never say that a statistical test *accepts* the null hypothesis, rather we say that it is not able to reject it. In this paper we abuse the *accepts* notation for clarity purposes. More details on this are available in Section 3.4.

By exploiting the test result function, we can formally define the three-dimensional cloud of points in the GEV parameters space where the test accepts the distributions:

Definition 3.2 (Region of Acceptance). Given a time trace $\mathscr{X}$, the Region of Acceptance for a GEV distribution of a statistical test with test result function T is the cloud of points R:

$$R(\mathscr{X}) := \{(\mu, \sigma, \xi) \in \mathbb{R}^3 : T(\mathscr{X}, (\mu, \sigma, \xi)) = 0\}. \quad \triangleleft$$

A visual example of the Region of Acceptance is depicted in Figure 1. To shorten the notation, we sometimes avoid to write the measurements parameter $R = R(\mathscr{X})$. In the same parameters space, it is possible to identify the tuple $(\overline{\mu}, \overline{\sigma}, \overline{\xi})$ as the point that represents the output of the EVT estimator. We call this point the Best Fit Point (BFP). This point may or may not be inside the region R, i.e. it may be accepted ($T(\cdot) = 0$) or rejected by the GoF test ($T(\cdot) = 1$). If this point is outside the region, the estimator fails to provide a valid distribution. The region may even not exist,

[3]For some GoF statistical tests, the critical value depends on the reference distribution (e.g. the Anderson-Darling test). We write it as a function of the reference distribution $CV(\mathcal{G})$ or simply CV.

e.g. when the original distribution is not in the domain-of-attraction of any GEV distribution. It is worth reminding that the GoF test has a *false positive rate* that is equivalent to the chosen level of significance α, i.e. the GoF test wrongly rejects a distribution with α probability. In a probabilistic real-time scenario, this means that with α probability a valid pWCET distribution is mistakenly rejected. This has however safe consequences, since this is a failure condition for the EVT analysis that does not produce any pWCET distribution.

From now on, we consider this point BFP as part of the region, i.e. we assume to have estimated a distribution $\mathcal{G}$ that successfully passed the statistical testing procedure. We will discuss this further in the experimental evaluation (Section 6), when we obtain a BFP rejected by the GoF test during the analysis of a real dataset. According to this assumption and Definition 3.1, the test assigns a statistic value $D(\mathscr{X}, (\overline{\mu}, \overline{\sigma}, \overline{\xi}))$ to the BFP, that is lower than the critical value CV. In general, the BFP point does not correspond to the point with the minimum statistic value provided by the test. In particular, we can define the following point as:

Definition 3.3 (Best Statistic Point, BSP). Given a time trace $\mathscr{X}$, the Best Statistic Point for a statistical test with statistic $D(\cdot)$ is:

$$(\mu^*, \sigma^*, \xi^*) := \arg\min_{\mu,\sigma,\xi} D(\mathscr{X}, (\mu, \sigma, \xi)). \qquad \triangleleft$$

Examples of BFP and BSP points are depicted in Figure 1b. According to the previous assumption, the region R has at least one point, i.e. the BFP. Thus the BSP point always exists with $D(\mathscr{X}, (\mu, \sigma, \xi)) < CV$. Before proceeding with the region analysis, we define the following property of the statistic of a statistical test:

Definition 3.4 (Correct statistic). Given a sample $\mathscr{X}$ of size n drawn from a distribution $\mathcal{A}$, we say that a statistic $D(\mathscr{X}, \mathcal{A})$ of a given statistical test is correct iff $D(\mathscr{X}, \mathcal{A}) \to \overline{K}$ for $n \to \infty$ with $\overline{K} \in \mathbb{R}$ and $D(\mathscr{X}, \mathcal{A}) \geq \overline{K}$ for any finite value of n. $\triangleleft$

To put it less formal, a statistic is correct if, when applied to the exact distribution of samples and having a sample of infinite size, it provides the minimal possible value (e.g. $D = 0$ in KS test). This property and the well-known *consistent estimator* property enable the following asymptotic result:

Lemma 3.5. *If the estimator is* consistent *and the statistic computed by the statistical test is* correct*, then both the best fit point and the best statistic point converge to the real unknown pWCET distribution point* $(\mu^{\circledast}, \sigma^{\circledast}, \xi^{\circledast})$*:*

$$(\overline{\mu}, \overline{\sigma}, \overline{\xi}) \to (\mu^{\circledast}, \sigma^{\circledast}, \xi^{\circledast}) \quad n \to \infty;$$
$$(\mu^*, \sigma^*, \xi^*) \to (\mu^{\circledast}, \sigma^{\circledast}, \xi^{\circledast}) \quad n \to \infty$$

where n is the size of the set $\mathscr{X}$ used for training or testing the pWCET distribution. $\triangleleft$

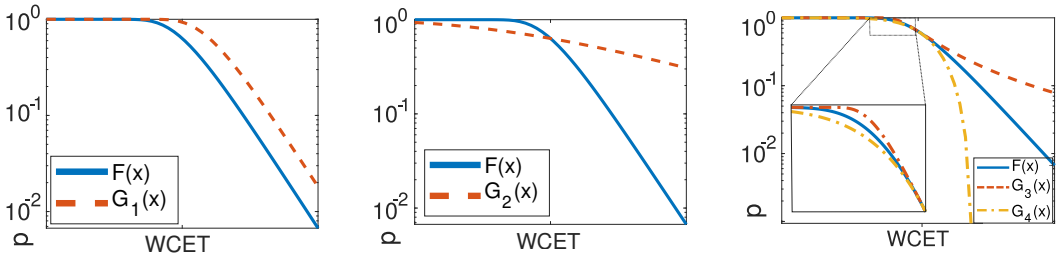
Proof. This result is an immediate consequence of the definitions of *consistent* estimator and *correct* statistic of the test. $\square$

This asymptotic result can be exploited to derive the following theorem.

Theorem 3.6. *Given a time trace $\mathscr{X}$, if the statistic of the considered statistical test is correct, the exact true distribution $P^{\circledast}$ is inside the acceptance region R:*

$$(\mu^{\circledast}, \sigma^{\circledast}, \xi^{\circledast}) \in R. \qquad \triangleleft$$

Proof. Let be n the size of the set $\mathscr{X}$ and $P^*_n \in R$ the best statistic point of Definition 3.3. We provide this proof by contradiction. Let assume that $P^{\circledast} \notin R$. It follows that $D(\mathscr{X}, P^{\circledast}) > CV$ and, consequently, $D(\mathscr{X}, P^{\circledast}) > D(P^*_n)$. When $n \to \infty$, $P^*_n \to P^{\circledast}$ and $D(P^*_n) \to D(P^{\circledast})$. Since the



(a) Varying μ parameter (b) Varying σ parameter (c) Varying ξ parameter

Fig. 2. Comparison of the complementary-cdfs $\bar{F}(x) = 1 - F(x) = P(x \geq X)$ varying the different parameters of the reference GEV distribution.

statistic is correct as Definition 3.4, $D(P_n^*) \to \overline{K}$, consequently $D(P^{\circledast}) = \overline{K}$. But $\forall n\ D(P_n^*) \geq \overline{K}$ and $K < D(P_n^*) < CV$, therefore $D(P^{\circledast}) < CV$ that is in contradiction with the hypothesis $P^{\circledast} \notin R$. □

This theorem states, from a real-time viewpoint, that the true – but unknown – pWCET distribution is always inside the region R. This has impact on the evaluation of the confidence of the pWCET result, as described below.

WCET over-estimation result. Each point in the region $R(\mathscr{X})$ corresponds to a pWCET distribution with parameter tuple (μ, σ, ξ). Thanks to the result of Theorem 3.6, we can state the following corollary:

Corollary 3.7. *Given a region $R(\mathscr{X})$, a violation probability $\overline{p}$, and a point $\hat{P} \in R(\mathscr{X})$ such that $\hat{P} = \arg\max_P F'_P(\overline{p})$, then either $\hat{P} = P^{\circledast}$ or the pWCET associated to $\hat{P}$ overestimates the real pWCET given by $P^{\circledast}$ at violation probability $\overline{p}$.*

In other words, given a fixed value for the violation probability $\overline{p}$ we can compute the WCET for each point of region R. Then, the maximum of these WCETs is either the true WCET or a safe overestimation of the WCET, at violation probability $\overline{p}$. Conversely, there is no point $P \in \mathbb{R}$ that, in general, overestimates the WCET for any probability $\overline{p}$. A possible solution to this issue is presented later in Section 4.

3.2 Exploring the Region of Acceptance

The region $R(\mathscr{X})$ describes the estimation uncertainty of the three parameters of the GEV distribution. Its size along the three axes depends on several factors, including the distribution of the input data, the chosen test statistic, the significance level α and the number of samples n. In particular, when increasing the sample size n, the ability of the test to detect invalid distributions improves, leading to, in general, a decrease of the region size. Moreover, the three dimensions are strictly correlated. For example, experimental evidences show that points inside the region representing a Fréchet distribution ($\xi > 0$) have usually lower values of σ than points inside the region representing a Weibull distribution ($\xi < 0$)[4]. To compare the distributions corresponding to the points inside the region, it is necessary to clearly define the order relations between two pWCETs.

pWCET ordering via statistical dominance. In previous articles on probabilistic real-time [44, 45], the ordering relation between pWCET has been defined by using the simplest form of partial ordering between distributions:

[4]This is neither a formal nor a general rule, but a recurring behaviour experienced by performing EVT estimations.

Definition 3.8 (First-order stochastic dominance [36]). A probabilistic-WCET pWCET_A *dominates* a pWCET_B iff the probability of observing a WCET larger than x is always equal or higher in pWCET_A with respect to pWCET_B, but the two distributions must not be exactly the same. In notation form:

$$\begin{aligned}\text{pWCET}_A \succ \text{pWCET}_B \leftrightarrow\ & [\forall x : F_A(x) \leq F_B(x) \\ & \wedge \exists y : F_A(y) < F_B(y)]\,,\end{aligned}$$

where $F_A(x)$ and $F_B(x)$ are respectively the cdf of pWCET_A and of pWCET_B. ◁

An example of first-order stochastic dominance is shown in Figure 2a[5]: the distribution $\overline{G_1}(x)$ dominates the distribution $\overline{F}(x)$. However, this is a very restrictive partial ordering: it is not possible to apply it to situations like the one depicted in Figure 2b. Econometrics analyses frequently overcome this problem using the so-called *second-order stochastic dominance* that has been studied in [47] for extreme value distributions. Even if this dominance is widely used in financial risk analysis, it does not provide the necessary guarantees for the distribution tail. This *non-applicability* to pWCET is described in details in Appendix A. Rather, we suggest to use a less restrictive dominance that keeps the safety of real-time requirements valid:

Definition 3.9 (Left tail-restricted first-order stochastic dominance [34]). A probabilistic-WCET pWCET_A *left dominates* a pWCET_B iff the probability of observing a WCET larger than x is always equal or higher in pWCET_A with respect to pWCET_B with $x \in [\overline{x}, +\infty)$, but the two distributions must not be exactly the same. In notation form:

$$\begin{aligned}\text{pWCET}_A \overset{\text{L}}{\succ} \text{pWCET}_B \leftrightarrow\ & \exists \overline{x}\,[\forall x > \overline{x} : F_A(x) \leq F_B(x) \\ & \wedge \exists y > \overline{x} : F_A(y) < F_B(y)]\,.\end{aligned}$$

Every first-order stochastic dominance is also a left tail-restricted first-order stochastic dominance:

$$\text{pWCET}_A \succ \text{pWCET}_B \implies \text{pWCET}_A \overset{\text{L}}{\succ} \text{pWCET}_B. \qquad \triangleleft$$

Although this definition is still a partial ordering, we can describe a larger set of relation, e.g. the scenario of Figure 2b[5]: the cdf $\overline{G_1}(x)$ left dominates the cdf $\overline{F}(x)$, i.e. fixed a WCET value, the pWCET related to $\overline{G_1}(x)$ provides an higher violation probability p for any $x > \overline{x}$ with $\overline{x} = 100$ in the depicted example.

Points dominance analysis. Having defined the previously described orders for pWCET, we can now formalize the dominance when we move along one direction from a chosen point inside the region. In particular, Figures 2a, 2b, 2c depict the simplest scenarios. Using the notation inside the figures, let be $\text{pWCET}_F \sim GEV(\mu, \sigma, \xi)$ and $\text{pWCET}_G \sim GEV(\mu', \sigma', \xi')$, hence:

- if $\mu' > \mu$ and $\sigma' = \sigma, \xi' = \xi$ then $\text{pWCET}_G \succ \text{pWCET}_F$
- if $\sigma' > \sigma$ and $\mu' = \mu, \xi' = \xi$ then $\text{pWCET}_G \overset{\text{L}}{\succ} \text{pWCET}_F$
- if $\xi' > \xi$ and $\mu' = \mu, \sigma' = \sigma$ then $\text{pWCET}_G \succ \text{pWCET}_F$

It is also possible to build more complex relations when two or more variables change:

- if $\xi' > \xi$ and σ' and/or μ' change in any directions then $\text{pWCET}_G \overset{\text{L}}{\succ} \text{pWCET}_F$;
- if $\sigma' > \sigma$ and $\xi' = \xi$ and μ' changes in any direction then $\text{pWCET}_G \overset{\text{L}}{\succ} \text{pWCET}_F$.

[5]As a reminder, the cumulative distribution function is defined as $F(x) = P(x < X)$. Instead, its complementary, $\bar{F} = 1 - F(x) = P(x \geq X)$, is depicted in the figures.

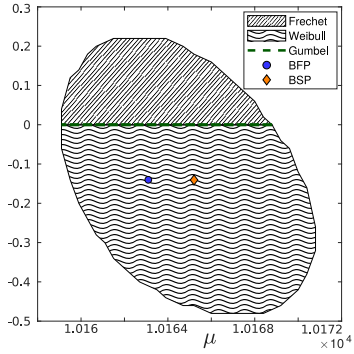


Fig. 3. The Region of Acceptance R generated from a Gaussian distribution sample and test CvM with the σ values collapsed. It is possible to notice that R includes all the three possible extreme value distributions.

Starting from these relations and according to the previous ordering definitions, it is possible to know if, by moving from a specified point inside the region to an another point inside the region, we are overestimating or underestimating the pWCET. When $\text{pWCET}_P > \text{pWCET}_{P'}$ the pWCET related to point P' is safely overestimated by P for any violation probability value p. Rather if $\text{pWCET}_P \overset{\text{L}}{>} \text{pWCET}_{P'}$ the pWCET related to point P' is safely overestimated by P for any violation probability value $p > \overline{p}$ with $\overline{p}$ that can be computed solving the equivalence equation between the iccdfs of both points. If $p < \overline{p}$, the distributions may potentially intersect in several points, making it impossible to conclude anything without further analyses.

3.3 EVT distribution classes

The Region of Acceptance may include more than one extreme value distribution classes. In order to show this, we have generated a random sample $X_1, X_2, ..., X_{50\,000}$ from a Gaussian distribution $\mathcal{N}(10000, 100)$. After applying BM, 80% of the sample is used to produce the pWCET estimation, while the remaining 20% is used to run the CvM test and build its Region of Acceptance R depicted in Figure 3. In this figure, we have not illustrated the scale parameter σ axis in order to clarify the variation of the ξ parameter. It can be noticed that the region includes all the three possible extreme value distributions: Fréchet ($\xi > 0$), Weibull ($\xi < 0$) and Gumbel ($\xi = 0$). Our estimator produced a Weibull distribution ($\xi < 0$) and also the BSP is in the same distribution class. From statistical theory, we know that any Gaussian is in the domain of attraction of a Gumbel, thus these two points, for $n \to \infty$, will converge to the Gumbel line of Figure 3.

Besides from the statistical interest on the distribution type, there is a significant effect on the pWCET:

- The Weibull distribution is a *truncated-tail* distribution, i.e. there exists a maximum iccdf $F'(p)$ value for $p \to 0$, thus upper-limiting the WCET.
- The Gumbel distribution is a *light-tail* distribution, i.e. the iccdf $F'(p) \to \infty$ for $p \to 0$, but the $F'(p)$ goes to zero faster than the exponential distribution, making the WCET unbounded. However, to obtain a linear increase of the WCET, the probability p should decrease faster than an exponential function.
- The Fréchet distribution is a *heavy-tail* distribution, i.e. the iccdf $F'(p) \to \infty$ for $p \to 0$, but the $F'(p)$ goes to zero slower than the exponential distribution: arbitrarily large WCET has a non-negligible probability to be observed. Moreover, if $\xi > 1$, the mean of the distribution is infinite: $E[X] = \infty, X \sim GEV(\mu, \sigma, \xi > 1)$.

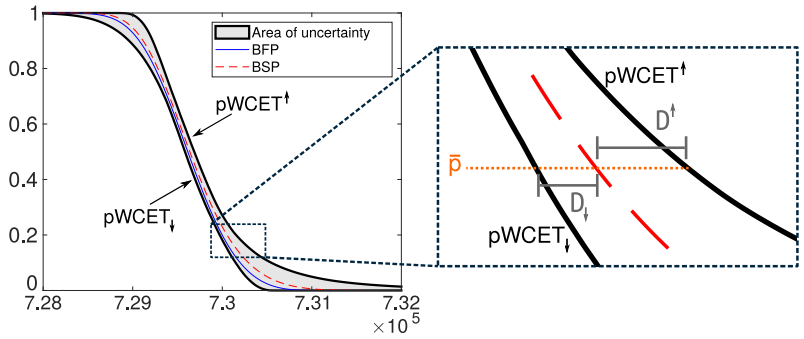


Fig. 4. The overlap of the CDFs of all distributions related to each point of the region of Figure 1. The upper-bound of this area is the curve $\text{pWCET}^{\uparrow}$, while the lower-bound the curve $\text{pWCET}_{\downarrow}$.

According to Section 3.2 we can define the following ordering: $\text{pWCET}_{\xi>0} \overset{L}{\succ} \text{pWCET}_{\xi=0} \overset{L}{\succ} \text{pWCET}_{\xi<0}$. The fact that a region spreads over different GEV distributions can then provide a way to increase the pWCET estimation quality. If some knowledge of the system and the task is available – e.g. it is known that the execution time must be bounded – then the exact distribution $P^{\circledast}$ cannot be in the Fréchet region. In this case, if the estimator generates a pWCET $\overline{P}$ with $\xi > 0$ and if we consider as the worst-case the point $P' \in R$ having $\xi = 0$, then P' is both safe and tighter than $\overline{P}$. Vice versa, if $\overline{P}$ has $\xi < 0$ and we known that the WCET can assume arbitrarily large values, then we must move the point $\overline{P}$ towards the Fréchet region to obtain a more pessimistic but robust estimation. The in-depth discussion and analysis of the system properties involving the concept of maximum domain of attraction is left as future work.

Identifying the correct model for pWCET. Some effort has been dedicated in previous works to identify the best model that fits the worst-case execution time values. The results are controversial: some researchers [2, 10, 11, 25] claim we should consider only the Weibull or Gumbel distribution, while others [30, 44] do not exclude the Fréchet too. The Gumbel distribution is often the only one taken into account because it is an upper-bound of the Weibull, however, it is not always possible to use such upper-bound strategy without hindering the result reliability [38]. All these works are mostly based on empirical considerations, which lack strong mathematical justifications, that are indeed difficult to achieve. For this reason, this work does not consider any GEV class restriction and it does not aim at identifying the best model for WCET estimation. Instead, it provides general methods, that work with any GEV class, to perform future investigations also in the direction of finding the best pWCET model. In the experimental part of this paper (Section 6) a relevant result on ξ uncertainty, intrinsic in the definition of Region of Acceptance, is discussed: it is not possible to directly rely on the pWCET estimation provided by the estimator because its inaccuracies can lead to unreliable or untight results.

3.4 The relation with statistical power

The Region of Acceptance is defined over the test result function of Definition 3.1. As we already said, the "acceptance" term is used in contrast with the standard nomenclature used in statistical tests. This is because, a statistical test can not prove the truthfulness of the null hypothesis but only its falseness. However, the *statistical power* of the test can be used to obtain a confidence on the test results when the null hypothesis is not rejected. In fact, the statistical power is the complement of the probability of a false negative result, i.e. of not rejecting an invalid distribution. The estimation

of statistical power for GEV distributions is already available in literature [41], allowing us to select a sample size for the input measurements that guarantees a certain confidence level [42].

4 ESTIMATION BOUNDS AND UNCERTAINTY

In Section 3.2 we discussed how to define the dominance between different points of the Region of Acceptance. The upper-bound for a single probability value was provided in Section 3.1. However, it does not exist, in general, one single point, i.e. one single valid distribution, able to dominate the overall region. In this section we propose a method to overcome this problem by estimating a pWCET curve that pessimistically bounds all the possible valid distributions of the region.

The pessimistic pWCET curve. The idea behind the pessimistic bound is to obtain a robust and safe estimation of the pWCET by taking the worst-case curve generated by overlapping the CDF of all the points inside the Region of Acceptance. The requirement for pessimism (safety of pWCET estimates) is illustrated in Figure 4, and formalized in the following definition:

Definition 4.1 (Pessimistic pWCET Curve). The curve $\text{pWCET}^{\uparrow}$ is defined as the locus of point (WCET, p) such that $\text{WCET} \in D$ and $p = \max_{P \in R}[1 - F_P(\text{WCET})]$, where D is the domain of the worst-case execution time and F_P is the cdf corresponding to the point P in the region R. ◁

As the definition clearly points out this locus of points first-order stochastic dominates all the other points $\text{pWCET}^{\uparrow} \succ \text{pWCET}_P \ \forall P \in R$, thus making the pessimistic pWCET curve a safe over-estimation of the real distribution. However, when the region is computed in a real scenario, the space of parameters cannot be explored continuously and the set of points inside the region must be discretized. The pessimistic pWCET curve reliability depends also on the resolution selected to build the region: in the unlucky case that $P^{\circledast}$ is in the proximity of the region boundaries, the resolution of μ, σ, ξ used to build the region may not be sufficient to include $P^{\circledast}$. More generally, to obtain a safe pessimistic pWCET curve it is sufficient to consider one more layer outside the region: $(\mu \pm \delta\mu, \sigma \pm \delta\sigma, \xi \pm \delta\xi)$ where $\delta\mu, \delta\sigma, \delta\xi$ are the parameter resolutions used in the region exploration.

The tightest pWCET curve. The same definition used for the pessimistic pWCET curve can be used to define its symmetrical tightest version i.e., the black lower curve in Figure 4:

Definition 4.2 (Tightest pWCET Curve). The curve $\text{pWCET}_{\downarrow}$ is defined as the locus of point (WCET, p) such that $\text{WCET} \in D$ and $p = \min_{P \in R}[1 - F_P(\text{WCET})]$, where D is the domain of the worst-case execution time and F_P is the cdf corresponding to the point P in the region R. ◁

It is important to remark that the $\text{pWCET}_{\downarrow}$ does not necessarily represent an optimistic bound to the real pWCET distribution. This locus of points has indeed passed the goodness-of-fit test, making it a valid extreme value estimation for the real pWCET. The $\text{pWCET}_{\downarrow}$ is, however, the less robust estimation in the whole set of possible distributions. We will recall and extend this concept later on in Section 5.

Uncertainty area. The area between $\text{pWCET}_{\downarrow}$ and $\text{pWCET}^{\uparrow}$ contains all the possible pWCET distributions according to our definition of Region of Acceptance. We informally call this space *area of uncertainty* and it is depicted as the gray area in Figure 4. This area is strictly correlated with the region size and parameters spread previously discussed in Section 3.2: the bigger the region R the bigger the area of uncertainty. Since we moved from the GEV parameters space to the ccdf space, the area of uncertainty provides a new metric to compare different possible estimations.

Definition 4.3 (Area of uncertainty). The *area of uncertainty* is defined as the area in the ccdf-space composed of the points of all ccdf curves of all pWCET distribution belonging to the Region of Acceptance. Equivalently, it is the area between the $\text{pWCET}^{\uparrow}$ and the $\text{pWCET}_{\downarrow}$ curves. Let be

$\text{pWCET}^{\uparrow}(x)$ and $\text{pWCET}_{\downarrow}(x)$ their respective curve functions $D \to [0;1]$, then the value of this area is:

$$A := \int_0^{\infty} [\text{pWCET}^{\uparrow}(x)]dx - \int_0^{\infty} [\text{pWCET}_{\downarrow}(x)]dx. \quad \triangleleft$$

The value of A can be easily computed numerically and it represents a novel metric to empirically evaluate the quality of the probabilistic analysis. Large values of A suggest that our region includes large values of uncertainty not only in the parameter space, but also in the pWCET space. This is the case discussed in Section 3.3 when in our region estimation all the three GEV models are plausible according to the considered statistical test. Vice versa, when the value of A is small, this is a clue that at least the distribution class is correct. The value of A may be also infinite: when at least one point P with $\xi \geq 1$, the mean value of the distribution and consequently the area under the ccdf are infinite. In this case, we can make two possible interpretations: either the analysis has been incorrectly performed or the system behaviour shows strong evidences of unbounded WCET. We could also exploit this to compare statistical tests: the area size is a direct measure of their quality, because a test with a smaller area is able to detect greater violations rather than another test executed with the same experimental setup but with a larger area.

5 TIGHTNESS AND PESSIMISM TRADE-OFF

This section exploits the previous theoretical results and tools to propose decision-making methods on the trade-off between tightness and pessimism of the estimated pWCET.

5.1 Best fit point vs best statistic point

According to Section 3.1, the BFP $\overline{P}$ provided by the estimator and the BSP P^* provided by the statistical test are good approximations of the unknown exact pWCET $P^{\circledast}$. In general, it is not possible to establish which of $\overline{P}$ or P^* is the closest to the exact pWCET. However, the closest point to $P^{\circledast}$ is not necessarily the *best* in a real-time context. We could in fact consider a different point that safely over-estimates the real distribution pWCET using one of the dominance definitions of Section 3.2. Accordingly, the first decision criterion is to select $\overline{P}$ if $\overline{P} \succ P^*$ or vice versa. In alternative, we can consider the left tail-restricted dominance and select $\overline{P}$ if $\overline{P} \overset{\text{L}}{\succ} P^*$ or vice versa; in this case, our decision remains valid if, in the evaluation of pWCET distribution, the computed WCET at a given probability level p is higher than the value $\overline{x}$ of Definition 3.9. These selection criteria with the statistical dominance concepts can be applied to any other point of the region R. Unfortunately, it is not always possible to select between two points with stochastic dominance, due to its partial ordering.

5.2 The robustness ratio

To evaluate the pessimism and tightness of a given pWCET distribution belonging to a point of the region R, we propose a metric based on an empirical formula. Nevertheless, the next Section 5.3 shows the existence of a relation between this formula and a well-defined statistical parameter.

Definition 5.1 (Robustness ratio). Let us assume that we select a probability $\overline{p}$ and a distribution corresponding to a point $P \in R$. At this probability, we can compute three WCETs: from P, from $\text{pWCET}_{\downarrow}$, and from $\text{pWCET}^{\uparrow}$. We call: (1) $D_{\downarrow}$ the absolute value of the distance between the WCET computed in P and the WCET computed with $\text{pWCET}_{\downarrow}$; (2) $D^{\uparrow}$ the absolute value of the distance between the WCET computed in P and the WCET computed with $\text{pWCET}^{\uparrow}$ (see right-side of Figure

4 for clarity). We can now define the *robustness ratio* as:

$$r = \frac{D_{\downarrow} - D^{\uparrow}}{D_{\downarrow} + D^{\uparrow}}. \qquad \triangleleft$$

This ratio r is always $r \in [-1; +1]$. When $r \to -1$, the distribution of point P is near the tightest one. When $r \to +1$, the distribution of point P is instead near the pessimist one. The robustness ratio r is then a metric representing the trade-off between tightness and pessimism. The experimenter can choose among all the valid distributions based on the value of this ratio, by knowing from the results of Section 3.1 that the true WCET value is inside this interval.

5.3 Confidence in the pWCET analysis

If we consider the previous definition of robustness ratio, by selecting a desired value of violation probability $\overline{p}$, we can derive the WCET interval from the curves $\text{pWCET}_{\downarrow}$ and $\text{pWCET}^{\uparrow}$. This interval is written as $I_W^{\overline{p}} = [\text{WCET}_{\downarrow}; \text{WCET}^{\uparrow}]$ and, according to Figure 4, its size is $D_{\downarrow} + D^{\uparrow}$. Since the real distribution related to point $P^{\circledast}$ is inside the region – from Theorem 3.6 – and since we know that this interval represents all the pWCET distribution values – from Definitions 4.1 and 4.2 –, then the real WCET value at the given probability $\overline{p}$ is inside this interval: $\text{WCET}_{\overline{p}}^{\circledast} \in [\text{WCET}_{\downarrow}; \text{WCET}^{\uparrow}]$.

Definition 5.2 (Confidence of the pWCET analysis). Given a probability $\overline{p}$ and a $\text{WCET} \in I_W^{\overline{p}}$, the confidence of the pWCET analysis c is defined as:

$$c = P\left[P(X > \text{WCET}) \leq \overline{p}\right]. \qquad \triangleleft$$

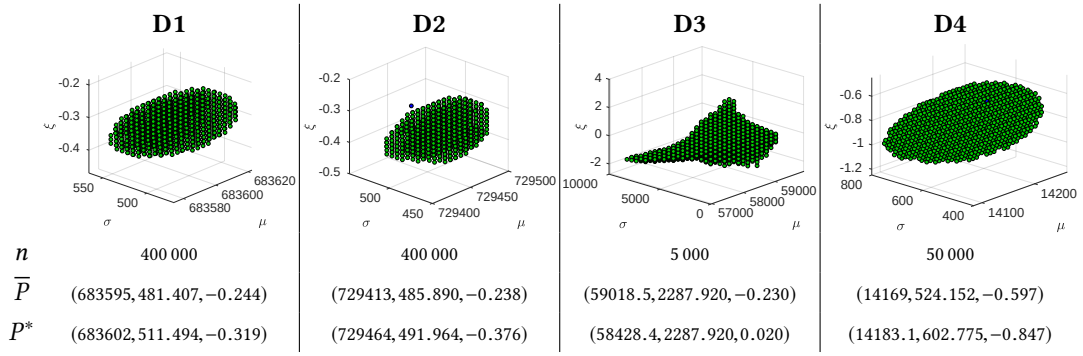
It is important to carefully dwell on this definition. The confidence is defined as the probability that the system violation probability is underestimated. If $c = 1$, the estimated couple (p, WCET) is surely safe. If $c < 1$, there exists a certain degree of uncertainty on the safety of $(\overline{p}, \text{WCET})$. The reader should not confuse the two probabilities: $\overline{p}$ is the chosen violation probability, a run-time property of the system, i.e. the probability to experience a larger WCET than the estimated one; c is the confidence, a property of the analysis, i.e. the probability to have estimated an unsafe couple $(\overline{p}, \text{WCET})$.

This confidence can be linked with the previously defined robustness ratio: if we select $\text{WCET}^{\uparrow}$, then $c = 1$ and $r = 1$, consequently $\text{WCET}^{\uparrow}$ surely upper-estimates the real WCET at probability $\overline{p}$. If the WCET value selected is not the right-most value, i.e. $r < 1$, the confidence is potentially less than the truth value: $c \leq 1$. In other words, selecting a less pessimistic, but still valid according to the chosen test, WCET may be potentially under-estimated. Clearly, c is a non-decreasing function of WCET, i.e. higher WCETs have higher confidence. The computation of the precise c value with respect to the variation of the chosen WCET is left as a future work. This is possible thanks to the previously cited statistical power of the selected GoF test.

6 EXPERIMENTAL EVALUATION

To evaluate the benefits of the previously introduced notations and techniques we considered four datasets representing different execution conditions and systems. The analysis of the proposed time traces showed the effectiveness of dealing with the uncertainty of the region of acceptance and related models.

The state-of-the-art *chronovise* tool [37] has been used to perform the MBPTA analysis on the following datasets:



	D1	D2	D3	D4
n	400 000	400 000	5 000	50 000
$\overline{P}$	(683595, 481.407, −0.244)	(729413, 485.890, −0.238)	(59018.5, 2287.920, −0.230)	(14169, 524.152, −0.597)
P^*	(683602, 511.494, −0.319)	(729464, 491.964, −0.376)	(58428.4, 2287.920, 0.020)	(14183.1, 602.775, −0.847)

Fig. 5. Region of Acceptance, number of samples, estimator best fit point (BFP) and best statistic point (BSP) of the datasets under analysis.

D1) An industrial safety-critical application from Airbus, running on a multi-core platform. The time measured are referred to a single task execution with other tasks running in other cores and interfering on shared resources[6].
D2) The same as the previous dataset, but considering a different task of the same safety-critical real-time application.
D3) A memory-intensive task running on a multi-core T4240, stressing the data-cache with interferences on the overall cache hierarchy, shared memory and bus [17].
D4) A time trace of a GPU application running under different execution conditions. The dataset is the same as in [3].

The paper of Nolte et al. [33] proposed a classification for real-time workloads in the context of probabilistic real-time. Some proposed constraints are however too strict for real applications. We guaranteed A.3.1 (*Avoid usage of shared services and drives in the software architecture.*) for D1 and D2, while D3 runs on PikeOS (so we can consider valid A.3.2 that requires predictability of services) and D4 runs on CUDA, so neither A.3.1 nor A.3.2 applies for D4. Regarding the hardware states (A4 group), the cache status was disregarded (A.4.2). Finally, the execution time of the tasks is not affected by the state of the environment (A.5.2).

Datasets D1 and D2 are composed of 400 000 time measurements, while the sample sizes of D3 and D4 are respectively 5 000 and 50 000. All the time traces in the aforementioned datasets are real measurements of the task execution time acquired with appropriate instrumentation of the applications. We verified the satisfaction of the iid hypothesis running a standard LjungBox test. As, the EVT was applicable, we could filter the data via the Block-Maxima method with a block size of $B = 20$, empirically chosen but in line with previous works [24] [30] [48].

6.1 Region of Acceptance

From the sample output of the BM approach, the GEV distribution is fitted using the well-known Maximum Likelihood Estimator (MLE). To build the Region of Acceptance we consider as goodness-of-fit test the *Cramér-von Mises criterion (CvM)* [8]. Another possible test is the Kolmogorov-Smirnov test (KS) [31]. Both tests have *correct statistics*, i.e. they satisfy Definition 3.4 as proved in the Appendix B.

[6]No more details on the Airbus use case can be provided since it is an actual industrial application currently under investigation.

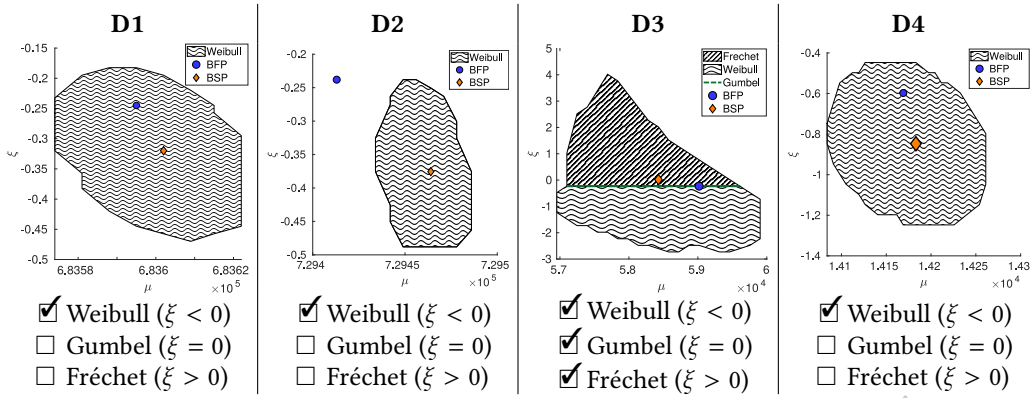


Fig. 6. The Regions of Acceptance for the datasets under analysis with the σ value collapsed.

To find the region R it is necessary to explore the parameter space around the estimated distribution $\bar{P}$. We have begun by uniformly exploring 40 points around each parameter, leading to a total number of $40^3 = 64\,000$ explorations. The initial interval has been set to ±10% of the estimated value and it was step-by-step increased to include the whole region. The CvM test has been applied to each point obtaining the set of accepted points, i.e. the region R, depicted in Figure 5. The time complexity depends on the number of points explored and how the statistic is computed for the chosen test. For CvM the computational complexity is $O(nm)$ where n is the number of time measurements and m is the number of explored points. Instead, for KS the time complexity becomes $O(n^2m)$. In this experimental evaluation, the total time required to build each region was less than 10 seconds on a standard workstation.

It is possible to notice that in the dataset D2 the estimated point is outside the region: the GEV distribution estimated by MLE is not a valid distribution according to the CvM test result. This is a violation of the initial assumption that the best fit estimator point $\overline{P}$ is inside the region. In this case, before beginning with the parameter space exploration not only we have no information on how many points should be explored, we also do not know whether the region exists or not. For example, assuming that the input time measurements are distributed according to a statistical distribution that is not in the domain of attraction of any generalized extreme value distribution, e.g. the Poisson distribution, then no point is expected to pass the test ($R = \varnothing$), whatever GEV distribution is estimated. In our lucky case, the region exists and we have been able to find it because it is in the ±10% interval of at least one parameter. For completeness, we checked why the MLE estimator failed to obtain a valid distribution fitting the data, and we discovered the presence of a local minimum of the MLE optimization function at the estimated $\overline{P}$. One possible solution to this problem, left as future work, is to initially use the Probabilistic Weighted Moment (PWM) estimator to obtain the point inside the region and then to improve the estimation with MLE. In the considered corner case of D2, the parameter space exploration guides us to find a set of points – the region – fulfilling the goodness-of-fit test, i.e. with a valid pWCET distribution otherwise impossible to find using only the estimated point $\overline{P}$. This is another accidental advantage of using the Region of Acceptance to evaluate the pWCET output of the estimator algorithm.

6.2 Distribution shapes

Figure 6 shows the region collapsing the σ axis to show the spread of the ξ parameter. D1, D2, and D4 regions contain only points from Weibull distribution. In real-time world it means that the

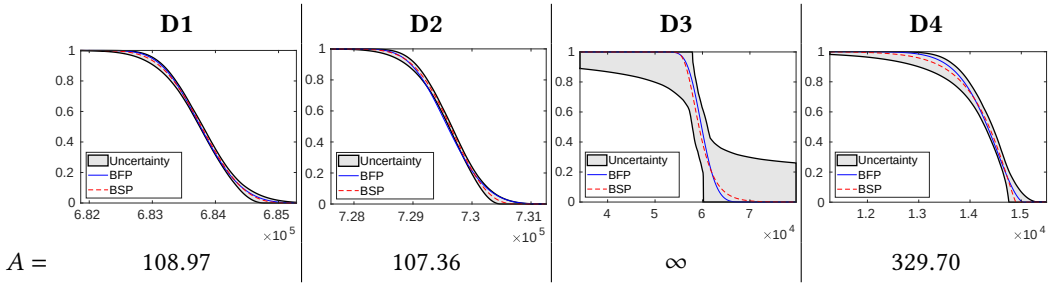


Fig. 7. Area of uncertainty, tightest upper-bound and pessimistic upper-bound curves, BFP and BSP distributions.

observed phenomenon, i.e. the execution times, has a finite maximum, i.e. a finite WCET. Instead, D3 is more problematic because it includes both Gumbel and Fréchet distribution classes. Even more, it includes Fréchet distributions with $\xi > 1$. This means that the WCET is not finite, but also its mean value is not finite, suggesting that either there is a problem in the execution time measurements or the system has actually an unbounded WCET.

It is worth noticing the position of the best fit point and the best statistic point in our datasets. The latter is in general at the center of the region. The D2 case, where the estimated point is outside the region, has been already discussed in the previous Section 6.1. The D3 case is interesting: the estimated point represents a Weibull distribution with $\xi = -0.23$, while the point that has the best statistic is near the Gumbel line and is actually a Fréchet distribution with $\xi = 0.02$. This leads to an important conclusion: the distribution estimated by the estimator is a light-tail distribution, but the test is able to accept another distribution with even better statistic having the Fréchet tail. Consequently, we can conclude that relying on the estimator result without a sensitivity analysis on ξ may lead to unreliable or untight results, since the real pWCET distribution can belong to another GEV class.

6.3 Result uncertainty

Referring to Figure 7, the effect of ξ parameter spreading is clear: it is immediately visible that D3 has the largest area of uncertainty. Moreover, the absolute value of this area is infinite due to the presence of valid Fréchet distributions with $\xi \geq 1$. For the other scenarios, the area has been computed by performing the numerical integration of Definition 4.3. D1 and D2 have lower uncertainties compared to D4. This is the consequence of a smaller region of acceptance and, in particular, less uncertainty on ξ. We notice that there is no general rule on the domination between $\overline{P}$ and P^*. For example, in D3 $\text{pWCET}_{P^*} \overset{\text{L}}{\succ} \text{pWCET}_{\overline{P}}$ while in D4 $\text{pWCET}_{\overline{P}} \overset{\text{L}}{\succ} \text{pWCET}_{P^*}$. Instead, as expected by their definition, $\text{pWCET}^{\uparrow} \succ \text{pWCET}_{\overline{P}}$, $\text{pWCET}^{\uparrow} \succ \text{pWCET}_{P^*}$, $\text{pWCET}_{\overline{P}} \succ \text{pWCET}_{\downarrow}$, $\text{pWCET}_{P^*} \succ \text{pWCET}_{\downarrow}$ for all cases. There is only one exception that is not visible in the figure: the D3 best fit point is outside the region and in this case there are some WCET values for which the $\text{pWCET}_{\downarrow}$ does not under-estimate the probability, i.e. $\text{pWCET}_{\overline{P}} \not\succ \text{pWCET}_{\downarrow}$. However, the relaxed version is still valid in this case: $\text{pWCET}_{\overline{P}} \overset{\text{L}}{\succ} \text{pWCET}_{\downarrow}$. The consequence on the WCET estimation is that the estimated distribution associated with the point $\overline{P}$ outside the region is potentially unsafe for some WCET or p values because it underestimates the curve representing the lower-bound on the distributions accepted by the goodness-of-fit test.

p	Distribution	D1	D2	D3	D4
10^{-3}	$\text{pWCET}^{\uparrow}$	685 525	731 054	$1.3 \cdot 10^{15}$	15 314
	$\overline{P}$	685 198	731 059	66 943	15 032
	$\text{pWCET}_{\downarrow}$	684 728	730 456	60 192	14 756
10^{-6}	$\text{pWCET}^{\uparrow}$	686 075	731 365	$1.5 \cdot 10^{26}$	15 368
	$\overline{P}$	684 773	731 378	68 566	15 046
	$\text{pWCET}_{\downarrow}$	685 494	730 490	60 192	14 757
10^{-9}	$\text{pWCET}^{\uparrow}$	686 231	731 425	$1.7 \cdot 10^{41}$	15 370
	$\overline{P}$	685 548	731 439	68 898	15 046
	$\text{pWCET}_{\downarrow}$	684 774	730 491	60 192	14 757
10^{-12}	$\text{pWCET}^{\uparrow}$	686 275	731 436	$2.0 \cdot 10^{50}$	15 370
	$\overline{P}$	685 558	731 451	68 966	15 046
	$\text{pWCET}_{\downarrow}$	684 774	730 491	60 192	14 757

Table 1. The computed WCET from the curves of Table 7 at different violation probability levels.

p	Distribution	D1	D2	D3	D4
10^{-1}	P^*	−0.093216	−0.066638	−0.994008	−0.310016
	$\overline{P}$	−0.017073	0.669008	−0.995131	−0.042482
10^{-3}	P^*	−0.255874	−0.265608	−1.000000	−0.513920
	$\overline{P}$	0.178674	1.017288	−1.000000	−0.012972
10^{-6}	P^*	−0.372082	−0.367926	−1.000000	−0.549735
	$\overline{P}$	0.107353	1.030012	−1.000000	−0.052946
10^{-9}	P^*	−0.417336	−0.395697	−1.000000	−0.551476
	$\overline{P}$	0.062515	1.031415	−1.000000	−0.055915
10^{-12}	P^*	−0.432052	−0.402085	−1.000000	−0.551556
	$\overline{P}$	0.044533	1.031659	−1.000000	−0.056071

Table 2. The robustness ratios of BFP $\overline{P}$ at different violation probability levels.

6.4 pWCET: tightness vs pessimism

Having computed $\text{pWCET}^{\uparrow}$ and $\text{pWCET}_{\downarrow}$, it is now possible to estimate the WCET according to a violation probability p. The WCET value for the different curves and for some values of violation probability are presented in Table 1. The effect of the presence of valid points with a Fréchet distribution in D3 is evident: the WCET of $\text{pWCET}^{\uparrow}$ is clearly too large to be considered feasible in any scheduling analysis. Instead, $\overline{P}$ and $\text{pWCET}_{\downarrow}$ provide valid approximations, but with less confidence: if we select $\overline{P}$ and $\text{pWCET}_{\downarrow}$, there potentially is a non-null probability that our WCET result is unsafe according to Definition 5.2. The WCET values of D1, D2, and D3 are distributed in a smaller interval and are all apparently feasible to be used for scheduling analysis.

To explore the differences between $\overline{P}$ and P^* we presented their robustness ratio in Table 2. In D1, D2, and in D4, $\overline{P}$ is more pessimist than P^* while in D3 it is the opposite. This is in line with our previous graphical result of Figure 7. The fact that one point is always pessimist with respect to the other point for all the considered probabilities must not be taken as a general rule: the robustness ratio is a value at a fixed probability and it may behave differently for different values of it. The presence of valid pWCET distributions is clear in D3: both points are definitely tighter than the pWCET$^{\uparrow}$, as it is also experimentally verified in Table 1.

6.5 Summary of the experimental results

To summarize the experimental evaluation, we recap the major steps and the conclusions that can be drawn from the four datasets under analysis:

(1) The Region of Acceptances of test CvM has been generated by exploring the space around the GEV distribution parameters provided by MLE estimator;
(2) Even if the estimated point $\overline{P}$ for dataset D2 does not pass the GoF test, we have been able to find the Region of Acceptance in the ±10% of $\overline{P}$ parameter space. Despite there are no guarantees that this exploration is always successful, in our scenario it was useful to find valid pWCET distributions that we would not have otherwise found;
(3) From the analysis of shape parameter uncertainty, we noticed that one dataset (D3) includes all the three distribution types. This has significant impact on the pWCET uncertainty: the most robust pWCET curve leads to an unrealistic WCET at small probability values;
(4) If a potential reduction in the pWCET confidence is acceptable, the issue of the previous point can be easily solved by selecting another point inside the region according to its robustness ratio value and test statistical power;
(5) The estimated distribution $\overline{P}$ is not necessarily the best one, neither in terms of safety nor in terms of tightness. The same is valid for the BSP P^*. A careful evaluation must be performed by exploiting one of the decision making tools provided.

7 CONCLUSIONS

Any distribution estimation routine suffers from estimation errors caused by the necessarily finite number of input samples. In probabilistic WCET analyses the safety of the results depends on several conditions imposed by the EVT conditions. Even considering all the open challenges on these conditions solved, the estimated pWCET distribution is still affected by uncertainty. This article discussed this problem by providing a set of mathematical tools to deal with the parameter uncertainty, with the goal to be a step towards a more reliable pWCET estimation. In particular, the *region of acceptance* has been defined on the GEV distribution parameters space. By exploring this region, it is possible to move the estimated pWCET distribution to more reliable or to tighter distributions. The advantages on both the safety and the tightness of the pWCET distribution have been showed by performing the analysis on real time traces of different nature, including real industrial datasets. Several possible future works on improving the pWCET methodology are needed to increase the reliability on MBPTA methods. In particular, the representativity problem is still the most crucial barrier to the introduction of MBPTA in critical systems. The use of the statistical power, already cited in Section 5.3, can help in improving the estimation uncertainty quantification and it is currently an ongoing work.

ACKNOWLEDGMENTS

This research was partially funded by EU project RECIPE H2020 (grant no. 801137) [21].

REFERENCES

[1] J. Abella, C. Hernandez, E. Quiñones, F. J. Cazorla, P. R. Conmy, M. Azkarate-askasua, J. Perez, E. Mezzetti, and T. Vardanega. 2015. WCET analysis methods: Pitfalls and challenges on their trustworthiness. In *10th IEEE International Symposium on Industrial Embedded Systems (SIES)*. IEEE, 1–10. https://doi.org/10.1109/SIES.2015.7185039

[2] J. Abella, E. QuiÃśones, F. Wartel, T. Vardanega, and F. J. Cazorla. 2014. Heart of Gold: Making the Improbable Happen to Increase Confidence in MBPTA. In *2014 26th Euromicro Conference on Real-Time Systems*. IEEE, 255–265. https://doi.org/10.1109/ECRTS.2014.33

[3] Kostiantyn Berezovskyi, Fabrice Guet, Luca Santinelli, Konstantinos Bletsas, and Eduardo Tovar. 2016. Measurement-Based Probabilistic Timing Analysis for Graphics Processor Units. In *Architecture of Computing Systems - ARCS 2016 - 29th International Conference, Nuremberg, Germany, April 4-7, 2016, Proceedings*. Springer International Publishing, 223–236.

[4] G. Bernat, A. Colin, and S. M. Petters. 2002. WCET analysis of probabilistic hard real-time systems. In *23rd IEEE Real-Time Systems Symposium, 2002. RTSS 2002.* IEEE, 279–288. https://doi.org/10.1109/REAL.2002.1181582

[5] Francisco J. Cazorla, Leonidas Kosmidis, Enrico Mezzetti, Carles Hernandez, Jaume Abella, and Tullio Vardanega. 2019. Probabilistic Worst-Case Timing Analysis: Taxonomy and Comprehensive Survey. *ACM Comput. Surv.* 52, 1, Article 14 (Feb. 2019), 35 pages. https://doi.org/10.1145/3301283

[6] Francisco J. Cazorla, Eduardo Quiñones, Tullio Vardanega, Liliana Cucu, Benoit Triquet, Guillem Bernat, Emery Berger, Jaume Abella, Franck Wartel, Michael Houston, Luca Santinelli, Leonidas Kosmidis, Code Lo, and Dorin Maxim. 2013. PROARTIS: Probabilistically Analyzable Real-Time Systems. *ACM Trans. Embed. Comput. Syst.* 12, 2s, Article 94 (May 2013), 26 pages. https://doi.org/10.1145/2465787.2465796

[7] Xavier Civit, Joan del Castillo, and Jaume Abella. 2018. A Reliable Statistical Analysis of the Best-Fit Distribution for High Execution Times. In *21st Euromicro Conference on Digital System Design, DSD 2018, Prague, Czech Republic, August 29-31, 2018*. IEEE, 727–734. https://doi.org/10.1109/DSD.2018.00012

[8] Harald Cramér. 1928. On the composition of elementary errors. *Scandinavian Actuarial Journal* 1928, 1 (1928), 13–74. https://doi.org/10.1080/03461238.1928.10416862

[9] Sandor Csorgo and Julian J. Faraway. 1996. The Exact and Asymptotic Distributions of Cramer-von Mises Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (01 1996), 221–234. https://doi.org/10.2307/2346175

[10] L. Cucu-Grosjean, L. Santinelli, M. Houston, C. Lo, T. Vardanega, L. Kosmidis, J. Abella, E. Mezzetti, E. Quiñones, and F. J. Cazorla. 2012. Measurement-Based Probabilistic Timing Analysis for Multi-path Programs. In *2012 24th Euromicro Conference on Real-Time Systems*. IEEE, 91–101. https://doi.org/10.1109/ECRTS.2012.31

[11] Corentin Damman, Gregory Edison, Fabrice Guet, Eric Noulard, Luca Santinelli, and Jerome Hugues. 2016. Architectural Performance Analysis of FPGA Synthesized LEON Processors. In *Proceedings of the 27th International Symposium on Rapid System Prototyping: Shortening the Path from Specification to Prototype*. ACM, New York, NY, USA, 33–40. https://doi.org/10.1145/2990299.2990306

[12] D. Dasari, B. Akesson, V. Nélis, M. A. Awan, and S. M. Petters. 2013. Identifying the sources of unpredictability in COTS-based multicore. In *Int. Symp. on Industrial Embedded Systems*. IEEE, 39–48. https://doi.org/10.1109/SIES.2013.6601469

[13] Robert Davis and Liliana Cucu-Grosjean. 2019. A Survey of Probabilistic Schedulability Analysis Techniques for Real-Time Systems. *Leibniz Transactions on Embedded Systems* 6, 1 (2019), 04–1–04:53. https://doi.org/10.4230/LITES-v006-i001-a004

[14] Robert I. Davis, Alan Burns, and David Griffin. 2017. On the Meaning of pWCET Distributions and their use in Schedulability Analysis. (2017). https://www-users.cs.york.ac.uk/~robdavis/papers/RTSOPS2017pWCET.pdf

[15] Robert I. Davis, Luca Santinelli, Sebastian Altmeyer, Claire Maiza, and Liliana Cucu-Grosjean. 2013. Analysis of Probabilistic Cache Related Pre-emption Delays. In *25th Euromicro Conference on Real-Time Systems, ECRTS 2013, Paris, France, July 9-12, 2013*. IEEE, 168–179. https://doi.org/10.1109/ECRTS.2013.27

[16] Darinka Dentcheva and Andrzej Ruszczyński. 2003. *Portfolio optimization with stochastic dominance constraints*. Elsevier. https://doi.org/10.18452/8306

[17] Julien Durand, Youcef Bouchebaba, and Luca Santinelli. 2019. Statistical Analysis for Shared Resources Effects with Multi-Core Real-Time Systems. In *13th IEEE International Symposium on Embedded Multicore/Many-core Systems-on-Chip, MCSoC 2019, Singapore, Singapore, October 1-4, 2019*. IEEE, 362–371. https://doi.org/10.1109/MCSoC.2019.00058

[18] S. Edgar and A. Burns. 2001. Statistical analysis of WCET for scheduling. In *Proceedings 22nd IEEE Real-Time Systems Symposium (RTSS 2001)*. IEEE, 215–224. https://doi.org/10.1109/REAL.2001.990614

[19] M. Fernandez, D. Morales, L. Kosmidis, A. Bardizbanyan, I. Broster, C. Hernandez, E. Quinones, J. Abella, F. Cazorla, P. Machado, and L. Fossati. 2017. Probabilistic Timing Analysis on Time-randomized Platforms for the Space Domain. In *Proceedings of the Conference on Design, Automation & Test in Europe*. ACM and IEEE, 738–739.

[20] R. A. Fisher and L. H. C. Tippett. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* 24, 2 (1928), 180âĂŞ190. https://doi.org/10.1017/S0305004100015681

[21] William Fornaciari, Giovanni Agosta, David Atienza, Carlo Brandolese, Leila Cammoun, Luca Cremona, Alessandro Cilardo, Albert Farres, José Flich, Carles Hernandez, Michal Kulchewski, Simone Libutti, José Maria Martínez, Giuseppe Massari, Ariel Oleksiak, Anna Pupykina, Federico Reghenzani, Rafael Tornero, Michele Zanella, Marina Zapater, and Davide Zoni. 2018. Reliable Power and Time-Constraints-Aware Predictive Management of Heterogeneous Exascale Systems. In *Proceedings of the 18th International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS '18)*. Association for Computing Machinery, New York, NY, USA, 187–194. https://doi.org/10.1145/3229631.3239368

[22] S. Jiménez Gil, I. Bate, G. Lima, L. Santinelli, A. Gogonel, and L. Cucu-Grosjean. 2017. Open Challenges for Probabilistic Measurement-Based Worst-Case Execution Time. *IEEE Embedded Systems Letters* 9, 3 (Sept 2017), 69–72. https://doi.org/10.1109/LES.2017.2712858

[23] B. Gnedenko. 1943. Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics* 44, 3 (1943), 423–453. http://www.jstor.org/stable/1968974

[24] Fabrice Guet, Luca Santinelli, and Jerome Morio. 2017. 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) (OpenAccess Series in Informatics (OASIcs))*, Jan Reineke (Ed.), Vol. 57. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 3:1–3:13. https://doi.org/10.4230/OASIcs.WCET.2017.3

[25] Jeffery Hansen, Scott Hissam, and Gabriel A. Moreno. 2009. Statistical-Based WCET Estimation and Validation. In *9th International Workshop on Worst-Case Execution Time Analysis (WCET'09) (OpenAccess Series in Informatics (OASIcs))*, Niklas Holsti (Ed.), Vol. 10. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 1–11. https://doi.org/10.4230/OASIcs.WCET.2009.2291

[26] Jun-Haeng Heo, Hongjoon Shin, Woosung Nam, Juseong Om, and Changsam Jeong. 2013. Approximation of modified Anderson–Darling test statistics for extreme value distributions with unknown shape parameter. *Journal of hydrology* 499 (2013), 41–49.

[27] R. Kirner and P. Puschner. 2008. Obstacles in Worst-Case Execution Time Analysis. In *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)*. IEEE, 333–339. https://doi.org/10.1109/ISORC.2008.65

[28] A. Kolmogorov. 1933. Sulla determinazione empĂŋrica di uma legge di distribuzione. *Inst. Ital. Attuari, Giorn.* 4 (1933), 83–91.

[29] L. Kosmidis, E. Quińones, J. Abella, T. Vardanega, I. Broster, and F. J. Cazorla. 2014. Measurement-Based Probabilistic Timing Analysis and Its Impact on Processor Architecture. In *2014 17th Euromicro Conference on Digital System Design*. IEEE, 401–410. https://doi.org/10.1109/DSD.2014.50

[30] G. Lima, D. Dias, and E. Barros. 2016. Extreme Value Theory for Estimating Task Execution Time Bounds: A Careful Look. In *Euromicro Conference on Real-Time Systems (ECRTS)*. IEEE, 200–211. https://doi.org/10.1109/ECRTS.2016.20

[31] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[32] Daniel McFadden. 1978. Modeling the choice of residential location. *Transportation Research Record* 1, 673 (1978), 72–77.

[33] Thomas Nolte, Meng Liu, and Bjorn Lisper. 2014. Challenges with Probabilities in Response-Time Analysis of Real-Time Systems. In *5th Real-Time Scheduling Open Problems Seminar, RTSOPS, Spain*. 3–4.

[34] Edgar Elias Osuna. 2013. Tail-restricted stochastic dominance. *IMA Journal of Management Mathematics* 24, 1 (2013), 21–44. https://doi.org/10.1093/imaman/dpr023

[35] Stefan M Petters. 2003. Comparison of trace generation methods for measurement based WCET analysis. In *In Proceedings of the 3rd International Workshop on Worst Case Execution Time Analysis*.

[36] James P. Quirk and Rubin Saposnik. 1962. Admissibility and Measurable Utility Functions. *The Review of Economic Studies* 29, 2 (1962), 140–146. https://doi.org/10.2307/2295819

[37] F. Reghenzani, G. Massari, and W. Fornaciari. 2018. chronovise: Measurement-Based Probabilistic Timing Analysis framework. *J. Open Source Software* 3, 28 (2018), 711. https://doi.org/10.21105/joss.00711

[38] F. Reghenzani, G. Massari, and W. Fornaciari. 2018. The Misconception of Exponential Tail Upper-Bounding in Probabilistic Real-Time. *IEEE Embedded Systems Letters* 11, 3 (2018), 77–80. https://doi.org/10.1109/LES.2018.2889114

[39] F. Reghenzani, G. Massari, and W. Fornaciari. 2019. A Probabilistic Approach to Energy-Constrained Mixed-Criticality Systems. In *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE/ACM, 1–6. https://doi.org/10.1109/ISLPED.2019.8824991

[40] F. Reghenzani, G. Massari, W. Fornaciari, and A. Galimberti. 2019. Probabilistic-WCET Reliability: On the Experimental Validation of EVT Hypotheses. In *Proceedings of the International Conference on Omni-Layer Intelligent Systems (COINS '19)*. ACM, New York, NY, USA, 229–234. https://doi.org/10.1145/3312614.3312660

[41] F. Reghenzani, G. Massari, L. Santinelli, and W. Fornaciari. 2019. Statistical Power Estimation Dataset for External Validation GoF tests on EVT distribution. *Data in Brief* 25 (jun 2019), 104071. https://doi.org/10.1016/j.dib.2019.104071

[42] Federico Reghenzani, Luca Santinelli, and William Fornaciari. 2019. Why Statistical Power Matters for Probabilistic Real-Time: Work-in-Progress. In *Proceedings of the International Conference on Embedded Software Companion (EMSOFT '19)*. Association for Computing Machinery, New York, NY, USA, Article Article 3, 2 pages. https://doi.org/10.1145/3349568.3351555

[43] Michael Thomas Rolf-Dieter Reiss. 2007. *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser Basel. https://doi.org/10.1007/978-3-7643-7399-3.

[44] L. Santinelli, F. Guet, and J. Morio. 2017. Revising Measurement-Based Probabilistic Timing Analysis. In *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*. IEEE, 199–208. https://doi.org/10.1109/RTAS.2017.16

[45] Luca Santinelli and Zhishan Guo. 2017. On the Criticality of Probabilistic Worst-Case Execution Time Models. In *Dependable Software Engineering. Theories, Tools, and Applications*, Kim Guldstrand Larsen, Oleg Sokolsky, and Ji Wang (Eds.). Springer, 59–74.

[46] K. P. Silva, L. F. Arcaro, and R. S. d. Oliveira. 2017. On Using GEV or Gumbel Models When Applying EVT for Probabilistic WCET Estimation. In *2017 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 220–230. https://doi.org/10.1109/RTSS.2017.00028

[47] Ganghuai Wang, J. H. Lambert, and Y. Y. Haimes. 1999. Stochastic ordering of extreme value distributions. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 29, 6 (Nov 1999), 696–701. https://doi.org/10.1109/3468.798077

[48] F. Wartel, L. Kosmidis, C. Lo, B. Triquet, E. QuiÃśones, J. Abella, A. Gogonel, A. Baldovin, E. Mezzetti, L. Cucu, T. Vardanega, and F. J. Cazorla. 2013. Measurement-based probabilistic timing analysis: Lessons from an integrated-modular avionics case study. In *8th IEEE Int. Symp. on Industrial Embedded Systems*. IEEE, 241–248. https://doi.org/10.1109/SIES.2013.6601497

A NON-APPLICABILITY OF SECOND-ORDER STOCHASTIC DOMINANCE TO PWCET PROBLEM

To overcome the limitation of *first-order stochastic dominance* as described in Section 3.2, in financial risk analysis the *second-order stochastic dominance* is often used. This dominance is defined as:

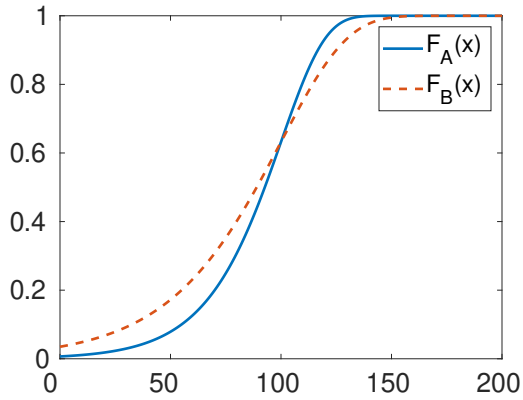
Definition A.1. A cumulative distribution function $F(x)$ second-order stochastic dominates a cumulative distribution function $G(x)$ iff:

$$\int_{-\infty}^{c} F(x)dx \leq \int_{-\infty}^{c} G(x)dx \quad \forall c \in \mathbb{R}$$

with the strict inequality holding for some c. ◁

In financial applications, the cdf $F(x)$ is usually preferred being less risky than the other option. This is because the average value of a random variable defined under $G(x)$ is greater or equal to the one defined under $F(x)$ for any possible non-decreasing function of them [16]. However, this dominance is not sufficient for probabilistic real-time systems. In order to show this, we provide a counterexample exploiting the following property of a Gumbel distribution [47]: Given a random variable A distributed according to a Gumbel distribution $GEV(\mu_A, \sigma_A, 0)$ and a random variable B distributed according to a Gumbel distribution $GEV(\mu_B, \sigma_B, 0)$, if $\mu_A = \mu_B$ and $\sigma_A < \sigma_B$, then $F_A(x)$ (i.e. the cdf of A) second-order stochastic dominates $F_B(x)$ (i.e. the cdf of B). The two cdfs are depicted in the following figure:

As it is possible to see, even if $F_A(x)$ second-order stochastic dominates $F_B(x)$, after the intersection point (around $x \approx 100$) the complementary cdf is no more upper-bounding, leading $F_A(x)$ to underestimate the WCET w.r.t. $F_B(x)$. Looking with the other axis, selecting a value for the probability, the $F_A(x)$ provides a smaller value compared to the $F_B(x)$, potentially underestimating the WCET in our application.



B CORRECT STATISTICS

B.1 Proof: KS has a correct statistic

The test statistic of Kolmogorov-Smirnov test is [28]:

$$D(\mathscr{X}, \mathcal{A}) = \sup_x |F_n(x) - F_{\mathcal{A}}(x)|$$

where $F_n(x)$ is the *empirical cumulative distribution function (ecdf)* and $F_{\mathcal{A}}$ is the cumulative distribution function of the reference distribution. The ecdf is in turn defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[-\infty;x]}(X_i)$$

where $1_A(x)$ is the *characteristic function* defined as:

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

Thanks to the strong law of large numbers, the ecdf converges to the real cdf almost surely: $P(\lim_{n\to\infty} F_n(x) = F(x)) = 1$. Since in our definition the $\mathscr{X}$ sample is drawn from the reference distribution $\mathcal{A}$, then $D(\mathscr{X}, \mathcal{A})$ converges almost surely to zero, i.e. $\overline{K} = 0$. The second part of the definition is easily proved: the KS statistic is always positive, thus $\forall n \in \mathbb{N}$ it is true that $D(\mathscr{X}, \mathcal{A}) \geq \overline{K}$. □

B.2 Proof: CvM has a correct statistic

This proof is similar to the proof on KS test. The Cramér-von Mises discrete test statistic [9] is:

$$D(\mathscr{X}, \mathcal{A}) = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F_{\mathcal{A}}(x_i) \right)^2$$

This statistic has been derived from the discretization of the continuous CvM statistic:

$$D(\mathscr{X}, \mathcal{A}) = \int_{-\infty}^{\infty} n * (F_n(x) - F(x))^2 \, dF(x)$$

Thanks to the strong law of large numbers, we know that the ecdf converges to the real cdf almost surely: $P(\lim_{n\to\infty} F_n(x) = F(x)) = 1$. Moreover, the rate of convergence is $\sqrt{n}$, thus $(F_n(x) - F(x))^2$ converges to zero with a linear rate. The product with n is then finite and constant with respect to the integration variable. For this reason, the final value of the integral is 1. The statistic is consequently converging to a constant value. For any finite value of n, the difference $(F_n(x) - F(x))^2$ is not zero

but for sure positive, thus the value of the integral is greater than 1. Finally, n is a positive integer and the multiplication with the integral is still greater than 1, proving the second part of the definition. □