Probabilistic-WCET Reliability: On the experimental validation of EVT hypotheses

Federico Reghenzani
federico.reghenzani@polimi.it
Politecnico di Milano - DEIB
Milano, Italy

Giuseppe Massari
giuseppe.massari@polimi.it
Politecnico di Milano - DEIB
Milano, Italy

William Fornaciari
william.fornaciari@polimi.it
Politecnico di Milano - DEIB
Milano, Italy

Andrea Galimberti
andrea6.galimberti@mail.polimi.it
Politecnico di Milano - DEIB
Milano, Italy

ABSTRACT

The interest in probabilistic real-time is increasing, in response to the lack of traditional static WCET analysis methods for applications running on complex systems, like multi/many-cores and COTS platforms. However, the probabilistic theory is still immature and, furthermore, it requires strong guarantees on the timing traces, in order to provide safe probabilistic-WCET estimations. These requirements can be verified with appropriate statistical tests, as described in this paper, and tested with synthetic and realistic sources, to assess their ability to detect unreliable results. In this work, we identified also the challenges and the problems of using statistical test based procedures for probabilistic real-time computing.

CCS CONCEPTS

• **Computer systems organization → Real-time systems**; *Embedded systems.*

KEYWORDS

Probabilistic Real-Time, pWCET, Extreme Value Theory

ACM Reference Format:
Federico Reghenzani, Giuseppe Massari, William Fornaciari, and Andrea Galimberti. 2019. Probabilistic-WCET Reliability: On the experimental validation of EVT hypotheses. In *INTERNATIONAL CONFERENCE ON OMNI-LAYER INTELLIGENT SYSTEMS (COINS), May 5–7, 2019, Crete, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3312614.3312660

1 INTRODUCTION

The estimation of the *Worst-Case Execution Time (WCET)* is essential for hard real-time systems, in which the timing constraints of the tasks must be guaranteed under any condition. Failing to meet these constraints leads the system to misbehave with possible unacceptable consequences, especially in the case of mission-/safety-critical applications. Consequently, a timing analysis requires the estimated WCET value for a critical task to be greater or equal to the real WCET. On the other hand, this estimation must be as tight as possible to the real WCET, in order to minimize the resource assignment over-provisioning.

Recently, getting a safe but tight WCET has become a challenging problem. The growing computational power demand of embedded systems, in addition, but opposed to, the reaching of technology limits, is increasing the hardware complexity of processors – such as the introduction of many-cores, multi-level caches, complex pipelines, etc… This leads to hindering the use of traditional WCET estimation techniques [17] [19] [5]. The problem is even magnified when dealing with Commercial-Off-The-Shelf (COTS) components, mixed-criticality and general-purpose operating systems [26] [27].

1.1 Probabilistic Real-Time

Given the aforementioned scenarios, *probabilistic (hard) real-time* has been proposed as a possible solution to WCET estimation problem. This approach is founded on the well-known *Extreme Value Theory (EVT)*, which is typically applied to natural disaster prediction, For example, to estimate the probability of unseen catastrophic floods. The theory is briefly described in Section 2. The use of EVT in real-time systems has been proposed at the beginning of 2000s by Burns et al. [7] and Bernat et al. [4]. The first paper described EVT and how it can be used for probabilistic real-time analysis, while the latter focused on the algebraic properties needed to combine several probabilistic-WCET estimations. A few years later, the EVT has been applied with measurement-based methodologies [32].

Probabilistic real-time based approaches can be divided into two classes [1]: *Static Probabilistic Time Analysis (SPTA)* and the *Measurement-Based Probabilistic Time Analysis (MBPTA).* MBPTA, the subject of this work, has been proposed to estimate the so-called probabilistic-WCET (pWCET), by directly sampling the execution times of the tasks. Unlikely the classical WCET estimations, the pWCET is not a single value. Rather it is a statistical distribution, characterized by the following *cumulative distribution function (cdf)*:

$$p = P(X > \overline{WCET})$$

where X is the random variable representing the task execution time. By using this distribution, it is possible to compute the probability

COINS, May 5–7, 2019, Crete, Greece

ACM ISBN 978-1-4503-6640-3/19/05…$15.00
https://doi.org/10.1145/3312614.3312660

of violation (p) of a given $\overline{WCET}$ or, vice versa, the $\overline{WCET}$ given the probability of violation (p). The pWCET is considered *safe* if the estimated distribution "upper-bounds" the worst-case execution time with a probability value equal or higher than the real one [1].

Lately, some researchers on probabilistic analyses focused their effort on creating architectures, from which to generate time traces fulfilling the EVT requirements (see Section 2.1) [18] [9]. These results are however considered controversial [30]. Others focused on the theoretical aspects of MBPTA [23] [32] [2], which still presents several challenges to address [13] [29].

Contributions. Several articles in literature assume the truth of the EVT hypotheses or do not systematically assess their validity: some works applied improper hypothesis tests, erroneously run multiple tests on the same data, or reached conclusions without a proper evaluation of the statistical effects. Strategies based on expert knowledge instead, like graphical plot analysis, do not offer a systematic approach and thus quantitative information on the pWCET reliability.

In this article, we aim at (1) analyzing and making a selection of the statistical tests fitting the probabilistic real-time computing case; (2) clearly stating the problems and the statistical aspects affecting the pWCET reliability. Moreover, we highlight some common errors recurring when statistical tests are applied to MBPTA.

2 EXTREME VALUE THEORY IN REAL-TIME COMPUTING

The statistical theory of extremes has been developed to study the "tails" of a distribution. In this regard, the aforementioned *Extreme Value Theory (EVT)* is opposite to the central limit theorem, which focuses on the behaviour of the distribution around its mean value.

Given a sequence of independent and identically distributed random variables $X_1, X_2, ..., X_n$, the EVT provides the limit distribution at the extremes, i.e. the $\max(X_1, X_2, ..., X_n)$ or $\min(X_1, X_2, ..., X_n)$. In a real-time computing scenario, $X_1, X_2, ..., X_n$ is a sequence of execution times of a given task. Consequently, since for the WCET estimation we are interested in the maximum value, we can formalize the probability of not incurring in a execution time longer than a certain value x as follows:

$$\begin{aligned} &P(\max(X_1, X_2, ..., X_n) \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) \\ &\overset{\text{iid}}{=} P(X_1 \leq x)P(X_2 \leq x) \cdots P(X_n \leq x) = F^n(x) \end{aligned} \tag{1}$$

As we mentioned above, $F(x)$ is the cumulative distribution function (cdf) of $X_1, X_2, ..., X_n$. Without entering in statistical details, it is possible to demonstrate that [8]:

$$\exists a_n, b_n \text{ s.t. } \lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{2}$$

where $G(x)$ is the cdf of the so-called *Extreme Value Distribution*. The form of this distribution can be generalized as subsequently described and its parameters can be estimated from data.

The parameters of $G(x)$ can be estimated by grouping the data using the *Block-Maxima (BM)* or the *Peak-over-Threshold (PoT)* approach. In the first case, the time values are grouped inside blocks

[1] Formal definitions for pWCET comparison are available in [32].

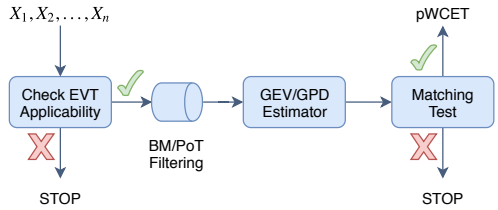

Figure 1: pWCET estimation flow based on the EVT.

of constant size B, to then compute the maximum value for each block. Formally:

$$\begin{aligned} X^{BM} &= \{X_1^{BM}, X_2^{BM}, ..., X_{n/B}^{BM}\} \\ X_i^{BM} &= \max(X_{B\cdot(i-1)+1}, X_{B\cdot(i-1)+2}, ..., X_{B\cdot i}) \end{aligned} \tag{3}$$

The latter instead, discards values by removing any sample lower than a predefined threshold P:

$$X^{PoT} = \{X_i \text{ s.t. } X_i > P\} \tag{4}$$

According to the Extreme Value Theory [11], X^{BM} and X^{PoT} converge respectively to the *Generalized Extreme Value Distribution (GEVD)* and to *Generalized Pareto Distribution (GPD)*. These distributions can be then exploited to compute the pWCET, following the flow depicted in Figure 1 [28].

The pWCET obtained is thus representative of the real distribution of extremes, and consequently safe for real-time computing, if and only if the following EVT hypotheses hold:

(1) the input measurements must be *identically and independently distributed (i.i.d.)*
(2) the X^{BM} or X^{PoT} set must be in the *domain of attraction* of an extreme distribution

As explained in the next sections, both the hypotheses are necessary to obtain a reliable pWCET.

2.1 The i.i.d. hypothesis

As for other statistical theories, the classical formulation of EVT requires the random samples to be identically and independently distributed. In real-time computing, this hypothesis is mainly dependent on the processor and the system architecture. For example, a processor with a standard cache would not be able to fulfill the independence requirement, because subsequent executions of the same task will be affected by the status of the cache. In practice, the i.i.d. requirement can be relaxed in favor of the stationary property and weaker independence properties [21] [33]. Such hypotheses must hold [32] and can be formalized as follows:

Stationarity. Given a random sequence $X_1, X_2, ..., X_n$ of size n, the process is said to be *strict stationary* iff for any choice of k, l, m with $0 < k + l + m < n$ the following condition is true: $F(X_k, X_{k+1}, ..., X_{k+l}) = F(X_{k+m}, X_{k+m+1}, ..., X_{k+m+l})$, where F is the cdf of the joint distribution. This condition implies identical distribution of the random variables. In real-time computing, the stationary hypothesis indicates a flat distribution of execution times, with constant variance. A task that drastically changes the job execution time after some runs, for instance, violates this property.

Short-range independence. Given a sequence of random variables $X_1, X_2, ..., X_n$ of size n, the sequence is said to be short-range

independent if for any $i_1 < i_2 < \cdots < i_p < j_1 < \cdots < j_p \leq n$ s.t. $j_1 - i_p \geq s > 1$, defining F_{IJ} the cdf of $i_1, ..., i_p, j_1, ..., j_n$, F_I the cdf of $i_1, ..., i_p$, F_J the cdf of $j_1, ..., j_p$ we have $|F_{IJ} - F_I F_J| \leq \alpha_{n,s}$ where α is non-decreasing in s and $\lim_{n\to\infty} \alpha_{n,[n\sigma]} = 0, \quad \forall \sigma > 0$. In real-time computing, an example of cause of short-independence property violation is the presence of processor cache effects between two job instances.

Long-range independence. According to this property, the time series does not show a significant correlation across large time-spans. We define this property by defining its opposite. A long-range dependent sequence can be defined as: a random sequence $X_1, X_2, ..., X_n$ of size n is said to have long-range dependence if its auto-correlation function $\rho(\tau)$ decays exponentially: $\rho(\tau) \sim \frac{L(\tau)}{\tau^{1-2d}}$ with $0 < d < \frac{1}{2}$ where $L(\tau)$ verifies $\lim_{t\to\infty} \frac{L(at)}{L(t)} = 1$ for some $a > 0$.

2.2 The domain-of-attraction hypothesis

The last property, introduced in [32], is called *matching* and it is related to the *domain of attraction* hypothesis, i.e. the convergence to an EVT distribution of the random output sequence of BM (or PoT). This depends on several factors, including the BM (or PoT) procedure itself and how the timing samples are measured.

When the timing samples are represented with continuous variables, such as directly time measure the system, the domain of attraction hypothesis is verified in the overwhelming majority of the times [31]. This is not necessarily true instead when we consider discrete distributions.

The matching property is usually checked with *a posteriori* statistical tests, that verify whether the resulting distribution actually matches the input data. Typical tests are the Kolmogorov-Smirnov [16] and Anderson-Darling [34].

This hypothesis is, on the one hand, not very generalizable, on the other hand, often true for continuous data, thus not interesting with respect to the scope of this paper. This does not mean that this hypothesis should not be considered in probabilistic real-time analyses, but since it depends on several factors, we need specific procedures for each scenario.

3 STATISTICAL TESTS IN MBPTA

All the hypotheses previously described can be verified through suitable statistical tests. The results are reject/not-reject responses that corresponds to the adherence or not to the EVT hypothesis. This, in turn, can represent a true/false boolean response to the problem of verifying the pWCET reliability. Therefore, performing proper statistical tests in a correct way is fundamental, other than being a necessary step towards the certifiability of the probabilistic approaches.

3.1 Assessing the EVT hypotheses via hypothesis testing

In this paragraph, we discuss about the choice of the of statistical tests needed to verify the previously described hypotheses: stationary, short-range independence and long-range independence.

A statistical test is typically described by its hypothesis scheme. Usually the symbol H_0 represents the null hypothesis, while the symbol H_1 or H_a the alternative hypothesis. The result of a test can be "reject the null hypothesis" or "unable to reject the null hypothesis". In the first case, the test has detected strong evidences that the null hypothesis is probably false, while the alternative hypothesis is probably true.

The outcome of a statistical test (reject/not-reject) comes from the evaluation of the *p-value* or the *critical value*. As the two approaches are exactly equivalent, we have decided to consider only the second one. The critical value is a constant value derived from the significance level α. It is compared against the *statistic* computed over the data to take the reject/not-reject decision. The computation of both critical value and the statistic depends on the specific test.

Stationarity.

In statistical literature several studies on stationary processes are available together with several test procedures. In particular, there is a large availability of unit-root tests – a particular case of non-stationarity – but less availability of general stationary tests. Given a time series $X = \{X_1, X_2, ..., X_n\}$ we are searching a test with the following hypothesis scheme:

H_0 : the time series X is stationary
H_1 : the time series X is not stationary

In this regard, the most used one is the Kwiatkowsky, Phillips, Schmidt and Shin (KPSS) test [20]. While, a variant of KPSS considers a relaxed null hypothesis "the time series is stationary or trend stationary". For the EVT hypothesis of stationarity, we are interested in the most strict one, thus we do not consider this variant.

Short-Range dependence. To test the short-range dependence of data, we selected the Brock, Dechert, Scheinkman and LeBaron (BDS) test [6]. For probabilistic real-time, we decided to select this test, since being a *portmanteau test*, i.e. the null hypothesis is well specified, but the alternative hypothesis it is not.

Given a time series $X_1, X_2, ..., X_n$:

H_0 : the time series x is independent
H_1 : the time series X has some sort of dependency

Most of the other available tests detect specific sort of dependency (e.g. serial correlation or deterministic chaos). Therefore, we decided to maintain the test with the most general detection capability. This increases the time trace rejections, i.e. erroneous rejections of safe probabilistic-WCET, but it also reduces the false negative results, i.e. missing rejections of unsafe probabilistic-WCET.

Long-Range dependence. The *Hurst Exponent* (H) is the traditional index used to measure the long-term memory of a time series in financial applications [25]. H is a number in the range $[0; 1]$ indicating the degree of long-term dependency: $H = 0.5$ means a perfectly random and uncorrelated time series, while $H < 0.5$ or $H > 0.5$ indicates a negative or positive correlated time series, respectively. However, performing a statistical test on H is nontrivial [10] and, to the best of our knowledge, it does not exist a well-assessed test. The Hurst index is computed from the R/S statistic equation [15] instead, that can be directly used as a test:

H_0 : the time series has no long-range dependency
H_1 : the time series has long-range dependency

This test is sensitive to long-range dependency but also to short-range dependency.

Test result interpretation. An improper test selection, a too small sample size or the continuous re-sampling to obtain a time trace that passes the tests are common examples of not carefully analyzed scenarios, that often led to misleading conclusions.

The execution of multiple tests on the same data is another case where the shift of the α critical level is not often taken into account, as described in Section 3.2.

3.2 Sample size and significance level

Once the statistical test procedure is defined, the next critical step is to set the sample size and the significance level α.

The sample size, i.e. the number of time measurements composing a time trace, is a parameter affecting both WCET estimation precision and safety. Using a limited number of samples may cause the failure of the extreme value distribution estimator or, even worse, the estimation of an incorrect distribution. The sample size affects also the reliability of the statistical test results. However, in previous works, this was often a not considered or not well justified aspect. Most of the times, the number of measurements was empirically established. In this section and in the following one, we will argue about the importance of the sample size in probabilistic real-time and how it affects the reliability of KPSS, BDS, and R/S.

Moreover, when the sub-hypotheses presented in the previous section has to be checked, the experimenter performs a sequence of three statistical tests, usually on the same data. In general, executing multiple hypothesis tests on the same data increases the false-positive rate on the null hypothesis rejection of the overall test [3]:

$$\alpha_{\text{global}} = 1 - (1 - \alpha)^n \tag{5}$$

where n is the number of tests (in our case $n = 3$).

For common values $\alpha = 0.05$ and $\alpha = 0.01$, the resulting global significance levels are respectively $\alpha_{\text{global}} \approx 0.14$ and $\alpha_{\text{global}} \approx 0.03$. The real significance level is thus higher than the single test levels, entailing a higher false-positive rate in rejection. Rejecting a sample implies that the pWCET estimation process stops, avoiding unsafe pWCET estimation. This makes it difficult to characterize an architecture according to its capability of fulfill the EVT hypothesis: it is not possible to use a single test result. The test needs to be run several times and we need to consider the overall ratio reject/not-reject: a rejection ratio near α identifies an architecture that verifies the EVT hypotheses, while an higher ratio represents a violation of EVT hypotheses.

3.3 Safety considerations

When the result of a hypothesis test is evaluated, the experimenter can incur in two possible errors: (1) reject the null hypothesis when it is actually true (*Type I* error) and (2) retain the null hypothesis when it is actually false (*Type II* error). The experimenter can control the Type I error by changing the significance level α. Type II error depends rather on the statistical power of the test. Unfortunately, the statistical power is neither simple to control nor to estimate.

Let W be the statistical power, its analytical definition is: $W = 1 - \beta$ where $\beta = P(\text{Accept } H_0 | H_0 \text{ is false})$. The statistical power W depends on several parameters, including the significance level, the input data distribution, the test statistic and the sample size. It can however easily be increased, by enlarging the sample size.

	Reject_{any}	Reject_{KPSS}	Reject_{BDS}	$\text{Reject}_{R/S}$
A1	13.9%	6.8%	5.5%	4.5%
A2	12.3%	5.3%	4.9%	4.3%
A3	11.4%	4.9%	5.3%	3.4%
B1	100%	4.7%	100%	6.9%
B2	100%	100%	100%	100%
B3	100%	83.1%	99.2%	98%
B4	100%	100%	5.5%	100%

Table 1: Tests rejections of synthetic time traces. The 'any' column represents rejection percentage of at least one test.

In our scenario, the Type II error represents the inability to detect a violation of EVT hypotheses, which consequently generates an incorrect extreme-value distribution, that may lead to unsafe pWCET computation. For hard real-time systems, a preliminary study on the statistical power is therefore necessary, to both select the proper sample size and to estimate the statistical power. The latter can then be used in the evaluation of the pWCET reliability and, consequently, in the safety analysis of the overall system.

4 EXPERIMENTAL EVIDENCES

In this section, we present the experimental evaluation of the chosen statistical tests. The expectation is to get high rejection rates for time traces that do not satisfy the conditions described in Section 2.1. On the other hand, if the source of the samples is a distribution that verifies the EVT hypotheses, then the rejection rate should settle around the significance value α.

4.1 Time trace sources

For characterizing the properties of the proposed test, we used both synthetic time samples and real benchmarks executions. The first class of time traces has been designed to stress the detection capability of each statistical test. The real benchmarks are instead executed on different hardware platforms, with well-known real-time capabilities, to show an evaluation of the probabilistic predictability of the target system.

Without losing generality, we evaluated the tests with a level of significance $\alpha = 0.05$. This means that we expected for each test a type I error (i.e. false-positive rate) of 5%. In our scenario, this is a conservative error: each test excludes 5% of the times a dataset that is actually valid for EVT estimation. The overall type I error can be computed using Equation 5, obtaining 14% ($\alpha_{global} = 0.14$).

Synthetic sources. Let $X_{a:b}$ be an ordered subset of the full time trace $X_{1:n}$. For synthetic and controlled time traces we used both i.i.d. and non i.i.d. sources. For the former, we selected the following EVT-compliant distributions:

A1 $X_{1:n} \sim \mathcal{N}(10, 1)$: Gaussian (normal)
A2 $X_{1:n} \sim \mathcal{P}(10)$: Poisson
A3 $X_{1:n} \sim \Gamma(10, 1)$: Gamma

Then we tested three non-compliant distributions:

B1 $X_{1:\frac{n}{2}} \sim \mathcal{N}(10,1); X_{\frac{n}{2}+1:n} \sim \mathcal{P}(1)$: a normally distributed time trace for the first half part and then a Poisson distribution; it represents a sequence of independent but not identically distributed samples.

B2 $X_{1:n} \sim AR(2)$: an auto-regressive model of order 2, with constant 10 and auto-regressive coefficients $(0.7, 0.25)$. This class represents a short-range dependent time source.

B3 $X_{1:n} \sim ARFIMA(\frac{1}{2}, 0, 0, 0, \frac{1}{4})$: an auto-regressive fractionally integrated moving average model with AR, MA, and I coefficients zero, constant $\frac{1}{2}$ and $d = \frac{1}{4}$. This class represents a time source with long memory.

B4 $X_{1:n} = \{\forall i \in [1;n] | X_i \sim N(10 + 0.001 \cdot i, 1)\}$: non identically distributed samples with long-range dependence, but short-range independent.

We have drawn a total of 1 000 000 samples for each distribution and then we split in groups of size 1 000 for a total of 1 000 evaluations.

Real sources. Concerning the experimental evaluation on real platforms, we run four state-of-the-art benchmarks of the WCET Mälardalen suite [14]: `sqrt`, `minver`, `fdct`, `complex`. We implemented each benchmark onto five different platforms, whose well-known architecture characteristics introduce different degrees of unpredictability:

R1 `PIC`: time-deterministic and simple processor: a PIC18F45K50 microcontroller without operating system;

R2 `STM`: time-deterministic platform with a L1D and L1I cache: STM32F7 board programmed bare-metal without operating system;

R3 `MIO`: time-deterministic platform with a real-time operating system: the STM32F4 with Miosix operating system [2];

R4 `ODR`: partially unpredictable platform: multi-core Odroid XU-3 with a Linux OS (vanilla kernel);

R5 `INT`: completely unpredictable platform: multi-core Intel i7 with a Linux OS (vanilla kernel).

The benchmarks have been slightly modified to add: (1) a PRNG for input data generation (with the exception of `complex` where the input is constant), (2) an external loop to run the benchmark multiple times, (3) a toggling mechanism for a GPIO to signal the start and stop of a benchmark execution. To maintain consistency among all platform, the PRNG has been initialized with the same seed. This way each platform generates the same sequence of pseudo-random inputs to the benchmarks. The time measurements have been acquired by measuring the GPIO interval between the rising edge (start of the computation) and the falling edge (end of the computation), using a commercial logical analyzer with a $10ns$ resolution. Each benchmark then has been executed 100 000 times by using time series of size 1 000 for statistical testing, for a total number of 100 estimations for each benchmark.

4.2 Results

Synthetic samples. The results on time traces from synthetic sources are shown in Table 1. For i.i.d. datasets (A1-A3) it is possible to notice a rejection rate based on evaluations of single tests around 5%, that actually matches the chosen significance level α. The rejection rate of all the tests is slightly below 14%, that is the significance level value computed by using Equation 5. This value represents the false-positive error rate, i.e. the percentage of time series that is discarded even if they are generated by compliant sources.

[2] http://miosix.org/

		Reject$_{any}$	Reject$_{KPSS}$	Reject$_{BDS}$	Reject$_{R/S}$
sqrt	R1	23%	7%	15%	6%
	R2	24%	2%	20%	3%
	R3	29%	6%	23%	7%
	R4	14%	1%	13%	1%
	R5	34%	20%	12%	21%
minver	R1	16%	6%	6%	6%
	R2	15%	9%	5%	6%
	R3	17%	10%	7%	8%
	R4	44%	1%	44%	1%
	R5	78%	61%	31%	66%
fdct	R1	8%	2%	5%	2%
	R2	20%	4%	15%	4%
	R3	100%	99%	100%	99%
	R4	62%	16%	48%	15%
	R5	81%	67%	45%	71%
complex	R1	79%	19%	72%	16%
	R2	56%	5%	52%	3%
	R3	100%	0%	100%	0%
	R4	92%	5%	92%	1%
	R5	100%	60%	95%	71%

Table 2: Tests rejections of R1-R5 time traces. The 'any' column represents rejection percentage of at least one test.

Regarding the results of time traces that do not satisfy at least one EVT condition (B1-B4), we can notice the rejection rate is always 100%. We can observe that the power of BDS is high for B1-B3 but it is not for B4, where KPPS and R/S are able to reject the hypothesis. On the contrary, for B1 only BDS results appears to be sufficiently powerful. Moreover, it is worth highlighting that B1 is a non-identically distributed time series, but KPSS is not able to detect it, while BDS provides for it. This is due to the lack of statistical power of KPSS in case of weak stationary, but not strict stationary time series [24].

Real platforms samples. Table 2 shows the results when time traces are generated by executing benchmark applications on the aforementioned platforms. It is possible to observe the expected trend of generating *less-compliant* time traces, with the increasing of the hardware complexity. The traces generated by the `complex` benchmark are hardly analyzable for all platforms, due to the lack of variability. This is in contrast with the common logic behind the WCET analysis for which a more stable timing is preferrable. The statistical tests described and the EVT in general instead, require a minimal degree of variability, as also shown by Lima et al. [22]. The benchmark `complex` lacks of variability as it is the only benchmark one – out of the four benchmarks – that performs simple computation on the same input data for each iteration. For example, in the PIC microcontroller case, the variability of time measurements of `complex` benchmark is due only to the measurement errors of the instrumentation.

For all the other benchmarks, the simple PIC microcontroller generates deterministic time traces that lead to a low rejection rate,

close to the significance level, i.e. the false positive rate. The MIO and STM platforms present higher values of rejection, caused by the presence of the operating system and cache memories, respectively. As expected, and with the only exception of `sqrt` case, the probabilistic theory cannot be used for the Odroid, and least of all, the Intel CPU based machine. The only unexpected outlier is the `fdct` benchmark on the Miosix board. Here the rejection rate is 100% without a clear reason. By observing the time traces, we hypothezise that the instruction prefetcher, the board is equipped with, causes large recurrent variations compared to the intrinsic variability of the `fdct` benchmark, that triggers the detection of a short-range dependence. However, this conclusion requires a more in-depth analysis on the specific system, that falls outside the scope of this paper.

5 CONCLUSION

Statistical hypothesis testing plays a key role on the reliability of probabilistic real-time estimation and the consequently safety of critical systems. However, some state-of-the-art works do not follow a systematic procedure in performing statistical tests. This work proposed three tests to assess the i.i.d. hypothesis, that have been tested on synthetic dataset and real-time traces. Moreover, we discussed which factors affect the reliability of statistical test procedures applied to probabilistic real-time computing.

ACKNOWLEDGMENTS

Work funded by the RECIPE H2020 Project [12] (Grant no. 801137).

REFERENCES

[1] J. Abella, D. Hardy, I. Puaut, E. Qui nones, and F. J. Cazorla. 2014. On the Comparison of Deterministic and Probabilistic WCET Estimation Techniques. In *2014 26th Euromicro Conference on Real-Time Systems*. IEEE, 266–275. https://doi.org/10.1109/ECRTS.2014.16

[2] Jaume Abella, Maria Padilla, Joan Del Castillo, and Francisco J. Cazorla. 2017. Measurement-Based Worst-Case Execution Time Estimation Using the Coefficient of Variation. *ACM Trans. Des. Autom. Electron. Syst.* 22, 4, Article 72 (June 2017), 29 pages. https://doi.org/10.1145/3065924

[3] R. Bender and S. Lange. 2001. Adjusting for multiple testing - when and how? *Journal of Clinical Epidemiology* 54, 4 (April 2001), 343–349.

[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] C. Brandolese, S. Corbetta, and W. Fornaciari. 2011. Software energy estimation based on statistical characterization of intermediate compilation code. In *IEEE/ACM International Symposium on Low Power Electronics and Design*. 333–338. https://doi.org/10.1109/ISLPED.2011.5993659

[6] W. A. Broock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron. 1996. A test for independence based on the correlation dimension. *Econometric Reviews* 15, 3 (1996), 197–235. https://doi.org/10.1080/07474939608800353

[7] A. Burns and S. Edgar. 2000. Predicting computation time for advanced processor architectures. In *Proceedings 12th Euromicro Conference on Real-Time Systems. Euromicro RTS 2000*. 89–96. https://doi.org/10.1109/EMRTS.2000.853996

[8] Enrique Castillo, Ali S Hadi, Narayanaswamy Balakrishnan, and José-Mariá Sarabia. 2005. *Extreme value and related models with applications in engineering and science*. Wiley Hoboken, NJ.

[9] F. J. 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. 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

[10] Michel Couillard and Matt Davison. 2005. A comment on measuring the Hurst exponent of financial time series. *Physica A: Statistical Mechanics and its Applications* 348 (2005), 404 – 418. https://doi.org/10.1016/j.physa.2004.09.035

[11] 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), 180–190.

[12] William Fornaciari, Giovanni Agosta, David Atienza, and Carlo et al. Brandolese. 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)*. ACM, New York, NY, USA, 187–194. https://doi.org/10.1145/3229631.3239368

[13] 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.

[14] Jan Gustafsson, Adam Betts, Andreas Ermedahl, and Björn Lisper. 2010. The Mälardalen WCET Benchmarks – Past, Present and Future. In *10th International Workshop on Worst-Case Execution Time Analysis, WCET 2010, July 6, 2010, Brussels, Belgium*, Björn Lisper (Ed.). OCG, Brussels, Belgium, 137–147.

[15] H. E. HURST. 1951. Long term storage capacity of reservoirs. *ASCE Transactions* 116, 776 (1951), 770–808. https://ci.nii.ac.jp/naid/10011004012/en/

[16] Frank J. Massey Jr. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Statist. Assoc.* 46, 253 (1951), 68–78.

[17] 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)*. 333–339.

[18] L. Kosmidis, C. Curtsinger, E. QuiÃśones, J. Abella, E. Berger, and F. J. Cazorla. 2013. Probabilistic timing analysis on conventional cache designs. In *2013 Design, Automation Test in Europe Conference Exhibition (DATE)*. 603–606. https://doi.org/10.7873/DATE.2013.132

[19] O. Kotaba, J. Nowotsch, M. Paulitsch, S. M. Petters, and H. Theiling. 2013. Multicore in real-time systems–temporal isolation challenges due to shared resources. In *Workshop on Industry-Driven Approaches for Cost-effective Certification of Safety-Critical, Mixed-Criticality Systems*. Grenoble, France, 6.

[20] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. 1992. 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 (1992), 159 – 178. https://doi.org/10.1016/0304-4076(92)90104-Y

[21] M. R. Leadbetter and Holger Rootzen. 1988. Extremal Theory for Stochastic Processes. *Ann. Probab.* 16, 2 (04 1988), 431–478.

[22] G. Lima and I. Bate. 2017. Valid Application of EVT in Timing Analysis by Randomising Execution Time Measurements. In *2017 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*. 187–198. https://doi.org/10.1109/RTAS.2017.17

[23] G. Lima, D. Dias, and E. Barros. 2016. Extreme Value Theory for Estimating Task Execution Time Bounds: A Careful Look. In *2016 28th Euromicro Conference on Real-Time Systems (ECRTS)*. 200–211. https://doi.org/10.1109/ECRTS.2016.20

[24] L.R. Lima and B. Neri. 2013. A Test for Strict Stationarity. In *Uncertainty Analysis in Econometrics with Applications*, Van-Nam Huynh, Vladik Kreinovich, Songsak Sriboonchitta, and Komsan Suriya (Eds.). Springer Berlin Heidelberg, 17–30.

[25] Bo Qian and Khaled Rasheed. 2004. Hurst exponent and financial market predictability. In *Proceedings of the Second IASTED International Conference on Financial Engineering and Applications*.

[26] Petar Radojković, Sylvain Girbal, Arnaud Grasset, Eduardo Quiñones, Sami Yehia, and Francisco J. Cazorla. 2012. On the Evaluation of the Impact of Shared Resources in Multithreaded COTS Processors in Time-critical Environments. *ACM Trans. Archit. Code Optim.* 8, 4, Article 34 (Jan. 2012), 25 pages. https://doi.org/10.1145/2086696.2086713

[27] F. Reghenzani, G. Massari, and W. Fornaciari. 2017. Mixed Time-Criticality Process Interferences Characterization on a Multicore Linux System. In *2017 Euromicro Conference on Digital System Design (DSD)*. IEEE, Wien, 427–434. https://doi.org/10.1109/DSD.2017.18

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

[29] F. Reghenzani, G. Massari, and W. Fornaciari. 2018. The Misconception of Exponential Tail Upper-Bounding in Probabilistic Real-Time. *IEEE Embedded Systems Letters* (2018), 1–1. https://doi.org/10.1109/LES.2018.2889114

[30] Jan Reineke. 2014. Randomized Caches Considered Harmful in Hard Real-Time Systems. *Leibniz Transactions on Embedded Systems* 1, 1 (2014), 03–1–03:13. https://doi.org/10.4230/LITES-v001-i001-a003

[31] R.D. Reiss and M. Thomas. 2007. *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser Basel. https://books.google.it/books?id=I-g-I_I2OZIC

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

[33] L. Santinelli, J. Morio, G. Dufour, and D. Jacquemart. 2014. On the Sustainability of the Extreme Value Theory for WCET Estimation. In *14th International Workshop on Worst-Case Execution Time Analysis (OpenAccess Series in Informatics (OASIcs))*, Vol. 39. 21–30. https://doi.org/10.4230/OASIcs.WCET.2014.21

[34] M. A. Stephens. 1974. EDF Statistics for Goodness of Fit and Some Comparisons. *J. Amer. Statist. Assoc.* 69, 347 (1974), 730–737. https://doi.org/10.1080/01621459.1974.10480196