On the Reliability of the Probabilistic Worst-Case Execution Time Estimates

Fabrice Guet, Luca Santinelli, Jérôme Morio

On the Reliability of the Probabilistic Worst-Case Execution Time Estimates

Fabrice Guet, Luca Santinelli, and Jérôme Morio
ONERA - The French Aerospace Lab, Toulouse France
{fabrice.guet|luca.santinelli|jerome.morio}@onera.fr

1. INTRODUCTION

The execution of software tasks within real-time systems needs to be analysed with respect to both functional and non-functional constraints. In particular, real-time systems require strict timing evaluations of the tasks execution behavior, especially their Worst-Case Execution Time (WCET).

Safety-critical embedded systems exhibit execution time variability, although classical real-time modeling and analyses account only for the worst-case. The systemic complexity of real-time systems comes from the hardware complexity (e.g., current multi-core architectures, shared resources such as memory, and speculative mechanisms like cache memories and pipelines [13, 23]), the software complexity (e.g., multiple embedded functionalities, wide interoperability, co-existence of functional and non-functional constraints), complex system component interactions and dependences, and diverse environments. All of them participate to the variability in the temporal behavior of the tasks.

Regarding this systemic complexity, probabilistic approaches are emerging as effective alternative to deterministic approaches for WCETs estimate. Their objective is to characterize system execution variabilities with probability distributions that associate to multiple possible WCETs their probability of occurence within a system execution trace, on contrary to deterministic approaches that provide a single WCET estimate. The challenge is to ensure the predictability based on the probabilities. So far, the probabilistic approaches are less costly in modeling task execution behavior and more accuracy with regard real-time systems average performances compared to the deterministic approaches.

This paper focuses on Measurement-Based approaches for Timing Analyses (MBPTA). MBPTA relies on both execution time measurements and the application of the Extreme Value Theory (EVT). Thus an exact model of both hardware and software is not required, contrary to deterministic approaches, as the measurement of the actual system behavior is sufficient. The MBPTA provides probabilistic Worst-Case Execution Time (pWCET) estimates[1] [6, 18]. Currently, the main problem of the MBPTA is the lack of mathematical robustness since EVT actual application relies on non systematic statistical approaches.

Hardware systemic effects in real-time systems [33] make EVT applicability difficult with regard to its required theoretical hypotheses. It is necessary then to ensure the applicability of the EVT to realistic embedded systems (non time-randomized embedded systems). Moreover, real-time systems require strong guarantees on the pWCET estimates thus, diagnostic tests have to be introduced to check hypotheses for generalizing the EVT applicability to any embedded systems [37].

In this paper we propose the logical workflow that checks the applicability of the EVT for the pWCET estimation problem. The proposed framework is a DIAGnostic tool for the eXTReMe value theory, named DIAGXTRM. The tool applies tests and makes a decision on the reliability of the resulting pWCET estimate without human intervention. The objective is to establish a systematic and reproducible process for estimating the pWCET which is able to cope with both performance and safety of existing as well as future real-time systems.

Organization of the paper. In Section 2 we relate the WCET problem, especially for the MBPTA, by depicting existing approaches and stressing the novelty of the proposed framework. In Section 3, we set the basics of the real-time probabilistic modeling and focus on the theoretical aspects of the EVT applicability. Section 4 presents the main steps of the DIAGXTRM tool, and Section 5 develops the tests that compose the tool. In Section 6 DIAGXTRM is applied to a realistic hardware platform running a set of tasks. Section 7 is for conclusions and future work.

2. RELATED WORK

Estimating the WCET of a task for hard real-time systems has been addressed in many ways [40]; all differ depending on the kind of hardware architecture.

Platforms are said to be deterministic whenever the execution time of a task is the same for the same input data. They are said to be non deterministic instead, whenever the execution time varies for the same input data. The non determinism comes from hardware components like cache memories, pipelines, etc. [35].

Static deterministic timing analysis and measurement-based deterministic timing analysis are effective for deterministic platforms. Static approaches provide safe WCET estimates as they are proved to be the worst. They rely either on an exact modeling of the system and a complete exploration of all its state or on a simplified version where some conditions are respected or even enforced. Measurement-based approaches provide timing behavior upper-bounds as distributions that overcome most of the possibilities. That is the reason why static approaches are preferred on measurement-based ones to give high guarantees on the system constraints. Nevertheless, when it comes to non deterministic architectures, static approaches produce pessimistic WCET estimates due to the overall systemic complexity; the analytical modeling phase is more and more difficult, the models confidence decreases, and the resulting estimates deteriorate [7]. However, tools based on the static modeling of both hardware and software aspects

[1] pWCETs are alternative to deterministic Worst-Case Execution Times as distributions with multiple extreme execution times, each with a probability of happening.

are able to provide safe but pessimistic WCET estimates because they take into account the worst-case at every modeling step. The estimates could be far from actual measurements and hardware performances.

The non determinism resulting from enhanced performance and the consequent execution time variability question the deterministic approaches. Facing this new challenge, probabilistic approaches tend to emerge: they can be either Static-based [18, 9] or Measurement-Based Timing Analyses (MBPTA). DIAGXTRM is a MBPTA approach and is able to capture well the systemic effects together with the coherence mechanisms between shared resources. As it relies on end-to-end measurements of the task execution time, it does not require a huge amount of information or exact hardware nor software models. The probabilistic worst-case profiles are derived on the basis of the set of execution time measurements. Nevertheless, as MBPTA relies on measurements, the lack of completeness of experimental conditions can lead to unsafe pWCET estimates due to unobserved execution conditions.

2.1 MBPTA approaches

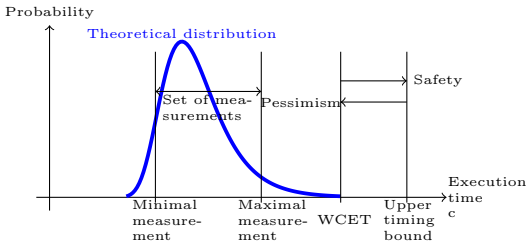


Figure 1: Overview of the WCET problem. Example of a timing probabilistic profile of a task.

The objective of MBPTA approaches is to derive probabilistic profiling of the timing behavior of a task, like in Figure 1, through a statistical modeling. Such a profile has to tend to the true theoretical distribution of execution times. In particular, MBPTA approaches are interested in modeling extreme execution times, for characterizing the worst-case, i.e. the values far from the average execution times, and potentially not measured. The probabilistic theory that focuses on extreme values and large deviations from the average values is the Extreme Value Theory (EVT) [20].

The EVT is a probabilistic paradigm that aims at predicting the improbable, i.e. it enables to derive the probability of rare events without requiring too many simulations. The EVT for estimating the WCET of a task in a scheduling analysis is first used in [19] where the Gumbel distribution is applied for modeling the distribution of execution times. A first algorithm for applying the EVT appeared in [25]. It extracts values from a sequence of execution time measurements according to the block maxima paradigm[2] and fits a Gumbel distribution to the measurements. Then the fitted distribution is compared to the measurements through a χ^2-test to confirm the model, otherwise the process is applied again for another number of extracted values.

The EVT applicability for embedded systems is first questioned in [33] and in particular about the statistical independence and the continuity of the execution time measurements. Two directions emerged from those questions: 1) the randomization of the hardware for solving the independence problem. Within the PROARTIS project first and then the PROXIMA project [1, 2], the EVT is applied to artificial (ad-hoc) random systems (Random Replacement policies in cache memories),[4, 15, 17]; 2) the elaboration of a methodology for guaranteeing the applicability of the EVT from the strong fundamentals of its mathematical hypotheses [5, 8, 31, 37] and derive reliable pWCET estimates from any real-time system (time-randomized as well as non time-randomized).

The approach proposed in the DIAGXTRM tool is a MBPTA approach and aims at solving the problem of the EVT applicability for real-time systems (both time-randomized and non time-randomized systems) by pursuing the works in [19, 25, 33, 37]. It represents the first structured and formal approach designed as a logical workflow that evaluates the EVT hypotheses for guaranteeing the MBPTA estimates. DIAGXTRM applies tests proved to be efficient for the considered analyses and an automatic parameter estimate process to provide pWCET estimates with an associated confidence for the EVT hypotheses.

3. PROBABILISTIC MODELING

The EVT relies on measurements of the system performance parameters, here the execution time of a task τ, for estimating extreme behaviors, where the worst-case should lie. The variability of the execution time of a task motivates its definition as a random variable[3], denoted by $\mathcal{C}$, which picks different possible values within the set[4] $\Omega \subseteq \mathbb{N}$, i.e. the distribution support of execution times the task τ can take to complete with a certain probability. The definition of a random variable stands for the uncertainty that lies on the uncertain systemic effects that occur in real-time systems. Each measurement C_i at discrete time instant i, is stored in a trace $\mathcal{T}$ such that $\forall i, \mathcal{T}(i) = C_i$. The length of $\mathcal{T}$ is denoted by n.

Three equivalent representations are used for $\mathcal{C}$, each is a probability distribution function: for all possible execution time $c \in \Omega$, it exists i) the cumulative distribution function (CDF) $F_{\mathcal{C}}(c) = P(\mathcal{C} \leq c)$, ii) the complementary cumulative distribution function function (CCDF) $\overline{F}_{\mathcal{C}}(c) = P(\mathcal{C} > c) = 1 - F_{\mathcal{C}}(c)$, and iii) the probability mass function $f_{\mathcal{C}}(c) = P(\mathcal{C} = c)$ (for a continuous random variable it is $f_{\mathcal{C}}(c) = \frac{d}{dc} F_{\mathcal{C}}(c)$). The discrete random variable $\mathcal{C}$, based on the execution time measurements is said to be the Execution Time Profile[5] (ETP) of the task τ.

One key element about the pWCET relates to its theoretical existence: *the pWCET exists but cannot be observed* since it is the distribution of extreme execution times that are very hard to measure and potentially impossible to observe. To measure execution times with very low probability (e.g., 10^{-9}), it would require a large amount of simulations and well defined experimental conditions. Moreover, the worst-case conditions have to be guaranteed to be explored making such approach very costly in terms of time and exploration conditions. The lack of completeness of the experimental conditions cannot ensure the existence of pWCET estimates directly from ETPs.

[2]The sequence is divided into blocks of same size and only the maximum value of each block is retained.

[3]A random variable is a variable whose value is subject to variations due to chance, i.e. randomness, in a mathematical sense. Generalizing, also non-variable execution time could be represented as random variables, with only one value and the probability of happening equal to 1. Since execution times are from measurements, they results into empirical random variables.

[4]Execution time can assume only discrete values as multiple of the system clock.

[5]ETPs are discrete distributions since task execution times can only be a multiple of the system clock.

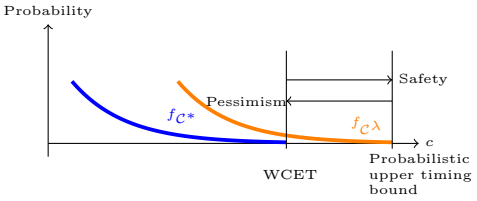


Figure 2: The pWCET estimate problem with the relationship between exact pWCET and pWCET estimate. An example of safe estimate $\mathcal{C}^\lambda$ with respect to the exact $\mathcal{C}^*$.

For formally defining the exact pWCET, we introduce the partial ordering of random variables by comparing their CCDF. Thus a random variable $\mathcal{C}_i$ is greater than or equal to a $\mathcal{C}_j$, $\mathcal{C}_i \succeq \mathcal{C}_j$, iff $P(\mathcal{C}_i > c) \geq P(\mathcal{C}_j > c)$, for every $c \in \Omega_{\mathcal{C}_i} \cap \Omega_{\mathcal{C}_j}$. Thus, the exact pWCET is the least upper random variable over all the ETPs for every execution condition. We denote the exact pWCET of a task by the random variable $\mathcal{C}^*$. Since exact pWCETs are impossible to obtain, as for any timing analysis approaches, we focus on pWCET estimates $\mathcal{C}^\lambda$. A pWCET estimate $\mathcal{C}^\lambda$ has to be safe i.e. has to be greater than $\mathcal{C}^*$, so $\mathcal{C}^\lambda \succeq \mathcal{C}^*$ like in Figure 2. The statistical modeling method from the EVT is the process we apply to achieve $\mathcal{C}^\lambda$.

3.1 Reliable pWCET estimates

The EVT is a widely used theory for predicting the improbable, i.e. giving probabilities of occurrence to extreme behaviors.

Under the hypothesis of *independent* and *identically distributed* (iid) execution time measurements $C_1, \ldots, C_n$ from an average discrete cumulative distribution function $F_{\mathcal{C}}$. The EVT ensures that the limit law of the maxima, i.e. the extreme execution times, denoted by $M_n = \max(C_1, \ldots, C_n)$ is a Generalized Extreme Value Distribution (GEV) H_ξ under norming constants such as the shape parameter $\xi \in \mathbb{R}$, the mean $\mu \in \mathbb{R} > 0$ and the variance $\sigma^2 \in \mathbb{R} > 0$ of the extreme execution times, with the Fisher-Tippett-Gnedenko theorem [20, 24].

This result implies that $F_{\mathcal{C}}$ belongs to the *Maximum Domain of Attraction* of the GEV H_ξ, denoted by $F_{\mathcal{C}} \in MDA(H_\xi)$. Given $\mathcal{C}$, whenever the iid hypothesis is respected and under good norming constants, the GEV is an appropriate distribution for the extreme execution times.

Depending on the value of ξ, the GEV can be either the Frechet ($\xi > 0$), the Gumbel ($\xi = 0$), or the Weibull ($\xi < 0$) distribution. In previous works the pWCET distribution has been assumed to be Gumbel, while here no assumption is made about the resulting GEV distribution and so there is no restriction on the values that ξ can take. The objective of the study is to get reliable pWCET estimates so that the distribution has to best-fit the measurements: the Gumbel can result from the best-fit or it can be imposed afterwards.

Considering $\mathcal{C}$ and $F_{\mathcal{C}}$, the CDF of the peaks $\mathcal{C} - u$ above the threshold u knowing $\mathcal{C} > u$ is

$$F^u(c) = P\{\mathcal{C} \leq u + c \mid \mathcal{C} > u\} = 1 - \frac{1 - F_{\mathcal{C}}(u+c)}{1 - F_{\mathcal{C}}(u)}. \quad (1)$$

If $F_{\mathcal{C}} \in MDA(H_\xi)$ then the limit law of the peaks is given by the Pickands theorem [36]:

THEOREM 3.1 (PICKANDS THEOREM). *$F_{\mathcal{C}} \in MDA(H_\xi)$ iff*

$$\lim_{u \to c_0} \sup_{0 \leq c \leq c_0 - u} |F^u(c) - GPD_\xi(c)| = 0, \quad (2)$$

where c_0 is the potential WCET of τ. GPD_ξ the Generalized Pareto Distribution with the same shape parameter ξ as H_ξ, and F^u from Equation (1).

The Pickands Theorem states that for values above a threshold, the nearest the threshold is to the actual WCET (which is the task execution time right end-point for increasing values) the more the distribution of execution times tends to a Generalized Pareto Distribution.

DEFINITION 3.2 (GENERALIZED PARETO DISTRIBUTION). *The distribution function GPD_ξ is the Generalized Pareto Distribution (GPD) defined as:*

$$GPD_\xi(c) = \begin{cases} 1 - (1 + \xi \times (c-u)/\alpha_u)^{-1/\xi} & if\ \xi \neq 0 \\ 1 - \exp(-(c-u)/\alpha_u) & if\ \xi = 0, \end{cases} \quad (3)$$

with $\alpha_u = \mu - \xi(u - \sigma^2)$, and defined on $\{c, 1 + \xi(c-u)/\alpha_u > 0\}$.

This fixes the basis of the EVT POT approach which consists in extracting the execution time measurements from $\mathcal{T}$ above a threshold u and fitting the experimental CDF with the continuous distribution function P_ξ. By applying the POT approach to the trace of execution times, the pWCET estimate which is the distribution of extreme execution times $\mathcal{C}^\lambda$ is a GPD.

For applying the EVT, one needs i) independent and ii) identically distributed execution time measurements from iii) a distribution in the Maximum Domain of Attraction of a GEV of shape parameter ξ. Those three elements are the hypotheses to check for having reliable pWCET estimates.

In practice, the independence hypothesis is difficult to assume because of history dependence in memory components as explained in Section 2.1. Moreover, the true distribution of the execution times is unknown and prevents from proving that execution times are identically distributed from a distribution in the Maximum Domain of Attraction of a GEV.

Further researches in the EVT domain proved the convergence of the Fisher-Tippett-Gnedenko theorem for stationary execution time measurements under two conditions [29, 30], and so the applicability of the EVT in the more general stationary case. The conditions especially relax the strict independence of the measurements and it is not necessary to know precisely the probabilistic law of the execution times as soon as they are stationary.

The strict hypotheses that prevented EVT applicability to non time-randomized embedded systems (todays systems), notably the independence, are so released allowing to apply the EVT to the pWCET estimate problem for any real-time system (both time-randomized and non time-randomized).

4. A DIAGNOSTIC TOOL FOR ESTIMATING THE PWCET WITH THE EXTREME VALUE THEORY

The main challenge of the MBPTA is the definition of a systematic approach that provides reliable pWCET estimates with the EVT. The reliability of a process comes from its definition: it is crucial to well identify the hypotheses and to choose both powerful tests and a proper parameter estimate process. A test is said to be powerful if it is able to reject a hypothesis when it is known to be false but also not reject it when it is known to be true. The reliability of the pWCET estimates holds if every hypothesis of the EVT is verified. Making use of the well defined tests and a proper estimate of

the distribution parameters, here ξ and α_u, the reliability can be guaranteed.

The DIAGXTRM, by construction tends to reduce the sources of uncertainty that lie on the execution time measurements to fulfill the EVT hypotheses and the selection of the threshold [38]; moreover it quantify the estimates confidence. In that sense, the tool is a diagnostic of the stastical modeling with the EVT.

The tool is designed as a logical workflow which checks the applicability of the EVT with specific tests. For an input trace of execution times, DIAGXTRM provides a pWCET estimate $\mathcal{C}^{\lambda}$ and an associated confidence with regard to the EVT applicability hypotheses. The hypotheses to check on the trace of execution time measurements are: 1) stationarity, 2) short range dependence, 3) local independence of the peaks, 4) empirical peaks over the threshold follow a GPD. The four hypotheses define the hypothesis testing blocks includes in the main steps of the tool, described in Figure 3. In this section the DIAGXTRM is presented at a high level; the tests that compose it will be detailed in the following section.

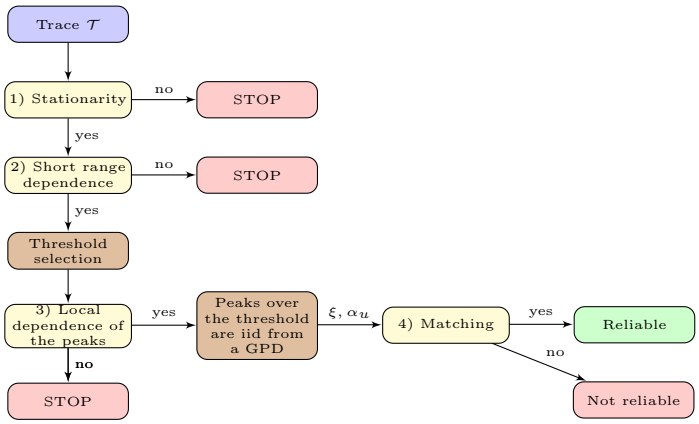


Figure 3: Decision diagram of the DIAGXTRM enlisting the tests and actions applied.

4.1 Design of the tests

DIAGXTRM is mainly based on the hypothesis testing theory that studies the rejection of a null hypothesis H_0. If H_0 is not rejected it is a necessary but not sufficient condition to satisfy H_0. The first step is to select an appropriate metric that evaluates the possibility of rejecting H_0, then the metric is applied to the trace $\mathcal{T}$ of execution time measurements returning a *result* through which making a decision about H_0.

In the design phase of the test, training sets are used to quantify the power of the metric for detecting H_0. The focus is on the conditional probability to reject H_0 knowing that H_0 is true $p-value = P(\overline{H_0}|H_0)$, which is the false positive rate of the test. The arbitrary threshold to reject H_0 is the value α defining the confidence interval for the test, hence for the hypothesis testing. A test may have a symmetric confidence interval, a two-sided test, otherwise this is a one-sided test. If the *result* of the applied metric to $\mathcal{T}$ is within the confidence interval, then H_0 is not rejected. Usually, α is chosen near 0 e.g., 0.01, 0.05 or 0.1, and corresponds to as many critical values *cv*s like in the two-sided test illustrated in Figure 4. If the *result* is in the darkest area, then H_0 is rejected but one has $\alpha \times 100\%$ of rejecting wrongly H_0.

We consider the possibility to fulfill H_0 [26], and use a fuzzy logic approach to test hypotheses. As the possibility to fulfill H_0 increases and so the confidence in H_0, the necessity to reject it decreases. Fuzzy logic is widely used to build decision making processes and is called robust statistics [12, 22, 26] when applied to statistics by quantifying the uncertainties associated to classical statistical approaches.

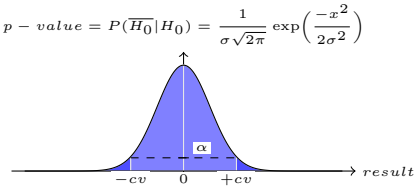


Figure 4: Hypothesis testing with a metric following a Gaussian law. (cv: the associated critical value to the α false positive rate)

For instance, instead of having or not a stationary trace of ETs, fuzzy logic quantifies whether the trace is near or far from the stationary model. The nearest it is the more confidence there is in the EVT applicability. Instead of one α level, 4 values are selected so that it is possible to either reject H_0 or accept H_0 with low, medium, high and full confidence level, corresponding to the $p-values$ 0.01, 0.025, 0.05 and 0.1. To resume, the approach we are formalizing for the pWCET estimation with EVT defuzzifies the statistical test by associating fuzzy $p-values$ to human-understandable confidence levels, depicted in Figure 5, and ease the decision making.

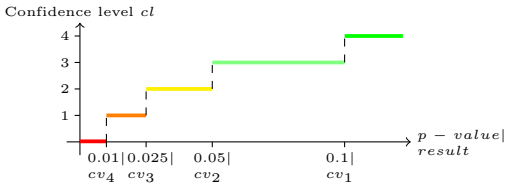


Figure 5: The DEFUZZIFICATION is a function from fuzzy $p-values$ (or equivalently their associated critical value (cv)) to confidence levels in $\{0;1;2;3;4\}$. For increasing confidence levels, H_0 is rejected or H_0 is accepted with low, medium, high and full confidence.

4.2 Decision making process

Each hypothesis testing block, blocks 1), 2), 3) and 4) in Figure 3, provides a *result* about the trace of execution times and so a confidence level as in Figure 5. Those confidence levels aims at reliable pWCET estimates with the EVT with regard to its hypothses applicability. One purpose of the fuzzy approach is to have a common scale for every test in order to aggregate each confidence level and to get a final confidence level on the pWCET estimate with the EVT. There exist many ways to aggregate the confidence levels, but one requirement is to have an aggregation in agreement with the tool specifications. In particular, the reliability is ensured when every hypothesis is guaranteed.

In the proposed process, there are four hypotheses to check: 1) the stationarity, 2) the short range dependence, 3) the local dependence of the peaks and 4) the matching with a GPD model. The final confidence level is denoted by $cl_{reliability}$ as a possibility metric to fulfill the whole process. Consequently, the confidence levels associated to each hypothesis are: cl_1, cl_2, cl_3 and cl_4. To statisfy the reliability requirement, if one confidence level is zero then the reliability has to be zero too. The confidence in the model is the barycenter of all the confidence levels so that it leads to Algorithm 1.

Algorithm 1 AGGREGATION algorithm of the confidence levels in the DIAGXTRM

1: $confidence\_levels \leftarrow [cl_1, cl_2, cl_3, cl_4]$ ▷ Previous analyses
2: **procedure** AGGREGATION($confidence\_levels$)
3: **if** $\min(confidence\_levels) \geq 1$ **then** ▷ Reliability
4: $cl_{reliability} \leftarrow (cl_1 + cl_2 + cl_3 + cl_4)/4$ ▷ Reliability levels
5: **else**
6: $cl_{reliability} \leftarrow 0$
7: **end if**
8: **end procedure**

Let H_0*: the EVT is applicable to* $\mathcal{T}$ be a null hypothesis, then $cl_{reliability}$ gives the confidence in fulfilling H_0. With regard to Algorithm 1, either H_0 is rejected for a null $cl_{reliability}$, or H_0 is accepted and in this case the higher $cl_{reliability}$ is, the more confidence in the model there is and thus in the pWCET estimates. The power of the tool to fulfill H_0 and to provide reliable pWCET estimates, depends also on the selected tests for each hypothesis. The DIAGXTRM is a high level methodology to provide reliable pWCET estimates, and one may easily replace a selected test in its respective hypothesis testing block by a better one thanks to new research works in time series (trace) analysis.

5. TESTS DETAILS

5.1 Stationarity analysis

Stationarity is an essential property in statistical analyses but it is usually assumed. The problem is even more diffcult because there is no systematic way to study stationarity and it often relies on subjective analyses [34].

DEFINITION 5.1 (STRICTLY STATIONARY TRACE). *A trace $\mathcal{T} = (C_1, C_2, \ldots)$ is a strictly stationary trace if for all j, k, l, the set of execution times $C_j, \ldots, C_{j+k}$ has the same probabilistic law as the set $C_{l+j}, \ldots, C_{l+j+k}$.*

If the execution time measurements in $\mathcal{T}$ are such that they respect Definition 5.1, then there is strong evidence that measurements are identically distributed (id) from the same probabilistic law (e.g., Gaussian, Gumbel, Weibull etc). As probabilistic laws are continuous, the stationarity analysis also addresses the problem of continuous execution times, even though execution times are discrete variables (see footnote 4). The stationarity analysis in the DIAGXTRM applies a test to check Definition 5.1.

As in practice the law of the execution times is unknown, we consider that a trace of execution times, at each discrete time instant t, can be written as the sum of a deterministic trend $f(t)$, a random walk r_t and a stationary residual ϵ_t [27]:

$$\mathcal{T}(t) = f(t) + r_t + \epsilon_t. \tag{4}$$

r_t is a random walk and may be written $r_t = r_{t-1} + u_t$ where u_t is a noise following a Gaussian distribution of mean 0 and unknown standard deviation σ_u. The Kwiatowski Philips Schmidt Shin (KPSS) test [27] checks whether $\mathcal{T}$ has a null deterministic trend and a null random walk for stating $\mathcal{T}$. In the case of a null deterministic trend, the KPSS test consists in testing the null hypothesis $H_0 : \sigma_u = 0$.

The KPSS test is applied to $\mathcal{T}$ and its confidence level is evaluated on the basis of the KPSS *result* and the associated $p - values$ as in Section 4.1, of the test detailed in [27].

5.2 Independence analysis

The independence analysis focuses on the short range dependence that refers to Berman's condition, or condition D in [29, 30]:

CONDITION 1 ($D(u_n)$). *For any integers p,q and n: $1 \leq i_1 < \ldots < i_p < j_1 < \ldots < j_p \leq n$ such that $j_1 - j_p \geq l$ we have*

$$|P(\{C_i, i \in A_1 \cup A_2\} \leq u_n) - P(\{C_i, i \in A_1\} \leq u_n) P(\{C_i, i \in A_2\} \leq u_n)| \leq \alpha_{n,l}, \tag{5}$$

where $A_1 = \{i_1, \ldots, i_p\}$, $A_2 = \{j_1, \ldots, j_p\}$ and $\alpha_{n,l} \to 0$ as $n \to \infty$ for some sequence $l = l_n = o(n)$.

For *distant enough* measurements, here l as the distance, and with u_n a sequence in the Fisher-Tippett-Gnedenko theorem, Condition 1 assures that the limit law of the maxima is still a GEV. In this view, blocks of execution time measurements of length l are considered, and their degree of correlation is evaluated with the Brock Dechert Scheinkman (BDS) test [11]. By choosing different values of length l, the degree of correlation varies and enables to identify particular patterns within the trace of execution times; Condition 1 holds for decorrelated blocks. The BDS test consists in testing the null hypothesis $H_0 : \mathcal{T}$ *is an iid trace* [11, 34] on the basis of the correlation integral. It allows to evaluate the statistical relationship between consecutive measurements (independence) and if they belong to the same distribution (identical distribution).

DEFINITION 5.2 (CORRELATION INTEGRAL). *The correlation integral $CI_{l,n}(\epsilon)$ at embedding dimension l for a distance ϵ is*

$$CI_{l,n}(\epsilon) = \frac{1}{\binom{n}{2}} \sum_{1 \leq s < t \leq n} \sum \chi_\epsilon(||C_s^l - C_t^l||). \tag{6}$$

For an iid trace $\mathcal{T}$:

$$\forall\ l,\ \epsilon,\ CI_l - CI_1^l \simeq 0 \ for\ n \to \infty. \tag{7}$$

The correlation integral measures the degree of correlation between patterns (C_s^l and C_t^l) of different lengths l within $\mathcal{T}$ depending the absolute distance ϵ and if it tends to the correlation integral of 1-length patterns (CI_1) to the power of l then the short range dependence is accepted. The *result* of the BDS test follows a Gaussian law of mean 0 and standard deviation 1 giving the critical values [11] and so the associated confidence levels as in Section 4.1. The BDS test is applied in practice as in Algorithm 2.

Algorithm 2 Application of the BDS test [10]

1: **procedure** INDEPENDENCEANALYSIS($\mathcal{T}$)
2: **for** $\epsilon \in \{\frac{1}{2}sd(\mathcal{T}), sd(\mathcal{T}), \frac{3}{2}sd(\mathcal{T})\}$ **do** ▷ Correlation distance
3: **for** l from 2 to $\frac{length(\mathcal{T})}{200}$ by 1 **do** ▷ Embedding dimension
4: $results.append$(DEFUZZIFICATION $(BDS(\mathcal{T}, \epsilon, l))$) ▷ *results* is a list of the BDS test results
5: **end for**
6: $cl_2 \leftarrow \frac{sum(results)}{length(results)}$ ▷ Aggregation of the results
7: **end for**
8: **end procedure**

5.3 Extreme independence

The reliability of the statistical model of the extreme execution time measurements depends on their independence. The extreme independence analysis depends on the selected threshold u like in Figure 3 that gives the peaks of execution time and stresses the presence of unreliable peaks that directly impact the GPD law. For instance, if an extreme burst of measurements occur, like many tasks running in parallel on different cores trying to access a memory unit at the same time, then all the measurements in or close to the burst depends on this same rare event. The peaks close to the burst are dependent endangering the pWCET estimates reliability, as formalized in condition D' [29, 30]:

CONDITION 2 $(D'(u_n))$. *The relation*

$$\lim_{i\to\infty} \limsup_{n\to\infty} n \sum_{j=1}^{[n/i]} P(C_1 > u_n, C_j > u_n) = 0, \qquad (8)$$

has to be verified.

Condition 2 imposes that for one measurement over the threshold then the probability the following execution times are over the same threshold has to tend to zero. If the relation holds for the trace of execution times then the peaks over the threshold are independent i.e. there is no cluster of extreme execution times. It is important to note that the relation highly depends on the selected threshold.

The extreme index (EI) θ, $\theta \in]0;1]$ [20], indicates the degree of clustering of the peaks over the threshold. The EI expresses the probability to have a peak distant enough from another peak to be independent. In presence of bursts of peaks this probability decreases in function of the bursts size. The more the peaks are distant from each other the more the probability and so the more EI is near 1. According to the GPD idea, the proability of occurence of a peak decreases exponentially leading to an estimator of θ [21].

We build the set of inter-arrival times T_i, defined by the amount of measurements between two peaks, that follow an exponential process of intensity θ. The distribution of the inter-arrival times provides the unbiased estimator [21]:

$$\theta = \frac{2\left(\sum_{i=1}^{k-1}(T_i - 1)\right)^2}{(k-1)\sum_{i=1}^{k-1}(T_i - 1)(T_i - 2)}, \qquad (9)$$

with k the number of peaks over the threshold also called the tail sample fraction standing for the number of execution time measurements that belong to the tail distribution.

Condition D' is a fuzzy condition, such that θ has to be near 1 in order to accept it. Hence, there is no systematic condition to conclude about the value of θ so that we define the confidence levels in Table 1 as in Section 4.1.

$\theta \in$	[1.00;0.95[	[0.95;0.90[	[0.90;0.85[	[0.85;0.80[	[0.80;0.00[
cl_3	4	3	2	1	0

Table 1: Confidence Levels of the Extreme Index.

As θ is the inverse of the mean size of the clusters, choosing 0.80 as the least bound on θ prevents from big clusters, so from unreliable pWCET estimates.

5.4 GPD parameter estimate and model matching

Independent peaks are extracted from the trace of execution times relatively to the selected threshold u, and are stored in a list *peaks*. It rests to estimate the parameters ξ and α_u of $F_{\mathcal{C}^\lambda}$ in Equation (3). For this purpose, we choose the *Maximum Likelihood Estimation* [20] method that performs well as it converges to convenient parameters.

Considering the set $\lambda = \{\xi, \alpha_u\}$ of the parameters to estimate according to a GPD, the *Maximum Likelihood Estimation* is an optimization problem that consists in exploring values of ξ and α_u and find λ that maximizes:

$$\ell(\lambda, exc) = \begin{cases} \ln \prod_{i=1}^{k} \frac{1}{\alpha_u}\left(\xi \times \frac{peaks[i]-u}{\alpha_u} + 1\right)^{-\frac{\xi+1}{\xi}} & if\ \xi \neq 0, \\ \ln \prod_{i=1}^{k} \frac{1}{\alpha_u} \exp\left(-\frac{(peaks[i]-u)}{\alpha_u}\right) & if\ \xi = 0. \end{cases} \qquad (10)$$

Once the extreme execution time measurements have been fitted with a GPD it is necessary to check whether they really come from a GPD. To do so we introduce a matching test based on a quadratic statistic because it measures the square distance between the empirical CDF of the extreme measurements and $F_{\mathcal{C}^\lambda}$ estimated previously. The *Cramer Von Mises criterion* (CVM) performs well in the case of Extreme Value Distributions [28, 39] because it detects well whether the measurements come from the chosen distribution. The *result* of the CVM test measures the *distance* between the empirical distribution of the measurements and the pWCET estimate according to a GPD; a *distance* is defined as the *result* of the CVM test. Thus, the nearer zero the *distance* it is the better the GPD fits the extreme measurements, hence the more reliable the model it is. For applying the CVM test, the extreme measurements are sorted increasingly in a list *uom* for upper-ordered measurements such that:

$$distance = \sum_{i=1}^{k} \left(F_{\mathcal{C}^\lambda}(uom[i]) - \frac{2i-1}{2\times k}\right)^2 + \frac{1}{12\times k}. \qquad (11)$$

Critical values of the CVM test are detailed in [16].

From the reliable pWCET estimate $\mathcal{C}^\lambda$ it is possible to derive WCET thresholds for a desired risk probability p. Formally, WCET thresholds are defined as the tuple $\langle WCET; p\rangle$ such that $p = P\left(\mathcal{C}^\lambda > WCET\right)$. For a desired risk probability e.g., 10^{-9} in aeronautics certification, the WCET threshold is directly given by [20]:

$$WCET = \begin{cases} u + \frac{\alpha_u}{\xi}\left(\frac{n}{k}p^{-\xi} - 1\right) & if\ \xi > 0, \\ u - \alpha_u \log(\frac{n}{k}p) & if\ \xi = 0, \\ u - \frac{\alpha_u}{\xi} & if\ \xi < 0. \end{cases} \qquad (12)$$

The rationale of the WCET thresholds lies on two pilars: measurements on the real-time system, and the applicability of the EVT in order to infer the probabilistic law of extreme execution times. For very low risk probabilities e.g., 10^{-9}, WCET thresholds may not appear in reality, they only exist on the basis of the mathematical rationale of the EVT, which is more rationale than adding a percentage to the maximal execution time measurement. In the case $\xi < 0$, the risk probability zero exists, so that the WCET is deduced. In static analyses the difficulty is to have a complete model of the system; wrong or non complete models endangers the estimate confidence, while the proposed approach directly faces the real system. Furthermore, the probability of the WCET threshold also depends on the probability of the execution conditions e.g., input task parameters, that lead to the execution time measurements.

5.5 Threshold selection

The peaks of execution time highly affect the pWCET estimate because the estimate has to fit the peaks according to a GPD. The threshold is a great source of uncertainties as

for different values correspond to different pWCET estimates increasing the uncertainty around the best estimate. Reviews for the threshold selection refer to many approaches [38] and still there is no systematic process for the selection. The threshold u is then a critical parameter because it directly provides the tail sample fraction k used for the parameter estimate, and impacting the reliability of the pWCET estimate.

We focus on the tail sample fraction to select the peaks such that the pWCET estimates uncertainty is minimized. To ensure tail convergence, k has to verify the two conditions, $\lim_{n\infty} k = \infty$ and $\lim_{n\infty} k/n = 0$ [38]. Hence, the tail sample fraction has to be small relatively to all the measurements in the trace and big enough to ensure the convergence of the limit law of the maxima. Moreover, for small values of k the GPD parameters vary a lot in function of k, whereas for greater values of k the parameters are biased by the amount of measurements, this is the bias-variance problem [20]. The threshold selection problem is resumed in Figure 6, where the central law of the measurements is a Gaussian distribution while the tail distribution is a GPD. The problem is then to estimate the right amount of execution time peaks by selecting the right threshold which lies in the uncertain threshold area. Thus we can only consider the tail distribution and not the central one. The existence of the right threshold relies on the hypothesis that the tail distribution of execution times converges well to a GPD.

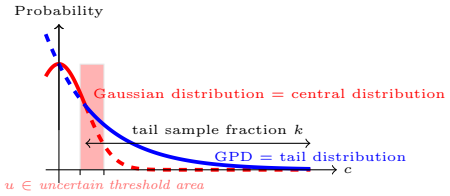
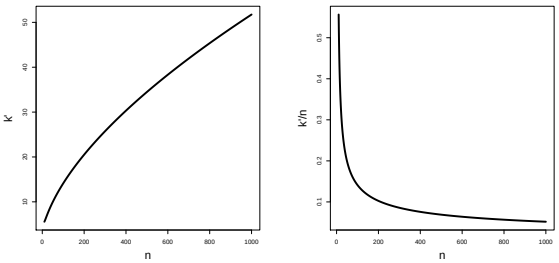


Figure 6: The threshold selection problem, [14].

One specification of the DIAGXTRM tool is to be fully automatic so that we choose to apply a computational method, based on the respect of the CVM criterion. To converge to the right amount of execution time peaks, we first scope a potential area based on a rule of thumb $k' = \frac{n^{2/3}}{\log(\log(n))}$ [32] ensuring above conditions of convergence as showed in Figures 7(a) and 7(b).



(a) k' in function of n. (b) k'/n in function of n.

Figure 7: Plots of the rule of thumb k' in function of the length n of the trace of execution time measurements.

An uncertain area is drawn around k', where the right threshold should lie. The number of execution time peaks varies in the interval $k \in [k_{low} = \lfloor 0.5 \times k' \rfloor; k_{up} = \lfloor 1.5 \times k' \rfloor]$ to explore other thresholds and still to cope with conditions of convergence. The threshold u is a function of the tail sample fraction k given by the quantile function q. q is a function of a percentage (between 0 and 1) and returns the execution time such that the desired percentage of measurements is below the returned execution time. Hence, $u = q(1 - k/n)$. Then, the peaks over u are fitted with a GPD giving the pWCET estimate and the distance between the experimental peaks and the pWCET estimate is evaluated with the CVM test. Finally, we iterate on k to cover the whole uncertain area.

The matching *result* given by the CVM test is a good indicator for selecting the right threshold because it indicates whether the execution time peaks really come from a GPD. Consequently, if the matching test gives a high confidence level for a threshold then we this threshold should be selected. To cope with conditions of convergence, a preference is added for thresholds given by a tail sample fraction close to k'. The matching test is so reduced from 4 to 3 confidence levels and a bonus in $[0; 1]$ is added depending on the value of k. The bonus evolves linearly from 0 to 1 in $[k_{low}; k']$ and from 1 to 0 in $[k'; k_{up}]$. To sum up, the computation of the confidence level for solving the threshold selection problem is presented in Algorithm 3.

Algorithm 3 Confidence level algorithm for the threshold selection

1: **procedure** THRESHOLDSELECTION($k, distance$)
2: $cl_4 \leftarrow$ DEFUZZIFICATION ($distance$) ▷ $\in \{0; 1; 2; 3\}$
3: **if** $k \geq k'$ **then**
4: $cl_4 \leftarrow cl_4 + \frac{1}{1-\frac{k'}{k_{up}}} - \frac{\frac{1}{1-\frac{k'}{k_{up}}}}{k_{up}} \times k$ ▷ $\in [0; 4]$
5: **else**
6: $cl_4 \leftarrow cl_4 + \frac{1}{1-\frac{k'}{k_{low}}} - \frac{\frac{1}{1-\frac{k'}{k_{low}}}}{k_{low}} \times k$ ▷ $\in [0; 4]$
7: **end if**
8: **end procedure**

As main contributions of the paper focus on the logical workflow and the decision making process regarding the applicability of the EVT, evaluations of Algorithm 3 will be the subject further investigations.

6. APPLICATION

In this section, DIAGXTRM is applied to a case study where the considered task is the *fibcall* from the Mälardelen benchmark [3]. To it, we intend finding its pWCET estimates in different execution conditions. The defined case study represents an example that could be done in an industrial declination. The *fibcall* task computes the i^{th} term in the Fibonacci sequence by a *for* loop implementation, with $i \in [2; 30] \cap \mathbb{N}$ so that there is no infinite loop. The set of possible inputs is denoted by $IN = \{i,\ i \in [2; 30] \cap \mathbb{N}\}$. The whole DIAGXTRM tool is implemented in R.

Hardware Platform.

The platform running the *fibcall* task has two Intel®Xeon®E5620 2.4 GHz sockets, each one with four cores and three levels of cache. The first two levels (L1 and L2) are private to each core, while the last level (LLC, equivalently L3) is shared to the cores belonging to the same socket.

Execution Conditions.

The task is implemented in C and runs periodically on one core; no interrupt (Irq) are present on the running core as they are redirected to other cores with Python system programming. To guarantee the real-time task execution, we set its scheduling policy to the Linux SCHED_FIFO policy. The

fibcall task is executed under different conditions to explore systemic effects (congestion and interference from shared resources) that can affect extreme execution times:

Scenario 1 $S1$**:** *fibcall* is executed in isolation, it represents the case with no task interference and the reference scenario to compare with.

Scenario 2 $S2$**:** *fibcall* is executed with the task *cnt* [3] on the same core, one after the other. *cnt* counts non negative numbers in a $10^4 \times 10^4$ matrix. Such a large data structure is applied to create interferences at different cache memory levels to *fibcall*.

Scenario 3 $S3$**:** *fibcall* is executed with the task *cnt* in parallel on a different core that shares a LLC with the core where *fibcall* runs. Thus no interference within the same core but interference through shared resources.

Scenario 4 $S4$**:** a combination of $S2$ and $S3$ with two *cnt* tasks. One *cnt* on the core where *fibcall* runs, and another *cnt* that runs in parallel on core sharing a LLC with the core where *fibcall* runs. Each *cnt* task explores its own matrix to create interferences at different cache memory levels and avoid concurrent problems.

The scenarios may correspond to different choices of tasks repartition in a safety-critical embedded system, and the objective is to cope with both aspects of timing performance and safety by respecting strictly given timing constraints. The experiment consists in executing 500 times the fibcall task according to each execution condition presented above. The longest experiment time is approximately 20 minutes due to the execution of *cnt* that has to be allowed to explore the whole $10^4 \times 10^4$ matrix in scenarios $S2$, $S3$ and $S4$. Task inputs in IN are imposed iid according to a Uniform law during the experiment as they are generated randomly by the random C function at each time instant.

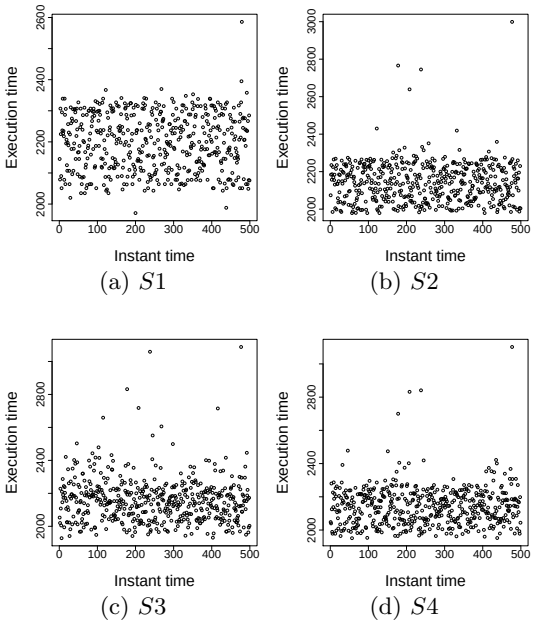
Results.

We now present the results of the experiments where execution times are measured in number of CPU cycles.

A first look at the traces in Figure 8 shows the repartition of the measurements and their randomness because there is no deterministic pattern over the time instants. Approximately, average execution times are between 2000 and 2300 CPU cycles. Measurements in the $S1$ case are concentrated in the average interval, while in the other cases, some measurements randomly deviate from the average interval.

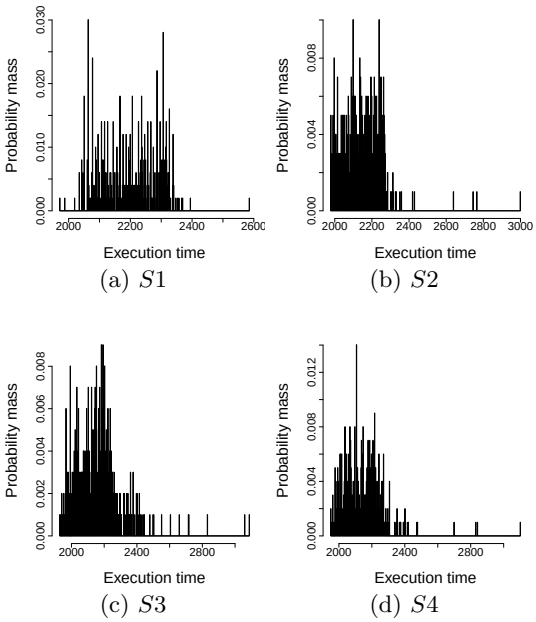
ETPs in Figure 9 confirm the different repartitions observed in each trace, and a more important presence of extreme execution times in $S2$, $S3$ and $S4$ than in $S1$. Each execution differs from another one by only a few *for* loops, at most 28, explaining the concentration of measurements in an average interval. The interferences introduced with task *cnt* appear clearly in the ETPs as some measurements deviate from the average interval.

DIAGXTRM is applied to every trace of execution times for deriving the pWCET estimate and evaluating its reliability for each scenario. The tool gives the modeling results of the extreme execution times in Table 2 and also the EVT applicability results in Table 3 for the reliability of the estimates. The selected thresholds u are between 2500 and 2400 CPU cycles right at the frontier with the average interval highlighted with the ETPs and traces. Hence, only the extreme execution times, which are outside the average interval, are used for the GPD parameter estimate. Every shape parameter ξ is strictly greater than 0, and the minimal one is $S1$'s which



(a) $S1$ (b) $S2$ (c) $S3$ (d) $S4$

Figure 8: Trace of execution time measurements for every execution condition.



(a) $S1$ (b) $S2$ (c) $S3$ (d) $S4$

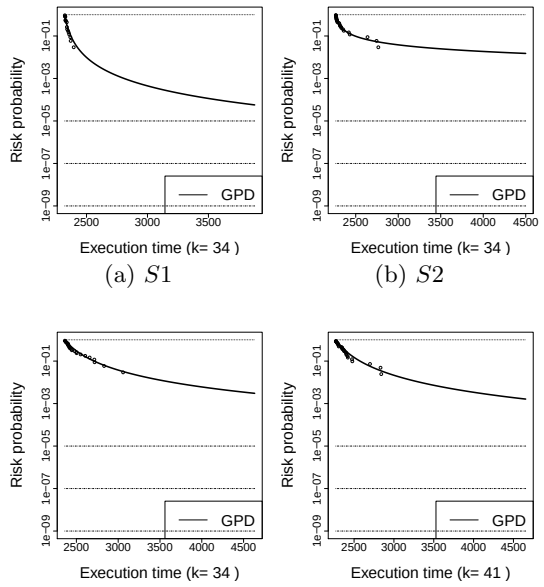
Figure 9: Experimental ETP for every execution condition.

is close to zero. By setting the risk probability p at 10^{-9}, as in aeronautics certification, the WCET threshold is deduced on the basis of parameters ξ and α_u for each respective scenario as in Equation (12). The greater ξ is, the more the WCET threshold diverges from the measurements; the greatest WCET threshold is $S2$'s which is around 10^7 times the respective maximal measurement. Estimated distributions of the extreme execution times are presented in Figure 10, showing the distribution convergence for each scenario.

All the traces are stationary (cl_1), with at least a high confidence and short-range independence (cl_2) is also verified for all the traces, as well as extreme independence (cl_3). Extreme

Trace $\mathcal{T}$	ξ	α_u	max	u	$\langle WCET; 10^{-9} \rangle$
S1	0.388	13.959	2586	2319.204	41719.696
S2	1.18	18.845	2999	2265.068	27960307288.975
S3	0.425	89.866	3088	2360.136	453591.23
S4	0.394	81.389	3102	2269.164	272528.444

Table 2: EVT results for every execution condition.



(a) $S1$ (b) $S2$

(c) $S3$ (d) $S4$

Figure 10: pWCET estimate for every execution condition.

independence is less obvious is S_4 but still accepted. The threshold selection criterion (cl_4) indicates also a high level of confidence for all the traces, because all matching confidence levels are strictly greater than 3. Algorithm 3 has the advantage to select the right threshold, if it exists, and to provide a confidence level about the distribution chosen for modeling the extreme execution times. Finally, the aggregation of all the confidence levels gives the reliability ($cl_{reliability}$) of the pWCET estimates, which are all strictly greater than 3, except for $S4$ that is quite near 3, then the pWCET estimates are highly reliable for all the scenarios.

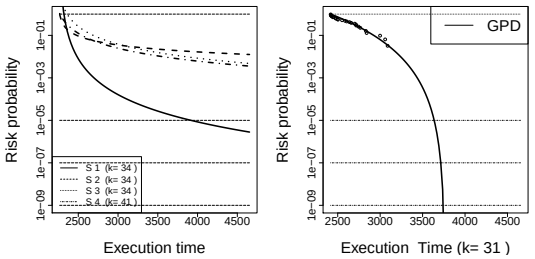
Trace $\mathcal{T}$	cl_1	cl_2	cl_3	cl_4	$cl_{reliability}$
S1	4	2.667	4	3.975	3.66
S2	4	4	3	3.975	3.744
S3	3	3.333	4	3.975	3.577
S4	4	2.333	1	3.604	2.734

Table 3: Confidence levels for every execution condition.

Originally, the distribution used for modeling the pWCET estimate is the Gumbel distribution ($\xi = 0$) by applying the block maxima approach. Within the DIAGXTRM this original approach may be evaluated by selecting block maxima instead of peaks and fitting a Gumbel GEV instead of a GPD. In this case study, given Theorem 3.1, the Gumbel distribution would be acceptable for the first scenario where ξ is near 0, and, as the Gumbel distribution converges to 0 faster than the Frechet ($\xi > 0$) distribution, it would decrease the pessimism of the WCET thresholds for the first scenario.

DIAGXTRM derives the pWCET estimate that best fits the peaks of execution time measurements for each scenario and all are diagnosed as reliable regarding the EVT applicability, however, the WCET thresholds of *fibcall* are all different. As pWCET estimates $\mathcal{C}^{\lambda,S2}$, $\mathcal{C}^{\lambda,S3}$ and $\mathcal{C}^{\lambda,S4}$ of scenarios $S2$, $S3$ and $S4$ converge slower than $\mathcal{C}^{\lambda,S1}$, they are more pessimistic due to the introduced interferences. Safety considerations would retain $\mathcal{C}^{\lambda,S4}$ as *fibcall* pWCET because $\mathcal{C}^{\lambda,S2} \succeq \mathcal{C}^{\lambda,S4} \succeq \mathcal{C}^{\lambda,S3} \succeq \mathcal{C}^{\lambda,S1} \succeq \mathcal{C}^{\lambda,S4}$ as shown in Figure 11(a), giving a WCET threshold of 2.796×10^{10} CPU cycles. However, the retained WCET threshold is more than 10^5 times greater than the WCET threshold in isolation which is 3500 CPU cycles questioning the rationale of this estimate. In scenarios $S2$, $S3$ and $S4$ the WCET threshold in isolation has more chances to be exceeded and respective probabilities to exceed it are 10^{-3}, 10^{-8} and 10^{-6}.

As interferences foster the appearance of extreme execution times, we gather the extreme execution times of all the scenarios. Let $\mathcal{T}^{\cap S}$ be the trace of extreme execution times of all the scenarios, then measurements are independent and stationary according to the diagnostic results and it is then possible to apply the EVT to $\mathcal{T}^{\cap S}$. Length of $\mathcal{T}^{\cap S}$ is 143 execution times so that the ideal tail sample fraction is 17 extreme execution times. The distribution of the extreme execution times of $\mathcal{T}^{\cap S}$ is deduced by applying the THRESHOLDSELECTION as in Algorithm 3 as shown in Figure 11(b). The final number of extreme execution times (31) is greater than the number given by Algorithm 3 (17), so that the distribution converges up to a risk probability of 10^{-9}. The matching confidence level (cl_4) of the model is equal to 4, so that the pWCET estimate $\mathcal{C}^{\lambda,\cap S}$ is fully accepted. As in this case $\xi < 0$, the WCET threshold for a null risk probability exists and is equal to 3764($= \lceil 3763.446 \rceil$) CPU cycles.



(a) Comparison of scenarios. (b) The tail of the tail distributions.

Figure 11: Plots including the four scenarios.

Under the hypothesis of non infinite blocking time of the task, the case $\xi < 0$ makes sense because the execution of *fibcall* has to end. By gathering larger amounts of extreme execution times from scenarios with different interference sources, estimates will refine the worst-case estimate. The degree of convergence of the pWCET estimate, given by ξ, indicates the impact of the introduced interferences, such that $S2$ is the scenario that impacts the most *fibcall* compared to the others. As a conclusion, *fibcall* WCET for the considered hardware platform is 3764 CPU cycles.

7. CONCLUSION

This paper presents the first sytematic and reproducible process for MBPTA approaches, with a logical workflow named DIAGXTRM, for applying the EVT to traces of execution times and deriving the pWCET of a task as well as its associated reliability. The systemic complexity of real time systems with non deterministic platforms (both time-randomized and non time-randomized) requires the use of MBPTA approaches to derive the pWCET of a task. The reliability of the pWCET

estimates in MBPTA approaches depends on the theoretical hypotheses of the EVT that have to be tested. Results of statistical tests are often fuzzy and it becomes hard to make a decision on their basis requiring the introduction of a metric that indicates the fulfillment of a hypothesis. Execution conditions that provide the execution time measurements directly impacts the pWCET estimate so that MBPTA requires conditions that foster extreme execution times to refine the task pWCET.

8. REFERENCES

[1] http://www.proartis-project.eu/.

[2] http://proxima-project.eu/.

[3] *WCET project/ Benchmarks*, 2013.

[4] J. Abella, J. del Castillo, F. Cazorla, and M. Padilla. Extreme value theory in computer sciences: The case of embedded safety-critical systems. In *Proceedings 26th Euromicro Conference on Real-Time Systems (ECRTS14)*. IEEE, 2014.

[5] J. Abella, J. del Castillo, M. Padilla, and F. Cazorla. 68. extreme value theory in computer sciences: The case of embedded safety-critical systems. In *Current Topics on Risk Analysis: ICRA6 and RISK 2015 Conference*, page 579.

[6] S. Altmeyer, L. Cucu-Grosjean, and R. I. Davis. Static probabilistic timing analysis for real-time systems using random replacement caches. *Real-Time Systems*, 51(1):77–123, 2015.

[7] S. Altmeyer, B. Lisper, C. Maiza, J. Reineke, and C. Rochange. WCET and mixed-criticality: What does confidence in WCET estimations depend upon? In *15th International Workshop on Worst-Case Execution Time Analysis, WCET 2015, July 7, 2015, Lund, Sweden*, pages 65–74, 2015.

[8] K. Berezovskyi, L. Santinelli, K. Bletsas, and E. Tovar. WCET measurement-based and extreme value theory characterisation of CUDA kernels. In *22nd International Conference on Real-Time Networks and Systems, RTNS '14, Versaille, France, October 8-10, 2014*, page 279, 2014.

[9] G. Bernat, A. Colin, and S. Petters. pWCET: A tool for probabilistic worst-case execution time analysis of real-time systems. Technical report, 2003.

[10] A. Bonache and K. Moris. Chaos dans les ventes de biens à la mode et implication pour le contrôle de gestion. Post-print, HAL, 2011.

[11] W. A. Brock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron. A Test for Independence based on the Correlation Dimension. *Econometric Reviews*, 15(3):197–235, 1996.

[12] J. J. Buckley. Fuzzy statistics: hypothesis testing. *Soft Comput.*, 9(7):512–518, 2005.

[13] A. Burns and B. Littlewood. Reasoning about the reliability of multi-version, diverse real-time systems. In *Proceedings of the 2010 31st IEEE Real-Time Systems Symposium*, RTSS '10, 2010.

[14] J. Carreau and Y. Bengio. A hybrid pareto mixture for conditional asymmetric fat-tailed distributions. *IEEE Transactions on Neural Networks*, 20(7):1087–1101, 2009.

[15] F. Cazorla, E. Quiñones, T. Vardanega, L. Cucu, B. Triquet, G. Bernat, E. Berger, J. Abella, F. Wartel, M. Houston, L. Santinelli, L. Kosmidis, C. Lo, and D. Maxim. PROARTIS: Probabilistically analysable real-time systems. *ACM Transactions on Embedded Computing Systems*, 2011.

[16] S. Csorgo and J. J. Faraway. The exact and asymptotic distributions of cramér-von mises statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 221–234, 1996.

[17] L. Cucu-Grosjean, L. Santinelli, M. Houston, C. Lo, T. Vardanega, L. Kosmidis, J. Abella, E. Mezzeti, E. Quinones, and F. J. Cazorla. Measurement-Based Probabilistic Timing Analysis for Multi-path Programs. In *23nd Euromicro Conference on Real-Time Systems (ECRTS)*. IEEE, 2012.

[18] R. I. Davis, L. Santinelli, S. Altmeyer, C. Maiza, and L. Cucu-Grosjean. Analysis of probabilistic cache related pre-emption delays. *Proceedings of the 25th IEEE Euromicro Conference on Real-Time Systems (ECRTS)*, 2013.

[19] S. Edgar and A. Burns. Statistical Analysis of WCET for Scheduling. In *RTSS*, pages 215–224. IEEE Computer Society, 2001.

[20] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events for insurance and finance*. Applications of mathematics. Springer, Berlin, Heidelberg, New York, 1997.

[21] C. Ferro and J. Segers. *Automatic Declustering of Extreme Values Via an Estimator for the Extremal Index*. EURANDOM, 2002.

[22] J. C. F. García and J. J. S. Méndez. A fuzzy logic approach to test statistical hypothesis on means. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, 4th International Conference on Intelligent Computing, ICIC 2008, Shanghai, China, September 15-18, 2008, Proceedings*, pages 316–325, 2008.

[23] M. Gardner and J. Lui. Analyzing stochastic fixed-priority real-time systems. In *5th International Conference on Tools and Algorithms for the Construction and Analusis of Systems*, 1999.

[24] E. Gumbel. *Statistics of Extremes*. Columbia University Press, 1958.

[25] J. Hansen, S. Hissam, and G. A. Moreno. Statistical-Based WCET Estimation and Validation. In *9th International Workshop on Worst-Case Execution Time Analysis (WCET'09)*, pages 1–11, 2009.

[26] Y. Kato, M. Takahashi, R. Ohtsuki, and S. Yamaguchi. A proposal of fuzzy test for statistical hypothesis. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 4, pages 2929–2934 vol.4, 2000.

[27] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root : How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3):159–178, 00 1992.

[28] F. Laio. Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters. *Water Resources Research*, 40, 2004.

[29] M. Leadbetter. On a basis for peaks over threshold modeling. *Statistics & Probability Letters*, 12(4):357–362, 1991.

[30] M. R. 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.

[31] M. Liu, M. Behnam, and T. Nolte. An evt-based worst-case response time analysis of complex real-time systems. In *8th IEEE International Symposium on Industrial Embedded Systems, SIES 2013, Porto, Portugal, June 19-21, 2013*, pages 249–258, 2013.

[32] M. Loretan and P. C. Phillips. Testing the covariance stationarity of heavy-tailed time series: An overview of the theory with applications to several financial datasets. *Journal of empirical finance*, 1(2):211–248, 1994.

[33] Y. Lu, T. Nolte, I. Bate, and L. Cucu-Grosjean. A Trace-Based Statistical Worst-Case Execution Time Analysis of Component-Based Real-Time Embedded Systems. In *16th IEEE International Conference on Emerging Technology and Factory Automation (ETFA11), WiP session*, September 2011.

[34] R. Manuca and R. Savit. Stationarity and nonstationarity in time series analysis. *Phys. D*, 99(2-3):134–161, Dec. 1996.

[35] V.-A. Paun, B. Monsuez, and P. Baufreton. On the Determinism of Multi-core Processors. In C. Choppy and J. Sun, editors, *1st French Singaporean Workshop on Formal Methods and Applications (FSFMA 2013)*, volume 31 of *OpenAccess Series in Informatics (OASIcs)*, pages 32–46, Dagstuhl, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[36] M. Piera-Martinez. *Modélisation des comportements extrêmes en ingénierie*. Theses, Université Paris Sud - Paris XI, Sept. 2008.

[37] 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*, pages 21–30, 2014.

[38] C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60, 2012.

[39] M. A. Stephens. Goodness-Of-Fit for the Extreme Value Distribution. *Biometrika*, 64(3):583–8, 1977.

[40] R. Wilhelm, J. Engblom, A. Ermedahl, N. Holsti, S. Thesing, D. B. Whalley, G. Bernat, C. Ferdinand, R. Heckmann, T. Mitra, F. Mueller, I. Puaut, P. P. Puschner, J. Staschulat, and P. Stenström. The worst-case execution-time problem - overview of methods and survey of tools. *ACM Trans. Embedded Comput. Syst.*, 7(3), 2008.